Terry D. Oswalt
Editor-in-Chief

Linda M. French
Paul Kalas
Volume Editors

Planets, Stars and Stellar Systems

VOLUME 3

Solar and Stellar Planetary Systems

Springer Reference

Planets, Stars and Stellar Systems

Solar and Stellar Planetary Systems

Terry D. Oswalt (Editor-in-Chief)
Linda M. French · Paul Kalas (Volume Editors)

Planets, Stars and Stellar Systems

Volume 3:
Solar and Stellar Planetary Systems

With 238 Figures and 20 Tables

Springer

Editor-in-Chief
Terry D. Oswalt
Department of Physics & Space Sciences
Florida Institute of Technology
University Boulevard
Melbourne, FL, USA

Volume Editors
Linda M. French
Professor of Physics and
Physics Department Chair
Wesleyan University
IL, USA

Paul Kalas
Astronomy Department
University of California
Berkeley, CA, USA

ISBN 978-94-007-5605-2 ISBN 978-94-007-5606-9 (eBook)
ISBN 978-94-007-5607-6 (print and electronic bundle)
DOI 10.1007/978-94-007-5606-9

This title is part of a set with
Set ISBN 978-90-481-8817-8
Set ISBN 978-90-481-8818-5 (eBook)
Set ISBN 978-90-481-8852-9 (print and electronic bundle)

Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012953926

© Springer Science+Business Media Dordrecht 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Series Preface

It is my great pleasure to introduce "Planets, Stars, and Stellar Systems" (PSSS). As a "Springer Reference", PSSS is intended for graduate students to professionals in astronomy, astrophysics and planetary science, but it will also be useful to scientists in other fields whose research interests overlap with astronomy. Our aim is to capture the spirit of 21st century astronomy – an empirical physical science whose almost explosive progress is enabled by new instrumentation, observational discoveries, guided by theory and simulation.

Each volume, edited by internationally recognized expert(s), introduces the reader to a well-defined area within astronomy and can be used as a text or recommended reading for an advanced undergraduate or postgraduate course. Volume 1, edited by Ian McLean, is an essential primer on the tools of an astronomer, i.e., the telescopes, instrumentation and detectors used to query the entire electromagnetic spectrum. Volume 2, edited by Howard Bond, is a compendium of the techniques and analysis methods that enable the interpretation of data collected with these tools. Volume 3, co-edited by Linda French and Paul Kalas, provides a crash course in the rapidly converging fields of stellar, solar system and extrasolar planetary science. Volume 4, edited by Martin Barstow, is one of the most complete references on stellar structure and evolution available today. Volume 5, edited by Gerard Gilmore, bridges the gap between our understanding of stellar systems and populations seen in great detail within the Galaxy and those seen in distant galaxies. Volume 6, edited by Bill Keel, nicely captures our current understanding of the origin and evolution of local galaxies to the large scale structure of the universe.

The chapters have been written by practicing professionals within the appropriate subdisciplines. Available in both traditional paper and electronic form, they include extensive bibliographic and hyperlink references to the current literature that will help readers to acquire a solid historical and technical foundation in that area. Each can also serve as a valuable reference for a course or refresher for practicing professional astronomers. Those familiar with the "Stars and Stellar Systems" series from several decades ago will recognize some of the inspiration for the approach we have taken.

Very many people have contributed to this project. I would like to thank Harry Blom and Sonja Guerts (Sonja Japenga at the time) of Springer, who originally encouraged me to pursue this project several years ago. Special thanks to our outstanding Springer editors Ramon Khanna (Astronomy) and Lydia Mueller (Major Reference Works) and their hard-working editorial team Jennifer Carlson, Elizabeth Ferrell, Jutta Jaeger-Hamers, Julia Koerting, and Tamara Schineller. Their continuous enthusiasm, friendly prodding and unwavering support made this series possible. Needless to say (but I'm saying it anyway), it was not an easy task shepherding a project this big through to completion!

Most of all, it has been a privilege to work with each of the volume Editors listed above and over 100 contributing authors on this project. I've learned a lot of astronomy from them, and I hope you will, too!

January 2013

Terry D. Oswalt
General Editor

Preface to Volume 3

Twenty years ago this volume could not have been written. For generations humans studied *the* Solar System and theorized about its origin. Speculation about other planetary systems dates back at least as far back as Democritus from antiquity, and Giordano Bruno and Galileo at the dawn of modern scientific inquiry. The discovery of, and the surge of research on, exoplanets in the last decade or so have shown that the Solar System in which we live is but one of thousands, perhaps millions, of planetary systems in the Milky Way galaxy. The study of Solar System bodies, both large and small, has enabled researchers to develop models of how planetary systems form and evolve. Discoveries and investigations of extrasolar planets, in turn, have helped improve understanding of how our own Solar System may have originated and evolved.

By combining solar system and extrasolar planet science into a single volume, the present work advocates a perspective that the two disciplines are converging and can be viewed as a single topic of study. Unfortunately, this merger makes the task of thoroughly covering the subject matter in a single volume impossible. Here the reader will find a diverse selection of material related to the Solar System and extrasolar planetary systems. The opening chapter by Youdin and Kenyon describes current paradigms for the formation of planetary systems, developed from observations of nascent circumstellar disks of gas and dust and more mature planetary systems. The chapter by Morbidelli then gives a model for the subsequent dynamical evolution of the Solar System, aiming at answering the question, "Why is there such a great diversity in the architectures of planetary systems?"

Of course, the Solar System continues to provide the most detailed data about planets. Important groups of objects are discussed in subsequent chapters: terrestrial planets (Barlow), Jovian planets (Chanover and Stevenson), asteroids and meteorites (Rivkin), and planetary rings (Tiscareno). Planetary magnetospheres are addressed in the chapter by Bagenal. The volume concludes with chapters describing direct observations of the two primary components of exoplanetary systems: the dusty debris belts or disks that dominate the cumulative surface area (Moro-Martin), and the exoplanets that comprise the majority of the mass in the system (Wright and Gaudi).

Carl Sagan wrote, "In all the history of mankind, there will be only one generation which will be the first to explore the solar system, one generation for which, in childhood, the planets are distant and indistinct discs moving through the night, and for which, in old age, the planets are places, diverse new worlds in the course of exploration" (London lecture at Eugene, Oregon, 1970). Less than a generation after those words were written, humanity has begun discovering planets beyond the solar system we live in and investigating the nature and evolution of planetary systems.

We owe the authors of these chapters a great deal for their dedicated scholarship and their willingness to put so much effort into these contributions. Their community spirit, cooperation, and patience are much appreciated.

Editor-in-Chief

Dr. Terry D. Oswalt
Department Physics & Space Sciences
Florida Institute of Technology
150 W. University Boulevard
Melbourne, Florida 32901
USA
E-mail: toswalt@fit.edu

Dr. Oswalt has been a member of the Florida Tech faculty since 1982 and was the first professional astronomer in the Department of Physics and Space Sciences. He serves on a number of professional society and advisory committees each year. From 1998 to 2000, Dr. Oswalt served as Program Director for Stellar Astronomy and Astrophysics at the National Science Foundation. After returning to Florida Tech in 2000, he served as Associate Dean for Research for the College of Science (2000–2005) and interim Vice Provost for Research (2005–2006). He is now Head of the Department of Physics & Space Sciences. Dr. Oswalt has written over 200 scientific articles and has edited three astronomy books, in addition to serving as Editor-in-Chief for the six-volume Planets, Stars, and Stellar Systems series.

Dr. Oswalt is the founding chairman of the Southeast Association for Research in Astronomy (SARA), a consortium of ten southeastern universities that operates automated 1-meter class telescopes at Kitt Peak National Observatory in Arizona and Cerro Tololo Interamerican Observatory in Chile (see the website www.saraobservatory.org for details). These facilities, which are remotely accessible on the Internet, are used for a variety of research projects by faculty and students. They also support the SARA Research Experiences for Undergraduates (REU) program, which brings students from all over the U.S. each summer to participate one-on-one with SARA faculty mentors in astronomical research projects. In addition, Dr. Oswalt secured funding for the 0.8-meter Ortega telescope on the Florida Tech campus. It is the largest research telescope in the State of Florida.

Dr. Oswalt's primary research focuses on spectroscopic and photometric investigations of very wide binaries that contain known or suspected white dwarf stars. These pairs of stars, whose separations are so large that orbital motion is undetectable, provide a unique opportunity to explore the low luminosity ends of both the white dwarf cooling track and the main sequence; to test competing models of white dwarf spectral evolution; to determine the space motions, masses, and luminosities for the largest single sample of white dwarfs known; and to set a lower limit to the age and dark matter content of the Galactic disk.

Volume Editors

Linda M. French
Professor of Physics and
Physics Department Chair
Wesleyan University
Illinois
USA

Linda French earned her A.B. degree in astronomy from Indiana University and her M.S. and Ph.D. in planetary astronomy from Cornell University. She has been a member of the Illinois Wesleyan faculty since 2002, where she is currently chair of the Physics Department. Dr. French has served as education officer and as secretary of the Division for Planetary Sciences of the American Astronomical Society. Her primary astronomical research focuses on the physical properties of primitive Solar System objects such as distant asteroids and comets. She has a particular interest in the nature and provenance of the Jovian Trojan asteroids. Dr. French is a frequent observer at telescopes around the world and involves undergraduate students in her research.

Paul Kalas
Astronomy Department
University of California
Berkeley, CA 94720-3411
USA
E-mail: kalas@berkeley.edu

Dr. Kalas is an associate adjunct professor in University of California Berkeley's Department of Astronomy since 2000. Previously, he was a postdoctoral scholar at the Space Telescope Science Institute in Baltimore, Maryland, and the Max-Planck Institute for Astronomy in Heidelberg, Germany. He received his Ph.D. in 1996 from the University of Hawaii, conducting a survey for dusty debris disks with one of the first stellar coronagraphs ever made. He continues to use state-of-the-art, high-contrast imaging instruments for the direct detection of planetary systems around nearby stars. Using ground-based coronagraphy and the Hubble Space Telescope, he has discovered several dusty debris disks as well as Fomalhaut b, which earned recognition as one of the most important scientific discoveries of 2008. He is also a lead coinvestigator of the Gemini Planet Imager science team. He founded the Spirit of Lyot conference series and developed the first comprehensive course on science ethics for graduate students in astronomy.

Table of Contents

Series Preface ..... **v**
Preface to Volume 3 ..... **vii**
Editor-in-Chief ..... **ix**
Volume Editors ..... **xi**
List of Contributors ..... **xv**

Volume 3

1 From Disks to Planets **1**
Andrew N. Youdin · Scott J. Kenyon

2 Dynamical Evolution of Planetary Systems **63**
Alessandro Morbidelli

3 Terrestrial Planets **111**
Nadine G. Barlow

4 Gas and Ice Giant Interiors **195**
David J. Stevenson

5 Atmospheres of Jovian Planets **223**
Nancy Chanover

6 Planetary Magnetospheres **251**
Fran Bagenal

7 Planetary Rings **309**
Matthew S. Tiscareno

8 An Overview of the Asteroids and Meteorites **377**
Andrew Rivkin

9 Dusty Planetary Systems **431**
Amaya Moro-Martín

10 Exoplanet Detection Methods **489**
Jason T. Wright · B. Scott Gaudi

Index **541**

List of Contributors

Fran Bagenal
Astrophysical and Planetary Sciences
Dept. and Laboratory for Atmospheric and Space Physics
University of Colorado
Boulder, CO
USA

Nadine G. Barlow
Department of Physics and Astronomy
Northern Arizona University
Flagstaff, AZ
USA

Nancy Chanover
Astronomy Department
New Mexico State University
Las Cruces, NM
USA

B. Scott Gaudi
Department of Astronomy
The Ohio State University
Columbus, OH
USA

Scott J. Kenyon
Smithsonian Astrophysical Observatory
Cambridge, MA
USA

Alessandro Morbidelli
Observatory of Nice
Nice
Cedex 4
France

Amaya Moro-Martín
Department of Astrophysics
Centro de Astrobiología
INTA-CSIC
Instituto de Tecnica Aeroespacial
Madrid
Spain
and
Department of Astrophysical Sciences
Princeton University
Princeton, NJ
USA

Andrew Rivkin
Johns Hopkins University/Applied Physics Laboratory
Laurel, MD
USA

David J. Stevenson
Division of Geological and Planetary Sciences
California Institute of Technology
Pasadena, CA
USA

Matthew S. Tiscareno
Center for Radiophysics and Space Research
Cornell University
Ithaca, NY
USA

Jason T. Wright
Department of Astronomy and Astrophysics
Penn State University
University Park, PA
USA

Andrew N. Youdin
Smithsonian Astrophysical Observatory
Cambridge, MA
USA

1 From Disks to Planets

Andrew N. Youdin · Scott J. Kenyon
Smithsonian Astrophysical Observatory, Cambridge, MA, USA

1 Introduction ***3***

2 Observational Constraints on Planet Formation Theories ***3***
2.1 Lessons from the Solar System 3
2.1.1 The Solar Nebula 3
2.1.2 Isotopic Timescales 5
2.1.3 Water 6
2.2 Disks Surrounding the Youngest Stars 7
2.3 The Exoplanet Revolution 8

3 Disk Properties and Evolution ***9***
3.1 Basic Disk Dynamics 10
3.2 Transport Mechanisms and the α Disk Model 13
3.3 Viscously Heated Disks 15
3.4 Steady Irradiated Disks 17
3.5 Time Dependence 18
3.6 Disk Instabilities and Fragmentation 19

4 From Dust to Planetesimals ***20***
4.1 The "Meter-Sized" Barrier 21
4.1.1 Radial Drift and the Basics of Disk Aerodynamics 22
4.1.2 Early Collisional Growth 25
4.2 Gravitational Collapse of Solids into Planetesimals 26
4.3 Aerodynamic Particle Concentration 28
4.4 Observational Constraints on Planetesimal Formation 30

5 Planetesimals to Planets ***32***
5.1 Growth of Solid Protoplanets 33
5.1.1 Basic Length and Velocity Scales 33
5.1.2 Gravitationally Focused Collisions 35
5.1.3 Planetesimal Velocity Evolution 37
5.1.4 Fragmentation 39
5.1.5 Planetesimal Accretion with Gas Damping 41
5.1.6 Numerical Simulations of Low-Mass Planet Formation 43
5.2 Accretion of Atmospheres 46
5.2.1 Static Protoplanet Atmospheres 46
5.2.2 Enhanced Planetesimal Accretion 47
5.2.3 The Core Accretion Instability 48

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_1, © Springer Science+Business Media Dordrecht 2013

5.2.4 Direct Accretion of Disk Gas (and How it Stops) 51
5.2.5 Numerical Simulations of Gas Giant Planet Formation 51
5.3 Direct Formation of Brown Dwarfs and Gas Giants 55

6 Summary ***57***

Acknowledgments ***57***

References ***58***

1 Introduction

Theories for the formation and evolution of planets depended primarily on geophysical data from the solar system (see Brush 1990, and references therein). Starting in the 1940s, astrophysical data began to provide new insights. Discoveries of pre-main-sequence stars in Taurus-Auriga, Orion, and other regions led to the concept that stars form in giant clouds of gas and dust (see Kenyon et al. 2008b, and references therein). Because nearly every young star has a circumstellar disk with enough mass to make a planetary system, theorists began to connect the birth of stars to the birth of planets. Still, the solar system remained unique until the 1990s, when the first discoveries of exoplanets began to test the notion that planetary systems are common. With thousands of (candidate) planetary systems known today, we are starting to have enough examples to develop a complete theory for the origin of the Earth and other planets.

Here, we consider the physical processes that transform a protostellar disk into a planetary system around a single star. Instead of discussing the astrophysical and geophysical data in detail (e.g., Dauphas and Chaussidon 2011), we focus on a basic introduction to the physical steps involved in building a planet. To begin, we discuss several observational constraints from the wealth of astrophysical and geophysical material in ❯ Sect. 2. We then describe the global physical properties and evolution of the disk in ❯ Sect. 3. Aside from the special conditions required for fragments in the disk to collapse directly into giant planets (❯ Sect. 5.3), most planets probably grow from micron-sized dust grains. Thus, we consider how a turbulent sea of grains produces the building blocks of planets and planetesimals (❯ Sect. 4) and how ensembles of planetesimals collide and merge into planets which may later accrete a gaseous atmosphere (❯ Sect. 5). We conclude with a brief summary in ❯ Sect. 6. ❯ *Table 1-1* defines frequently used symbols for reference.

2 Observational Constraints on Planet Formation Theories

We begin with the main observational constraints on planet formation processes, including raw materials, timescales, and outcomes. Detailed studies of the solar system and the disks around the youngest stars yield strict limits on the mass available and the time required to make a planetary system. The diverse population of exoplanets illustrate the many outcomes of planet formation.

2.1 Lessons from the Solar System

Until the discovery of exoplanets, the solar system was the only known planetary system. The solar system will continue to provide the most detailed data on planet formation, despite obvious issues of statistical significance and anthropic bias.

2.1.1 The Solar Nebula

The alignment of major planets in the ecliptic plane suggests that they formed within a flattened disk or "nebula." The philosopher Immanuel Kant and the mathematician Pierre-Simon

Laplace are often credited for this "nebular hypothesis." However, the scientist and theologian Emanuel Swedenborg first recorded this insight in his 1734 *Principia*. For a long time, the nebular hypothesis competed with the theory, proposed by the naturalist Buffon, that planets were tidally extracted from the Sun during an encounter. Though Laplace dismissed the encounter theory as being inconsistent with the circular orbits of the planets, it survived to reach peak popularity in the early twentieth century as the Chamberlin-Moulton hypothesis (Jeffreys 1929). Once Payne (1925) identified hydrogen as the most abundant element in stars, the nebular hypothesis regained favor. Adding the tidal theory's idea of planetesimals – small solid particles that condense out of hot gas – the nebular hypothesis began to develop into a robust theory for planet formation.

The minimum-mass solar nebula (MMSN) provides a simple estimate of the mass available in planet-forming disks. The recipe for the MMSN is to distribute the mass currently in the solar system's planets into abutting annuli, adding volatile elements (mainly hydrogen gas) until the composition is solar. Kuiper (1951) and Cameron (1962, see his Table 5) estimated the mass of the MMSN as a few percent of a solar mass. Weidenschilling (1977b) and Hayashi (1981) fit the now canonical $R^{-3/2}$ surface density law, bravely smoothing the mass deficits in the regions of Mercury, Mars, and the asteroid belt and the abundance uncertainties for the giant planets. The roughness of the fit is immaterial: the MMSN is not a precise initial condition but a convenient fiducial for comparing disk models.

We use the same MMSN as Chiang and Youdin (2010) with disk surface density profiles:

$$\Sigma_{\rm g} = 2,200\, F \left(\frac{R}{\rm AU}\right)^{-3/2} \ {\rm g\,cm^{-2}} \tag{1.1}$$

$$\Sigma_{\rm p} = 33\, F\, Z_{\rm rel} \left(\frac{R}{\rm AU}\right)^{-3/2} \ {\rm g\,cm^{-2}}\,, \tag{1.2}$$

where subscripts "g" and "p" respectively denote gas and particles (condensed solids) and R is the distance from the Sun. The parameter F scales the total mass; $F = 1$ is a reference MMSN. Integrated out to 100 AU, the MMSN disk mass is $0.03 M_\odot$.

The parameter $Z_{\rm rel}$ scales the ratio of solids to gas, or disk metallicity, as

$$Z_{\rm disk} = \Sigma_{\rm p}/\Sigma_{\rm g} = 0.015 Z_{\rm rel}\,, \tag{1.3}$$

evolves during the planet formation process, as evidenced by the enrichment of heavy elements relative to H in Jupiter and Saturn. Our fiducial value is normalized to the Lodders (2003) analysis of (proto-)solar abundances, which can be approximately summarized as

$$Z_{\rm rel} \simeq \begin{cases} 1 & T \lesssim 40\ {\rm K} \\ 0.78 & 40\ {\rm K} \lesssim T \lesssim 180\ {\rm K} \\ 0.33 & 180\ {\rm K} \lesssim T \lesssim 1{,}300 - 2{,}000\ {\rm K} \\ 0 & T \gtrsim 1{,}300 - 2{,}000\ {\rm K} \end{cases}. \tag{1.4}$$

The abundance of solids decreases with increasing temperature due to the sublimation of (most significantly) methane ice above 40 K, water ice above 180 K, and dust over a range of temperatures from roughly 1,300 to 2,000 K, covering the condensation temperatures for different minerals. Note that this definition of disk metallicity ignores the heavy elements in the gas phase, which are at least temporarily unavailable to produce planetary cores. For the disk temperature, this work adopts (❯ 1.38), the result for an irradiated disk with a self-consistently flared surface (Chiang and Goldreich 1997).

In the solar system today, the temperatures of solids are set by an equilibrium between heating and cooling. Although gravitational contraction (Jupiter) and tidal heating (many satellites, including the Moon) contribute some heating in a few objects, the Sun is the primary source of heating. Radiation from the Sun peaks at a wavelength of $\lambda_{\max,\odot} \approx 0.5\,\mu\mathrm{m}$. Objects with radius r, peak wavelength $\lambda_{\max,g} < r$, and no atmosphere radiate as nearly perfect blackbodies. Equating the energy they receive from a central star ($\pi r^2 \mathrm{L}_* / 4\pi R^2$) with the energy they emit ($4\pi r^2 \sigma T_{\mathrm{eq,bb}}^4$) leads to an equilibrium temperature ,

$$T_{\mathrm{eq,bb}} = 278 \left(\frac{\mathrm{L}_*}{\mathrm{L}_\odot}\right)^{1/4} \left(\frac{R}{\mathrm{AU}}\right)^{-1/2} \mathrm{K}\,, \tag{1.5}$$

where $\mathrm{L}_\odot$ is the luminosity of the Sun. Small grains with $r \gtrsim 1\,\mu\mathrm{m}$ and $r \lesssim \lambda_{\max,g}$ emit radiation inefficiently. In most cases, the radiative efficiency is

$$\epsilon = \begin{cases} 1 & \lambda \leq \lambda_0 \\ (\lambda/\lambda_0)^q & \lambda > \lambda_0 \end{cases}\,, \tag{1.6}$$

where λ_0 is the critical wavelength and $q \approx 1$–2 depends on grain properties. Usually λ_0 is roughly equal to the grain radius r. Because they can only radiate efficiently at short wavelengths, these grains have much larger temperatures. For $q = 1$,

$$T_{\mathrm{eq},s} = 468 \left(\frac{\mathrm{L}_*}{\mathrm{L}_\odot}\right)^{1/5} \left(\frac{R}{\mathrm{AU}}\right)^{-2/5} \left(\frac{\lambda_0}{\mu\mathrm{m}}\right)^{-1/5} \mathrm{K}\,. \tag{1.7}$$

To derive (1.7), include the efficiency in the grains' emitted radiation ($\propto 1/\epsilon$), and relate the wavelength of peak emission to the grain temperature with Wein's law ($\lambda \propto 1/T_{\mathrm{eq},s}$).

Coupled with the condensation temperatures in (1.4), these definitions allow us to identify the "snow line." Also known as the "frost line," this annulus in the disk[1] separates an inner region of rocky objects from an outer region of icy objects (Kennedy and Kenyon 2008). For blackbody grains, the water condensation temperature of 180 K implies $R_{\mathrm{snow}} \approx 2.7$ AU, roughly coincident with the asteroid belt. Similarly, the methane condensation temperature of 40 K yields another region beyond the outer edge of the Kuiper belt at 48 AU where solid objects have a combination of water and methane ice. Inside of ~0.1 AU, dust evaporates; rocky grains cannot exist so close to the Sun.

2.1.2 Isotopic Timescales

Meteorites delivered to Earth from the asteroid belt provide the most detailed chronology of the early solar system (e.g., Dauphas and Chaussidon 2011). Primitive meteorites from asteroids that did not undergo differentiation (or other significant alteration) preserve the best record of their formation. These primitive meteorites are called chondrites because they contain many chondrules. Chondrules are glassy inclusions, with a typical size ~0.1 – 1 mm. They provide evidence for high-temperature melting events in the solar nebula. The nature of these melting events is debated and beyond our scope (Connolly et al. 2006). Calcium-aluminum-inclusions (CAIs) are also present in primitive meteorites. CAIs experienced even more extreme heating than chondrules.

[1] Or, more generally, a spherical shell surrounding the central star

With ages up to 4567.11 ± 0.16 Myr (Russell et al. 2006), CAIs are the oldest known objects in the solar system. This age is consistent with current results for the main-sequence age of the Sun (Bonanno et al. 2002). The absolute ages of CAIs are measured by lead-lead dating, which makes use of half-lives, $t_{1/2}$, of uranium isotopes that are conveniently long. The decay chain of $^{235}U \rightarrow ^{207}Pb$ has $t_{1/2} = 0.704$ Gyr, while $^{238}U \rightarrow ^{206}Pb$ has $t_{1/2} = 4.47$ Gyr.

Radioisotopes with short half-lives yield accurate relative ages of meteorite components. These isotopes are "extinct"; they have decayed completely to daughter products whose abundances relative to other isotopes result in an age. The extinct isotope ^{26}Al decays to ^{26}Mg in $t_{1/2} = 0.73$ Myr. The abundance of ^{26}Mg relative to ^{27}Al and to ^{24}Mg yields an age for the sample. Aside from its use as a chronometer, ^{26}Al is a powerful heat source in young protoplanets.

Both absolute (lead-lead) and relative (^{26}Al) ages support a planet formation timescale of a few Myr. Most CAIs formed in a narrow window of $1 - 3 \times 10^5$ year; chondrule formation persisted for ~4 Myr or longer. Russell et al. (2006) discuss systematic uncertainties. Here, we emphasize the remarkable agreement of the few Myr formation time derived from primitive meteorite analyses and protoplanetary disk observations (Sect. 2.2).

The assembly of terrestrial planets from planetesimals requires tens of millions of years. Isotopic analysis of differentiated solar system bodies (including Earth, Mars, and meteorites) probes this longer timescale. The decay of radioactive hafnium into tungsten, $^{182}Hf \rightarrow ^{182}W$ with $t_{1/2} = 9.8$ Myr, dates core-mantle segregation. Tungsten is a siderophile (prefers associating with metals), while hafnium is a lithophile (prefers the rocky mantle); thus, Hf-W isotope ratios are the primary tool to date differentiation (See Chap. 3 by Barlow). Studies of Hf-W systematics indicate that asteroid accretion continued for ~10 Myr, Mars' core formed within 20 Myr, and the Earth's core grew over 30–100 Myr (Kleine et al. 2009). Astronomical observations of debris disks (Sect. 2.2) and dynamical studies of terrestrial planet accretion (Sect. 5) support these longer timescales.

2.1.3 Water

After hydrogen, water is the most abundant molecule in disks (e.g., Najita et al. 2007). Inside the snow line, water exists in the gaseous phase, though it dissociates at $T \gtrsim 2{,}500$ K. As water vapor interior to the ice line diffuses past the snow line, it condenses into icy grains. The snow line thus acts as a cold trap, where the enhanced mass in water ice (1.4) should accelerate the growth of planetesimals and perhaps gas giant planets (Stevenson and Lunine 1988).

Water is abundant throughout the solar system (e.g., Rivkin et al. 2002; Jewitt et al. 2007, and references therein). Outside of the Earth, water appears in spectra of comets, Kuiper belt objects, and satellites of giant planets (including the Moon) and bound within minerals on Mars, Europa, and some asteroids. Because they can be analyzed in great detail, meteorites from asteroids provide a wealth of information on water in the inner solar system. Many meteorites show evidence for aqueous alteration prior to falling onto the Earth. Most groups of carbonaceous chondrites and some type 3 ordinary chondrites contain hydrated minerals, suggesting association with liquid water. Despite some evidence that grains might react with water prior to their incorporation into larger solids, most analyses of the mineralogy suggest hydration on scales of mm to cm within chondrites and other meteorites (Zolensky and McSween 1988).

The water content of solar system bodies helps trace the evolution of the snow line (Kennedy and Kenyon 2008). Hydration within meteorite samples demonstrates that the snow line was

at least as close as 2.5–3 AU during the formation of the asteroids (Rivkin et al. 2002). Radiometric analyses suggest hydration dates from 5 to 10 Myr after the formation of the Sun, close to and perhaps slightly after the formation of chondrules. Ice on the Earth and on Mars suggests the possibility that the snow line might have been as close as 1 AU to the proto-Sun, a real possibility for passively irradiated disks (Sect. 3.4). Within the terrestrial zone, the rise in water abundance from Venus (fairly dry) to Earth (wetter) to Mars and the asteroids (wetter still) points to processes that either distributed water throughout the inner solar system (with a preference for regions near the snow line) or inhibited accretion of water (either vapor or ice) from the local environment. Abundance analyses, including D/H and noble gases, help to probe the (still imperfectly known) history of water in the inner solar system.

2.2 Disks Surrounding the Youngest Stars

Observations of young stars provide additional constraints on the early evolution of planetary systems. In nearby star-forming regions (Orion, Taurus, etc.), nearly every star with an age of 1 Myr or less has an optically thick circumstellar disk (Williams and Cieza 2011). The disk frequency seems independent of stellar mass. The data suggest the disks are geometrically thin, with a vertical extent of roughly 10–20% of their outer radius. They are composed of molecular gas and dust grains with sizes ranging from a few microns up to several mm. Around solar-type stars, young disks have typical luminosities of order $L_\odot$ and radii of order 100 AU.

Estimating disk masses is challenging. Aside from a few transitions possible only in warm material near the central star, H_2 is undetectable. The next most abundant molecule, CO, is optically thick and provides a crude lower limit to the total mass. Current estimates rely on a conversion from the dust emission at mm wavelengths to a dust mass and then to a mass in gas. These estimates are highly uncertain due to ignorance of dust-to-gas ratios and grain size distributions. The typical assumptions give disk masses ~0.01 $M_\odot$, a factor of 2–4 smaller than the MMSN. The dispersion for ensembles of a few hundred systems is roughly an order of magnitude (Andrews and Williams 2005; Williams and Cieza 2011). The size of the typical disk is similar to the semimajor axes of orbits in the Kuiper belt.

Current data demonstrate that more massive young stars have more massive disks. Recent observational programs concentrate on whether the ratio of disk mass to stellar mass is roughly constant or increases with stellar mass. Despite the larger scatter in this ratio at every stellar mass, the relation is probably linear (Williams and Cieza 2011).

High-resolution radio observations reveal interesting limits on the distributions of surface density and temperature within the brightest and most massive disks around nearby stars. Although disks with a broad range of surface density gradients are observed ($\Sigma \propto R^{-n}$ with $n \approx -0.6$ to 1.5), most observations indicate a typical $n \approx 0.5$–1.5 (Andrews and Williams 2005; Isella et al. 2009; Williams and Cieza 2011). Thus, the surface density gradient in the MMSN is steeper than the average protostellar disk but within the range observed in disks around other young stars.

The evolution of protostellar disks sets severe limits on the timescale for planet formation. Optical and ultraviolet spectra of young stars show that material from the disk flows onto the central star. The rate of this flow, the mass accretion rate, drops from well in excess of 10^{-8} $M_\odot$ year^{-1} at 1 Myr to much less than 10^{-11} $M_\odot$ year^{-1} at 10 Myr (Williams and Cieza 2011). Declining mass accretion rates implies much less gas near the young star. At the same time, the fraction of young stars with opaque dusty disks declines from nearly 100%

to less than 1%. Fewer dusty disks imply that the solid material has been incorporated into large (km-sized or larger) objects, accreted by the central star, or driven out of the system by radiation pressure or a stellar wind. Direct constraints on the amount of gas left in 10-Myr-old systems without opaque disks are limited to a few systems.

Among older stars with ages of 3–10 Myr, many disks have substantial mass beyond 30–50 AU but have inner holes apparently devoid of much gas or dust. The frequency of these "transition" disks suggests that the evolution opaque disk → opaque disk with inner hole → no opaque disk takes from 0.1–0.3 Myr up to 1–2 Myr (Currie and Sicilia-Aguilar 2011; Espaillat et al. 2012).

Once the opaque disk disappears, many pre-main- and main-sequence stars remain surrounded by 1–10-μm dust grains (Wyatt 2008). This material lies in a belt with radial extent $\delta R \approx 0.1 - 0.5\, R$ and vertical height $\delta z \approx 0.05 - 0.1\, R$, with $R \approx 1$–100 AU. Infrared spectroscopy suggests that the grains have compositions similar to material in the comets and the dust (zodiacal light) of the solar system. The total mass, a few lunar masses, exceeds the mass of dust in the inner solar system by factors of 100–1,000. These properties are independent of stellar metallicity and many other properties of the central star. However, the frequency of and the amount of dusty material in these dusty disks peak for stars with ages of 10–20 Myr and then decline approximately inversely with time (Currie et al. 2008).

These disks place interesting limits on the reservoir of large objects around stars with ages of 10–20 Myr. Among several possible grain removal mechanisms, the most likely are radiation processes and collisions (Backman and Paresce 1993). Grains orbiting the star feel a headwind from the incoming radiation from the central star, which causes the grain to spiral into the star (Burns et al. 1979). If the grains have a mass density $\rho_\bullet$, the orbital decay time for this Poynting-Robertson drag is

$$t_{\rm pr} = \left(\frac{4\pi r \rho_\bullet}{3}\right)\left(\frac{c^2 R^2}{\mathrm{L}_*}\right) = 710\ \rho_\bullet \left(\frac{r}{\mu\mathrm{m}}\right)\left(\frac{R}{\mathrm{AU}}\right)^2 \left(\frac{\mathrm{L}_*}{\mathrm{L}_\odot}\right)^{-1} \text{ year} . \tag{1.8}$$

The decay time, $t_{\rm pr} \sim 1$ Myr, is much shorter than the 100 Myr to 10 Gyr main-sequence lifetime of the central star. The large frequency of dusty disks among 50–500-Myr-(1–10 Gyr)-old A-type (G-type) stars suggests the grains are continually replenished over the main-sequence lifetime. By analogy with the solar system, where trails of dust result from collisions of asteroids (Nesvorný et al. 2003), high-velocity collisions among large (but undetectable) objects can replenish the dust. Adopting typical sizes for asteroids, ~1–10 km, a reservoir containing ~10 $\mathrm{M}_\oplus$ of material can explain the amount of dust around young stars and the time evolution of the dust emission among older stars. Because this mass is between the initial mass of solids in protostellar disks (~100–1,000 $\mathrm{M}_\oplus$) and the dust mass in the solar system ($\lesssim 10^{-4}\ \mathrm{M}_\oplus$), these systems are often called "debris disks" (Wyatt 2008, also See Chap. 9 by Moro-Martin).

2.3 The Exoplanet Revolution

The current pace of exoplanet discovery is extraordinarily rapid. Here, we highlight the main insights exoplanets bring to theories of planet formation.

Within the discovery space accessible with current techniques, exoplanets fill nearly all available phase space (e.g., Cumming et al. 2008; Gould et al. 2010; Howard et al. 2012; Johnson et al. 2011, and references therein). Among ~10–30% of middle-age solar-type stars, exoplanets lie within a few stellar radii from the central star out to several AU. Because there are many

multiplanet systems, the sample of short-period planets implies more planets than stars (Youdin 2011b). Though detection becomes more difficult, the frequency of exoplanets seems to grow with increasing distance from the parent star. The orbits have a broad range of eccentricities, with a peak at $e \sim 0.2$. Planet masses range from a rough upper limit at 10–20 M_J to a few Earth masses. Around stars with the same mass, lower mass planets are more frequent than higher mass planets. More massive stars tend to have more massive planets.

There is some evidence that exoplanets are more likely around metal-rich stars (Gonzalez 1997; Johnson et al. 2010). With the large samples available, this "planet-metallicity correlation" is now unambiguous for gas giants with masses ranging from the mass of Saturn up to ~10 M_J. Among lower mass planets, small samples currently prevent identifying a clear correlation. Larger samples with the *Kepler* satellite will yield a better test of the planet-metallicity correlation as a function of planet mass.

The origin of any planet-metallicity correlation establishes some constraints on formation theories. If the initial metallicity of the disk is identical to the current metallicity of the star, then gas giants – and perhaps other planets – form more frequently in more metal-rich disks. However, planets could pollute the stellar atmosphere after the rest of the gaseous disk disperses, raising the metallicity of the star above the initial metallicity of the disk. In this case, enhanced metallicity would be a result, not a cause of planet formation. Current data contradicts the pollution hypothesis (Fischer and Valenti 2005; Pasquini et al. 2007), but more study of the diverse exoplanet population is warranted.

To quantify the planet-metallicity correlation, Johnson et al. (2010) fit the frequency, f, of giant planets as a joint power law in stellar mass, M_*, and metallicity, $Z_{\rm Fe} \propto \log_{10}([{\rm Fe/H}])$,

$$f \propto M_*^{\alpha} Z_{\rm Fe}^{\beta} \,. \tag{1.9}$$

For giant planets in the California Planet Survey, $\alpha = 1.0 \pm 0.2$ and $\beta = 1.2 \pm 0.2$. Ignoring the dependence with stellar mass ($\alpha \equiv 0$) introduces a bias but yields a stronger relation with metallicity, ($\beta = 1.7 \pm 0.3$). For more massive stars with $M_* > 1.4 M_\odot$, $\alpha = 1.5 \pm 0.4$ and $\beta = 0.73 \pm 0.35$. Thus, the formation of giant planets around more massive stars is more sensitive to stellar mass and less sensitive to metallicity than for lower mass stars.

With exoplanet samples growing so rapidly, new analyses will change at least some of these conclusions. The most firm conclusions – that exoplanetary systems are common and have nearly as much diversity as possible – provide a good counterpoint to the wealth of detail available from in situ analyses of the solar system.

3 Disk Properties and Evolution

Stars form within collapsing clouds of gas and dust. When a cloud collapses, most infalling material has too much angular momentum to fall directly onto the nascent protostar. This gas forms a rotationally supported circumstellar disk (Cassen and Moosman 1981; Terebey et al. 1984). If the angular momentum in the disk is transported radially outward, gas can accrete onto the central star. Although jets launched near the protostar (Shu et al. 2000) or from the disk (Pudritz et al. 2007) or both can remove significant angular momentum from the disk, most analyses concentrate on how angular momentum flows through the disk.

At early times, the disk mass $M_{\rm disk}$ is similar to the stellar mass M_*. For a circumstellar disk with surface density Σ and radial flow velocity v_R, the rate material flows through the disk as a

function of radial coordinate R is

$$\dot{M} = -2\pi R v_R \Sigma \,. \tag{1.10}$$

Positive $\dot{M}$ corresponds to gas flowing toward, and eventually draining onto, the central star. If the mass infall rate from the surrounding molecular cloud $\dot{M}_i$ exceeds $\dot{M}$, the disk mass grows. If $M_{\rm disk}$ exceeds ~$0.3M_*$, gravitational instabilities within the disk can produce a binary companion (Adams et al. 1989; Kratter et al. 2010a). Smaller instabilities may form brown dwarfs or giant planets (Sect. 5.3). At late times, several processes – including the clearing action of protostellar jets and winds (Shu et al. 1987) – stop infall. Without a source of new material, the disk mass gradually declines with time.

In general, all of the physical variables characterizing the cloud and the disk change with radius and time. However, $\dot{M}_i$, $\dot{M}$, and Σ often change slowly enough that it is useful to construct steady disk models with a constant $\dot{M}$ throughout the disk. Here, we develop the basic equations governing the evolution of the disk and use steady disks to show the general features of all circumstellar disks.

An evolving gaseous disk sets the physical conditions in which small particles grow into planets. Physical conditions within the disk limit how planetesimals can form (Sect. 4) and how solid planets grow out of planetesimals (Sect. 5.1.5). Once solid planets form, the gaseous disk provides the mass reservoir for giant planet atmospheres (Sect. 5.2) and drives planet migration (see See Chap. 2 by Morbidelli).

3.1 Basic Disk Dynamics

To introduce basic concepts in disk dynamics, we describe orbital motion in a gas disk and the radial flow induced by viscosity. The orbital velocity of the gas v_ϕ is set by the balance of radial forces – centrifugal, pressure, and gravitational[2] – as

$$-\frac{v_\phi^2}{R} + \rho^{-1}\frac{\partial P}{\partial R} + \frac{GM_*}{R^2} = 0 \,, \tag{1.11}$$

where ρ is the gas density and P is the gas pressure. Away from the immediate vicinity of protoplanets, gravity from the central star typically dominates. Even a self-gravitating thin disk has $M_{\rm disk} \ll M_*$. For $P = P_0(R/R_0)^{-n}$ (the index n need only be locally valid), the orbital motion is

$$v_\phi = v_K\left(1 - \frac{nc_s^2}{\gamma v_K^2}\right)^{1/2} \,, \tag{1.12}$$

in terms of the Kepler velocity, $v_K = \sqrt{GM_*/R}$, and the sound speed, $c_s = \sqrt{\gamma P/\rho}$, with γ the adiabatic index. Plausible disk models have an outwardly decreasing pressure ($n > 0$) almost everywhere. Thus, pressure support typically gives sub-Keplerian rotation, but the correction is quite small, $v_\phi - v_K \sim -10^{-3}v_K$. It is often safe to ignore the pressure correction to orbital motions but not when studying the motion of solids relative to gas (Weidenschilling 1977a; Youdin 2010, Sect. 4.1.1).

Accretion disks also have radial inflow, which is constrained by the laws of mass, angular momentum, and energy conservation. Indeed accretion disks can be considered as machines

[2]The radial speeds associated with accretion produce negligible advective accelerations, $Dv_R/Dt \sim v_R^2/R$.

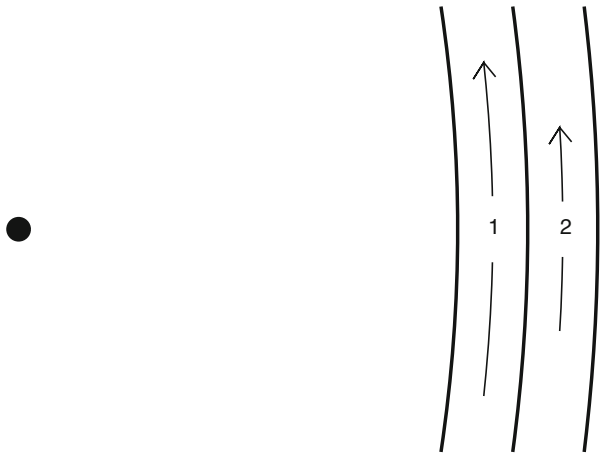


Fig. 1-1
Schematic view of two adjacent annuli in a disk surrounding a star (*black point at left*). Annulus 1 lies inside annulus 2; material in annulus 1 orbits the star more rapidly than material in annulus 2 ($\Omega_1 > \Omega_2$)

that radiate energy as they transport mass inward and angular momentum outward. The viscous disk model offers the simplest means to understand how a disk manages this feat. Consider a thin ring with two adjacent annuli at distances R_1 and R_2 from a star with mass M_* (Fig. 1-1). Material orbits the star with angular velocities, $\Omega_1 = \sqrt{GM_*/R_1^3}$ and $\Omega_2 = \sqrt{GM_*/R_2^3}$. If the gas has viscosity, the differential rotation, $\Omega_1 - \Omega_2 < 0$ for $R_2 > R_1$, produces a frictional (shear) force that attempts to equalize the two angular velocities. The resulting torques produce an outward flow of angular momentum. By moving a small amount of disk material onto distant, high-angular momentum orbits, large amount of mass can fall inward to low angular momentum orbits. Mass accretion is biased toward inflow because specific (i.e., per unit mass) angular momentum can increase to very large values but cannot fall below zero. The heat generated by viscous dissipation affects the disk temperature and the predicted spectra as described in Sect. 3.3.

For a disk with surface density Σ and viscosity ν, mass continuity and conservation of angular momentum lead to a nonlinear diffusion equation for Σ (e.g., Lynden-Bell and Pringle 1974; Pringle 1981),

$$\frac{\partial \Sigma}{\partial t} = 3R^{-1} \frac{\partial}{\partial R} \left(R^{1/2} \frac{\partial}{\partial R} \{ \nu \Sigma R^{1/2} \} \right) + \dot{\Sigma}_{\rm ext}. \tag{1.13}$$

The first term on the right-hand side corresponds to viscous evolution; the second is a source term which is positive for infall from the cloud ($\dot{\Sigma}_{\rm ext} = \dot{\Sigma}_i$) and negative for mass loss due to photoevaporation, disk winds, or accretion onto giant planets. Consider a simple model with ν = constant, $\dot{\Sigma}_{\rm ext} = 0$, and $\Sigma(R, t = 0) = m\delta(R - R_0)/2\pi R_0$. That is, the initial mass m is in a narrow ring at radius R_0. The time evolution of the surface density is

$$\Sigma(x, \tau) = \frac{m}{\pi R_0^2} \tau^{-1} x^{-1/4} e^{-(1+x^2)/\tau} I_{1/4}(2x/\tau) , \tag{1.14}$$

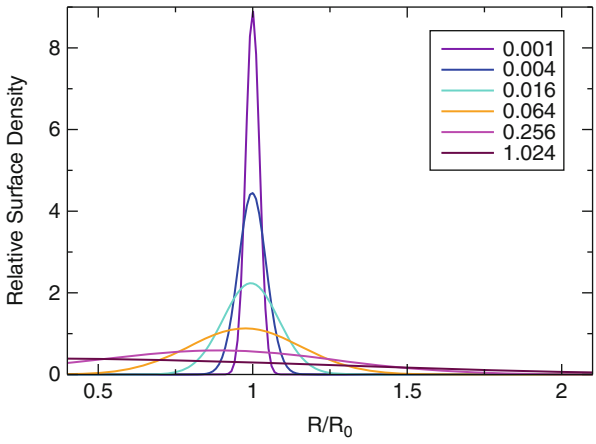


Fig. 1-2
Time evolution of the surface density for a ring with constant viscosity (1.14). Over time, viscous diffusion spreads the ring into a disk. The legend indicates the scaled time, τ, for each curve

where the scaled distance and time are $x = R/R_0$ and $\tau = t/t_0 = 12\nu t/R_0^2$ and $I_{1/4}$ is the modified Bessel function. *Figure 1-2* shows how this viscous solution asymptotically transports all of the mass to $R = 0$ and all of the angular momentum to infinity. This evolution occurs on the viscous timescale τ.

We now consider the evolution of disks with a more general viscosity law, $\nu = \nu_0(R/R_0)^\beta$. With a constant power law β, exact similarity solutions to (1.13) exist (Lynden-Bell and Pringle 1974). To develop intuition, we instead physically derive the approximate solution. After several viscous timescales have passed, the disk "forgets" the initial conditions (as seen in *Fig. 1-2*). Conservation of angular momentum alone (without applying mass conservation needed to derive (1.13)) gives

$$\frac{\partial \Sigma}{\partial t} = \frac{1}{R^{3/2}} \frac{\partial}{\partial R} \left[\frac{\dot{M}\sqrt{R}}{2\pi} - \frac{3}{2} \nu \Sigma \sqrt{R} \right] . \tag{1.15}$$

The two terms in square brackets represent the advection of angular momentum and viscous torques. For a steady disk with $\partial\Sigma/\partial t = 0$, the term in square brackets must equal a constant (independent of R). This constant represents the torque at the inner boundary, which can be neglected far from that boundary (as we show explicitly in (1.26)). Thus, angular momentum conservation gives

$$\Sigma \simeq \frac{\dot{M}}{3\pi\nu} \propto R^{-\beta} , \tag{1.16}$$

where the final proportionality assumes the power-law viscosity and constant $\dot{M}$. From (1.10), the accretion speed follows as

$$v_R = -\frac{3\nu}{2R} . \tag{1.17}$$

To derive the evolution of disk mass (and $\dot{M}$) in this limit, we adopt R_0 as the outer edge of the disk. The outer radius changes on the local viscous timescale

$$R_0 \sim \sqrt{\nu t} \sim (\nu_0 t)^{1/(2-\beta)} . \tag{1.18}$$

To conserve angular momentum and energy, the disk expands, requiring $\beta < 2$. Conservation of angular momentum, $J \sim M_{\rm disk}\Omega(R_0)R_0^2$, drives the evolution of the disk mass as

$$M_{\rm disk} \sim \frac{J}{\sqrt{GM_*}}(\nu_0 t)^{-1/(4-2\beta)} . \tag{1.19}$$

The accretion rate through the disk is

$$\dot{M} = \frac{dM_{\rm disk}}{dt} \propto t^{-(5-2\beta)/(4-2\beta)} . \tag{1.20}$$

This physical derivation emphasizes the role of conservation laws in setting the global viscous evolution of the disk. The results are consistent with similarity solutions derived using Green's functions (Lynden-Bell and Pringle 1974).

3.2 Transport Mechanisms and the α Disk Model

The source of a physical mechanism to drive disk accretion is a long-standing problem. The molecular viscosity, $\nu_{\rm mol} = c_s\lambda$ – the product of sound speed, c_s, and collisional mean free path, λ – is far too small. For the MMSN, $\lambda = \mu m_H c_s/\Omega\Sigma\sigma_{H_2} \approx 0.8R^{11/4}$ cm, where $\mu = 2.4$ is the mean molecular weight, m_H is the mass of a hydrogen atom, and σ_{H_2} is the collision cross section for the dominant constituent of the gas, H_2. The resulting accretion timescale

$$t_{\rm acc,mol} = \frac{M_{\rm disk}}{\dot{M}} \sim \frac{R^2}{\nu_{\rm mol}} \sim 7 \times 10^{13} \left(\frac{R}{\rm AU}\right)^{-4/7} \text{ year} \tag{1.21}$$

vastly exceeds the age of the universe. Viable mechanisms for angular momentum transport are sometimes identified as "anomalous" sources of viscosity. Because long-range interactions are often important, the analogy between a transport mechanism and the local viscosity is inexact. For sufficiently small-scale fluctuations, however, even self-gravitating disks are well approximated by the viscous prescription (Lodato and Rice 2004).

Anomalous viscosity is most often associated with turbulence in the disk. Because the size of turbulent eddies can greatly exceed λ, the effective turbulent viscosity can vastly exceed the molecular one, even for slow, subsonic turbulence. However, turbulence by itself does not explain accretion; velocity fluctuations must correlate for angular momentum to be transported outward. Moreover, it is difficult (if not impossible) for Keplerian shear to drive turbulence; the angular momentum gradient in disks is quite stable according to the Rayleigh criterion (Stone and Balbus 1996).

Over the past 50 years, theorists have considered a wide range of transport mechanisms: convective eddies, gravitational instabilities, internal shocks, magnetic stresses, orbiting planets, sound waves, spiral density waves, and tidal forces. Currently, the "magnetorotational instability" (MRI) is the leading candidate for a transport mechanism in low-mass disks (Balbus and Hawley 1991; Papaloizou and Nelson 2003). In this mechanism, modest magnetic fields thread ionized material orbiting the central star. An outward (or inward) perturbation of fluid stretches and shears the magnetic fields. The resulting torque amplifies the original perturbation and, crucially, transports angular momentum outward. Although this mechanism is attractive, the

low ionization fraction of protostellar disks restricts the MRI to surface layers of the disk at many radii. Disks may then contain extensive "dead zones" (Gammie 1996), where levels of transport and turbulence are reduced. In massive protostellar disks, gravitational waves and gravitoturbulence are another likely source of angular momentum transport (Lin and Pringle 1987).

To sidestep fundamental uncertainties of transport mechanisms, it is convenient to adopt a simple viscosity model (Shakura and Sunyaev 1973; Lynden-Bell and Pringle 1974). Setting $\nu = \alpha c_s H$ – where $H = c_s/\Omega$ is the vertical scale height of the disk – leads to the popular "α-disk" model. The dimensionless parameter $\alpha < 1$ (even $\ll 1$) since large values would lead to rapid shock dissipation and/or gravitational fragmentation. Similar to the mixing length parameterization of convection, α-disk models allow progress despite ignorance of the underlying dynamics. Detailed simulations typically quote measured transport rates in terms of effective α values.

This definition allows us to define the three important timescales in a viscous disk (e.g., Lynden-Bell and Pringle 1974; Pringle 1981). The shortest disk timescale is the dynamical (orbital) timescale, $t_d \sim \Omega^{-1}$:

$$t_d \approx 0.17 \text{ year} \left(\frac{M_*}{\mathrm{M}_\odot}\right)^{-1/2} \left(\frac{R}{\mathrm{AU}}\right)^{3/2} . \tag{1.22}$$

The disk establishes hydrostatic equilibrium in the vertical direction on the same timescale $t_\nu \approx H/c_s \approx \Omega^{-1} \approx t_d$.

The disk cooling time, $t_c \approx U/D$, is the ratio of the thermal energy content (per unit area), $U = C_\nu \Sigma T$, to the energy generation or dissipation rate, D. For a viscous disk, $D = 9\nu\Sigma\Omega^2/4$ (Pringle 1981); C_ν is the specific heat at constant volume. With $C_\nu T = (\gamma - 1)^{-1} P/\rho$ for an ideal gas, the cooling time is

$$t_c \approx \frac{4}{9\gamma(\gamma-1)} \alpha^{-1} \Omega^{-1} . \tag{1.23}$$

In this classical result, the thermal timescale depends only on the local dynamical timescale, the dimensionless viscosity (α), and the equation of state (γ). For a molecular gas, $\gamma \approx 7/5$. The cooling time is a factor of roughly α^{-1} larger than the dynamical time:

$$t_c \approx 0.08 \text{ year } \alpha^{-1} \left(\frac{M_*}{\mathrm{M}_\odot}\right)^{-1/2} \left(\frac{R}{\mathrm{AU}}\right)^{3/2} . \tag{1.24}$$

The viscous timescale – $t_\nu = R^2/3\nu$ – measures the rate matter diffuses through the disk. Using our expressions for the sound speed and the viscosity, $t_\nu \approx (\alpha\Omega)^{-1}(R/H)^2$. Thus, the viscous timescale is

$$t_\nu \gtrsim 0.17 \text{ year } \alpha^{-1} \left(\frac{R}{H}\right)^2 \left(\frac{M_*}{\mathrm{M}_\odot}\right)^{-1/2} \left(\frac{R}{\mathrm{AU}}\right)^{3/2} . \tag{1.25}$$

Typically the disk is thin, $H/R \approx 0.03$–0.1; thus, the viscous timescale is 100–1,000 times longer than the cooling time.

The radial velocity in (❯ 1.17) becomes $v_R \approx \alpha c_s H/R \approx 0.1\alpha(H/R)^2 v_\phi$. With $\alpha < 1$ and $H/R < 1$, v_R is much smaller than both the orbital velocity and the sound speed.

3.3 Viscously Heated Disks

As shown above, $t_\nu \gg t_c > t_d$ for $\alpha < 1$, so the thermal properties of the disk adjust rapidly to changes in the surface density distribution. This property is very useful in describing the thermal properties of many astrophysical disks, including active galactic nuclei, interacting binaries, and pre-main-sequence stars (Pringle 1981). We now describe the basic energetics of accretion disks with constant $\dot{M}$.

Since gas remains on Keplerian orbits as it accretes, the specific energy release infalling from $R + dR$ to R is $GM_* dR/(2R^2)$, since half the potential energy goes into the increase in kinetic energy. The energy release per unit area for a disk accreting at $\dot{M}$ is $F_K(R) = G\dot{M}M_*/(4\pi R^3)$. This result is completely independent of how the energy was released.

The energy released by viscous dissipation does not simply match the local change in kinetic energy. To fully describe the energetics of steady viscous disks, we must keep the integration constant when integrating (1.15) over radius to get

$$\nu\Sigma = \frac{\dot{M}}{3\pi}\left(1 - \sqrt{\frac{R_J}{R}}\right). \tag{1.26}$$

The integration constant, R_J, represents the torque $\dot{J}$ exerted at the inner boundary, $R_{\rm in}$ as $\dot{J} = \dot{M}\sqrt{GM_*}(\sqrt{R_{\rm in}} - \sqrt{R_J})$. The standard choice $R_J = R_{\rm in}$ is a zero-torque boundary condition. Negative torques are not allowed for steady disks, as $R_J > R_{\rm in}$ would require $\nu\Sigma < 0$. For $R_J = 0$, the maximum torque $\dot{J}_{\rm max} = \dot{M}\sqrt{GM_* R_{\rm in}}$ matches the flow of angular momentum past the inner boundary.

The laws of fluid dynamics in cylindrical coordinates (Shu 1992) give the viscous dissipation as[3]

$$D(R) = \nu\Sigma\left(R\frac{\partial\Omega}{\partial R}\right)^2 = \frac{3GM_*\dot{M}}{4\pi R^3}\left(1 - \sqrt{\frac{R_J}{R}}\right). \tag{1.27}$$

As advertised, this expression does not simply match the local release of Keplerian orbital energy; far from the boundary ($R \gg R_J$), $D(R) \approx 3F_K(R)$. Viscous disks transport energy (in addition to angular momentum) from the inner disk to the outer disk. Nevertheless, the rapid falloff with R means that most energy is dissipated close to the disk's inner edge.

Now consider the total energy release from large R to the inner boundary. The Keplerian energy release is just $L_K = GM_*\dot{M}/(2R_{\rm in})$. The total viscous luminosity is

$$L_d = 2\pi\int_{R_{\rm in}}^{\infty} D(R)2\pi R dR = \frac{3}{2}\frac{GM_*\dot{M}}{R_{\rm in}}\left(1 - \frac{2}{3}\sqrt{\frac{R_J}{R_{\rm in}}}\right) \tag{1.28}$$

For the zero-torque boundary condition ($R_J = R_{\rm in}$), the luminosity simply matches the release of Keplerian energy. However, the disk's luminosity increases due to work done by torques at the inner edge, up to $L_d = 3L_K$ for $R_J = 0$. For typical parameters in protostellar disks

$$L_d = \frac{f_d GM_*\dot{M}}{2R_{\rm in}} \approx 0.16 f_d L_\odot \left(\frac{\dot{M}}{10^{-8}\ {\rm M}_\odot\ {\rm year}^{-1}}\right)\left(\frac{M_*}{M_\odot}\right)\left(\frac{R_{\rm in}}{{\rm R}_\odot}\right)^{-1}, \tag{1.29}$$

where f_d ranges from 1 (no torque) to 3 (maximum torque).

[3] Note that Eq. 3.10 in Pringle (1981) has a factor of 2 typo in the intermediate result (involving ν) but reaches the correct final result (in terms of $\dot{M}$).

The maximum disk luminosity occurs for a disk that extends to the stellar surface, $R_{\rm in} = R_*$. The total accretion luminosity,

$$L_{\rm acc} = \frac{f_* G M_* \dot{M}}{2 R_*}, \tag{1.30}$$

with $1 \lesssim f_* \lesssim 2$, includes all the energy loss needed to come to rest on the rotating stellar surface. For a star rotating at breakup, $f_* \approx 1$. For a slowly rotating star, the damping of the orbital kinetic energy gives twice the energy release $f_* \approx 2$. Any difference $L_{\rm acc} - L_d \geq 0$ is emitted at the stellar surface. This difference must be positive (accretion should not cool the star), further constraining f_d. As a consistency check, note that a disk with an inner boundary at the surface of a star, $R_{\rm in} = R_*$, that rotates at breakup must satisfy the zero-torque boundary condition to avoid $L_{\rm acc} < L_d$.

In many cases, the accreting star has a magnetosphere that truncates the disk at $R_{\rm in} > R_*$ (e.g., Ghosh and Lamb 1979). Material then flows onto the star along magnetic field lines, collimated onto hot spots, which are hot because the accretion energy is emitted from a small fraction of the stellar photosphere. In most young stars, $R_{\rm in} \approx 3$–$5\ {\rm R}_*$ (e.g., Kenyon et al. 1996; Bouvier et al. 2007). Thus, the hot-spot luminosity

$$L_{\rm hot} = L_{\rm acc} - L_d = L_{\rm acc} \left(1 - \frac{f_d R_*}{f_* R_{\rm in}} \right) \tag{1.31}$$

can easily reach 60–80% of the total accretion luminosity. For star that corotates with the disk's inner edge, $f_d \approx f_* \approx 1$ is expected (Shu et al. 1994).

To calculate the temperature structure of viscously heated disks, note that the upper and lower halves of the disk each radiates half of $D(R)$. If the vertical optical depth $\tau > 1$, the disk photosphere then has effective temperature T_e,

$$\sigma T_e^4 = \frac{3 G M_* \dot{M}}{8 \pi R^3} \left(1 - \sqrt{\frac{R_J}{R}} \right). \tag{1.32}$$

Though T_e is calculated as if the disk radiates as a blackbody, the disk's atmosphere will radiate as a stellar atmosphere with spectral lines, especially when $T_e \gtrsim 1{,}000$–$1{,}500$ K. The effective temperature declines as $T_e \propto R^{-3/4}$ for $R \gtrsim$ a few R_*:

$$T_e = 85\ {\rm K} \left(\frac{\dot{M}}{10^{-8}\ {\rm M}_\odot\ {\rm year}^{-1}} \right)^{1/4} \left(\frac{M_*}{M_\odot} \right)^{1/4} \left(\frac{R}{\rm AU} \right)^{-3/4} \left(1 - \sqrt{R_J/R} \right). \tag{1.33}$$

In a simple, gray-atmosphere approach, the midplane temperature T_d is a factor of $\tau^{1/4}$ larger than T_e and is used to derive the scale-height H, the viscosity ν, and other physical variables. More rigorous approaches calculate T_e and T_d using a self-consistent prescription for the opacity throughout the disk.

The temperature distribution in (❯ 1.32) allows us to derive the surface density of an α disk. Assuming the midplane temperature scales like the effective temperature ($T_d \approx T_e \propto R^{-3/4}$), $\nu \propto c_s^2 \Omega^{-1} \propto r^{3/4}$ for the α prescription. Then (❯ 1.16) gives

$$\Sigma(R) = \Sigma_0 \left(\frac{R}{\rm AU} \right)^{-3/4} \tag{1.34}$$

where Σ_0, the surface density at 1 AU, depends on the mass accretion rate. Solving for the midplane temperature with realistic opacities yields $\Sigma \propto R^{-\beta}$ with $\beta \approx 0.6$–1 (e.g., Stepinski 1998; Chambers 2009).

3.4 Steady Irradiated Disks

Although viscous dissipation drives the overall evolution, radiation from the central star also heats the disk (Friedjung 1985; Adams and Shu 1986; Kenyon and Hartmann 1987). If the disk is an infinite, but very thin, sheet, it absorbs roughly 25% of the light radiated by the central star. For a 1-$L_\odot$ central star and disk accretion rates $\dot{M} \lesssim 10^{-8}$ $M_\odot$ year^{-1}, emission from irradiation exceeds emission from viscous dissipation (1.29). If the disk reradiates this energy at the local blackbody temperature, the radial temperature gradient of a flat, irradiated disk follows the gradient for a viscous disk, $T_d \propto R^{-3/4}$ (Friedjung 1985; Adams and Shu 1986).

Disks with vertical scale-height H absorb and reradiate even more starlight (Kenyon and Hartmann 1987). Chiang and Goldreich (1997) derive a general formalism for H and T_d in a "passive" disk with negligible $\dot{M}$. Defining θ as the grazing angle that starlight hits the disk, the temperature of a disk that emits as a blackbody is

$$T_d \approx \left(\frac{\theta}{2}\right)^{1/4} \left(\frac{R_*}{R}\right)^{1/2} T_* \,, \tag{1.35}$$

where T_* is the stellar temperature. The grazing angle is

$$\theta \approx 0.4\frac{R_*}{R} + R\frac{d}{dR}\left(\frac{h}{R}\right) \,, \tag{1.36}$$

where h is the height of the photosphere above the disk midplane. For a blackbody disk in vertical hydrostatic equilibrium, the grazing angle is the sum of a nearly flat component close to the star and a flared component far from the star:

$$\theta \approx 0.005\left(\frac{R}{\mathrm{AU}}\right)^{-1} + 0.05\left(\frac{R}{\mathrm{AU}}\right)^{2/7} \,. \tag{1.37}$$

The disk temperature beyond a few tenths of an AU is then

$$T_d \approx 155 \mathrm{\ K} \left(\frac{R}{\mathrm{AU}}\right)^{-3/7} \left(\frac{R_*}{R_\odot}\right)^{1/2} \left(\frac{T_*}{T_\odot}\right) \,. \tag{1.38}$$

For $\dot{M} \approx 10^{-8}$ $M_\odot$ year^{-1}, the irradiated disk is roughly twice as hot as a viscous accretion disk.

This temperature relation leads to a steeper surface density gradient in α disks. With $\nu \propto c_s^2 \Omega^{-1}$ and $c_s^2 \propto T_d$, $\nu \propto r^{15/14}$. Using this viscosity in (1.16),

$$\Sigma(R) = \Sigma_{0,\mathrm{irr}} \left(\frac{R}{\mathrm{AU}}\right)^{-15/14} \,, \tag{1.39}$$

where

$$\Sigma_{0,\mathrm{irr}} = \frac{2\,\mathrm{g\,cm^{-2}}}{\alpha} \left(\frac{\dot{M}}{10^{-8}\ M_\odot\ \mathrm{year}^{-1}}\right) \,. \tag{1.40}$$

The surface density gradient for an irradiated disk is steeper than the gradient for a viscous disk and is reasonably close to the gradient for the MMSN.

For identical α, hotter irradiated disks have larger viscosity and smaller surface density than cooler viscous disks. Integrating (1.39) over radius, the mass of an irradiated disk is

$$M_d \approx \left(\frac{10^{-4} M_\odot}{\alpha}\right) \left(\frac{\dot{M}}{10^{-8}\ M_\odot\ \mathrm{year}^{-1}}\right) \left(\frac{R_d}{100\ \mathrm{AU}}\right)^{15/14} \,. \tag{1.41}$$

When $\alpha \approx 10^{-3} - 10^{-2}$, this estimate is close to the observed masses of protostellar disks around young stars.

3.5 Time Dependence

Deriving more robust estimates of disk evolution requires a direct solution of (1.13). This exercise requires a model for the viscosity and a prescription for the thermodynamics and opacity of disk material. Analytic approaches assume a constant mass accretion rate through the disk. If α and τ are simple functions of the local variables Σ and T, then the diffusion equation can be solved exactly for $\Sigma(t)$, $\dot{M}(t)$, and other disk properties (Stepinski 1998; Chambers 2009). Numerical approaches allow α and $\dot{M}$ to vary with radius. Some solutions consider iterative solutions to the temperature and vertical structure (e.g., Hueso and Guillot 2005); others solve for the vertical structure directly using techniques developed for the atmospheres of stars (e.g., Bell and Lin 1994; D'Alessio et al. 1998).

To compare these approaches, we consider a simple model for a viscous disk irradiated by a central star. We assume that the optical depth of cool disk material is dominated by dust grains with a constant opacity κ_0; warmer dust grains evaporate and have a smaller opacity:

$$\kappa = \begin{cases} \kappa_0 , & T_d \leq T_{\rm evap} \\ \kappa_0 \left(\frac{T_d}{T_{\rm evap}} \right)^n , & T_d > T_{\rm evap} \end{cases} \tag{1.42}$$

For material with roughly solar metallicity, typical values are $\kappa_0 \approx 2$, $T_{\rm evap} \approx 1{,}380$ K, and $n = -14$ (e.g., Chambers 2009). With this opacity, we derive a self-consistent disk temperature and scale height (as in Hueso and Guillot 2005) and solve for the time evolution of Σ using an explicit solution to the diffusion equation (as in Bath and Pringle 1982).

Figure 1-3 compares analytic and numerical results for a disk with $\alpha = 10^{-2}$, initial mass $M_{d,0} = 0.04\,{\rm M}_\odot$, and initial radius $R_0 = 10$ AU surrounding a star with $\rm M_* = 1\,M_\odot$. The numerical solution tracks the analytic model well. At early times, the surface density declines steeply in the inner disk ($\Sigma \propto R^{-1.2}$; where dust grains evaporate) and more slowly in the outer disk

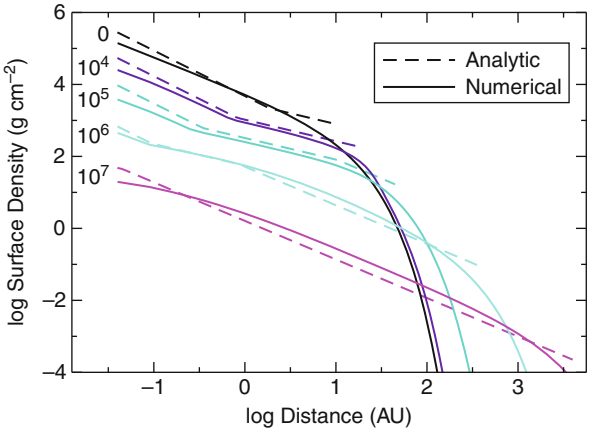


Fig. 1-3

Time evolution of the surface density of a gaseous disk surrounding a 1-$M_\odot$ star. *Dashed lines* show results for the analytic disk model of Chambers (2009); *solid lines* show results for our numerical solution of the diffusion equation. Despite small differences in the initial conditions, the numerical solution tracks the analytic model

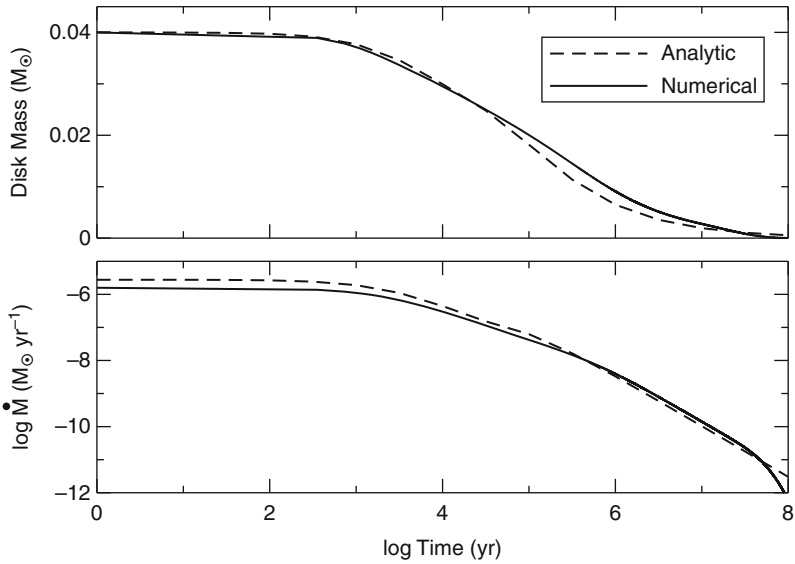


Fig. 1-4

Time evolution of the disk mass (*upper panel*) and disk accretion rate onto the central star (*lower panel*) for the analytic and numerical solutions in *Fig. 1-3*

($\Sigma \propto R^{-0.6}$; where viscous transport dominates). At late times, irradiation dominates the energy budget; the surface density then falls more steeply with radius, $\Sigma \propto R^{-1}$.

Other approaches lead to similar time evolution in the surface density. Early on, a massive disk is dominated by viscous heating. For these conditions, the simple analytic estimate of the surface density yields $\Sigma \propto R^{-n}$ with $n = 3/4$ (1.34), close to results for the numerical solution ($n = 0.6$) and other analytic and numerical ($n = 0.6 - 1$) approaches (Bath and Pringle 1982; Lin and Pringle 1990; Stepinski 1998; Chambers 2009; Alexander and Armitage 2009). As the disk ages, it evolves from a viscous-dominated to an irradiation-dominated system. Thus, the exponent n in the surface density relation approaches the limit ($n = 15/14$) derived in (1.39).

Figure 1-4 compares the evolution of the disk mass and accretion rate at the inner edge of the disk. In both solutions, the disk mass declines by a factor of roughly 2 in 0.1 Myr, a factor of roughly 4 in 1 Myr, and a factor of roughly 10 in 10 Myr. Over the same period, the mass accretion rate onto the central star declines by roughly four orders of magnitude.

3.6 Disk Instabilities and Fragmentation

In addition to evolution on the viscous timescale shown in *Fig. 1-3*, all disks vary their energy output on much shorter timescales. In compact binary systems, these fluctuations range from small, 10–20%, amplitude flickering on the local dynamical timescale to large-scale eruptions, factors of 10–100, that can last for several times the local viscous timescale (Warner 1995). Although many pre-main-sequence stars also display distinctive brightness variations (Joy 1945; Herbig 1962), the FU Ori variables provide the cleanest evidence for large-scale variations of the disk, rather than the environment or the central star (Hartmann and Kenyon 1996).

Theory suggests several types of instabilities in viscous disks (see Pringle 1981). In standard derivations of the structure of steady disks, radiative cooling balances heating from viscous stresses. However, radiative losses are set by local disk parameters; local parameters and an input accretion rate set viscous energy generation. Usually radiative losses can keep up with changes in disk structure; sometimes, radiation cannot balance viscous stresses, leading to a thermal instability. A limit cycle arises, where regions of the disk alternate between states where radiative losses exceed (and then fall below) the viscous energy input. This mechanism may produce FU Ori and other eruptions in the disks of pre-main-sequence stars (Hartmann and Kenyon 1996).

Viscous instabilities occur when changes in the local surface density do not produce parallel increases in the local mass transfer rate. From (❯ 1.16), $\dot{M} \propto \nu\Sigma$. In a steady disk, ν is fairly independent of Σ; thus, $\dot{M}$ changes in step with Σ. For the MRI viscosity mechanism, however, larger Σ leads to larger optical depths, less ionization, and smaller α. Thus, an MRI disk with growing (falling) surface density can produce a smaller (larger) viscosity, leading to an ever greater over- or underdensity in the surface density.

Although thermal and viscous instabilities change the temperature and surface density throughout the disk, they evolve on timescales much longer than the local orbital period. Massive disks can evolve on shorter timescales. If the local gravity in a region with size λ overcomes rotational support ($G\Sigma \gtrsim \Omega^2\lambda$) and thermal support ($G\Sigma \gtrsim c_s^2/\lambda$), this region can (begin to) collapse (Safronov 1960; Toomre 1964; Goldreich and Lynden-Bell 1965; Paczynski 1978). Together, these conditions require $c_s^2/(G\Sigma) \lesssim \lambda \lesssim G\Sigma/\Omega^2$. Collapse at any wavelength requires the disk satisfy the "Toomre instability criterion,"

$$Q \equiv \frac{c_s \Omega}{\pi G \Sigma} \lesssim 1 \,. \tag{1.43}$$

Setting the disk mass $M_d \approx \Sigma R^2$, a stable disk has

$$c_s \gtrsim \frac{M_d}{M_*}\, v_\phi \,. \tag{1.44}$$

When the disk first forms, $M_d \approx M_*$. Such "disks" cannot be thin because $H/R \sim c_s/v_\phi \gtrsim 1$.

In a viscous accretion disk, the stability criterion can be rewritten in terms of the accretion rate (Gammie 2001). With $\dot{M} = 3\pi\nu\Sigma$ and $\nu = \alpha c_s^2 \Omega^{-1}$, an unstable disk has $\dot{M}_Q \gtrsim 3\alpha c_s^3/G$. To evaluate the temperatures of unstable disks, we use $c_s = (\gamma kT/\mu m_H)^{1/2}$ and set $\gamma = 7/5$ and $\mu = 2.4$ for molecular gas:

$$\dot{M}_Q \gtrsim 2.4\, 10^{-4}\, \alpha T^{3/2}\, \mathrm{M}_\odot\ \mathrm{year}^{-1} \,, \tag{1.45}$$

with T in Kelvins. For the observed accretion rates in very young stars, $\dot{M} \sim 10^{-7}\ \mathrm{M}_\odot\ \mathrm{year}^{-1}$, unstable disks have $\alpha T^{3/2} \lesssim 0.4$. If α is large (10^{-2}), only very cold disks are unstable (T ~ 10–15 K); smaller α (e.g., 10^{-3}) allows instability in warmer disks (T ~ 50–60 K).

4 From Dust to Planetesimals

The accumulation of dust grains into planetesimals – solids greater than a kilometer in size – is the first step in the formation of terrestrial planets and giant planet cores. Several observational and theoretical reasons suggest that the formation of planetesimals is a separate step. Observationally, remnant planetesimals in the solar system and in extrasolar debris disks show that growth sometimes stalls before planets accumulate all planetesimals. Comets from the Oort cloud also suggest an intermediate stage between dust grains and planets (See ❯ Chap. 9 by Moro-Martin).

Theoretically, the physical processes responsible for the growth of planetesimals differ from those relevant to the final stages of planet formation. As Sect. 5 describes, few-body gravitational encounters – both scattering and gravitationally focused collisions – establish the rates of growth for icy and terrestrial planets. By contrast, the sticking of dust grains involves electrostatic forces. During planetesimal formation, particle dynamics is qualitatively different. Drag forces exerted by the gas disk dominate the motions of small solids. Though not negligible, the drag exerted on kilometer-sized or larger planetesimals is weaker than gravitational interactions (Sect. 5.2).

While planetesimal formation is a common occurrence in circumstellar disks, understanding how it happens has proved elusive. Observations of planetesimal formation in action are indirect. Particles beyond centimeter sizes contribute negligibly to images and spectra of circumstellar disks. Primitive meteorites record the conditions during planetesimal formation, but the implications for formation mechanisms are difficult to interpret – we need a better instruction manual.

Especially beyond millimeter sizes, experiments show that particle collisions often result in bouncing or breaking instead of sticking (Blum and Wurm 2000; Zsom et al. 2010; Weidling et al. 2012). Inefficient growth by coagulation is further complicated by the rapid infall of centimeter- to meter-sized solids into the star. These difficulties – often termed the "meter-sized barrier" – are explained in more detail in Sect. 4.1.

Gravitational collapse is one way to overcome the growth barrier. The mutual gravitational attraction of a collection of small solids could lead to a runaway collapse into planetesimals – even when sticking is inefficient and radial drift is fast. While appealing, this path encounters theoretical difficulties when stirring by turbulent gas is included. Section 4.2 describes the current status of the gravitational collapse hypothesis.

Even when self-gravity is weak, aerodynamic effects can concentrate solids in the disk. Particles tend to seek high-pressure regions in the disk. This tendency causes the inward drift mentioned above. Particles can also concentrate in localized pressure maxima. Predicting the sizes and lifetimes of pressure maxima is a difficult (and currently unsolved) problem of disk meteorology.

In addition to the passive response of solids to the gas disk, active particle concentration occurs when particles cause their own clumping by altering the flow of gas. Instabilities caused by gas drag, notably the streaming instability, provide a clumping mechanism that is both powerful and amenable to study by direct numerical simulations. The strong clumping driven by the streaming instability is capable of triggering gravitational collapse into ~100-km planetesimals. Both passive and active particle concentration mechanisms are reviewed in Sect. 4.3.

Theories based on complex nonlinear dynamics must be tested against, and refined by, observations. We discuss observational consequences of planetesimal formation models in Sect. 4.4. Unless stated otherwise, the numerical estimates in the section use the passively heated MMSN disk (Sect. 3). For more detailed reviews of planetesimal formation, see Chiang and Youdin (2010) and Youdin (2010). For a thorough review of collision experiments and their relevance to planetesimal formation, see Blum and Wurm (2008).

4.1 The "Meter-Sized" Barrier

We discuss in more detail the two components of the "meter-sized" barrier to planetesimal formation. The review of radial drift timescales in Sect. 4.1.1 also serves as an introduction

to the dynamics of solids in a gas disk. The discussion of collisional growth and destruction in ❯ Sect. 4.1.2 couples dynamical models of collision speeds to the complex physics of contact mechanics.

4.1.1 Radial Drift and the Basics of Disk Aerodynamics

The aerodynamic migration of small solids imposes the most stringent timescale constraint on planet formation: ≈100 years. Aerodynamic radial drift arises because solids encounter a headwind as they orbit through the gas disk (❯ 1.12). This headwind removes angular momentum from particle orbits, causing their inspiral. Infall speeds are fastest for solids near roughly meter sizes. The critical size is actually below a meter in standard disk models – especially in their outer regions. So the "meter-sized" barrier is a slight misnomer, but it has a better ring than the "millimeter-to-tens-of-centimeters-sized" barrier.

Radial pressure gradients in gas disks set the speed of the headwind. Plausible disk models are hotter and denser in the inner regions; on average, the radial pressure gradient is directed outward. If the radial pressure gradient is directed inward, a tailwind – and outward particle migration – results.

We express the headwind speed as the difference between the Keplerian velocity, $v_K = \sqrt{GM_*/R} = \Omega R$, and the orbital speed of the gas, $v_{g,\phi}$:

$$\eta v_K \equiv v_K - v_{g,\phi} \approx -\frac{\partial P/\partial \ln R}{2\rho_g v_K} \approx 25 \left(\frac{R}{\mathrm{AU}}\right)^{1/14} \mathrm{m\ s^{-1}}, \tag{1.46}$$

where P and ρ_g are the pressure and density of the gas and $\eta \sim c_s^2/v_K^2 \sim (H/R)^2 \sim 10^{-3}$ is a dimensionless measure of pressure support. In disks hotter than our passive model, headwinds and drift speeds are faster.

To derive (❯ 1.46), compute radial force balance assuming (correctly) that the radial pressure acceleration, $f_{P,R} = -\rho_g^{-1}\partial P/\partial R$, is weak compared to the centrifugal acceleration. Equivalently we can reproduce (❯ 1.46) by balancing the pressure and Coriolis forces, $f_{P,R} + f_{\mathrm{Cor},R} = 0$ with $f_{\mathrm{Cor},R} = -2\Omega\eta v_K$.

Drag forces set the response of particle orbits to the gas headwind. We express the drag acceleration felt by a solid particle as

$$\boldsymbol{f}_{\mathrm{drag}} = -\Delta\boldsymbol{v}/t_s\,, \tag{1.47}$$

where $\Delta\boldsymbol{v}$ is the particle velocity relative to the gas and t_s is the aerodynamic damping timescale for this relative particle motion.

The value of t_s depends on particle properties – such as the internal density, $\rho_\bullet$, and spherical radius, s – and on properties of the gas disk – ρ_g and c_s – as

$$t_s = \begin{cases} t_s^{\mathrm{Ep}} \equiv \rho_\bullet s/(\rho_g c_s) & \text{if } s < 9\lambda/4 \quad (1.48a) \\ t_s^{\mathrm{Stokes}} \equiv t_s^{\mathrm{Ep}} \cdot 4s/(9\lambda) & \text{if } 9\lambda/4 < s < \lambda/(4\mathrm{Ma}) \quad (1.48b) \\ t_s^{\mathrm{int}} \cdot (s/\lambda)^{3/5}\,\mathrm{Ma}^{-2/5}/4 & \text{if } \lambda/(4\mathrm{Ma}) < a < 200\lambda/\mathrm{Ma} \quad (1.48c) \\ t_s^{\mathrm{turb}} \equiv t_s^{\mathrm{Ep}} \cdot 6/\mathrm{Ma} & \text{if } s > 200\lambda/\mathrm{Ma} \quad (1.48d) \end{cases}$$

where $\mathrm{Ma} \equiv |\Delta\boldsymbol{v}|/c_s$, $\lambda \propto 1/\rho_g$ is the gas mean free path and $\mathrm{Re} \equiv 4s\mathrm{Ma}/\lambda$ is the Reynolds number of the flow around the particle. The cases are written in order of increasing particle size: Epstein's law of drag from molecular collisions, Stokes' law for viscous drag when

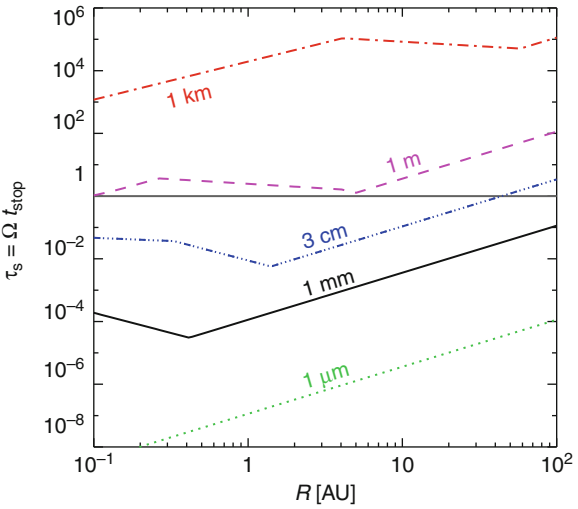


Fig. 1-5

Aerodynamic stopping time normalized to the Keplerian orbital frequency for a range of particle sizes in our reference minimum mass disk model. Small (large) values of τ_s indicate strong (weak) coupling of solids to the gas disk. The breaks in the curves are due to transitions between different drag laws, as described by (1.48). An internal density of $\rho_\bullet = 1\ \mathrm{g\ cm^{-3}}$ is assumed for the solids

Re < 1, an approximate intermediate Re case, and the drag from a fully developed turbulent wake for Re > 800. The turbulent drag force is more relevant for fully formed planetesimals and is commonly expressed as

$$F_{\mathrm{drag}} = -m\frac{|\Delta \boldsymbol{v}|}{t_s^{\mathrm{turb}}} = -\frac{C_D}{2}\pi s^2 \rho_g |\Delta \boldsymbol{v}|^2\,, \tag{1.49}$$

where the drag coefficient, $C_D \approx 0.44$ (Adachi et al. 1976; Weidenschilling 1977a).

The dynamical significance of drag forces is measured by comparing the stopping time and the orbital frequency, via the parameter

$$\tau_s \equiv \Omega t_s\,. \tag{1.50}$$

For $\tau_s \ll 1$, particles are carried along with the gas; for $\tau_s \gg 1$, gas drag is a small correction to Keplerian orbits. *Figure 1-5* plots τ_s for a range of particle sizes in our passively heated MMSN. At least in the inner disk, objects near meter sizes have $\tau_s \approx 1$. In the outer disk, where gas densities are lower, smaller solids have the critical $\tau_s = 1$. As we now show, $\tau_s = 1$ solids have the fastest radial drift speeds.

To derive the particle drift caused by the gas headwind, we consider the equations of motion for a particle in cylindrical coordinates, r and ϕ,

$$\ddot{R} - R\dot{\phi}^2 = -v_K^2/R - \dot{R}/t_s \tag{1.51}$$

$$R\ddot{\phi} + 2\dot{R}\dot{\phi} = -\left(R\dot{\phi} - v_{g,\phi}\right)/t_s\,. \tag{1.52}$$

To find the steady drift solutions, we make several approximations that are valid when drag forces are strong. We neglect the radial inertial acceleration, $\ddot{R}$, and express the azimuthal

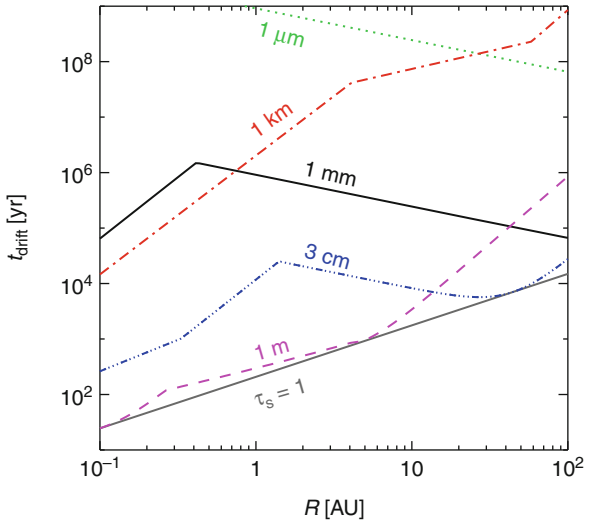


Fig. 1-6

Radial drift timescales, $R/\dot{R}$, for the same disk models and particle sizes as in Fig. 1-5. The fastest drift timescale is for the particle size that has $\tau_s = 1$ as indicated by the *gray curve*. Drift timescales are much faster than disk lifetimes of a few Myr, especially near the "meter-sized" barrier

motion as a small deviation from the Keplerian frequency, $\dot{\phi} = \Omega + \delta v_\phi / R$ where $|\delta v_\phi| \ll \Omega R$. The azimuthal acceleration is then $\ddot{\phi} \approx \dot{\Omega} \approx -3\Omega\dot{R}/(2R)$.

The radial drift speed follows from (1.46), (1.50), (1.51) and (1.52) as

$$\dot{R} \approx -2\eta v_K \left(\frac{\tau_s}{1+\tau_s^2} \right) . \tag{1.53}$$

Solids with $\tau_s = 1$ have the fastest infall speed, $-\dot{R} = \eta v_K$. The corresponding timescale

$$\min(t_R) \sim (\eta\Omega)^{-1} \sim 200 (R/\mathrm{AU})^{13/14} \text{ year} \tag{1.54}$$

is a very strong constraint on growth. This constraint is the main element of the meter-sized growth barrier. *Figure 1-6* plots the radial drift timescales for a range of particle sizes.

To complete this brief introduction to particle aerodynamics, we give the azimuthal drift speed of solids through the (sub-Keplerian) gas as

$$R\dot{\phi} - v_{g,\phi} = \delta v_\phi + \eta v_K = \eta v_K \frac{\tau_s^2}{1+\tau_s^2} . \tag{1.55}$$

Large, $\tau_s \gg 1$ solids experience the full ηv_K headwind, yet their radial drift is slow because their inertia is so large. Small, $\tau_s \ll 1$ solids are dragged by the gas and only feel a mild headwind. We thus see why radial drift is fastest near $\tau_s \approx 1$. For these intermediate sizes, drag forces are strong enough to overcome particle inertia but not so strong as to cause perfect coupling.

These idealized calculations explain the basics of radial drift. A pressing question is whether ignored effects could mitigate the radial drift problem. The existence of a headwind is the most crucial assumption, and it can vanish in localized pressure maxima as addressed below. Even when these maxima exist, headwinds still prevail in the majority of the disk. We also assume

that aerodynamic drag only affects the solids and not the gas component of the disk. When the distributed mass density of solids ρ_p becomes comparable to the gas density ρ_g, then it is no longer acceptable to ignore the feedback of drag forces on the gas. Nakagawa et al. (1986) showed how drift speeds become slower when $\rho_p \gtrsim \rho_g$. This feedback is also the source of powerful drag instabilities – both shearing and streaming – that we address below. Thus, there is no simple way to ignore the radial drift problem – its resolution has consequences for how planetesimals form.

4.1.2 Early Collisional Growth

Planetesimal formation begins with the collisional agglomeration of dust grains into larger solids, a process that is observed to proceed up to millimeter sizes in T Tauri disks (Williams and Cieza 2011). The conceptually simplest mechanism to form planetesimals is for this collisional growth to proceed past kilometer sizes. However, both direct experiment and theoretical arguments show that coagulation beyond millimeter sizes is inefficient at best. This inefficiency is particularly problematic due to the timescale constraints imposed by radial drift. The combination of inefficient sticking and rapid infall together comprises the formidable "meter-sized" barrier.

Although collision rates do not rule out rapid growth, they place tight constraints on the sticking efficiency To make this conclusion, we approximate the mean collision time as $t_{\rm coll} \sim 1/(\Omega\tau)$, where $\tau \sim \Sigma_p/(\rho_\bullet s)$ is the vertical optical depth. This approximation for the collision rate is good when $\tau_s \gg 1$, and we show below that it suffices for $\tau_s = 1$. The ratio of collision to drift timescales for $\tau_s = 1$ solids is thus roughly

$$\left.\frac{t_{\rm coll}}{t_{\rm drift}}\right|_{\tau_s=1} \sim \frac{\eta}{Z} \sim 0.3\left(\frac{R}{10\,{\rm AU}}\right)^{4/7}\left(\frac{0.01}{Z}\right), \tag{1.56}$$

where we assume Epstein drag, appropriate for the outer disk. We use the result of hydrostatic balance that $\Sigma_g \sim \rho_g c_s/\Omega$. When the collision time exceeds the drift time, collisional growth is ruled out. Even when the two are close, growth requires an efficient rate of sticking per collision. This constraint is most severe in the outer disk.

While turbulent motions increase collision speeds, they do not increase the collision rate above the geometric estimate used to derive (❯ 1.56). The reason is that turbulence also increases the particle layer thickness, H_p, thereby decreasing the mean particle density, ρ_p. We can compute the collision rate due to turbulence as $t_{\rm coll}^{-1} \sim n\sigma v$. Here the particle number density is $n \sim \Sigma_p/(H_p m_p)$, where m_p is the particle mass. The particle layer thickness due to turbulent stirring is[4]

$$H_p = H_\alpha = \sqrt{\frac{\alpha_D}{\tau_s}}H_g\,. \tag{1.57}$$

This well-known result (Cuzzi et al. 1993; Carballido et al. 2006; Youdin and Lithwick 2007) normalizes the turbulent diffusion, D, to the dimensionless parameter, $\alpha_D \equiv D/(c_s H_g)$. The cross section is $\sigma \sim s^2$, and the relative velocity due to turbulent motions is $v \sim \sqrt{\alpha\tau_s/(1+\tau_s^2)}c_s$ (Markiewicz et al. 1991; Chiang and Youdin 2010). For $\tau_s > 1$, the collision rate necessarily

[4]To accommodate the even mixing of small grains (not our current concern), we require $H_p \le H_g$.

agrees with the optical depth estimate. For $\tau_s < 1$, the collision rate is also independent of the strength of turbulence as $t_{coll}^{-1} \sim Z\Omega$. These cases agree at $\tau_s \sim 1$ and confirm the constraint set by (1.56).

Collision rates are not the only concern. Collisions can also result in bouncing or fragmentation that stalls, or even reverses, growth. Below speeds of ~1 m s^{-1}, small dust grains stick efficiently as a consequence of van der Waals interactions and the efficient dissipation of kinetic energy (Chokshi et al. 1993; Blum and Wurm 2000). As particles grow and as collision speeds increase, the collisional kinetic energy increases. Short-range sticking forces cannot match this increase in kinetic energy, because they are surface area limited (Youdin 2004). Experimental work confirms that collisions between equal-mass objects do not produce growth beyond ~ millimeter sizes (Blum and Wurm 2008).

Collisions between lower mass projectiles and higher mass targets offer another route to growth. In this scenario, impact speeds exceed the meters-per-second value expected to produce growth. As explained above, since small solids are tied to the gas flow, they impact larger solids (which decouple from the gas) at the full headwind speed, $\eta v_K \gtrsim 25$ m s^{-1}. Indeed when the latest experimental results are combined with dynamical estimates of collision speeds for a dispersion of particle sizes, growth stalls at only millimeter sizes (Zsom et al. 2010).

Based on observed SEDs, disks likely find a way to overcome these obstacles and grow solids beyond millimeter sizes (Williams and Cieza 2011). The mechanisms responsible for enhanced coagulation remain unclear. The particle concentration mechanisms discussed below could augment particle sticking. As shown explicitly in Johansen et al. (2009b), collision speeds are reduced in dense particle clumps.

Most experimental work on grain-grain collisions uses porous silicates. If ices are stickier, growth beyond millimeter sizes is possible. In the low pressure of disks, ices sublimate; there is no liquid available to make the equivalent of wet snow. Saturn's rings are an excellent laboratory to explore the outcomes of gentle, ~ mm s^{-1}, collisions between ices (Youdin and Shu 2002). Here, sticking forces are constrained by their inability to overcome tidal shear and produce growth beyond ~5-m objects. Terrestrial experiments on low-temperature ices suggest that centimeter-sized frosty objects stick at collision speeds below ~0.1 m s^{-1} (Supulver et al. 1997). While possibly a crucial ingredient, this limit appears insufficient to allow icy surfaces to bridge the meter-sized barrier.

4.2 Gravitational Collapse of Solids into Planetesimals

Self-gravity provides a qualitatively different route to the formation of planetesimals. Instead of bottom-up growth, the gravitational instability (GI) hypothesis of Safronov (1969) and Goldreich and Ward (1973) offers a top-down approach. In this theory, a sea of small solids collapses coherently into a gravitationally bound planetesimal. This collapse does not rely on sticking forces, proceeds faster than radial drift, and bypasses the meter-sized barrier.

The gravitational collapse hypothesis encounters several theoretical difficulties. The crucial issue is the ability of turbulent gas to prevent collapse (Weidenschilling 1995). Until recently, these theoretical obstacles seemed insurmountable. Progress in coupled particle-gas dynamics has led to a revival (Youdin and Shu 2002; Johansen et al. 2006, 2007; Cuzzi et al. 2008; Youdin 2011a). Some of these mechanisms use aerodynamic concentration as the initial concentration mechanism (Sect. 4.3), but all eventually rely on self-gravity for the final collapse to solid densities.

We focus in this subsection on "pure" gravitational collapse from a relatively smooth background. Although separating these processes from aerodynamic concentration is artificial, this historical approach allows us to isolate the main issues of each mechanism. We first discuss the standard model of gravitational collapse of a disk of solids, which has many similarities to gravitational instabilities in a gas disk (Sect. 5). We then briefly describe how gas drag changes this standard picture, a research area where progress is still being made.

The simplest criterion for gravitational collapse requires self-gravity to overcome the tidal distortion of the central star. This condition is met when the particle density exceeds the Roche limit,

$$\rho_\mathrm{p} > \rho_\mathrm{R} \simeq 0.6 \frac{M_*}{R^3} \simeq 130 \frac{\sqrt{m_*}}{F} \left(\frac{R}{\mathrm{AU}} \right)^{-3/14} \rho_\mathrm{g} \,, \tag{1.58}$$

where $m_* = M_*/M_\odot$ and F is the mass enhancement factor for the MMSN from Sect. 2. Sekiya (1983) derives this result for the case of a disk midplane with solids perfectly coupled to an incompressible gas, making use of the powerful formalism of Goldreich and Lynden-Bell (1965). The relation of this critical density to the Toomre (1964) Q criterion for GI is discussed (in the context of planetesimals) by Chiang and Youdin (2010) and Youdin (2011a).

For a given particle surface density, (1.58) implies that planetesimals form with a mass $M_\mathrm{ptml} \sim \Sigma_\mathrm{p}^3 / \rho_\mathrm{R}^2$. After contraction to solid densities, the planetesimal size is

$$R_\mathrm{ptml} \sim \frac{\Sigma_\mathrm{p}}{\rho_\bullet^{1/3} \rho_\mathrm{R}^{2/3}} \approx 5 \frac{F Z_\mathrm{rel}}{m_*} \sqrt{\frac{R}{\mathrm{AU}}} \, \mathrm{km} \,. \tag{1.59}$$

Though the current relevance is not so clear, this kind of estimate played a key role in defining the canonical planetesimal size to be near a kilometer.

To satisfy the density criterion of (1.58), solids must settle vertically to a midplane layer with thickness $H_\mathrm{R} = \Sigma_\mathrm{p}/\rho_\mathrm{R}$. Even the faintest whiff of turbulence probably produces a much thicker layer. Although the disk midplane could be a "dead zone" devoid of magnetized turbulence (Gammie 1996), interactions among particles can drive enough turbulence to halt settling (Youdin 2010).

Vertical shear instabilities usually prevent the sedimentation of small particles into a layer thinner than $H_\eta \simeq \eta R$ (Weidenschilling 1980; Youdin and Shu 2002). As particle inertia in the midplane increases from sedimentation, solids begin to drag the midplane gas toward the full Keplerian speed. As in the Kelvin-Helmholtz instability, the vertical shear with the overlying particle-poor gas drives overturning. With $H_\eta / H_\mathrm{R} \sim 200$, GI seems ruled out.

The revival of the GI hypothesis requires abandoning two faulty assumptions. The surface density of solids can increase above MMSN – or any initial – values. The evolution of solid and gas components decouples due to drift motions (Stepinski and Valageas 1996). The radial drift of small solids from the outer disk generically leads to "particle pileups," a snowplow effect that increases the surface density in the inner disk (Youdin and Shu 2002; Youdin and Chiang 2004). The local concentration mechanisms discussed in Sect. 4.3 can be even more powerful.

The critical Roche density is also too stringent. Planetesimal formation can be triggered when $\rho_\mathrm{p} \gtrsim \rho_\mathrm{g}$, a criterion about a 100 times less severe than the Roche limit in (1.58). Several interesting effects arise when the particle density approaches the gas density. Vertical shear instabilities lose their ability to overturn a layer that is so heavy (Sekiya 1998; Youdin and Shu 2002). When disk rotation is included, the case for particle inertia halting vertical overturning is less clear (Gómez and Ostriker 2005; Lee et al. 2010). However, when perfect

coupling is relaxed, and streaming instabilities appear, the relevance of $\rho_p \gtrsim \rho_g$ reemerges as the threshold for strong clumping, as described below.

When gas drag is present, the Roche density is not the relevant criteria for GI. Ward (1976, 2000) investigated a dissipative mode of GI that has no formal stability threshold. Collapse always occurs in principle, but it becomes slower and spans a larger radial extent for small particles. Youdin (2011a) included radial turbulent diffusion and showed that radial spreading – not vertical stirring – is the dominant stabilizing influence for dissipative GI. When vertical stirring is accounted for, it turns out that $\rho_p \gtrsim \rho_g$ is typically required for dissipative GI to proceed faster than radial drift. The result that dissipative GI depends so simply on particle inertia is mostly a numerical coincidence and relies on the fact that $Q_g\sqrt{2\eta} \sim 1$ in the MMSN, see Eq. 55 of Youdin (2011a). The important point is that the relevance of particle inertia – specifically the $\rho_p \gtrsim \rho_g$ criterion – has been established for a range of mechanisms.

Although this lesser degree of particle settling is substantial, it may require a local enrichment of the disk metallicity, $Z = \Sigma_p/\Sigma_g$. If the particle scale height is set by particle-driven turbulence to H_η, then the particle density exceeds the gas density if

$$Z > \frac{\eta R}{\sqrt{2\pi} H_g} \simeq 0.014 \frac{1}{\sqrt{m_*}} \left(\frac{R}{\mathrm{AU}}\right)^{2/7}, \tag{1.60}$$

again with $m_* = M_*/M_\odot$. The near agreement with solar abundances is remarkable and could be related to the correlation of giant planets with host star metallicity (Youdin and Shu 2002). Assuming that the stellar photospheres reflect the abundance of solids in the disk (Fischer and Valenti 2005), the early formation of planetesimals could be a crucial factor in the formation of gas giants (Johansen et al. 2009b).

Though poorly constrained, the role of external (not particle-driven) midplane turbulence may be interesting. For small solids, constraints on the level of turbulence that allows settling to $\rho_p \gtrsim \rho_g$ are quite stringent. Using (❯ 1.57), sedimentation to $\rho_p > \rho_g$ requires that midplane turbulence satisfy

$$\alpha_D \lesssim 2\pi Z^2 \tau_s \approx 10^{-4} Z_{\mathrm{rel}}^2 \frac{s}{\mathrm{cm}} \left(\frac{R}{10\ \mathrm{AU}}\right)^{3/2}. \tag{1.61}$$

Thus, when trying to form planetesimals via GI, it helps to have some combination of weak turbulence, particle growth, and enriched metallicity (Σ_p/Σ_g). These requirements become more stringent toward the inner disk (Youdin 2011a).

Thus, even in the GI hypothesis, particle growth by coagulation plays a crucial role. Particles must grow until they decouple from the gas. Provided this growth occurs, GI – likely aided by other concentration mechanisms – provides a plausible way past the meter-sized barrier.

4.3 Aerodynamic Particle Concentration

We now consider aerodynamic processes that can concentrate particles even when self-gravity is negligible. Many of these processes rely on the presence of turbulence in the disk. This connection raises a general question: does turbulence help or hinder planetesimal formation? By stirring particles, turbulence increases their collision speeds which can lead to more destructive collisions. Furthermore, the diffusive effects of turbulence oppose particle settling and concentration. On the other hand, turbulence can concentrate particles in a variety of ways. Which tendency wins depends on details, notably particle size. Since small solids with $\tau_s \ll 1$ drift and settle slowly, they require much weaker turbulence to participate in aerodynamic concentration.

Localized pressure maxima are very powerful particle traps. When the pressure bump takes the form of an axisymmetric ring, the trap is very effective (Whipple 1972). Solids migrate into these rings and accumulate at the center where they encounter no headwind. The MRI naturally produces axisymmetric pressure bumps, via the generation of zonal flows that are somewhat analogous to the surface winds of Jupiter (Johansen et al. 2009a; Fromang and Stone 2009). The relevance of MRI-induced pressure maxima is subject to two caveats: turbulent stirring associated with MRI may lead to destructive collisions, and the disk midplane may be insufficiently ionized for the MRI to operate (Turner et al. 2010).

Nonaxisymmetric pressure maxima can also trap particles. When the disk is young and massive, spiral arms in the gas probably provide an important source of turbulence (Rice et al. 2006). However, this phase of disk evolution may be too turbulent and/or brief for significant planetesimal formation. Isolated pressure maxima take the form of anticyclonic vortices (Chavanis 2000). Vortices are embedded in, and thus flow with, the sub-Keplerian gas (Youdin 2010). Although the vortex center is not a stable point for particle concentration, a point upstream (in the direction of orbital motion) is. The implications for vortex size are discussed in Youdin (2010). The formation and survival of vortices is a topic of ongoing research (Lithwick 2009).

We have so far focused on particle concentration over many orbits where disk rotation and Coriolis forces play a central role. Small turbulent eddies have short turnover times, $t_{\rm eddy} \ll 1/\Omega$, and are unaffected by rotation. In this regime, pressure maxima occur not at the centers of anticyclonic vortices but between vortices of either sign. The concentration of heavy particles in these regions of low enstrophy (vorticity squared) was first described in the fluid dynamics community (Maxey 1987).

Cuzzi et al. (2001) applied small-scale concentration to protoplanetary disks. They showed that ~1-mm solids – specifically the chondrules that are discussed in ❯ Sect. 4.4 – can concentrate at the "inner" or dissipation scale of turbulence. These are the smallest eddies that have the shortest turnover time $t_{\rm i}$. Particles with a matching stopping time, $t_{\rm s}, \sim t_{\rm i} \sim 30$ s, are preferentially flung from the eddies and concentrated between them. Chondrules can plausibly satisfy this condition. The ability to concentrate such small particles makes this mechanism unique. The relevance of such brief concentrations is unclear. The characteristic mass involved is also quite small, at most that of a 10-cm rock (Chiang and Youdin 2010).

To overcome these issues, Cuzzi et al. (2008) developed a model that concentrates chondrules on larger scales that contain enough mass to form ~100-km planetesimals. This model involves a somewhat speculative extrapolation. In particular, it assumes that all scales of a turbulent cascade contribute equally to the concentration of chondrule-sized particles. This assumption is a significant deviation from the original mechanism that requires eddy and stopping times to match. See Chiang and Youdin (2010) for further discussion, which concludes that more study of this intriguing mechanism is required.

Clearly particle concentration mechanisms are fraught with uncertainties in the detailed dynamical behavior of gas in protoplanetary disks. Some – certainly not all – of these uncertainties are overcome by the realization that particles can cause their own concentration by collectively altering the gas dynamics (Goodman and Pindor 2000). In the streaming instability of Youdin and Goodman (2005), particle concentrations arise spontaneously from radial drift motions. As described in ❯ Sect. 4.1.1, these drift motions are an inevitable consequence of pressure support in disks. The linear growth of streaming instabilities is strongest for $\rho_{\rm p} > \rho_{\rm g}$, because particle inertia must be large for drag feedback to influence gas motions. When $\rho_{\rm p} < \rho_{\rm g}$, growth is fastest for $\tau_{\rm s} \approx 1$, when drift speeds are fastest (Youdin and Johansen 2007). While streaming instabilities involve complex dynamics – 3D motions of both the gas and solid

components – simplified toy models (Goodman and Pindor 2000; Chiang and Youdin 2010) and considerations of geostrophic balance (Jacquet et al. 2011) help explain how particle density perturbations self-reinforce.

Numerical simulations show that the nonlinear clumping from the streaming instability can be quite strong (Johansen and Youdin 2007; Johansen et al. 2009b; Balsara et al. 2009; Bai and Stone 2010a). Particle densities $\gtrsim 10^3 \rho_g$ are achieved in the absence of self-gravity, and clumping tends to increase with numerical resolution. The conditions for strong clumping are similar to those giving rapid linear growth: partial decoupling, $\tau_s \gtrsim 0.1$, and large particle inertia $\rho_p \gtrsim 0.2\rho_g$.

When vertical stratification is included, the midplane particle density evolves consistently due to settling and stirring by both streaming and vertical shearing instabilities. In these simulations, there is a critical disk metallicity for particle clumping which is slightly supersolar (Johansen et al. 2009b; Bai and Stone 2010a), consistent with (❯ 1.60). This metallicity varies with the radial pressure gradient, η; smaller gradients promote clumping (Johansen et al. 2007; Bai and Stone 2010b).

In ❯ Sect. 4.2, we noted that GI depends on particle growth by coagulation. Since particle sedimentation to $\rho_p \gtrsim \rho_g$ is a crucial prerequisite, particle growth remains essential when the streaming instability provides the initial particle concentration. However, growth need not result in a single particle size or a very narrow size distribution. Though the smallest solids participate less in clumping by streaming instabilities, including a dispersion in particle sizes does not prevent strong clumping (Johansen et al. 2007).

The particle concentration produced by the streaming instability is more than sufficient to trigger gravitational collapse. Johansen et al. (2007) formed gravitationally bound objects equivalent to ~500-km planetesimals within only a few orbits of initial collapse. More recent simulations suggest the formation of lower mass objects with equivalent ~100–200-km radii (Johansen et al. 2009b). The crucial differences are the inclusion of the MRI in the earlier study and smaller particle sizes in the second. The inclusion of particle collisions also has a modest effect on the resulting planetesimal masses (Johansen et al. 2012). A more thorough investigation of parameter space, combined with resolution studies, is required. The massive planetesimals produced in these simulations exceed the standard estimate of kilometer-sized planetesimals because gravitational collapse occurs not from a smooth background but from aerodynamically concentrated clumps.

4.4 Observational Constraints on Planetesimal Formation

We now discuss how observations constrain dynamical theories of planetesimal formation. The solar system provides the most detailed information on planetesimals and allows comparison between the inner asteroidal reservoir and the Kuiper belt objects and comets of the outer solar system. The crucial issue is to what extent today's planetesimals reveal the clues of their formation, especially after ~4 Gyr of dynamical, collisional, and thermal evolution.

Primitive, undifferentiated meteorites give us a hands-on view of the composition of planetesimals. The most common of these are the aptly named "ordinary chondrites." With filling factors up to 90%, they are primarily composed of 0.1–1-mm chondrules. Chondrules are glassy spheres, poetically referred to as "fiery drops of rain" (Sorby 1863). The origin of chondrules – in particular their source of heating – is debated and beyond our scope, see Hewins (1996). The prevalence of chondrules in ordinary chondrites strongly motivates further investigation of the mechanisms that could concentrate solids this small (Cuzzi et al. 2001, 2008).

Despite this attractive conclusion, chondrules may not be the universal building blocks of all planetesimals. Because their abundances most closely match solar, the CI class of chondrites is considered the most primitive (Lodders 2003). Yet CI chondrites contain no chondrules. It is also unclear whether chondrules were present in the first generation of planetesimals. Most chondrules formed at least 1.5 Myr and up to 4 Myr after the rarer CAIs (calcuim-aluminum inclusions; Connelly et al. 2007; Krot et al. 2007). Thus, planetesimals probably formed before major chondrule forming events, especially the planetesimals that formed the cores of Jupiter and Saturn. Since planetesimals that form early will trap more radioactive heat and differentiate, it seems likely that the undifferentiated chondrites represent a later phase of planetesimal formation (Kleine et al. 2005). The relation between chondrules, meteorites, and planetesimal formation continues to be the focus of intense interdisciplinary research.

Planetesimals that remain in the asteroid belt can also provide clues to their formation. The radially banded zonation of different spectral classes of asteroids is well known (Gradie and Tedesco 1982). This observation suggests separate formation epochs, with each event creating a "clan" of chemically and spectrally similar planetesimals. Youdin (2011a) proposed large-scale, drag-mediated GI as the cause of these events.

The size distribution of objects within planetesimal belts provides other clues to their formation. Breaks in the size distribution – that is, changes in its power-law slope – point to shifts in formation and/or erosion processes. The asteroid belt has a break near a radius ~50 km. Morbidelli et al. (2009a) argue that the asteroids with radii $\gtrsim$50 km reflect their initial sizes. Specifically they assert that the largest asteroids have undergone minimal collisional evolution and could not have formed via collisional growth of smaller planetesimals. Since most of the mass is contained in the largest asteroids, their model plausibly produces the numerous small objects below the break via collisional disruption. That interpretation places GI as the preferred formation mechanism. By including streaming instabilities, the simulations of Johansen et al. (2007) predicted that large initial sizes were possible. Conclusively proving that a size distribution is unobtainable by collisional growth is rather difficult. Weidenschilling (2010) contends that collisional growth of asteroids can be accomplished starting with 0.1-km planetesimals – which themselves presumably grew by coagulation past the meter-sized barrier.

Curiously, the Kuiper belt also has break in its size distribution at ~50-km radii (Bernstein et al. 2004). This break is not measured directly; a combination of an observed luminosity distribution and an estimate of the albedo yields the distribution of radii (Petit et al. 2008). Ongoing surveys of the Kuiper belt seek to provide constraints on the size distribution for the various components of the Kuiper belt.

Understanding the origin of the break requires a model for KBO formation and dynamical interactions with gas giants. Reproducing the observed size distribution with collisional growth models requires an initially massive Kuiper belt followed by dynamical depletion; a break occurs when depletion excites erosive collisions among KBOs with radii below the break (Kenyon and Bromley 2004c). The break radius depends on excitation; more (less) excitation by more (less) massive gas giants yields a break at larger (smaller) radii. Matching the location of the break and the apparent slope of the KBO size distribution below the break requires numerical calculations with growth and depletion, which are an active area of research (Kenyon et al. 2008a; Morbidelli et al. 2008).

The relative roles of collisional and dynamical depletion affect the interpretation of the KBO size break. Pan and Sari (2005) argue that the break is not primordial but due to ongoing collisional erosion that continues to push the break to larger sizes. However, Nesvorný et al. (2011) claim that this collisional history is ruled out on two grounds. First, the collisional strengths

required for such destruction are too weak. Second, such an intense collisional bombardment would destroy the observed Kuiper belt binaries.

The observed binary fraction in the cold, classical Kuiper belt is ≳20% (Noll et al. 2008). The colors of the two components of Kuiper belt binaries are nearly identical, a fact interpreted as representing a common chemical composition (Benecchi et al. 2009). This observation provides the most compelling support for the GI hypothesis in the outer solar system or perhaps anywhere. Gravitational collapse can naturally produce binary planetesimals as a consequence of angular momentum conservation during the contraction of a swarm of small solids (Nesvorný et al. 2010). Binaries – and higher-order multiples – formed this way should have the same chemical composition since they formed from the same well-mixed clump of small solids. While mechanisms for the dynamical capture of KBO binaries are well developed (Goldreich et al. 2002; Noll et al. 2008), these models do not obviously explain matching colors. Moreover, the physical conditions require for capture, making the collisional survival of these binaries questionable (Nesvorný et al. 2011), especially for wide binaries (Parker and Kavelaars 2012).

Outside the solar system, exoplanets and debris disks inform the prevalence and consequences of planetesimal formation. The higher incidence of giant planets around stars with supersolar metallicities (discussed in ❯ Sect. 2.3) might be tied to planetesimal formation. As shown in (❯ 1.60), this connection is strongly suggested by the supersolar *disk* metallicity threshold for strong clumping by streaming instabilities. Since the disk metallicity can increase over time (Youdin and Chiang 2004), this threshold does not imply that lower metallicity stars can never form planetesimals.

Indeed, the streaming instability/GI model explains why lower metallicity and lower mass stars should form less massive planets (Johansen et al. 2009b). Either directly or by the passage of time, enriching the disk metallicity involves the loss of gas. Thus, the initially low metallicity systems that require enrichment are less likely to form giant planets. This conclusion is especially true in the lower mass disks thought to surround lower mass stars. These general trends are revealed by radial velocity surveys (Sousa et al. 2008; Johnson et al. 2010). The *Kepler* transit survey will test these trends, since it is finding striking numbers of small, short-period planets (Howard et al. 2012; Youdin 2011b). Characterization of the *Kepler* stars will thus powerfully constrain planetesimal formation models.

5 Planetesimals to Planets

Once planetesimals become larger than a few kilometers – potentially they are born much larger as discussed above – gravitationally focused collisions dominate growth into protoplanets. The size when "planetesimals" become "protoplanets" is vague. Although we use the terms interchangeably, ~1,000 km is a useful threshold. Depending on location and gas temperature, ~1,000-km protoplanets are the smallest planets capable of binding disk gas into an atmosphere.

We describe the accretion of solid protoplanets in ❯ Sect. 5.1. We start by discussing the processes that operate in a gas-free disk, including gravitationally focused collisions (❯ Sect. 5.1.2), velocity excitation (❯ Sect. 5.1.3), and collisional fragmentation (❯ Sect. 5.1.4). We then describe planetesimal interactions with the gaseous disk (❯ Sect. 5.1.5). ❯ Section 5.1.6 describes simulations of terrestrial planet formation that put these ingredients together. The accretion of a gaseous atmosphere (❯ Sect. 5.2) affects planetesimal accretion

(Sect. 5.2.2) and transforms a planetary core into a gas giant (Sects. 5.2.3 and 5.2.4). We discuss numerical simulations combining the growth of giant planet cores and atmospheres in Sect. 5.2.5. Finally, a young, massive gas disk might fragment directly into a gas giant or a brown dwarf. Section 5.3 describes this formation channel and whether it might explain some exoplanets, especially massive giants at large radial distances.

5.1 Growth of Solid Protoplanets

Unlike planetesimal formation, it is easy to understand why planetesimals grow into larger protoplanets, even if the details are complicated. For the largest planetesimal in any region of the disk, collisions essentially always result in growth. Planetesimal velocities cannot be locally excited above the escape speed of the largest protoplanet. Consequently, the kinetic energy of collisions does not exceed the gravitational potential at the surface of the largest protoplanet. Collisions dissipate a fraction, sometime quite large, of the impact kinetic energy. Even if the impacting planetesimal shatters, growth is assured.

Collisions among smaller planetesimals, however, often lead to erosion or catastrophic fragmentation. When an external, massive perturber stirs a belt of planetesimals, planetesimals collide at velocities larger than their escape velocity. These high-velocity collisions tend to erode or completely shatter planetesimals. The dust and changes in the planetesimal size distribution that result from these collisions are relevant for debris disks and for asteroids and Kuiper belt objects.

Unlike the planetesimal formation phase, aerodynamic drag no longer plays a starring role in protoplanet growth. However, drag can still help regulate planetesimal velocities. The accretion of atmospheres (see Sect. 5.2) also affects planetesimal capture.

Deriving the precise evolution of a swarm of planetesimals is a complex numerical problem being attacked from several angles (Sect. 5.1.6). However, we can develop a reasonably accurate picture of the evolution with the "two groups approximation" reviewed in greater detail by Goldreich et al. (2004). This approximation considers interactions between small, low-mass planetesimals with mass m_s and larger, more massive planetesimals with mass m_l. The planetesimal masses $m_{s,l} = 4\pi r_{s,l}^3 \rho_\bullet / 3$ are related to their radii $r_{s,l}$ and internal mass density, which we fix at $\rho_\bullet = 2\,\mathrm{g\,cm^{-3}}$ unless stated otherwise. Neighboring planetesimals have similar semimajor axes, a, and orbital frequencies, Ω. Though detailed treatments need not make this approximation, we equate orbital eccentricities $e_{s,l}$ and inclinations (in radians) but allow e_s and e_l to differ. The random velocities relative to a circular orbit are thus $v_{s,l} \approx e_{s,l}\,\Omega a$, and the vertical scale heights of the planetesimal disks are $H_{s,l} \approx v_{s,l}/\Omega$. When the nature of the planetesimal is unspecified, we drop the s and l subscripts.

5.1.1 Basic Length and Velocity Scales

A useful scale for studying interactions of planetesimals and protoplanets is the Hill (1878) radius

$$R_H = \left(\frac{m}{3M_*}\right)^{1/3} a\,. \tag{1.62}$$

Planetesimals separated by $\lesssim R_H$ are within the Hill sphere where their mutual gravitational attraction dominates the tidal gravity from the central star. While a mutual Hill radius can be defined, in practice it suffices to consider the more massive planetesimal.

The size of a planetesimal in Hill units defines the parameter

$$\psi \equiv \frac{r}{R_H} = \left(\frac{3\rho_*}{\rho_\bullet}\right)^{1/3} \frac{R_*}{a} \simeq 6 \times 10^{-3} \left(\frac{\mathrm{M}_*}{\mathrm{M}_\odot}\right)^{1/3} \left(\frac{\mathrm{AU}}{a}\right) , \tag{1.63}$$

where ρ_* is the mean mass density of the central star. Since $3\rho_*/\rho_\bullet \sim 1$, the parameter $\psi \sim R_*/a$ is roughly the angular size of the central star as observed from the planetesimal. The smallness of ψ represents the fact that physical collisions are rare compared to gravitational scattering.

At the Hill radius, the orbital speed about the protoplanet is the Hill velocity

$$v_H = \Omega R_H = \left(\frac{m}{3M_*}\right)^{1/3} v_K . \tag{1.64}$$

When planetesimal random velocities exceed v_H, two-body encounters are "dispersion dominated," negligibly affected by orbital shear. Random speeds below v_H cause "shear-dominated" encounters that involve (restricted) three-body dynamics. We describe below how the Hill velocity divides different accretion regimes.

The outcome of a shear-dominated encounter between planetesimals depends on the difference in semimajor axes δR, relative to the Hill radius (Petit and Henon 1986). When $\delta R \lesssim 1$–$2\, R_H$, the two planetesimals are deflected on a horseshoe orbit. More distant encounters with $\delta R \gtrsim 2\sqrt{3} R_H$ result in small-angle scattering. For intermediate separations, 1–$2\, R_H \lesssim \delta r \lesssim 2\sqrt{3} R_H$, planetesimals enter the Hill sphere, experience chaotic deflections, and (if no collision occurs) leave the Hill sphere with typical relative velocity v_H.

In Hill units, the escape speed from the surface of a protoplanet, $v_{\rm esc} = [2Gm/r]^{1/2}$, is $v_{\rm esc} \sim v_H/\psi^{1/2}$. Planetesimal velocities can be gravitationally excited up to the escape speed of the largest protoplanet. To estimate when a massive protoplanet might eject nearby planetesimals (or protoplanets), we compare the escape velocity of the protoplanet to the orbital escape velocity, $v_{\rm esc,*} \approx \sqrt{2} v_K$. When

$$\frac{v_{\rm esc}}{v_{\rm esc,*}} \approx 0.15 \left(\frac{m}{\mathrm{M}_\oplus}\right)^{1/3} \left(\frac{a}{\mathrm{AU}}\right)^{1/2} \left(\frac{\mathrm{R}_*}{\mathrm{R}_\odot}\right)^{-1/2} \tag{1.65}$$

exceeds unity, a planet of mass m can eject other nearby protoplanets. Terrestrial planets like Earth are too low mass and too close to the Sun to eject objects. The four solar system giants can all eject planetesimals; Jupiter is the most efficient at ejecting comets (and spacecraft) from the solar system (Fernandez and Ip 1984). Aside from collisional grinding, (❯ 1.65) implies that planetesimal formation is more efficient closer to a star.

The concepts of the Hill sphere and the Roche lobe are identical, though often used in different contexts. Both describe the region where the gravity of an object exceeds the tidal perturbation from its companion. Formally, both volumes are defined by the critical equipotential containing the L_1 and L_2 Lagrange points. The Roche lobe is more distorted than a sphere when it describes binary stars that are similar in mass. The Roche radius or Roche limit describes the distance from a primary object at which the secondary becomes tidally disrupted and might form planetary rings. Aside from order unity corrections due to fluid effects or internal strength, the concept of individual versus tidal gravity is again identical. To summarize, a secondary is at the Roche limit from the primary when it fills its own Hill sphere (or Roche Lobe). For an

ensemble of very small planetesimals trying to become a much larger planetesimal, the Roche limit sets the critical density for gravitational collapse (1.58).

5.1.2 Gravitationally Focused Collisions

In this section, we describe the growth rates of large protoplanets (subscripted by l) accreting either large protoplanets or smaller planetesimals (unsubscripted). Gravitational focusing is the most important aspect of growth. Smaller random velocities for accreted planetesimals yield larger gravitational focusing factors and shorter growth times. We defer to later sections the self-consistent calculation of planetesimal velocities and assume the standard case, $v_l < v_s$. Greenberg et al. (1984), Wetherill and Stewart (1993), Kenyon and Bromley (2008), and references in each paper describe more detailed expressions for growth rates.

We begin with the dispersion-dominated regime where $v > v_{H,l}$. The mass accretion rate results from the usual isotropic expression as $\dot{m}_l = mn\sigma v$. Adopting the surface mass density of planetesimals, Σ, the number density of planetesimals, n, is

$$mnv = \Sigma\Omega \,. \tag{1.66}$$

The cross section

$$\sigma = \pi(r_l + r)^2 f_{G,\mathrm{disp}} \tag{1.67}$$

is the product of the geometric area and the gravitational focusing factor $f_{G,\mathrm{disp}}$. If the velocity of the incoming planetesimal at infinity is $v > v_{\mathrm{esc},l}$, there is no gravitational focusing and $f_{G,\mathrm{disp}} = 1$. When $v \ll v_{\mathrm{esc},l}$, the speed on impact is roughly $v_{\mathrm{esc},l}$. Angular momentum conservation during a two-body encounter sets the impact parameter for a grazing collision as $r_l v_{\mathrm{esc},l}/v$. This expression yields $f_{G,\mathrm{disp}} \approx (v_{\mathrm{esc},l}/v)^2$. Including energy conservation gives both cases simultaneously as

$$f_{G,\mathrm{disp}} = 1 + \beta(v_{\mathrm{esc},l}/v)^2 \,, \tag{1.68}$$

where $\beta = 1$ for a pure two-body interaction and $\beta \approx 2.7$ accounts for anisotropic effects introduced by orbital dynamics (Greenzweig and Lissauer 1990; Spaute et al. 1991; Wetherill and Stewart 1993).

Putting these results together and ignoring order unity coefficients give the dispersion-dominated growth timescale $m_l/\dot{m}_l$ as

$$t_{\mathrm{disp}} \approx \frac{\rho_\bullet r_l}{\Sigma\Omega f_{G,\mathrm{disp}}} \,. \tag{1.69}$$

This result is just the geometric collision time over the focusing factor.

For shear-dominated encounters with $v < v_{H,l}$, collision rates are affected by chaotic trajectories inside the Hill sphere (Greenberg et al. 1991; Dones and Tremaine 1993). In this regime, planetesimal disks are thinner than the Hill radius, $H \sim v/\Omega < R_{H,l}$. Thus, planetesimals enter the Hill sphere at the 2D mass accretion rate, $\dot{m}_H \sim \Sigma R_{H,l}^2 \Omega$. The probability, P, of a collision within the Hill sphere has two cases. Both use the maximum impact parameter for gravitationally focused collisions $b_{\max} \sim r_l v_{\mathrm{esc},l}/v_{H,l} \sim \psi^{1/2} R_{H,l}$. If the scale height of the disk is (relatively) thick with $H \sim v/\Omega > b_{\max}$, the collision probability $P \sim b_{\max}^2/(R_{H,l} H)$ is the ratio of the collision cross section to the area of the accreting disk of planetesimals. For a thinner planetesimal disk, $P \sim b_{\max}/R_{H,l}$ is the ratio of the impact parameter to the Hill radius.

Combining the mass flow rate through the Hill sphere with both limits of the collision probability yields the shear-dominated growth timescale, $m_l/(P \cdot \dot{m}_H)$, as

$$t_{\rm shear} \sim \frac{\rho_\bullet r_l}{\Sigma \Omega f_{G,\rm shear}} \,. \tag{1.70}$$

This timescale is again expressed as the product of the geometric collision time and a gravitational focusing factor

$$f_{G,\rm shear} \sim \left(\psi \frac{v}{v_{H,l}} + \psi^{3/2} \right)^{-1} \,. \tag{1.71}$$

Thus, for $v < \psi^{1/2} v_{H,l}$, gravitational focusing reaches its maximum value of $f_G \sim \psi^{-3/2}$, resulting in the fastest possible growth rate. Because inclination excitation is weak in the shear-dominated regime, this fastest thin-disk accretion rate likely applies for all large protoplanets with $v_l < v_{H,l}$ (Goldreich et al. 2004). Aside from this issue of anisotropic velocities, the dispersion and shear-dominated focusing factors match at $v \sim v_{H,l}$ with $f_G \sim 1/\psi$.

For numerical estimates of growth timescales, we consider three cases. For the slow case, we take $v > v_{\rm esc,l}$ and no gravitational focusing. For the intermediate case, we identify $v \approx v_{H,l} \approx \psi^{1/2} v_{\rm esc,l}$ as the transition between shear and dispersion, dominated. The fast case considers the maximum focusing factor $f_G \approx \psi^{-3/2}$ appropriate for $v \lesssim \psi^{1/2} v_{H,l} \approx \psi v_{\rm esc,l}$ (and possibly for higher speeds when planetesimal $i \ll e$). Using (❯ 1.2) for the surface density of planetesimals, the growth times become

$$t_{\rm slow} \approx \frac{\rho_\bullet r_l}{\Sigma \Omega} \approx 10^7 \left(\frac{m_l}{\rm M_\oplus} \right)^{1/3} \left(\frac{1}{F Z_{\rm rel}} \right) \left(\frac{a}{\rm AU} \right)^3 \text{year} \,. \tag{1.72a}$$

$$t_{\rm int} \approx \frac{\rho_\bullet r_l}{\Sigma \Omega} \psi \approx 5 \times 10^4 \left(\frac{m_l}{\rm M_\oplus} \right)^{1/3} \left(\frac{1}{F Z_{\rm rel}} \right) \left(\frac{a}{\rm AU} \right)^2 \text{year} \,. \tag{1.72b}$$

$$t_{\rm fast} \approx \frac{\rho_\bullet r_l}{\Sigma \Omega} \psi^{3/2} \approx 4{,}000 \left(\frac{m_l}{\rm M_\oplus} \right)^{1/3} \left(\frac{1}{F Z_{\rm rel}} \right) \left(\frac{a}{\rm AU} \right)^{3/2} \text{year} \,. \tag{1.72c}$$

These mass doubling times increase with protoplanet mass. Thus, if gravitational focusing stays fixed or decreases (far from a certainty), these estimates also give the total accumulation time. These expressions omit the dependence on stellar mass and planet density for clarity but the growth times scale $\propto \rho_\bullet^{2/3}/{\rm M_*}^{1/2}$, $\rho_\bullet^{1/3}/{\rm M_*}^{1/6}$, and $\rho_\bullet^{1/6}{\rm M_*}^0$ for the three cases, respectively. Though stellar mass is not a major dynamical effect, it could correlate with disk mass or metallicity (here meaning planetesimal to gas ratio), normalized above by F and $Z_{\rm rel}$, respectively. Higher density protoplanets have smaller cross sections and grow more slowly, but this effect becomes much less significant as gravitational focusing increases.

Gravitational focusing dramatically speeds up the growth of protoplanets, especially in the outer disk. Without focusing, planets accumulate in $t_{\rm slow} \sim$ tens of Myr inside a few AU and more than 1 Gyr outside 5 AU. While a long growth time for terrestrial planets is acceptable, gas giants must form within a few Myr. Thus, formation of giant planet cores in the outer disk requires strong focusing, when the growth time for a 10 $\rm M_\oplus$ core at 50 AU, $t_{\rm fast} \sim 3$ Myr, is close to the lifetime of most gas disks. If strong focusing occurs, the formation of distant gas giants depends on the ability of cores to accrete gas (Rafikov 2011).

Protoplanet accretion also depends on how the velocity distribution evolves. In dispersion-dominated gravitational focusing, the growth time $t_{\rm disp} \propto 1/r_l$; larger planetesimals grow faster than smaller ones. Although it is not the fastest regime, this "runaway" growth requires

that gravitational focusing factors remain in the dispersion-dominated regime. With $v_{esc}/v_H \sim \psi^{-1/2} \propto a^{1/2}$, runaway growth persists longer in the outer disk (Greenzweig and Lissauer 1990).

When the largest protoplanets enter the shear-dominated regime, runaway growth ends. For the thick disk case, $t_{shear} \propto r_l^0$ is independent of size. With either no focusing, $t_{slow} \propto r_l$, or the fastest (thin-disk) shear-dominated accretion $t_{fast} \propto r_l$, smaller protoplanets grow faster and can catch up to the larger ones. In this "oligarchic" growth, many oligarchs compete to accrete small planetesimals, leading to an ensemble of oligarchs throughout the disk.

As the largest protoplanets grow, they try to accrete all solid material in their vicinity. Two planets on circular orbits separated by a little more than $2\sqrt{3}\ R_H$ are stable (Gladman 1993). However, a fairly stable system with more planets requires larger separations, $\sim B R_H$ with $B = 7$–10 (Lissauer 1987; Kokubo and Ida 1998). Planets that accrete all material within $B R_H$ are "isolated." Setting $m_{iso} = 2\pi\Sigma a B R_H$ leads to the isolation mass,

$$m_{iso} = \frac{(2\pi B\Sigma)^{3/2}}{(3M_*)^{1/2}} a^3 \approx 0.08 \left(\frac{B}{7}\right)^{3/2} \left(\frac{F Z_{rel}}{0.33}\right)^{3/2} \left(\frac{M_*}{M_\odot}\right)^{-1/2} \left(\frac{a}{AU}\right)^{3/4} M_\oplus \ . \qquad (1.73)$$

With $F = 1$ and $Z_{rel} = 0.33$, isolated objects in the terrestrial zone have masses comparable to Mercury and Mars. The MMSN has room for 30–50 isolated objects between the orbits of Mercury and Mars. Because their escape velocities are much smaller than their orbital velocities (1.65), isolated protoplanets eventually collide and merge to form Earth-mass planets (*Fig. 1-8*).

Outside the snow line, $Z_{rel} = 0.78$ at 5 AU yields an isolation mass of roughly 1 $M_\oplus$. As we show later, this mass is too small to bind the gas required for a gas giant. Increasing the mass of the MMSN ($F \approx 5$) increases the isolation mass to the "typical" core mass of 10 $M_\oplus$ needed for a massive atmosphere. Thus, the MMSN is fine for the terrestrial planets, but it is not massive enough to allow formation of gas giants similar to Jupiter and Saturn. The extra mass required is consistent with observations of disks around the youngest stars (Sect. 2).

5.1.3 Planetesimal Velocity Evolution

As the previous section makes clear, the evolution of planetesimal velocities establishes the rate protoplanets accrete smaller planetesimals. Gravitational scattering is more common than physical collisions; thus, planetesimal velocities rapidly adjust as large protoplanets grow.

Several processes modify the random velocities of planetesimals. The source of random kinetic energy is known as viscous stirring. This process uses planetesimal encounters – predominantly gravitational scattering – to extract energy from orbital shear. Dynamical friction redistributes kinetic energy among planetesimals of different masses, pushing them toward equipartition. Thus, smaller (larger) planetesimals damp (excite) the random velocities of the larger (smaller) planetesimals (Wetherill and Stewart 1989; Kokubo and Ida 1995; Kenyon and Luu 1998). Ignoring ejections and gas drag, physical collisions are the only source of kinetic energy damping. Collisional damping is especially effective for small planetesimals, $r \lesssim 1$–100 m, that collide frequently (Ohtsuki 1992; Kenyon and Luu 1998). When collisions produce small fragments that collide even more frequently, damping is very efficient. Goldreich et al. (2004) discuss order-of-magnitude derivations of these processes. As with accretion, behaviors vary between the dispersion- and shear-dominated regimes. It is common to refer to the excitation and damping of planetesimal velocities as "heating" and "cooling," respectively.

The main goal of this introduction to velocity evolution is to show that planetesimals cannot be heated above – and can sometimes be cooled significantly below – the escape velocity of the large protoplanets. We focus on dispersion-dominated encounters to explain this result, which is crucial for ensuring the gravitationally focused collisions required to make planets on reasonable timescales.

We first consider the simple case where all planetesimals have the same size. When $v < v_{\rm esc}$, viscous stirring is dominated by gravitational scattering and occurs on the scattering timescale. This heating timescale is well approximated by the two-body relaxation time from stellar dynamics (Binney and Tremaine 2008). For the $n\sigma v$ estimate of the gravitational scattering rate, we use (1.66) and compute the cross section $\sigma \sim b_{\rm scatt}^2$ from the impact parameter for strong gravitational scattering, $b_{\rm scatt} \sim Gm/v^2$. Together, these give the viscous stirring timescale (Ida and Makino 1993)

$$t_{\rm stir,disp} \simeq C_1 \frac{v^4}{G^2 m \Sigma \Omega} \sim \frac{\rho_\bullet r}{\Sigma \Omega} \left(\frac{v}{v_{\rm esc}} \right)^4 , \tag{1.74}$$

where the constant $C_1 \approx 1/40$ arises from a more detailed analysis (Ohtsuki et al. 2002) and is similar to the Coulomb logarithm, $\ln \Lambda$, in stellar dynamics (and plasma physics). The final approximate expression in (1.74) facilitates comparison with the collision rates.

The cooling rate is the gravitationally focused collision rate, which follows from (1.69) as

$$t_{\rm cool,disp} \sim \frac{\rho_\bullet r}{\Sigma \Omega} \left(\frac{v}{v_{\rm esc}} \right)^2 . \tag{1.75}$$

Balancing the stirring and cooling rates implies $v \sim v_{\rm esc}$. While the correct answer, the reasoning is incomplete. Gravitational focusing is weak for $v \gtrsim v_{\rm esc}$. In this regime, stirring and collisional cooling rates are comparable. A slight imbalance in favor of heating could lead to runaway growth of v and an eventual collisional cascade. This runaway requires nearly elastic physical collisions, as in the collision of two basketballs. In the idealized model of Goldreich and Tremaine (1978), collisions among planetesimals with coefficients of restitution $\gtrsim$0.63 (comparable to a baseball, but smaller than a basketball or a table tennis ball) bounce often enough to lead to runaway heating. Coefficients of restitution for planetesimals are probably much smaller than 0.5 (Porco et al. 2008); the velocity runaway is unlikely. Similarly sized planetesimals will excite random velocities to the surface escape speed, $v \sim v_{\rm esc}$.

Returning to the two groups approximation, we consider stirring of small planetesimals by large protoplanets. Dynamical friction ensures $v_s > v_l$ (confirmed below); planetesimals dominate the encounter speed. The stirring of small planetesimals by larger ones then occurs on a timescale

$$t_{\rm stir,disp} \sim \frac{\rho_\bullet r_l}{\Sigma_l \Omega} \left(\frac{v_s}{v_{{\rm esc},l}} \right)^4 . \tag{1.76}$$

Comparison with (1.74) shows that large planetesimals dominate the stirring of small planetesimals when $\Sigma_l m_l > \Sigma_s m_s$. Initially, $\Sigma_s > \Sigma_l$; small planetesimals contain enough mass to affect growth. To dominate stirring, however, large planetesimals can contain a minority of the mass.

Due to stronger stirring by large protoplanets, small planetesimals are excited to $v_s > v_{{\rm esc},s}$. At these speeds, collisions between small planetesimals generally cause collisional fragmentation or erosion. The resulting smaller planetesimals then collisionally cool more efficiently. Even without this extra cooling, gravitational focusing arises. When $v_s > v_{{\rm esc},s}$, small planetesimals

cool by colliding with other small planetesimals on the geometric timescale, $t_{\rm cool} \sim \rho_\bullet r_s/(\Sigma_s \Omega)$. Balancing these heating and cooling rates gives

$$v_s \sim \left(\frac{\Sigma_l}{\Sigma_s}\frac{r_s}{r_l}\right)^{1/4} v_{{\rm esc},l} \ . \tag{1.77}$$

When $\Sigma_s > \Sigma_l$, $v_s \ll v_{{\rm esc},l}$; small planetesimal accretion is strongly gravitationally focused.

We now consider whether the growth of large protoplanets is dominated by the accretion of small planetesimals or other large protoplanets. Planetesimals with the larger product of surface density and gravitational focusing, Σf_G, drive the fastest growth (1.69)–(1.70). A balance of viscous self-stirring and cooling by dynamical friction against small planetesimals then sets the velocity dispersion of large protoplanets. For dispersion-dominated encounters, this balance gives (for details, see Goldreich et al. 2004)

$$\frac{v_l}{v_s} \sim \left(\frac{\Sigma_l}{\Sigma_s}\right)^{1/4} \ . \tag{1.78}$$

Since $\Sigma f_{G,{\rm disp}} \propto \Sigma/v^2$, small planetesimals contribute more to the growth of large protoplanets, by a factor $(\Sigma_s/\Sigma_l)^{1/2} > 1$.

This introduction only begins to touch on the complexities of planetesimal velocity evolution. However, even these simple considerations show that gravitationally focused accretion of small planetesimals by large protoplanets is likely. Earlier, we explained that collisional erosion plays a key role in cooling small planetesimals to $v_s > v_{{\rm esc},s}$. Now, we turn to even more violent encounters, with $v_s \gg v_{{\rm esc},s}$, which lead to catastrophic disruption.

5.1.4 Fragmentation

As large planetesimals grow, they stir up the velocities of smaller planetesimals to the disruption velocity. Instead of mergers, collisions then yield smaller planetesimals and debris. Continued disruptive collisions lead to a collisional cascade, where leftover planetesimals are slowly ground to dust (Dohnanyi 1969; Williams and Wetherill 1994; Kobayashi and Tanaka 2010). Radiation pressure from the central star ejects dust grains with $r \lesssim 1$–$10\,\mu$m; Poynting-Robertson drag pulls larger grains into the central star (Burns et al. 1979; Artymowicz 1988; Takeuchi and Artymowicz 2001). Eventually, small planetesimals are accreted by the large planetesimals or ground to dust.

To understand the origin of the collisional cascade, we consider the outcome of a head-on collision between two identical planetesimals. During the impact, some kinetic energy heats up the planetesimals; the rest goes into the internal energy of material in the planetesimals. When the impact energy is small, the extra internal energy is small compared to the binding energy of either planetesimal; the two objects merge into a single, larger planetesimal. When the impact energy is larger than the binding energy, the collision shatters the planetesimals into a few smaller planetesimals and a lot of dust.

Estimating the binding energy of planetesimals relies on two approaches (Davis et al. 1985; Housen and Holsapple 1990, 1999; Holsapple 1994; Benz and Asphaug 1999; Leinhardt et al. 2008; Leinhardt and Stewart 2009). Sophisticated collision experiments yield the internal strengths of small rocky and icy objects, $r \lesssim 10$–100 cm. Theoretical investigations derive the strength from analytic or numerical models of the crystalline structure and the equation of state of the material. In both cases, investigators derive the energy Q_D^* required to remove half of

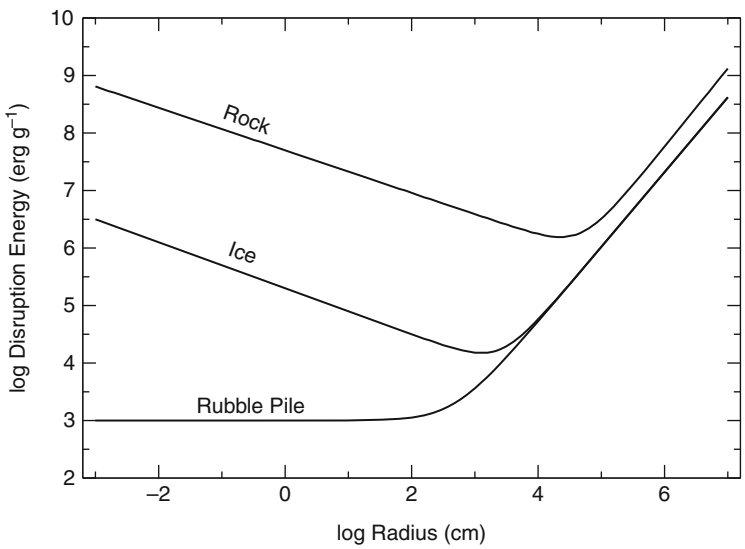
the combined mass of two colliding planetesimals and eject this mass to infinity. Although more sophisticated approaches include the impact velocity in Q_D^*, we focus on a simpler expression that depends only on radius,

$$Q_D^* = Q_b r_s^{\beta_b} + \rho Q_g r_s^{\beta_g} \,. \tag{1.79}$$

Here $Q_b r_s^{\beta_b}$ is the bulk (tensile) component of the binding energy, and $\rho Q_g r_s^{\beta_g}$ is the gravity component of the binding energy.

Laboratory experiments and detailed numerical collision simulations yield a broad range of results for Q_D^* (◉ *Fig. 1-7*; Housen and Holsapple 1990; Benz and Asphaug 1999; Holsapple et al. 2002; Leinhardt et al. 2008. In the strength regime at small sizes, the binding energy of a planetesimal depends on the number of flaws – cracks, fissures, etc. – in the material. Larger planetesimals have more flaws and smaller strengths. In the gravity regime at large sizes, the binding energy depends on the internal pressure. Larger planetesimals have larger internal pressures and larger strengths. The lower density and weaker crystalline structure of ice leads to smaller strengths than basalts and other rocks.

Models for the breakup of comet Shoemaker-Levy 9 suggest a smaller component of the bulk strength (Asphaug and Benz 1996), implying small disruption energies for small planetesimals (◉ *Fig. 1-7*; "Rubble Pile"). A low strength is consistent with numerical simulations of "rubble piles"; structures with countless flaws held loosely together. This structure probably



◉ Fig. 1-7

Disruption energy, Q_D^*, for icy objects. The *solid curves* plot typical results derived from numerical simulations of collisions (e.g., Benz and Asphaug 1999; Durda et al. 2004; Leinhardt and Stewart 2009) that include a detailed equation of state for basalt (rock) or crystalline ice (ice). In the strength regime ($r \lesssim 10^2$–10^4 cm), smaller particles are stronger. In the gravity regime ($r \gtrsim 10^5$ cm), larger objects are stronger. The "rubble pile" curve shows results consistent with model fits to comet breakups (e.g., Asphaug and Benz 1996)

results after icy or rocky planetesimals suffer numerous impacts which disrupt the internal structure (removing most of the tensile component of the binding energy) but do not destroy the object.

The collisional cascade begins when the impact energy of colliding small planetesimals equals Q_D^*. Because the random velocities of small planetesimals equal the escape velocities of large planetesimals, the impact energy depends only on the mass of a large planetesimal. Equating this energy to Q_D* allows us to derive the "disruption mass," the mass of a large planetesimal at the onset of the collisional cascade. With $v_{\mathrm{esc},s} \ll v_{\mathrm{esc},l}$, the impact energy per unit mass in the center-of-mass frame is roughly $v^2/8 \approx v_{\mathrm{esc},l}^2/8$. Setting this energy equal to Q_D^*, we solve for the disruption mass:

$$m_d = \left(\frac{3}{4\pi\rho_\bullet}\right)^{1/2} \left(\frac{8Q_D^*}{G}\right)^{3/2} \sim 3.5\times10^{-6}\rho_\bullet^{-1/2}\left(\frac{Q_D^*}{10^7\ \mathrm{erg\ g^{-1}}}\right)^{3/2}\ \mathrm{M}_\oplus. \tag{1.80}$$

When small planetesimals have sizes exceeding ~1 km, Q_D^* is fairly independent of their composition. For typical $Q_D^* \approx 10^7 - 10^9$ erg g^{-1}, the disruption mass is roughly 0.003–3 Pluto masses. Collisional cascades begin well before planets reach their final masses.

Once disruption commences, the final mass of a planet depends on the timescale for the collisional cascade (Kenyon and Bromley 2004a, 2008; Leinhardt and Richardson 2005). If disruptive collisions produce dust grains much faster than planets accrete planetesimals, planets cannot grow much larger than the disruption radius and have a maximum mass $m_{l,\mathrm{max}} \approx m_d$. However, if planets accrete grains and leftover planetesimals effectively, planets reach the isolation mass before collisions and radiation pressure remove material from the disk ($m_{l,\mathrm{max}} \approx m_{\mathrm{iso}}$; Goldreich et al. 2004).

In a gas-free environment, larger m_d in the inner disk enables planets to accrete much of the debris before destructive collisions and radiative processes remove it. Rocky planet masses then approach the isolation mass. In the outer disk, smaller planets cannot accrete debris before it is lost. Icy planets cannot grow much larger than m_d.

5.1.5 Planetesimal Accretion with Gas Damping

Gas slows the random velocities of smaller planetesimals. Larger protoplanets are less affected by drag and are damped by dynamical friction. The drag force, F_D, of (❯ 1.49) damps the kinetic energy of planetesimals (now with size r_s, not the size s of dust, pebbles, and boulders) at a rate $t_{\mathrm{gas}} \equiv v_s(dv_s/dt)^{-1} = v_s(F_D/m_s)^{-1}$. With $\rho_{\mathrm{g}} \sim \Sigma_{\mathrm{g}}\Omega/c_{\mathrm{s}}$, where c_{s} is the gas sound speed,

$$t_{\mathrm{gas}} \sim \frac{1}{C_D}\frac{\rho_\bullet r_s}{\Sigma_{\mathrm{g}}\Omega}\frac{c_{\mathrm{s}}}{v_s}\,. \tag{1.81}$$

To understand the impact of damping, we consider an ensemble of small planetesimals stirring themselves (❯ 1.74). Without gas, small planetesimals excite their velocities to $v_s \sim v_{\mathrm{esc},s}$. If $t_{\mathrm{gas}} < t_{coll} \sim \rho_\bullet r_s/(\Sigma_s\Omega)$, drag exceeds collisions as the dominant cooling mechanism. This switch happens when

$$r_s \gtrsim \frac{1}{C_D}\frac{\Sigma_s}{\Sigma_{\mathrm{g}}}\frac{c_{\mathrm{s}}}{\sqrt{G\rho_\bullet}} \sim 30\left(\frac{R}{\mathrm{AU}}\right)^{-3/14} Z_{\mathrm{rel}}\ \ \mathrm{km}\,. \tag{1.82}$$

To be damped by gas drag, planetesimals must exceed this minimum size. This somewhat counterintuitive result (drag is often more significant for smaller bodies) arises from (1) nonlinear

drag laws and (2) a velocity scale, $v_{\mathrm{esc},s}$, that increases with size. The numerical value of the size threshold decreases if Z_{rel} is reduced due to an inefficiency of turning dust into planetesimals.

When (1.82) holds, the stronger damping of self-stirred planetesimals ensures $v_s < v_{\mathrm{esc},s}$. Collisions are gravitationally focused, and runaway growth begins earlier than in a disk without gas.

As growth proceeds, larger protoplanets dominate the stirring of smaller planetesimals. With stronger stirring, smaller and smaller planetesimals are damped by nonlinear gas drag instead of collisions. Even if collisions are initially more significant, gas drag becomes the dominant coolant as growth proceeds. To compute the random speeds of small planetesimals, we assume dispersion-dominated encounters and balance the heating of (1.76) with the cooling of (1.81) to get

$$\frac{v_s}{v_{\mathrm{esc},l}} \sim \left(\frac{1}{C_D} \frac{r_s}{r_l} \frac{\Sigma_l}{\Sigma_{\mathrm{g}}} \right)^{1/5} . \tag{1.83}$$

With more mass in small planetesimals $\Sigma_l < \Sigma_s \ll \Sigma_{\mathrm{g}}$, gravitational focusing, $f_{G,\mathrm{disp}} \sim (v_{\mathrm{esc},l}/v_s)^2$, becomes strong. Accretion of small planetesimals can become shear dominated. As described in Sect. 5.1.2, runaway growth transitions to oligarchy (but does not slow down) in the transition to shear-dominated accretion. A self-consistent analysis of these processes is facilitated by the numerical calculations summarized in Sect. 5.1.6.

Details aside, the large oligarchs stir small planetesimals past their escape speed and up to the disruption velocity (Sect. 5.1.3). Disruptive collisions among small planetesimals produce a collisional cascade, which grinds planetesimals into smaller and smaller objects. Without gas, planetesimals are ground into small dust grains which are dragged into the star by Poynting-Robertson drag or ejected from the planetary system by radiation pressure. With gas damping, the collisional cascade halts at some intermediate size (~10 cm to 10 m), depending on factors such as the mass of the oligarchs, gas density, material strength, and orbital distance. The damped velocities are then slow enough that the oligarchs accrete these small rocks rapidly. This rapid accretion enables oligarchs to reach the isolation mass on short timescales, even in the outer disk (Rafikov 2004; Kenyon and Bromley 2009).

Collisional grinding a set of small planetesimals into small dust grains requires a very depleted gas disk. For the cascade to proceed down to 1–10-μm particles, these grains must decouple from the gas (on an orbital timescale, we assume for simplicity). Epstein drag applies for low gas densities (and long gas mean free paths). From (1.48a), particles with sizes $s \gtrsim \Sigma_{\mathrm{g}}/\rho_\bullet$ decouple from the gas. From (1.1), the depletion factor (relative to the MMSN) required to avoid entrainment is

$$F \lesssim 10^{-7} \frac{s}{\mu\mathrm{m}} \left(\frac{R}{\mathrm{AU}} \right)^{3/2} . \tag{1.84}$$

This low-mass disk may not last very long. Nevertheless, current observational limits only constrain gas surface densities in debris disks to $F \lesssim 1\%$ of the MMSN, not yet sufficient to asses the dynamical significance of gas. ALMA should place much tighter constraints (See Chap. 9 by Moro-Martin).

5.1.6 Numerical Simulations of Low-Mass Planet Formation

Analytic estimates provide a good understanding of each piece of the planet formation process. However, putting the whole set of processes into a coherent theory requires numerical calculations. Clusters of computers can now finish an end-to-end calculation in a reasonable amount of time. Several groups are building toward this simulation, but no complete calculation exists.

Constructing numerical simulations of planet formation involves identifying and solving a set of coupled differential equations which describe the evolution of the gaseous disk and the masses and orbital properties of solid objects. Selecting the proper approach depends on the nature of the problem. Hydrodynamics codes address the evolution of the gaseous disk and how planets accrete material and migrate within the disk (D'Angelo et al. 2003; Nelson and Papaloizou 2004). Smooth particle hydrodynamics allows detailed solutions to outcomes of binary collisions between large protoplanets (e.g., Earth-Moon and Pluto-Charon formation; Canup 2008, 2011). Solving the coagulation equation with a fragmentation algorithm yields the mass and time evolution of solid particles ranging in size from 1 μm up to roughly 1,000 km (Safronov 1969; Wetherill and Stewart 1993; Kenyon and Luu 1999; Birnstiel et al. 2010). To treat the dynamical evolution of large planets, N-body treatments provide accurate and often fast solutions (Chambers and Wetherill 1998; Chambers 2001; Raymond et al. 2004; Nagasawa et al. 2005; Kokubo et al. 2006).

Most investigations of terrestrial planet formation employ a coagulation code or an N-body code. An N-body code cannot possibly follow the trajectories of the $\gtrsim 10^{12}$ small planetesimals expected in a MMSN. Coagulation models, which treat planetesimals as a statistical ensemble of objects with a distribution of e and i, can solve for the time evolution of their masses and orbits throughout runaway and oligarchic growth (Wetherill and Stewart 1993). Once most of the solid mass is in a few protoplanets, the statistical approach fails. N-body codes can then follow the evolution during the late stages of oligarchic growth and throughout chaotic growth.

Several hybrid codes combine aspects of both approaches (Spaute et al. 1991; Weidenschilling et al. 1997; Bromley and Kenyon 2006; Charnoz and Morbidelli 2007; Raymond et al. 2011). To follow the evolution of the gaseous disk together with the solids, Bromley and Kenyon (2011) solve the radial diffusion equation for the gaseous disk (❯ 1.13) and employ a merged coagulation + N-body code for the solids. In these treatments, the coagulation code follows solids with masses smaller than the promotion mass, $m_{\rm pro}$; the N-body code tracks protoplanets with $m > m_{\rm pro}$. Comparisons with other simulations and with analytic theory provide tests of these techniques (e.g., Kenyon and Luu 1998; Fraser 2009; Morbidelli et al. 2009b).

To illustrate the formation process, we summarize results for several calculations of terrestrial planets and gas giant cores. Because this aspect of this field is growing so rapidly, we focus on a few simple examples.

Coagulation codes begin with an ensemble of planetesimals in place at $t = 0$ (Kenyon and Bromley 2010). Planetesimals are placed in concentric annuli according to a fixed initial surface density relation. These planetesimals often have a single size of 1–100 km; sometimes calculations begin with a distribution of sizes. Because dynamical friction efficiently damps the velocities of the largest planetesimals, planets grow faster in calculations with a size distribution of planetesimals. Starting with an ensemble of small planetesimals leads to faster growth than an ensemble of large planetesimals. The initial surface density sets the growth time. Planets grow faster in more massive disks. In many calculations, the planetesimals evolve in a gaseous disk which also evolves in time; the disk evolution may be proscribed in advanced or calculated along with the planetesimals.

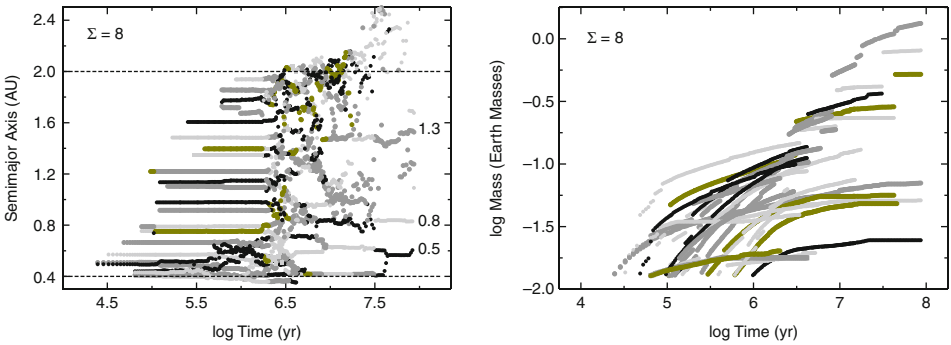


Fig. 1-8

Evolution of oligarchs in the terrestrial zone. The calculation starts with 1-km planetesimals ($\rho_\bullet = 3$ g cm^{-3}) in a disk with $\Sigma_s = 8$ g cm^{-2} $(a/1\ \text{AU})^{-1}$. *Left panel*: The time evolution of semimajor axis shows three phases that start at the inner edge of the grid and propagate outward: (1) after runaway growth, isolated oligarchs with $m \gtrsim 4 \times 10^{25}$ g form and continue to grow very rapidly; (2) oligarchs develop eccentric orbits, collide, and merge; and (3) a few massive oligarchs eventually contain most of the mass and develop roughly circular orbits. The legend indicates masses (in $M_\oplus$) for the largest oligarchs. *Right panel*: The mass evolution of oligarchs shows an early phase of runaway growth (steep tracks) and a longer phase of oligarchic growth (flatter tracks), which culminates in a chaotic phase where oligarchs grow by captures of other oligarchs (steps in tracks). Despite the steeper appearance of some of the mass tracks during runaway growth, protoplanets grow more rapidly during oligarchic growth

Figure 1-8 shows the evolution of oligarchs in an evolutionary sequence starting with an ensemble of 1-km planetesimals at 0.4–2 AU. Following a short runaway growth phase, protoplanets with $m \gtrsim m_{\text{pro}}$ appear in a wave that propagates out through the planetesimal grid. As these oligarchs continue to accrete planetesimals, dynamical friction maintains their circular orbits, and they evolve into "isolated" protoplanets. Eventually, large oligarchs start to interact dynamically at the inner edge of the grid; a wave of chaotic interactions then moves out through the disk until all oligarchs interact dynamically. Once a few large oligarchs contain most of the mass in the system, dynamical friction between the oligarchs and a few leftover planetesimals starts to circularize their orbits. This process excites the lower mass oligarchs and leftover planetesimals, which are slowly accreted by the largest oligarchs. At the end of the calculation, the masses, semimajor axes, and orbital eccentricities of stable planets are similar to those of the terrestrial planets in the solar system.

Comparisons between the results of hybrid and N-body calculations show the importance of including planetesimals in the evolution. Both approaches produce a few terrestrial mass planets in roughly circular orbits. Because dynamical friction between leftover planetesimals and the largest oligarchs is significant, hybrid calculations produce planets with more circular orbits than traditional N-body calculations. In most hybrid calculations, lower mass planets have more eccentric orbits than the most massive planets, as observed in the solar system. In both approaches, the final masses of the planets grow with the initial surface density; the number of planets is inversely proportional to the initial surface density of solid planetesimals. However, the overall evolution is faster in hybrid calculations: oligarchs start to interact earlier and produce massive planets faster.

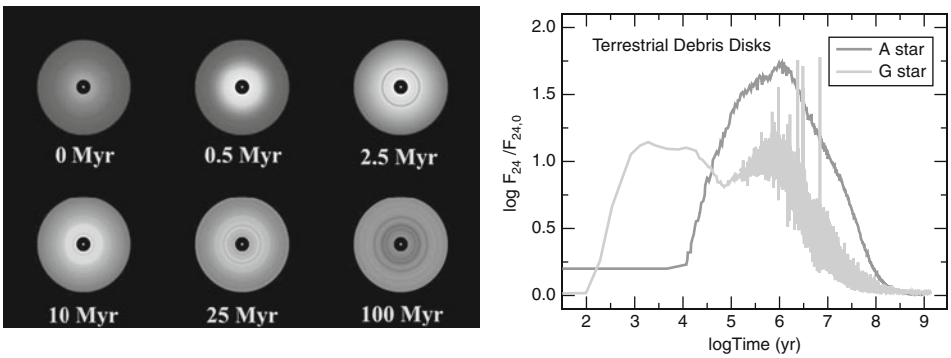


Fig. 1-9

Evolution of debris disks in the terrestrial zone (Kenyon and Bromley 2004b, 2005). For an A-type star with a luminosity of ~50 $L_\odot$, the range in blackbody temperatures of planetesimals at 3–20 AU (425–165 K) is similar to the range in the solar system at 0.4–2 AU (440–200 K). *Left panel*: Images of a disk extending from 3–20 AU around an A-type star. The intensity scale indicates the surface brightness of dust, with *black* the lowest intensity and *white* the highest intensity. *Right panel*: Mid-IR excess for two debris disk models. The *light gray line plots* the ratio of the 24-μm flux from a debris disk at 0.4–2-AU disk relative to the mid-IR flux from a G-type star. The *dark gray line* shows the evolution for the A-star disk shown in the *left panel*

In hybrid calculations, the isolation mass and the number of oligarchs are more important as local quantities than as global quantities. As waves of runaway, oligarchic, and chaotic growth propagate from the inner disk to the outer disk, protoplanets growing in the inner disk become isolated at different times compared to protoplanets growing in the outer disk. Thus, the isolation mass in hybrid models is a function of heliocentric distance, initial surface density, and *time*, which differs from the classical definition (1.73).

During oligarchic growth of the simulation in *Fig. 1-8*, viscous stirring excites leftover planetesimals to the disruption velocity. A series of separate simulations demonstrates that the collisional cascade produces copious amounts of dust, which absorb and scatter radiation from the central star. Following the growth of protoplanets, the cascade begins at the inner edge of the disk and moves outward. For calculations with a solar-type central star, it takes ~0.1 Myr for dust to form throughout the terrestrial zone (0.4–2 AU). The timescale is ~1 Myr for the terrestrial zone of an A-type star (3–20 AU). As the collisional cascade proceeds, protoplanets impose structure on the disk (*Fig. 1-9*, left panel). Bright rings form along the orbits of growing protoplanets; dark bands indicate where a large protoplanet has swept up dust along its orbit. In some calculations, the dark bands are shadows, where optically thick dust in the inner disk prevents starlight from shining on the outer disk (Grogan et al. 2001; Kenyon and Bromley 2004a; Durda et al. 2004).

In the terrestrial zones of A-type and G-type stars, the dust emits mostly at mid-IR wavelengths. In calculations with G-type central stars, formation of a few lunar mass objects at 0.4–0.5 AU leads to copious dust production in a few thousand years (*Fig. 1-9*, right panel). As protoplanets form farther out in the disk, the disk becomes optically thick and the mid-IR excess saturates. Once the orbits of oligarchs start to overlap (~1 Myr), the largest objects sweep the disk clear of small planetesimals. The mid-IR excess fades. During this decline, occasional

large collisions generate large clouds of debris that produce remarkable spikes in the mid-IR excess (Kenyon and Bromley 2002, 2005).

In A-type stars, the terrestrial zone lies at greater distances than in G-type stars. Thus, debris formation in calculations with A-type stars begins later and lasts longer than in models with G-type stars (Fig. 1-9, right panel). Because the disks in A-type stars contain more mass, they produce larger mid-IR excesses. At later times, individual collisions play a smaller role, which leads to a smoother evolution in the mid-IR excess with time. Although the statistics for G-type stars is incomplete, current observations suggest that mid-IR excesses are larger for A-type stars than for G-type stars (See Chap. 9 by Moro-Martin).

Collisional cascades and debris disk formation may impact the final masses of terrestrial planets. Throughout oligarchic growth, roughly ~25–50% of the initial mass in planetesimals is converted into debris. For solar-type stars, the disk is optically thick, so oligarchs probably accrete the debris before some combination of gas drag, Poynting-Robertson drag, and radiation pressure removes it. In the disks of A-type stars, the debris is more optically thin. Thus, these systems may form lower mass planets per unit surface density than disks surrounding less massive stars. Both of these assertions require tests with detailed numerical calculations.

5.2 Accretion of Atmospheres

Protoplanetary atmospheres have a rather different character than the mature planetary atmospheres of solar system planets and exoplanets. The crucial distinction is that protoplanets orbit within a gas disk. The disk supplies the atmosphere's gas and provides an external binding pressure until the planet opens a clean gap. The protoplanetary atmospheres inherit its composition from the disk. However, the fraction of heavy elements that wind up in the core – versus the dust and ablated planetesimals that remain in the atmosphere – is a key uncertainty. This uncertainty crucially affects the mean molecular weight, μ, and opacity, κ, of the atmosphere.

5.2.1 Static Protoplanet Atmospheres

As protoplanets grow, they become massive enough to bind a gaseous atmosphere. The atmosphere is significantly denser than the surrounding disk gas when the core's gravitational binding energy exceeds the thermal energy of the gas. For a solid core of mass m_c and radius r_c, this occurs when $r_B > r_c$, where the Bondi radius

$$r_B = Gm/c_s^2 \,, \tag{1.85}$$

where total planet mass $m = m_c + m_a$, including the gravitationally bound atmosphere's mass, m_a. Equivalently a core with an atmosphere must exceed the Bondi mass

$$m_c > m_B \equiv \sqrt{\frac{3}{4\pi\rho_\bullet}} \frac{c_s^3}{G^{3/2}} \simeq 10^{-3} \left(\frac{a}{\mathrm{AU}}\right)^{-9/14} \frac{l_*^{3/8}}{\tilde{\mu}^{3/2}} \ \mathrm{M}_\oplus \tag{1.86}$$

where we use the gas temperature in an irradiated disk (1.38) and normalize the stellar luminosity as $l_* = \mathrm{L}_*/\mathrm{L}_\odot$ and the gas mean molecular weight as $\tilde{\mu} = \mu/(2.4\ m_H)$.

As the core mass increases beyond m_B, the atmosphere becomes denser and more massive. The outer boundary of the atmosphere, $r_{\mathrm{out}} = \min(r_B, R_H)$, is set by the Bondi radius until the

atmosphere fills the Hill sphere. This transition occurs for massive protoplanets, as $r_B > R_H$ requires

$$m > m_{\rm trans} = \frac{c_s^3}{\sqrt{3}G\Omega} \simeq 3\left(\frac{a}{\rm AU}\right)^{6/7} \frac{l_*^{3/8}}{\sqrt{m_*}\tilde{\mu}^{3/2}}\ {\rm M}_\oplus\,. \tag{1.87}$$

Comparing to the disk's gas scale-height $H_g = c_s/\Omega$, the criteria $r_B > H_g$ and $R_H > H_g$ also reproduce (1.87) within order unity constants ($m_{\rm trans}$ increases by $3^{3/2}$ for the $R_H > H_g$ criterion). Thus, when $m > m_{\rm trans}$, the protoplanet is no longer uniformly embedded in the disk midplane. It can feel the top and bottom of the disk and start to open a gap (See Chap. 2 by Morbidelli). Outside ~5 AU, the core accretion instability generally occurs for $m < m_{\rm trans}$. Thus, the 3D structure of the gas disk can usually be ignored when describing the onset of core accretion (Sect. 5.2.3) but not its final evolution (Sect. 5.2.4).

The structure of an proplanetary atmosphere obeys the equations (which also govern stellar structure) of hydrostatic balance

$$\frac{dP}{dr} = -\frac{Gm}{r^2}\rho = -\rho g\,, \tag{1.88}$$

mass conservation

$$\frac{dm}{dr} = 4\pi r^2 \rho\,, \tag{1.89}$$

and energy transport by optically thick, $\tau \sim \kappa P/g > 1$, radiative diffusion,

$$\frac{16\sigma T^3}{3\kappa\rho}\frac{dT}{dr} = -\frac{L}{4\pi r^2}\,. \tag{1.90}$$

Wherever radiative diffusion would satisfy the Schwarzschild criterion, $d\ln T/d\ln P > \nabla_{\rm ad}$, the energy transport becomes convective. The adiabatic index $\nabla_{\rm ad} = 2/7$ for an ideal diatomic gas but in general must be determined from detailed equation-of-state calculations (Saumon et al. 1995; Saumon and Guillot 2004). Because convective transport is efficient, the temperature profile follows an adiabat, $T \propto P^{\nabla_{\rm ad}}$, instead of (1.90) in convective regions. For the ideal gas equation of state, $P = \rho\mathcal{R}T$, with the (specific) gas constant $\mathcal{R} = k_B/\mu m_H$.

The masses of stable atmospheres (that do not undergo the core accretion instability) place interesting constraints on planet formation. In the terrestrial zone, typical isolation masses (1.73) are roughly 0.1 $\rm M_\oplus$. For almost any accretion time, these planets have stable atmospheres with masses much smaller than the planet's mass. Icy planets formed at tens of AU, however, have much larger isolation masses of several $\rm M_\oplus$and can support much more massive atmospheres. In the solar system, the dichotomy between terrestrial planets with thin atmospheres and icy planets with massive atmospheres is consistent with our estimates. Once we have a large sample of rocky/icy exoplanets with well-characterized atmospheres, we will see if the same dichotomy persists.

5.2.2 Enhanced Planetesimal Accretion

For low-mass protoplanets with $r_{\rm out} = r_B$, the size of the atmosphere relative to the core is

$$\frac{r_{\rm out}}{r_c} \simeq 800\left(\frac{m_c}{10\ {\rm M}_\oplus}\right)^{2/3}\left(\frac{a}{5\ {\rm AU}}\right)^{3/7}\frac{\tilde{\mu}}{l_*^{1/4}}\,. \tag{1.91}$$

For Mars-mass planets and larger, the radius of the atmosphere is ten or more times larger than the radius of the core. Extended atmospheres can significantly enhance planetesimal accretion.

When a planetesimal encounters the atmosphere of a protoplanet, it experiences enhanced gas drag. Capture results if the planetesimal loses enough orbital energy. A thin atmosphere has little impact on very large planetesimals; the collisional cross section is still $\pi r^2 f_g$. For sufficiently small planetesimals, any encounter with the atmosphere allows the planet to capture the planetesimal; the collisional cross section is then $\pi r_{\rm out}^2 f_g$. For intermediate sizes, the effective cross section lies somewhere between $\pi r^2 f_g$ and $\pi r_{\rm out}^2 f_g$. To address this regime, Inaba and Ikoma (2003) define an "enhanced radius" r_e, where the collisional cross section is $\pi r_e^2 f_g$. In this approach, $r_e = r$ for accreting very large planetesimals, and $r_e = r_{\rm out}$ for accreting very small planetesimals. However, very small rocks and dust grains will be too tightly coupled to the gas to accrete.

Compared to the estimates for dispersion- and shear-dominated growth in ❯ Sect. 5.1, atmospheres can enhance accretion rates by 1–2 orders of magnitude (Chambers 2008). When leftover small planetesimals have typical radii of 0.1–10 km, an isolated terrestrial planet with a thin atmosphere has a small radius enhancement, $r_e \approx 1 - 2$. Thus, these isolated objects never experience rapid growth from an enhanced radius. Isolated icy planets are more massive and support massive atmospheres. These objects have $r_e \approx 10$ for accreting 0.1–10-km planetesimals and $r_e \approx 3$ for accreting 100+-km planetesimals. Once they develop atmospheres, icy isolated objects rapidly sweep up any leftover planetesimals along their orbits.

5.2.3 The Core Accretion Instability

Low-mass protoplanets (near m_B) have low-mass, optically thin atmospheres. More massive cores bind thicker atmospheres, which trap the heat of planetesimal accretion. As the heat is radiated away, the atmosphere becomes denser and more massive. When the atmosphere's mass exceeds roughly the core mass, the atmosphere cannot maintain hydrostatic equilibrium and collapses (Harris 1978; Mizuno 1980). This collapse – referred to as the "core accretion instability" – leads to rapid gas accretion and the birth of a gas giant. The critical core mass (sometimes called the "crossover mass" because collapse occurs when the core and atmosphere masses roughly match) for this instability is often, but not always, ~10 $M_\oplus$ (Ikoma et al. 2000; Rafikov 2006).

Near the crtical core mass, real protoplanetary atmospheres are convective in the interior and (for low enough planetesimal accretion rates) radiative in the exterior region that matches onto the disk (Rafikov 2006). However, an illustrative, and historically important, calculation by Stevenson (1982) demonstrates the essential features of the core accretion instability by (incorrectly) assuming the atmosphere is completely radiative with constant opacity κ. We summarize this calculation before comparing it to more detailed computations.

Following Stevenson (1982), we calculate the structure and mass in the atmosphere by keeping the mass, m, constant in the hydrostatic balance equation (❯ 1.88), that is, neglecting the detailed variation from $m = m_c$ at the core to $m = m_c + m_a$ at the top. ❯ Equation 1.88 and ❯ 1.90 then give (temporarily omitting order unity coefficients for clarity)

$$T^3 dT/dP \sim \kappa L/(\sigma G m)\,. \tag{1.92}$$

To integrate this equation, we assume (correctly) that P and T are significantly higher at the base of the atmosphere than in the disk. We further keep L constant, appropriate if accreted planetesimals release their kinetic energy at the core's surface. This assumption yields $L = Gm_c\dot{m}/r_c$, where $\dot{m}$ is the planetesimal accretion rate. The slowing of planetesimals as they fall through the atmosphere is a minor correction included in detailed numerical models.

With these approximations, (1.92) integrates to a simple $T - P$ profile

$$T \sim \left(\frac{\kappa L}{\sigma G m} P\right)^{1/4} . \tag{1.93}$$

To obtain the density profile, we use the fact that for a barotropic relation $P \propto T^{\nabla_\infty}$ ($\nabla_\infty = 1/4$ for our example of a constant opacity), the hydrostatic balance equation for an ideal gas integrates to

$$T = \nabla_\infty \frac{Gm}{\mathcal{R} r} \tag{1.94}$$

in the atmospheric interior. The density profile follows as

$$\rho \sim \frac{\sigma}{\kappa L}\left(\frac{Gm}{\mathcal{R}}\right)^4 \frac{1}{r^3} . \tag{1.95}$$

More generally, $\rho \propto r^{1-1/\nabla_\infty}$, which is left as an exercise.

The atmosphere's mass follows by integrating (1.89) from r_c to $r_{\rm out}$:

$$m_a = \frac{\pi^2}{3}\chi\frac{\sigma}{\kappa L}\left(\frac{Gm}{\mathcal{R}}\right)^4 \simeq 2.0\frac{\sigma\chi}{\kappa\mathcal{R}^4}\frac{G^3 m^4}{\rho_\bullet^{1/3} m_c^{5/3}} t_{\rm acc} . \tag{1.96}$$

with $\chi \equiv \ln(r_{\rm out}/r_c)$ and order unity coefficients reinstated (despite the overall approximate nature of the calculation). The final expression relates the protoplanet luminosity to the (current) core growth timescale, $t_{\rm acc} = m_c/\dot{m} = Gm_c^2/(r_c L)$. Setting $m = m_c$, we can numerically evaluate

$$m_a \approx 9.4\left(\frac{m_c}{10\ {\rm M}_\oplus}\right)^{7/3}\frac{\tilde{\mu}^4}{\kappa_1}\frac{t_{\rm acc}}{\rm Gyr}\ {\rm M}_\oplus, \tag{1.97}$$

with $\chi = 6.7$ (from 1.91) and $\kappa_1 = \kappa/(1\,{\rm cm^2\,g^{-1}})$. For the chosen parameters, we are near the crossover mass, $m_a \sim m_c \sim 10\ {\rm M}_\oplus$. *Figure 1-10* shows the behavior of this simple atmosphere model. We emphasize that the numerical values of this simple model are only meant to be illustrative. For instance, the core must actually accrete in $t_{\rm acc} \lesssim 10\ {\rm Myr} \ll {\rm Gyr}$.

The existence of an instability arises from the nonlinearities in m_a. Expressing

$$m = m_c + k\frac{m^4}{m_c^{5/3}} \tag{1.98}$$

where k incorporates all the constants in (1.96), we can show that beyond a critical core mass, the total mass (unphysically) declines as m_c increases. Using calculus, the turnover[5] occurs where $dm_c/dm = 0$ at $m = m_c^{5/9}(4k)^{-1/3}$ and $m_c = 0.19k^{-3/4}$.

[5]Our values differ slightly from Stevenson (1982) because we assume constant accretion time instead of constant mass accretion rate. Further the 3/4 exponent in his Eq. 15 is a typo that should be 3/7.

This simple derivation at least qualitatively explains many features of more detailed core accretion models. The strong dependence of the atmospheric mass on μ is supported by studies showing that envelope pollution lowers the critical core mass (Hori and Ikoma 2011). The opacity is very sensitive to the amount and sizes of dust grains (Pollack et al. 1985). One popular way to speed up core accretion is to reduce the opacity (Hubickyj et al. 2005). The "correct" choice of opacity likely varies between planets and is poorly constrained. It is unclear how much dust from ablated planetesimals will remain in the atmosphere. Small grains both contribute significantly to opacity and settle slowly.

High planetesimal accretion rates increase the critical core mass. For very high accretion rates, especially in the inner disk, the protoplanet atmosphere will be fully convective (Rafikov 2006). The atmosphere then matches onto the same adiabat (constant entropy curve) as the disk gas and thus has the lowest possible mass. In this case the formation of gas giants is quite unlikely. If planetesimal accretion stops (or becomes suitably small), the relevant luminosity comes from the Kelvin-Helmoltz contraction of the atmosphere. Models that omit planetesimal accretion (but correctly compute contraction) thus provide a meaningful lower limit on the critical core mass (Papaloizou and Nelson 2005).

It has long been postulated that the cores of solar system giants might correspond to an isolation mass, ~5–20 $M_\oplus$, at 5–10 AU, even though this would require a massive planetesimal disk (Pollack et al. 1996). To understand why this assumption is reasonable, we turn to *Fig. 1-10*. Initially a low-mass planet accretes planetesimals rapidly. With a short accretion time, the puffy atmosphere is well below the crossover mass. As the core grows in mass, two effects bring the atmosphere closer to instability, (1) extra compression of the gas and (2) fewer and fewer planetesimals to accrete and heat the atmosphere. As the atmosphere cools, the critical core mass drops until it reaches the actual core mass. This cooling is most likely to happen after most planetesimals have been accreted, that is, at the isolation mass.

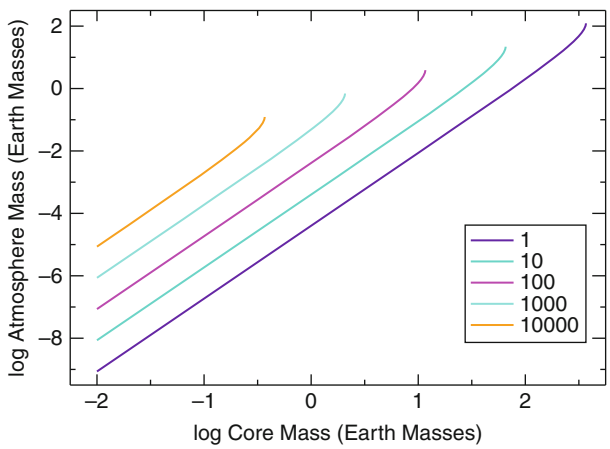


Fig. 1-10

Mass of planet atmosphere as a function of core mass and accretion time (1.97). The legend indicates the accretion time in Myr. At fixed core mass, the mass of the atmosphere grows with the accretion time. Longer accretion times allow more massive cores to have hydrostatic atmospheres

5.2.4 Direct Accretion of Disk Gas (and How it Stops)

When the envelope collapses, the planet starts to accrete gas directly from the protostellar disk. This dynamical process is not amenable to a stellar structure calculation. Direct accretion of gas is similar to accretion of planetesimals by oligarchs; the accretion rate is roughly the product of the planet's cross section (or impact parameter in the 2D limit), the local gas density (or surface density), and the encounter velocity, which is dominated by Keplerian shear for circular orbits.

Shortly after the core accretion instability, the planet crosses the transition mass (1.87), and the relevant accretion radius is the Hill radius, R_H. Since the transition mass also corresponds to the Hill radius exceeding the gas scale height (see discussion after (1.87)), accretion is not at the classic Bondi rate (e.g., Shu 1992). Instead the two-dimensional mass accretion rate applies,

$$\dot{m}_{\rm g} \sim \Sigma_{\rm g} \Omega R_H^2 \sim 10^5 \frac{F}{m_*^{1/6}} \left(\frac{m}{60\ {\rm M}_\oplus} \right)^{2/3} \left(\frac{5\ {\rm AU}}{a} \right) \frac{{\rm M}_\oplus}{\rm Myr} \,. \tag{1.99}$$

At $5 \times 10^{-7} {\rm M}_\odot\ {\rm year}^{-1}$, this rate exceeds the accretion rate onto most pre-main-sequence T Tauri stars!

Clearly, something must stop this influx of gas. The gaseous isolation mass is one natural stopping point. Applying (1.73) to the gas disk gives

$$m_{\rm iso,g} = \frac{(2\pi B \Sigma_{\rm g})^{3/2}}{(3{\rm M}_*)^{1/2}} a^3 \approx 2 \frac{F^{3/2}}{m_*} \left(\frac{a}{5\ {\rm AU}} \right)^{3/4} {\rm M}_{\rm J} \,. \tag{1.100}$$

While this result appears plausible for Jupiter, most disk models for Jupiter's core require $F \gtrsim 5$, requiring a second mechanism to halt gas accretion (Lissauer et al. 2009).

Opening a gap in the disk – which begins when (1.87) is satisfied – can slow down accretion so that the disk dissipates before the planet reaches the gaseous isolation mass. Since disk lifetimes are at least at least 1–3 Myr, accretion times must be at least this slow to halt growth (without invoking a fine-tuning of core accretion and disk dissipation timescales). Only a wide and relatively clean gap can slow accretion enough to explain final planet masses (D'Angelo and Lubow 2008). The effective viscosity of the disk must be low for a clean gap, which is an especially strong concern for self-gravitating disks (Kratter et al. 2010b, Sect. 5.3).

5.2.5 Numerical Simulations of Gas Giant Planet Formation

To conclude this section, *Fig. 1-11* illustrates the evolution of the semimajor axes of icy and gas giant planets from one simulation of material outside the terrestrial zone (e.g., Bromley and Kenyon 2011). The calculation begins with a single 1,000-km planetesimal and an ensemble of ~1-cm planetesimals in each of 96 annuli from 3 to 30 AU. Because the system has large gravitational focusing factors at $t = 0$, each large planetesimal rapidly sweeps up the small planetesimals along its orbit. With growth times proportional to the orbital period (1.72a), protoplanets at 3–7 AU grow much faster than those at larger a. Growth produces many isolated mass objects packed closely together. Early on, gravitational interactions among these objects jostle them around into overlapping orbits. After ~0.3 Myr, the most massive of these protoplanets begin to scatter lower mass protoplanets to smaller and larger a. Scattered protoplanets sweep up and scatter the large preexisting planetesimals in these orbits, accelerating the growth

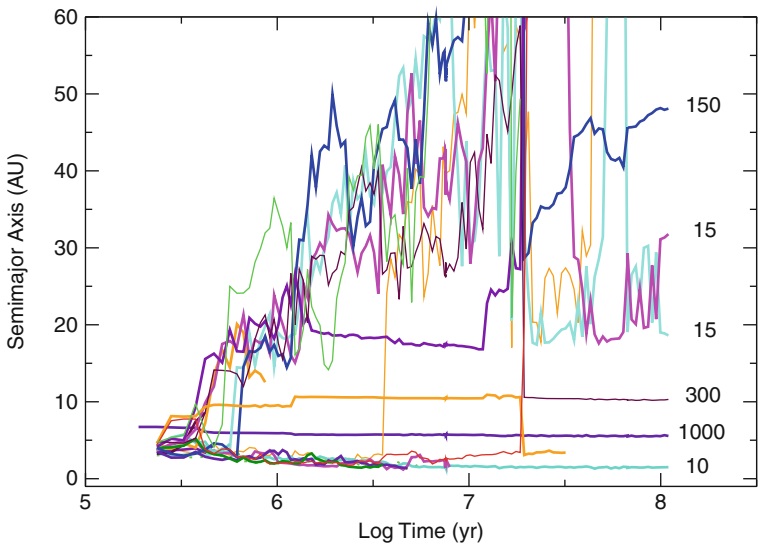


Fig. 1-11

Orbital evolution of icy oligarchs. The calculation starts with 1-cm and 1,000-km planetesimals ($\rho_\bullet = 1.5$ g cm^{-3}) in a disk with initial $\Sigma_s = 14$ g cm^{-2} $(a/3\text{ AU})^{-1} e^{-a/30\text{ AU}}$. During the first 3×10^5 year, ~20 oligarchs with $m \approx 0.05$ M$_\oplus$ form in a relatively narrow range of semimajor axis, 3–7 AU. As they grow, more massive oligarchs scatter lower mass oligarchs to large semimajor axes, $a \approx 1$–20 AU. At ~1 Myr, some oligarchs begin to accrete gas. Over the next 10 Myr, continued growth and scattering lead to collisions and mergers of oligarchs. Eventually, only a few oligarchs remain. The largest of these have masses comparable to the mass of Jupiter. The legend indicates masses (in $M_\oplus$) for the largest oligarchs in stable orbits around the central star

of all large protoplanets. At ~1 Myr, the largest protoplanets begin to accrete gas from the disk. As they grow, they scatter lower mass protoplanets to larger and larger a; eventually, they eject some of these low-mass protoplanets from the planetary system.

At the end of the calculation at 100 Myr, six planets remain on stable orbits. Two gas giants, with 1 and 3 Jupiter masses, have $a = 5$ and 10 AU. Inside these gas giants, a super-Earth with $m \approx 10$ M$_\oplus$ lies on a fairly circular orbit, $e \approx 0.02$ at $a \approx 1.5$ AU. Outside the gas giants, two more super-Earths occupy orbits in a 2:1 resonance. After many exchanges, the orbits of these two planets are likely stable. Finally, a planet with roughly 1.5 times the mass of Saturn rests in an orbit with modest eccentricity, $e \approx 0.1$ at $a \approx 50$ AU. The outcome of this simulation combines some properties of the solar system – four gas giants at 5–30 AU – along with some properties of known exoplanets, – a super-Earth at 1–2 AU.

This example illustrates several important differences between terrestrial and gas giant planet formation:

- Oligarchs form at 1 AU before they form at 5 AU. From (1.72a), the growth time scales with the orbital period and the enhancement of the surface density at the snow line. Comparing timescales at 0.4 and at 4 AU, we expect a factor of ~300 from the orbital period and a factor of 1/2.4 from the snow line enhancement. The factor of ~10 ratio of growth times at 0.4 and 4 AU in the simulations agrees well with expectations.

Table 1-1
Frequently Used Symbols

Symbol	Ref.	Meaning
General Physical Quantities		
AU	(1.1)	Astronomical unit
c	(1.8)	Speed of light
c_s	(1.12)	Sound speed
C_D	(1.49)	Drag coefficient
C_v	Sect. 3.2	Specific heat
P	(1.11)	Gas pressure
t		Time or timescale, often subscripted
T [T_{eq}]	(1.5)	[Equilibrium] temperature
v_K	(1.12)	Keplerian circular velocity
γ	(1.12)	Adiabatic index
κ	(1.42)	Radiative opacity
λ	(1.6), Sect. 3.6	Wavelength (light or other waves)
λ	Sect. 3.2, (1.48)	Gas mean free path
μ	Sect. 3.2	Mean molecular weight
σ	Sect. 3.2	Cross section
σ_{SB}	(1.32)	Stephan-Boltzmann constant
ρ		Density (mass per volume) of quantity, often subscripted
$\rho_\bullet$	(1.8)	Internal density of solid grain or planetesimal
Ω	Sect. 3.1	Keplerian orbital frequency
Stellar Quantities		
L_* [$L_\odot$]	(1.5)	Stellar [solar] luminosity
l_*	(1.87)	$L_*/L_\odot$, dimensionless stellar luminosity
M_* [$M_\odot$]	(1.9)	Stellar [solar] mass
m_*		$M_*/M_\odot$, dimensionless stellar mass
Z_{Fe}	(1.9)	Stellar metallicity
Disk Quantities		
D	(1.27)	Dissipation (i.e., heating) rate, per area
F	(1.1)	Disk mass relative to MMSN
H	(1.22)	Disk scaleheight
J [$\dot{J}$]	(1.19)	Angular momentum [torque]
L_{acc} [L_d]	(1.30) [(1.28)]	Accretion luminosity [part released in the disk]
$\dot{M}$	(1.10)	Accretion rate through disk
Q	(1.43)	Dimensionless measure of gravitational stability
R	(1.1)	Distance from star (cylindrical radius)
R_{in}	(1.28)	Radius of disk inner edge
R_J	(1.26)	Torque on inner disk edge, as length-scale
v_ϕ [v_R]	(1.10)	Orbital [radial] flow velocity
Z_{disk}	(1.2)	Disk "metallicity" as ratio of solids to gas
Z_{rel}	(1.4)	"Metallicity" relative to fiducial Solar value of 0.015
α	(1.23)	Dimensionless angular momentum transport coefficient
α_D	(1.57)	Dimensionless turbulent diffusion coefficient

Table 1-1
(Continued)

Symbol	Ref.	Meaning
η	(1.46)	Fraction by which rotation is slower than Keplerian
θ	(1.35)	Grazing angle of starlight on disk surface
ν	(1.13)	Viscosity, usually "anomalous"
ρ_R	(1.58)	Roche density for gravitational binding
Σ	(1.1), (1.10)	Surface density (mass per area) of disk, e.g., gas (Sect. 3) or planetesimals (Sect. 5.1). Subscripted as needed.
Σ_g [Σ_p]	(1.1) [(1.2)]	Surface density of gas [particle] disks.
$\dot{\Sigma}$	(1.13)	Inflow or outflow of mass from disk, per area
Protoplanet Quantities		
a	(1.62)	Semimajor axis (similar to disk R)
B	(1.73)	Width of feeding zone in R_H
f_G	(1.68) and (1.71)	Gravitational focusing factor for collisional cross section
m [m_c , m_a]	Sect. 5.1 [Sect. 5.2.1]	Protoplanet mass, total or [core, atmosphere]
m_s [m_l]	Sect. 5.1	Mass of small protoplanets, i.e., planetesimals [larger protoplanets]
m_{iso}	(1.73)	Isolation mass
$\dot{m}$ [$\dot{m}_l$]	Sect. 5.1.2	Accretion rate, for [large] protoplanet's mass growth
Q_D^*	(1.79)	Energy (per mass) threshold for catastrophic collisional disruption
r [r_s , r_l , r_c]	Sects. 5.1, Sect. 5.2.1	Radius of protoplanet [small, large, solid core]
r_B	(1.85)	Bondi radius for planet gravity to exceed thermal energy of disk gas
R_H	(1.62)	Hill radius for planet gravity to dominate stellar tides
s	(1.48)	Radius of small grain or planetesimal
t_s [τ_s]	(1.48) [(1.50)]	[Dimensionless] aerodynamic stopping timescale
v [v_s , v_l]	Sect. 5.1	Velocity dispersion of [small, large] protoplanets
v_{esc} [v_H]	(1.65) [(1.64)]	Escape speed from protoplanet's surface
Σ_s [Σ_l]	Sect. 5.1	Surface mass density of small [large] protoplanets
χ	(1.96)	Log ratio of atmosphere to core radius
ψ	(1.63)	r/R_H, radius in Hill units

- Oligarchs reach chaotic growth faster at 5 AU than at 1 AU. From the discussion in Sect. 5.1.2, icy oligarchs at 5 AU have larger isolation masses than rocky oligarchs at 1 AU (1.73). Thus, their gravitational interactions are stronger and lead to chaotic growth sooner.
- While rocky planets scatter low-mass protoplanets a few tenths of an AU, gas giant planets scatter some low-mass protoplanets close to the host star and eject many others from the planetary system. Fortunately, the solar system avoided either outcome. However, many exoplanets close to their host stars have large *e*. Although orbital migration probably accounts for exoplanets with small *a* and nearly circular orbits (See Chap. 2 by

Morbidelli), producing super-Earths with $a \lesssim 0.4$ AU and $e \gtrsim 0.1$ probably requires planet-planet scattering, as in *Fig. 1-11* (Jurić and Tremaine 2008; Raymond et al. 2011). In a few years, exoplanet statistics will allow critical tests of the ability of migration and scattering to explain the observed (a, e) distribution.

5.3 Direct Formation of Brown Dwarfs and Gas Giants

Though the core accretion is remarkably successful at explaining the diversity of planetary systems, it cannot be proved. It is thus useful to develop alternate theories. Ideally each theory would make testable predictions, but consistency with physics and existing astronomical observations also provides stringent constraints. Here we consider the leading alternate theory that a gravitationally unstable gas disk might fragment into bound objects that survive as gas giant planets (Kuiper 1951; Cameron 1978; Bodenheimer et al. 1980; Boss 2000). This theory should not be confused with the hypothesis that the solid component of the proplanetary disk gravitationally collapses into planetesimals (Safronov 1969; Goldreich and Ward 1973; Youdin and Shu 2002, Sect. 4.2).

Sect. 3.6 introduced the idea that disks are gravitationally unstable when the Toomre (1964) criterion,

$$Q = \frac{c_s \Omega}{\pi G \Sigma_g} \lesssim 1 \,, \tag{1.101}$$

is satisfied for a surface mass density, Σ. There, we outlined basic constraints on the accretion rate and temperature for stable disks. For a fragment to survive, it must cool quickly, so that it contracts on an orbital timescale. Analytic theory suggests bound fragments are possible but have the typical masses of brown dwarfs (Rafikov 2005; Kratter et al. 2010b); so far, numerical simulations are inconclusive (D'Angelo et al. 2010; Cai et al. 2010).

To outline the basic issues for gravitational instability, we consider a density perturbation in a disk which satisfies the Toomre Q criterion. For this fragment to continue to contract, it must cool on a sufficiently short timescale (Gammie 2001),

$$t_c < \xi \Omega^{-1} \,. \tag{1.102}$$

For likely conditions within a protoplanetary disk, numerical simulations suggest that the critical value of $\xi \approx 3$ (Gammie 2001; Rice et al. 2003), with uncertainties due to the equation of state and the opacity. Fragments that cannot cool on the orbital timescale will be sheared apart. For the cooling time in a viscous disk with an α-viscosity (1.23), this constraint places a limit on α:

$$\alpha \gtrsim 4/[9\xi\gamma(\gamma - 1)] \sim 0.3 \,. \tag{1.103}$$

Fragmenting disks also have very high accretion rates.

For the inner disk ($R < 10$ AU), combining the Toomre ($Q \lesssim 1$) and cooling criteria yields fragments only in hot, massive disks (Rafikov 2005; Matzner and Levin 2005). With disk masses comparable to the stellar mass, these conditions probably produce a bound stellar companion instead of a lower mass planet. Although colder, lower mass disks can have $Q \lesssim 1$; their slow cooling times do not allow the fragment to survive.

Forming planet-mass fragments in the outer disk, at 50–100 AU, is more attractive. Irradiation (instead of viscous transport) then dominates the energy budget of the disk

(D'Alessio et al. 1998). Although estimates for the cooling time are more complicated, an irradiated disk is thought to fragment more readily once it reaches $Q \lesssim 1$ (Rice et al. 2011; Kratter and Murray-Clay 2011).

In any part of the disk, the initial mass of a gravitational fragment is typically $\sim \Sigma_g H_g^2$ in terms of the disk scale-height H_g. For the most optimistic assumptions about cooling – specifically that the optical depth $\tau = 1$ – fragment masses are $\gtrsim 5$ M_J at 100 AU (Rafikov 2005). Less efficient cooling ($\tau \neq 1$) and closer orbital separations increase the fragment mass. These minimum masses require very cold disks ($T \lesssim 10$ K), which might be achievable in the outer portions of disks in low-mass star-forming regions like Ophiuchus (Andrews et al. 2009).

While making a fragment with a mass below ~10 M_J is challenging, keeping the fragment mass low is an even more difficult problem. Disks are likely to fragment before infall from the surrounding molecular cloud ceases. Preventing the disk fragment from accreting this infalling mass is a challenge. As explained in ❯ Sect. 5.2.4, stopping the flow of disk gas onto a planet requires a low-mass and fairly quiescent disk, exactly opposite to the conditions required for an unstable disk. Kratter et al. (2010b) quantify these issues and conclude that disks in nearby star-forming regions are more likely to produce 25–75-M_J brown dwarfs than 1–10-M_J planets. The mass problem might be helped if fragments cool just rapidly enough to remain bound. Then if they migrate inward, they would overflow their Roche lobes and be "tidally downsized" (Boley et al. 2010; Nayakshin 2010). This intriguing suggestion cannot yet be considered a solution.

Even if bound Jupiter-mass fragments form, understanding the time evolution of fragments within a $Q \sim 1$ disk requires sophisticated numerical calculations. In smooth particle hydrodynamics (SPH), an ensemble of particles represents the gas; each particle has a set of physical properties and responds to gravity, radiation pressure, and other forces (Benz et al. 1986; Monaghan 1992). Grid-based calculations lay out a set of points where the physical conditions of the gas are specified; solving a set of coupled differential equations yields the time evolution of the conditions at each point (Black and Bodenheimer 1975; Tohline 1980; Pickett et al. 1998).

In both approaches, disks close to the stability limit develop multiarm spiral structure (Kratter et al. 2010a; Mayer et al. 2002). Spiral modes generate turbulence throughout the disk (Nelson et al. 1998; Gammie 2001). The large amplitudes of these modes create a rippled surface above the disk midplane (Durisen et al. 2001). When the cooling rate is near the critical value ξ, the spiral structure maintains a rough equilibrium, where the amplitude of the spiral modes increases as the cooling rate declines (Mejía et al. 2005; Cai et al. 2006).

When the cooling time is smaller than the local rotational period, as in (❯ 1.102), the disk fragments (Gammie 2001; Johnson and Gammie 2003; Rice et al. 2003). For astrophysically relevant conditions, fragmentation requires the disk mass be at least 10% of the stellar mass. As the disk mass grows, spiral waves propagate throughout the disk. Within a few rotation periods, fragments form in the densest portions of the spirals (Boley et al. 2010). If the fragments continue to cool rapidly, they grow in mass and become bound objects; otherwise, they are sheared apart (Mayer et al. 2004).

The long-term evolution of bound fragments in self-gravitating disks is unclear. Because the numerical calculations involve such a wide range of scales, none can evolve a bound fragment long enough to determine its final fate. If the cooling time is short, and accretion from the disk is inefficient, bound fragments could become gas giant planets or more likely brown dwarfs. Such objects might also grow to full-fledged stellar companions (Bonnell and Bate 1994). Overcoming the numerical limits on the calculations requires faster computers and innovative techniques to follow the evolution of fragments in a time-dependent disk.

6 Summary

In the last decade or two, observations have revolutionized our understanding of planetary astronomy. In the 1990s, the thrilling discovery of the Kuiper belt nearly doubled the empirical size of the disk of the solar system (Jewitt and Luu 1993). A few years later, Mayor and Queloz (1995) discovered an extraordinary exoplanet orbiting only 0.05 AU from the solar-type star 51 Peg. Now, the number of known Kuiper belt objects easily exceeds 1,000, including a grand variety of dynamical and taxonomic classes that place interesting constraints on the origin and early evolution of the protosolar nebula (Barucci et al. 2008). Although there are only ~1,000 confirmed exoplanets, data from *Kepler* and many ground-based programs will certainly push the count past 10,000 in the next decade. Some of these will certainly challenge current ideas about planet formation.

Throughout this onslaught, theorists responded quickly with new ideas (dead zones in disks, migration, symplectic N-body codes) and variants on old ideas (collisional cascades, disk instabilities, multiannulus coagulation codes). Rapid developments in computing hardware fueled many advances; new analytical approaches drove others.

Today, we have a good basic theory of disk evolution. Despite uncertainties about the initial mass and temperature distribution and the origin of disk viscosity, analytic and numerical disk models provide a good framework for interpreting observations and for exploring the origin of planetary systems. Current research involves combining more elaborate versions of the basic theory (Sect. 3) with detailed models for the chemical evolution of disk material. Within the next decade, these investigations should improve our insight into the overall evolution of the disk and the growth and composition of dust grains with sizes ~1–10 mm.

We also understand how kilometer-sized or larger planetesimals become planets (Sect. 5). Although there are major uncertainties about the onset and the end of gas accretion and the interactions of massive planets with the gas disk, analytic theory and numerical simulations demonstrate that – on 1–10-Myr timescales – ensembles of planetesimals can evolve into terrestrial and gas giant planets within ~50 AU of the central star. Comparisons of observations with the predicted masses and orbital properties of planets and the predicted dust masses and luminosity evolution of debris disks are promising and will eventually produce stringent tests of the theory.

Despite these successes, we are still in search of a robust theory for planetesimal formation (Sect. 4). Excellent progress on the meter-sized barrier isolates the importance of radial drift and the problems associated with direct coagulation models. Some type of instability – either by direct gravitational collapse or a concentration mechanism such as the streaming instability – seems necessary to produce planetesimals on fast timescales. Larger numerical simulations will undoubtedly yield a better understanding of these instabilities. Exploring other physical mechanisms for particle growth and evolution is also essential.

Understanding fragmentation in protostellar disks promises to unify our understanding of star and planet formation. Although many physical and numerical issues remain unresolved, fragmentation is a promising way to produce gas giants and brown dwarfs at $\gtrsim$50–100 AU from the parent star. Despite the current lack of large samples of planets in this domain, direct imaging surveys are starting to discover 1–30-M_J objects with $a \sim$ 10–100 AU (Kalas et al. 2008; Marois et al. 2008; Lagrange et al. 2010; Currie et al. 2010; Kraus et al. 2011). With large uncertainties in model atmospheres, assessing the formation mechanism of these objects is difficult. Once large samples of planets and brown dwarfs at 10–100 AU are available, their properties will allow a robust assessment of the core accretion and disk instability mechanisms.

Acknowledgments

We thank Ben Bromley, Margaret Geller, Paul Kalas, and Kaitlin Kratter for advice and comments on the manuscript. Portions of this project were supported by *NASA's Astrophysics Theory Program* and the *Origin of Solar Systems Program* through grant NNX10AF35G and by endowment funds of the Smithsonian Institution.

References

Adachi, I., Hayashi, C., & Nakazawa, K. 1976, Prog. Theor. Phys., 56, 1756
Adams, F. C., & Shu, F. H. 1986, ApJ, 308, 836
Adams, F. C., Ruden, S. P., & Shu, F. H. 1989, ApJ, 347, 959
Alexander, R. D., & Armitage, P. J. 2009, ApJ, 704, 989
Andrews, S. M., & Williams, J. P. 2005, ApJ, 631, 1134
Andrews, S. M., Wilner, D. J., Hughes, A. M., Qi, C., & Dullemond, C. P. 2009, ApJ, 700, 1502
Artymowicz, P. 1988, ApJ, 335, L79
Asphaug, E., & Benz, W. 1996, Icarus, 121, 225
Backman, D. E., & Paresce, F. 1993, in Protostars and Planets III, ed. E. H. Levy & J. I. Lunine (Tucson: University of Arizona Press), 1253–1304
Bai, X., & Stone, J. M. 2010a, ApJ, 722, 1437
Bai, X., & Stone, J. M. 2010b, ApJ, 722, L220
Balbus, S. A., & Hawley, J. F. 1991, ApJ, 376, 214
Balsara, D. S., Tilley, D. A., Rettig, T., & Brittain, S. D. 2009, MNRAS, 397, 24
Barucci, M. A., Boehnhardt, H., Cruikshank, D. P., Morbidelli, A., & Dotson, R. 2008, The Solar System Beyond Neptune (Tucson: The University of Arizona Press)
Bath, G. T., & Pringle, J. E. 1982, MNRAS, 199, 267
Bell, K. R., & Lin, D. N. C. 1994, ApJ, 427, 987
Benecchi, S. D., Noll, K. S., Grundy, W. M., et al. 2009, Icarus, 200, 292
Benz, W., & Asphaug, E. 1999, Icarus, 142, 5
Benz, W., Slattery, W. L., & Cameron, A. G. W. 1986, Icarus, 66, 515
Bernstein, G. M., Trilling, D. E., Allen, R. L., et al. 2004, AJ, 128, 1364
Binney, J., & Tremaine, S. 2008, Galactic Dynamics (2nd ed.; Princeton: Princeton University Press)
Birnstiel, T., Dullemond, C. P., & Brauer, F. 2010, A&A, 513, A79+
Black, D. C., & Bodenheimer, P. 1975, ApJ, 199, 619
Blum, J., & Wurm, G. 2000, Icarus, 143, 138
Blum, J., & Wurm, G. 2008, ARA&A, 46, 21
Bodenheimer, P., Grossman, A. S., Decampli, W. M., Marcy, G., & Pollack, J. B. 1980, Icarus, 41, 293
Boley, A. C., Hayfield, T., Mayer, L., & Durisen, R. H. 2010, Icarus, 207, 509
Bonanno, A., Schlattl, H., & Paternò, L. 2002, A&A, 390, 1115
Bonnell, I. A., & Bate, M. R. 1994, MNRAS, 269, L45
Boss, A. P. 2000, ApJ, 536, L101
Bouvier, J., Alencar, S. H. P., Harries, T. J., Johns-Krull, C. M., & Romanova, M. M. 2007, in Protostars and Planets V (Tucson: University of Arizona Press), 479
Bromley, B. C., & Kenyon, S. J. 2006, AJ, 131, 2737
Bromley, B. C., & Kenyon, S. J. 2011, ApJ, 731, 101
Brush, S. G. 1990, Rev. Mod. Phys., 62, 43
Burns, J. A., Lamy, P. L., & Soter, S. 1979, Icarus, 40, 1
Cai, K., Durisen, R. H., Michael, S., et al. 2006, ApJ, 636, L149
Cai, K., Pickett, M. K., Durisen, R. H., & Milne, A. M. 2010, ApJ, 716, L176
Cameron, A. G. W. 1962, Icarus, 1, 13
Cameron, A. G. W. 1978, Moon Planet, 18, 5
Canup, R. M. 2008, Icarus, 196, 518
Canup, R. M. 2011, AJ, 141, 35
Carballido, A., Fromang, S., & Papaloizou, J. 2006, MNRAS, 373, 1633
Cassen, P., & Moosman, A. 1981, Icarus, 48, 353
Chambers, J. E. 2001, Icarus, 152, 205
Chambers, J. 2008, Icarus, 198, 256
Chambers, J. E. 2009, ApJ, 705, 1206
Chambers, J. E., & Wetherill, G. W. 1998, Icarus, 136, 304
Charnoz, S., & Morbidelli, A. 2007, Icarus, 188, 468
Chavanis, P. H. 2000, A&A, 356, 1089
Chiang, E. I., & Goldreich, P. 1997, ApJ, 490, 368
Chiang, E., & Youdin, A. N. 2010, Annu. Rev. Earth Planet. Sci., 38, 493
Chokshi, A., Tielens, A. G. G. M., & Hollenbach, D. 1993, ApJ, 407, 806
Connolly, H. C., Jr., Desch, S. J., Ash, R. D., & Jones, R. H. 2006, in Meteorites and the Early Solar System II, ed. D. S. Lauretta & H. Y. McSween (Tucson: University of Arizona Press), 383–397
Connelly, J. N., Amelin, Y., Krot, A. N., & Bizzarro, M. 2007, in Workshop on the Chronology of Meteorites and the Early Solar System (Houston: Lunar and Planetary Institute), 46–47

Cumming, A., Butler, R. P., Marcy, G. W., et al. 2008, PASP, 120, 531
Currie, T., & Sicilia-Aguilar, A. 2011, ApJ, 732, 24
Currie, T., Kenyon, S. J., Balog, Z., et al. 2008, ApJ, 672, 558
Currie, T., Bailey, V., Fabrycky, D., et al. 2010, ApJ, 721, L177
Cuzzi, J. N., Dobrovolskis, A. R., & Champney, J. M. 1993, Icarus, 106, 102
Cuzzi, J. N., Hogan, R. C., Paque, J. M., & Dobrovolskis, A. R. 2001, ApJ, 546, 496
Cuzzi, J. N., Hogan, R. C., & Shariff, K. 2008, ApJ, 687, 1432
D'Alessio, P., Canto, J., Calvet, N., & Lizano, S. 1998, ApJ, 500, 411
D'Angelo, G., & Lubow, S. H. 2008, ApJ, 685, 560
D'Angelo, G., Henning, T., & Kley, W. 2003, ApJ, 599, 548
D'Angelo, G., Durisen, R. H., & Lissauer, J. J. 2010, in Exoplanets, ed. S. Seager (Tucson: University of Arizona Press), 319–346
Dauphas, N., & Chaussidon, M. 2011, Annu. Rev. Earth Planet. Sci., 39, 351
Davis, D. R., Chapman, C. R., Weidenschilling, S. J., & Greenberg, R. 1985, Icarus, 63, 30
Dohnanyi, J. S. 1969, J. Geophys. Res., 74, 2531
Dones, L., & Tremaine, S. 1993, Icarus, 103, 67
Durda, D. D., Bottke, W. F., Enke, B. L., et al. 2004, Icarus, 170, 243
Durisen, R. H., Mejia, A. C., Pickett, B. K., & Hartquist, T. W. 2001, ApJ, 563, L157
Espaillat, C., Ingleby, L., Hernández, J., et al. 2012, ApJ, 747, 103
Fernandez, J. A., & Ip, W.-H. 1984, Icarus, 58, 109
Fischer, D. A., & Valenti, J. 2005, ApJ, 622, 1102
Fraser, W. C. 2009, ApJ, 706, 119
Friedjung, M. 1985, A&A, 146, 366
Fromang, S., & Stone, J. M. 2009, A&A, 507, 19
Gammie, C. F. 1996, ApJ, 457, 355
Gammie, C. F. 2001, ApJ, 553, 174
Ghosh, P., & Lamb, F. K. 1979, ApJ, 232, 259
Gladman, B. 1993, Icarus, 106, 247
Goldreich, P., & Lynden-Bell, D. 1965, MNRAS, 130, 97
Goldreich, P., & Tremaine, S. D. 1978, Icarus, 34, 227
Goldreich, P., & Ward, W. R. 1973, ApJ, 183, 1051
Goldreich, P., Lithwick, Y., & Sari, R. 2002, Nature, 420, 643
Goldreich, P., Lithwick, Y., & Sari, R. 2004, ARA&A, 42, 549
Gómez, G. C., & Ostriker, E. C. 2005, ApJ, 630, 1093
Gonzalez, G. 1997, MNRAS, 285, 403
Goodman, J., & Pindor, B. 2000, Icarus, 148, 537
Gould, A., Dong, S., Gaudi, B. S., et al. 2010, ApJ, 720, 1073
Gradie, J., & Tedesco, E. 1982, Science, 216, 1405
Greenberg, R., Weidenschilling, S. J., Chapman, C. R., & Davis, D. R. 1984, Icarus, 59, 87
Greenberg, R., Bottke, W. F., Carusi, A., & Valsecchi, G. B. 1991, Icarus, 94, 98
Greenzweig, Y., & Lissauer, J. J. 1990, Icarus, 87, 40
Grogan, K., Dermott, S. F., & Durda, D. D. 2001, Icarus, 152, 251
Harris, A. W. 1978, in Lunar and Planetary Institute Science Conference Abstracts, Vol. 9 (Houston: Lunar and Planetary Institute), 459–461
Hartmann, L., & Kenyon, S. J. 1996, ARA&A, 34, 207
Hayashi, C. 1981, Prog. Theor. Phys. Supp., 70, 35
Herbig, G. H. 1962, Adv. Astron. Astrophys., 1, 47
Hewins, R. H. 1996, in Chondrules and the Protoplanetary Disk, ed. R. H. Hewins, R. H. Jones, & E. R. D. Scott (New York: Cambridge University Press), 3–9
Hill, G. 1878, Am. J. Math., 1, 5, 129, 245
Holsapple, K. A. 1994, Planet. Space Sci., 42, 1067
Holsapple, K., Giblin, I., Housen, K., Nakamura, A., & Ryan, E. 2002, in Asteroids III, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 443–462
Hori, Y., & Ikoma, M. 2011, MNRAS, 416, 1419
Housen, K. R., & Holsapple, K. A. 1990, Icarus, 84, 226
Housen, K. R., & Holsapple, K. A. 1999, Icarus, 142, 21
Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, ApJS, 201, 15
Hubickyj, O., Bodenheimer, P., & Lissauer, J. J. 2005, Icarus, 179, 415
Hueso, R., & Guillot, T. 2005, A&A, 442, 703
Ida, S., & Makino, J. 1993, Icarus, 106, 210
Ikoma, M., Nakazawa, K., & Emori, H. 2000, ApJ, 537, 1013
Inaba, S., & Ikoma, M. 2003, A&A, 410, 711
Isella, A., Carpenter, J. M., & Sargent, A. I. 2009, ApJ, 701, 260
Jacquet, E., Balbus, S., & Latter, H. 2011, MNRAS, 415, 3591
Jeffreys, H. 1929, Observatory, 52, 173
Jewitt, D., & Luu, J. 1993, Nature, 362, 730
Jewitt, D., Chizmadia, L., Grimm, R., & Prialnik, D. 2007, in Protostars and Planets V (Tucson: University of Arizona Press), 863
Johansen, A., & Youdin, A. 2007, ApJ, 662, 627
Johansen, A., Klahr, H., & Henning, T. 2006, ApJ, 636, 1121
Johansen, A., Oishi, J. S., Mac Low, M.-M., et al. 2007, Nature, 448, 1022
Johansen, A., Youdin, A., & Klahr, H. 2009a, ApJ, 697, 1269
Johansen, A., Youdin, A., & Mac Low, M. 2009b, ApJ, 704, L75

Johansen, A., Youdin, A. N., & Lithwick, Y. 2012, A&A, 537, A125
Johnson, B. M., & Gammie, C. F. 2003, ApJ, 597, 131
Johnson, J. A., Aller, K. M., Howard, A. W., & Crepp, J. R. 2010, PASP, 122, 905
Johnson, J. A., Payne, M., Howard, A. W., et al. 2011, AJ, 141, 16
Joy, A. H. 1945, ApJ, 102, 168
Jurić, M., & Tremaine, S. 2008, ApJ, 686, 603
Kalas, P., Graham, J. R., Chiang, E., et al. 2008, Science, 322, 1345
Kennedy, G. M., & Kenyon, S. J. 2008, ApJ, 673, 502
Kenyon, S. J., & Bromley, B. C. 2002, ApJ, 577, L35
Kenyon, S. J., & Bromley, B. C. 2004a, AJ, 127, 513
Kenyon, S. J., & Bromley, B. C. 2004b, ApJ, 602, L133
Kenyon, S. J., & Bromley, B. C. 2004c, AJ, 128, 1916
Kenyon, S. J., & Bromley, B. C. 2005, AJ, 130, 269
Kenyon, S. J., & Bromley, B. C. 2008, ApJS, 179, 451
Kenyon, S. J., & Bromley, B. C. 2009, ApJ, 690, L140
Kenyon, S. J., & Bromley, B. C. 2010, ApJS, 188, 242
Kenyon, S. J., & Hartmann, L. 1987, ApJ, 323, 714
Kenyon, S. J., & Luu, J. X. 1998, AJ, 115, 2136
Kenyon, S. J., & Luu, J. X. 1999, AJ, 118, 1101
Kenyon, S. J., Yi, I., & Hartmann, L. 1996, ApJ, 462, 439
Kenyon, S. J., Bromley, B. C., O'Brien, D. P., & Davis, D. R. 2008a, in The Solar System Beyond Neptune, ed. M. A. Barucci, H. Boehnhardt, D. P. Cruikshank, A. Morbidelli, & R. Dotson (Tucson: University of Arizona Press), 293–313
Kenyon, S. J., Gómez, M., & Whitney, B. A. 2008b, in Handbook of Star Forming Regions, Vol. I, ed. B. Reipurth (San Francisco: Astronomical Society of the Pacific), 405–+
Kleine, T., Mezger, K., Palme, H., Scherer, E., & Münker, C. 2005, Geochim. Cosmochim. Acta, 69, 5805
Kleine, T., Touboul, M., Bourdon, B., et al. 2009, Geochim. Cosmochim. Acta, 73, 5150
Kobayashi, H., & Tanaka, H. 2010, Icarus, 206, 735
Kokubo, E., & Ida, S. 1995, Icarus, 114, 247
Kokubo, E., & Ida, S. 1998, Icarus, 131, 171
Kokubo, E., Kominami, J., & Ida, S. 2006, ApJ, 642, 1131
Kratter, K. M., & Murray-Clay, R. A. 2011, ApJ, 740, 1
Kratter, K. M., Matzner, C. D., Krumholz, M. R., & Klein, R. I. 2010a, ApJ, 708, 1585
Kratter, K. M., Murray-Clay, R. A., & Youdin, A. N. 2010b, ApJ, 710, 1375
Kraus, A. L., Ireland, M. J., Martinache, F., & Hillenbrand, L. A. 2011, ApJ, 731, 8
Krot, A., Amelin, Y., Bizzarro, M., et al. 2007, in Workshop on the Chronology of Meteorites and the Early Solar System (Houston: Lunar and Planetary Institute), 98–99
Kuiper, G. P. 1951, Proc. Natl. Acad. Sci., 37, 1
Lagrange, A.-M., Bonnefoy, M., Chauvin, G., et al. 2010, Science, 329, 57
Lee, A. T., Chiang, E., Asay-Davis, X., & Barranco, J. 2010, ApJ, 725, 1938
Leinhardt, Z. M., & Richardson, D. C. 2005, ApJ, 625, 427
Leinhardt, Z. M., & Stewart, S. T. 2009, Icarus, 199, 542
Leinhardt, Z. M., Stewart, S. T., & Schultz, P. H. 2008, in The Solar System Beyond Neptune, ed. M. A. Barucci, H. Boehnhardt, D. P. Cruikshank, A. Morbidelli, & R. Dotson (Tucson: University of Arizona Press), 195–211
Lin, D. N. C., & Pringle, J. E. 1987, MNRAS, 225, 607
Lin, D. N. C., & Pringle, J. E. 1990, ApJ, 358, 515
Lissauer, J. J. 1987, Icarus, 69, 249
Lissauer, J. J., Hubickyj, O., D'Angelo, G., & Bodenheimer, P. 2009, Icarus, 199, 338
Lithwick, Y. 2009, ApJ, 693, 85
Lodato, G., & Rice, W. K. M. 2004, MNRAS, 351, 630
Lodders, K. 2003, ApJ, 591, 1220
Lynden-Bell, D., & Pringle, J. E. 1974, MNRAS, 168, 603
Markiewicz, W. J., Mizuno, H., & Voelk, H. J. 1991, A&A, 242, 286
Marois, C., Macintosh, B., Barman, T., et al. 2008, Science, 322, 1348
Matzner, C. D., & Levin, Y. 2005, ApJ, 628, 817
Maxey, M. R. 1987, J. Fluid Mech., 174, 441
Mayer, L., Quinn, T., Wadsley, J., & Stadel, J. 2002, Science, 298, 1756
Mayer, L., Quinn, T., Wadsley, J., & Stadel, J. 2004, ApJ, 609, 1045
Mayor, M., & Queloz, D. 1995, Nature, 378, 355
Mejía, A. C., Durisen, R. H., Pickett, M. K., & Cai, K. 2005, ApJ, 619, 1098
Mizuno, H. 1980, Prog. Theor. Phys., 64, 544
Monaghan, J. J. 1992, ARA&A, 30, 543
Morbidelli, A., Levison, H. F., & Gomes, R. 2008, in The Solar System Beyond Neptune (Tucson: The University of Arizona Press), 275–292
Morbidelli, A., Bottke, W. F., Nesvorný, D., & Levison, H. F. 2009a, Icarus, 204, 558
Morbidelli, A., Bottke, W. F., Nesvorný, D., & Levison, H. F. 2009b, Icarus, 204, 558
Nagasawa, M., Lin, D. N. C., & Thommes, E. 2005, ApJ, 635, 578
Najita, J. R., Carr, J. S., Glassgold, A. E., & Valenti, J. A. 2007, in Protostars and Planets V (Tucson: University of Arizona Press), 507
Nakagawa, Y., Sekiya, M., & Hayashi, C. 1986, Icarus, 67, 375
Nayakshin, S. 2010, MNRAS, 408, L36
Nelson, R. P., & Papaloizou, J. C. B. 2004, MNRAS, 350, 849

Nelson, A. F., Benz, W., Adams, F. C., & Arnett, D. 1998, ApJ, 502, 342
Nesvorný, D., Bottke, W. F., Levison, H. F., & Dones, L. 2003, ApJ, 591, 486
Nesvorný, D., Youdin, A. N., & Richardson, D. C. 2010, AJ, 140, 785
Nesvorný, D., Vokrouhlický, D., Bottke, W. F., Noll, K., & Levison, H. F. 2011, AJ, 141, 159
Noll, K. S., Grundy, W. M., Chiang, E. I., Margot, J.-L., & Kern, S. D. 2008, in The Solar System Beyond Neptune, (Tucson: The University of Arizona Press), 345–363
Ohtsuki, K. 1992, Icarus, 98, 20
Ohtsuki, K., Stewart, G. R., & Ida, S. 2002, Icarus, 155, 436
Paczynski, B. 1978, Acta Astron., 28, 91
Pan, M., & Sari, R. 2005, Icarus, 173, 342
Papaloizou, J. C. B., & Nelson, R. P. 2003, MNRAS, 339, 983
Papaloizou, J. C. B., & Nelson, R. P. 2005, A&A, 433, 247
Parker, A. H., & Kavelaars, J. J. 2012, ApJ, 744, 139
Pasquini, L., Döllinger, M. P., Weiss, A., et al. 2007, A&A, 473, 979
Payne, C. H. 1925, PhD thesis, Radcliffe College
Petit, J., & Henon, M. 1986, Icarus, 66, 536
Petit, J.-M., Kavelaars, J. J., Gladman, B., & Loredo, T. 2008, in The Solar System Beyond Neptune (Tucson: The University of Arizona Press), 71–87
Pickett, B. K., Cassen, P., Durisen, R. H., & Link, R. 1998, ApJ, 504, 468
Pollack, J. B., McKay, C. P., & Christofferson, B. M. 1985, Icarus, 64, 471
Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, Icarus, 124, 62
Porco, C. C., Weiss, J. W., Richardson, D. C., et al. 2008, AJ, 136, 2172
Pringle, J. E. 1981, ARA&A, 19, 137
Pudritz, R. E., Ouyed, R., Fendt, C., & Brandenburg, A. 2007, in Protostars and Planets V (Tucson: University of Arizona Press), 277
Rafikov, R. R. 2004, AJ, 128, 1348
Rafikov, R. R. 2005, ApJ, 621, L69
Rafikov, R. R. 2006, ApJ, 648, 666
Rafikov, R. R. 2011, ApJ, 727, 86
Raymond, S. N., Quinn, T., & Lunine, J. I. 2004, Icarus, 168, 1
Raymond, S. N., Armitage, P. J., Moro-Martín, A., et al. 2011, A&A, 530, A62
Rice, W. K. M., Armitage, P. J., Bonnell, I. A., et al. 2003, MNRAS, 346, L36
Rice, W. K. M., Lodato, G., Pringle, J. E., Armitage, P. J., & Bonnell, I. A. 2006, MNRAS, 372, L9
Rice, W. K. M., Armitage, P. J., Mamatsashvili, G. R., Lodato, G., & Clarke, C. J. 2011, MNRAS, 418, 1356
Rivkin, A. S., Howell, E. S., Vilas, F., & Lebofsky, L. A. 2002, in Asteroids III (Tucson: University of Arizona Press), 235
Russell, S. S., Hartmann, L., Cuzzi, J., et al. 2006, in Meteorites and the Early Solar System II (Tucson: University of Arizona Press), 233–251
Safronov, V. S. 1960, Annales d'Astrophysique, 23, 979
Safronov, V. S. 1969, Evoliutsiia doplanetnogo oblaka. (Evolution of the Protoplanetary Cloud and Formation of the Earth and Planets), Nauka, Moscow (Translation 1972, NASA TT F-677)
Saumon, D., & Guillot, T. 2004, ApJ, 609, 1170
Saumon, D., Chabrier, G., & van Horn, H. M. 1995, ApJS, 99, 713
Sekiya, M. 1983, Prog. Theor. Phys., 69, 1116
Sekiya, M. 1998, Icarus, 133, 298
Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337
Shu, F. H. 1992, Physics of Astrophysics, Vol. II (Mill Valley: University Science Books)
Shu, F. H., Adams, F. C., & Lizano, S. 1987, ARA&A, 25, 23
Shu, F., Najita, J., Ostriker, E., et al. 1994, ApJ, 429, 781
Shu, F. H., Najita, J. R., Shang, H., & Li, Z. 2000, in Protostars and Planets IV (Tucson: University of Arizona Press), 789
Sorby, H. C. 1863, R. Soc. Lond. Proc. Ser. I, 13, 333
Sousa, S. G., Santos, N. C., Mayor, M., et al. 2008, A&A, 487, 373
Spaute, D., Weidenschilling, S. J., Davis, D. R., & Marzari, F. 1991, Icarus, 92, 147
Stepinski, T. F. 1998, Icarus, 132, 100
Stepinski, T. F., & Valageas, P. 1996, A&A, 309, 301
Stevenson, D. J. 1982, Planet. Space Sci., 30, 755
Stevenson, D. J., & Lunine, J. I. 1988, Icarus, 75, 146
Stone, J. M., & Balbus, S. A. 1996, ApJ, 464, 364
Supulver, K. D., Bridges, F. G., Tiscareno, S., Lievore, J., & Lin, D. N. C. 1997, Icarus, 129, 539
Takeuchi, T., & Artymowicz, P. 2001, ApJ, 557, 990
Terebey, S., Shu, F. H., & Cassen, P. 1984, ApJ, 286, 529
Tohline, J. E. 1980, ApJ, 235, 866
Toomre, A. 1964, ApJ, 139, 1217
Turner, N. J., Carballido, A., & Sano, T. 2010, ApJ, 708, 188
Ward, W. R. 1976, in Frontiers of Astrophysics, ed. E. H. Avrett (Cambridge: Harvard University Press), 1–40
Ward, W. R. 2000, in Origin of the Earth and Moon (Tucson: University of Arizona Press), 75–84
Warner, B. 1995, in Cambridge Astrophysics Series, 28 (Cambridge: Cambridge University Press)
Weidenschilling, S. J. 1977a, MNRAS, 180, 57
Weidenschilling, S. J. 1977b, Ap&SS, 51, 153

Weidenschilling, S. J. 1980, Icarus, 44, 172
Weidenschilling, S. J. 1995, Icarus, 116, 433
Weidenschilling, S. J. 2010, in Lunar and Planetary Institute Science Conference Abstracts, Vol. 41 (Houston: Lunar and Planetary Institute), 1453
Weidenschilling, S. J., Spaute, D., Davis, D. R., Marzari, F., & Ohtsuki, K. 1997, Icarus, 128, 429
Weidling, R., Güttler, C., & Blum, J. 2012, Icarus, 218, 688
Wetherill, G. W., & Stewart, G. R. 1989, Icarus, 77, 330
Wetherill, G. W., & Stewart, G. R. 1993, Icarus, 106, 190
Whipple, F. L. 1972, in From Plasma to Planet, ed. A. Elvius (Stockholm: Almqvist & Wiksell), 211–+
Williams, J. P., & Cieza, L. A. 2011, ARAA, 49, 67
Williams, D. R., & Wetherill, G. W. 1994, Icarus, 107, 117
Wyatt, M. C. 2008, ARA&A, 46, 339
Youdin, A. N. 2004, in ASP Conf. Ser. 323, Star Formation in the Interstellar Medium (San Francisco: Astronomical Society of the Pacific), 319
Youdin, A. N. 2010, in EAS Publications Series, Vol. 41, ed. T. Montmerle, D. Ehrenreich, & A.-M. Lagrange (Les Ulis: EDP Sciences), 187–207
Youdin, A. N. 2011a, ApJ, 731, 99
Youdin, A. N. 2011b, ApJ, 742, 38
Youdin, A. N., & Chiang, E. I. 2004, ApJ, 601, 1109
Youdin, A. N., & Goodman, J. 2005, ApJ, 620, 459
Youdin, A., & Johansen, A. 2007, ApJ, 662, 613
Youdin, A. N., & Lithwick, Y. 2007, Icarus, 192, 588
Youdin, A. N., & Shu, F. H. 2002, ApJ, 580, 494
Zolensky, M., & McSween, Jr., H. Y. 1988, in Meteorites and the Early Solar System, ed. J. F. Kerridge, & M. S. Matthews (Tucson: University of Arizona Press), 114–143
Zsom, A., Ormel, C. W., Güttler, C., Blum, J., & Dullemond, C. P. 2010, A&A, 513, A57

2 Dynamical Evolution of Planetary Systems

Alessandro Morbidelli
Observatory of Nice, Nice, Cedex 4, France

1 Introduction ***64***

2 The Gas-Disk Era ***65***
2.1 The Formation of the Giant Planets 65
2.2 Once Giant Planets are Formed: Type II Migration and Its Consequences 68
2.3 Planet–Planet Scattering as the Dominant Orbital Excitation Process 72
2.4 A Plausible Evolution of the Four Giant Planets of the Solar System 78

3 The Planetesimal-Disk Era ***81***
3.1 Brief Tutorial of Planetesimal-Driven Migration 81
3.2 Multi-resonant Planet Configurations and Planetesimal Scattering: The Solar System Case 85
3.3 The Late Heavy Bombardment as a Smoking Gun for a Late Instability of the Giant Planets 86
3.4 The Solar System Debris Disk: Are LHBs Common? 90

4 Terrestrial Planets ***93***
4.1 Linking Giant Planet Migration to Terrestrial Planet Accretion: The Grand Tack Scenario 95
4.2 Terrestrial Planets in Extrasolar Systems 97
4.3 Terrestrial-Planets Evolution During Giant Planets Instabilities 98

5 Conclusions ***101***

Acknowledgments ***102***

References ***102***

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_2, © Springer Science+Business Media Dordrecht 2013

Abstract: The apparent regularity of the motion of the giant planets of our solar system suggested for decades that said planets formed onto orbits similar to the current ones and that nothing dramatic ever happened during their lifetime. The discovery of extrasolar planets showed astonishingly that the orbital structure of our planetary system is not typical. Many giant extrasolar planets have orbits with semimajor axes of ~1 AU, and some have even smaller orbital radii, sometimes with orbital periods of just a few days. Moreover, most extrasolar planets have large eccentricities, up to values that only comets have in our solar system. Why is there such a great diversity between our solar system and the extrasolar systems, as well as among the extrasolar systems themselves? This chapter aims to give a partial answer to this fundamental question. Its guideline is a discussion of the evolution of our solar system, certainly biased by a view that emerges, in part, from a series of works comprising the "Nice model." According to this view, the giant planets of the solar system migrated radially while they were still embedded in a protoplanetary disk of gas and presumably achieved a multi-resonant orbital configuration, characterized by smaller interorbital spacings and smaller eccentricities and inclinations with respect to the current configuration. The current orbits of the giant planets may have been achieved during a phase of orbital instability, during which the planets acquired temporarily large-eccentricity orbits and all experienced close encounters with at least one other planet. This instability phase occurred presumably during the putative "Late Heavy Bombardment" of the terrestrial planets, approximately ~3.9 Gy ago (Tera et al. 1974). The interaction with a massive, distant planetesimal disk (the ancestor of the current Kuiper belt) eventually damped the eccentricities of the planets, ending the phase of mutual planetary encounters and parking the planets onto their current, stable orbits. This new view of the evolution of the solar system makes our system not very different from the extrasolar ones. In fact, the best explanation for the large orbital eccentricities of extrasolar planets is that the planets that are observed are the survivors of strong instability phases of original multi-planet systems on quasi-circular orbits. The main difference between the solar system and the extrasolar systems is in the magnitude of such an instability. In the extrasolar systems, encounters among giant planets had to be the norm. In our case, the two major planets (Jupiter and Saturn) never had close encounters with each other: They only encountered "minor" planets like Uranus and/or Neptune. This was probably just mere luck, as simulations show that Jupiter-Saturn encounters in principle could have occurred. Another relevant difference with the extrasolar planets is that, during the gas-disk phase, our giant planets avoided migrating permanently into the inner solar system, thanks to the specific mass ratio of the Jupiter/Saturn pair and the rapid disappearance of the disk soon after the formation of the giant planets. This chapter ends on a note on terrestrial planets. The structure of a terrestrial-planet system depends sensitively on the dynamical evolution of the giant planets and on their final orbits. It appears clear that habitable terrestrial planets, with moderate eccentricity orbits, cannot exist in systems where the giant planets became violently unstable and developed very elliptic orbits. Thus, our very existence is possible only because the instability phase experienced by the giant planets of our solar system was of "moderate" strength.

1 Introduction

It is now clear that observed planetary systems did not form in their current configuration, but they have been heavily modified by a nontrivial dynamical evolution. Processes like planet migration, resonant trapping, planet–planet scattering, mutual collisions, and hyperbolic

ejections have sculpted the structure of planetary systems since the formation of the planets; some of these processes might also have played a crucial role in the accretion of the planets themselves. The observational evidence that planetary systems can be very different from each other suggests that their dynamical evolutions have been very diverse, probably as a result of a strong sensitivity of the dynamics on environmental parameters or initial conditions.

The goal of this chapter is to review our current understanding of the possible dynamical evolutions of planetary systems and of the current open problems. The main focus will be on our solar system, because this is the system that is best modeled, thanks to the vast number of observational constraints. The dynamical history of our planetary system followed stepwise generic processes, but was also characterized by a number of specific "events" that act like bifurcation points in the evolution of planetary systems. Insights are gained by asking what would have happened if these events had occurred differently, thus helping explain the origin of diversity in planetary systems.

This chapter is divided in three parts. The first is devoted to the early evolution of giant planets when they are still embedded in the gas disk, i.e., during the first few millions of years following the formation of the central star. The second discusses the evolution of the giant planets after the disappearance of the gas, when they interact with a still massive planetesimal disk; this is the era of debris disks, which are commonly observed around stars even as old as 1 Gy. The third part will focus on terrestrial planets and on how their accretion and evolution depend on the evolution of the giant planets discussed before.

2 The Gas-Disk Era

2.1 The Formation of the Giant Planets

There are two possible mechanisms by which giant planets can form. The first is nicknamed the "core-accretion mechanism": The coagulation of solid particles forms a core typically of about ten Earth masses ($M_\oplus$) while the gas is still present in the protoplanetary disk; the core then traps by gravity a massive atmosphere of hydrogen and helium from the disk (Pollack et al. 1996) and becomes a giant planet. The second mechanism invokes the gravitational instability of the gaseous component of the disk (Cameron 1978): A cold, massive protoplanetary disk can break into a number of self-gravitating gas clumps, which then contract forming giant gaseous planets (Cassen et al. 1981; Boss 2000, 2001, 2002; see Durisen et al. 2007 for a review).

The debate to discriminate between these two models has been very intense over the last 10 years. Now, several direct or indirect observations suggest that the core-accretion mechanism is predominant for the formation of the planets detected so far. First, interior structure models of the giant planets of the solar system predict that all of them have massive solid cores (Guillot 2005; Militzer and Hubbard 2009; however, see Nettelmann et al. 2008 for a model arguing for a core-less Jupiter). Second, there is a clear correlation between the metallicity of stars and the probability that said stars have giant planets around them (Fischer and Valenti 2005). Third, transiting extrasolar planets are inferred to have solid cores whose relative mass is correlated with the metallicity of the host star (Guillot et al. 2006). All these features suggest that solids have a crucial role in giant planet formation, an aspect that is difficult to explain in the framework of the gravitational instability model (Boss 2002). Moreover, new hydrodynamical simulations that more accurately model thermodynamics in protoplanetary disks find that the formation of

long-lived, self-gravitating clumps of gas is likely only at large distances from the central star (>50–100 AU; Boley 2009). It is still unclear, though, whether the end products of these clumps can be giant planets or must be brown-dwarf-mass objects (Stamatellos and Whitworth 2008).

Thus, there is a growing consensus that the giant planets observed within a few AUs from their parent stars formed by the core-accretion process. The planets found at large distances from their parent stars (for instance, around HR 8799 – Marois et al. 2008 – or Fomalhaut – Kalas et al. 2008) are the best candidates to be the outcome of the gravitational instability process. There is still a possibility, though, that they are giant planets formed closer to the star by core accretion, which subsequently achieved large orbital distances through planet–planet scattering (Veras et al. 2009) or outward migration (Crida et al. 2009).

The core-accretion model, nevertheless, has its own problems. The main difficulty is understanding how a $\sim 10 M_\oplus$ core could form within a few million years (which is the typical survival time of a gas disk; Haisch et al. 2001). In the classical view, these cores form by collisional coagulation from a disk of planetesimals (small bodies of sizes and compositions similar to current asteroids or comets). In this environment, gravity starts to play a fundamental role, bending the trajectories of the colliding objects; this leads to an effective increase of the collisional cross section of the bodies by the so-called *gravitational focussing factor* (Greenzweig and Lissauer 1992). At the beginning, if the planetesimal disk is dynamically very cold (i.e., the orbits have tiny eccentricities and inclinations), the dispersion velocity of the planetesimals $v_{\rm rel}$ may be smaller than the escape velocity of the planetesimals themselves. In this case, a process of *runaway growth* begins, in which the relative mass growth of each object is an increasing function of its own mass M, namely,

$$\frac{1}{M}\frac{\mathrm{d}M}{\mathrm{d}t} \sim \frac{M^{1/3}}{v_{\rm rel}^2}$$

(Greenberg et al. 1978; Wetherill and Stewart 1989). However, as growth proceeds, the disk becomes dynamically heated by the scattering action of the largest bodies. When $v_{\rm rel}$ becomes of the order of the escape velocity from the most massive objects (i.e., $v_{\rm rel} \propto M_{\rm big}^{1/3}$), the runaway growth phase ends and the accretion proceeds in an *oligarchic growth* mode, in which the relative mass growth of the largest objects $\mathrm{d}\log M_{\rm big}/\mathrm{d}t$ is proportional to $M_{\rm big}^{-1/3}$ (Ida and Makino 1993; Kokubo and Ida 1998).

In principle, the combination of runaway and oligarchic growths should continue until the largest objects achieve an *isolation mass*, which is a substantial fraction of the initial total mass of local solids. In the outer solar system, beyond the so-called *snow line*[1] (Podolak and Zucker 2004), if the initial disk is sufficiently massive (about ten times the so-called minimal mass solar nebula or MMSN; Weidenschilling 1977; Hayashi 1981), it is expected that the end result is the formation of a few super-Earths (Thommes et al. 2003; Goldreich et al. 2004; Chambers 2006), as required in the core-accretion model for giant planet formation. N-body simulations, though, show that reality is not so simple. When the cores achieve a mass of about 1 $M_\oplus$, they start to scatter the planetesimals away from their neighborhood, instead of accreting them (Ida and Makino 1993; Levison et al. 2010), which slows their accretion rate significantly.

It has been proposed that gas drag (Wetherill and Stewart 1989) or mutual inelastic collisions (Goldreich et al. 2004) prevent the dispersion of the planetesimals by damping their

[1] The orbital radius beyond which temperature is cold enough that water condenses into ice. The snow line is situated at about 3–5 AU from a solar-mass star, depending on time and on disk models (Min et al. 2011).

orbital eccentricities, but in this case, the cores open gaps in the planetesimal disk (Levison and Morbidelli 2007; Levison et al. 2010), like the satellites Pan and Daphnis open gaps in Saturn's rings. Thus, the cores isolate themselves from the disk of solids. This effectively stops their growth. It has been argued that planet migration (Alibert et al. 2004) or the radial drift of sub-km planetesimals due to gas drag (Rafikov 2004) breaks the isolation of the cores from the disk of solids, but, again, N-body simulations show that the relative drift of planetesimals and cores simply collects the former in resonances with the latter (Levison et al. 2010); this prevents the planetesimals from being accreted by the cores. In fact, only planetesimals smaller than a few tens of meters drift in the disk fast enough to avoid trapping in any resonance with a growing core (Weidenschilling and Davis 1985). Lambrechts and Johansen (2012) argued that, if the mass in the disk is originally dominated by pebbles of a few decimeters, the largest planetesimals formed by the streaming instability process (Johansen et al. 2009) would keep accreting pebbles very efficiently, rapidly growing to several Earth masses.

In summary, the accretion of massive cores is still an open problem. A key uncertainty with synthetic models is that simple formulæ for the mass growth of the cores are made up to achieve the desired result, without any correspondence with the outcomes of N-body simulations (which should be considered as a sort of "ground truth" for dynamical processes).

Another problem of the core-accretion model originates from Type I migration. Type I migration is the label denoting the radial drift of planetary cores (with masses ranging from that of Mars to several Earth masses) due to their gravitational interaction with the gaseous component of the disk. Analytic and numerical studies have shown that a planetary core generates a spiral density wave in the disk (Goldreich and Tremaine 1979, 1980; Ward 1986, 1997; Tanaka et al. 2002). In the outer part of the disk, the wave trails the core. Thus, the gravitational attraction that the wave exerts on the core results in a negative torque that slows the core down. In the inner part of the disk, the wave leads the core, and therefore it exerts on it an acceleration torque. The net effect on the core depends on the balance between these two torques of opposite signs. Ward (1997) showed that in general cases, i.e., for disks with power-law radial density profiles, the negative torque exerted by the wave in the outer disk wins. Consequently, the core has to lose angular momentum, and its orbit shrinks: The planetary core migrates toward the central star, with a speed:

$$da/dt \propto M_p \Sigma_g (a/H)^2 ,$$

where a is the orbital radius of the planet, M_p is its mass, Σ_g is the surface density of the gas disk, and H is its height at the distance a from the central star.

Precise calculations show that an Earth-mass body at 1 AU, in a MMSN with scale height $H/a = 5\%$, migrates into the star in 200,000 year. Thus, planetary cores should fall onto the central star well before they can attain the mass required to capture a massive atmosphere and become giant planets.

Several mechanisms that might weaken or prevent Type I migration have been investigated. First, turbulence may turn inward Type I migration into a random walk (Nelson and Papaloizou 2003; Nelson 2005), which could save at least some of the cores. Second, if there is a steep positive $d\Sigma_g/da$ at some location in the disk, for instance, at the outer edge of a partially depleted central cavity, inward Type I migration should stop there (Masset et al. 2006). Finally, it has been recently shown that migration can be outward in the inner part of the disk, which transports and dissipates heat inefficiently due to its large opacity (Paardekooper and Mellema 2006; Baruteau and Masset 2008; Kley and Crida 2008; Paardekooper et al. 2010; Masset and

Casoli 2010; Bitsch and Kley 2011). In this case, all cores would migrate toward an intermediate region of the disk, where Type I migration is effectively erased (Lyra et al. 2010).

A particularly puzzling aspect of the core-accretion process is the evidence that in our solar system, massive cores of ~10$M_\oplus$ formed in the giant planets region, whereas in the inner solar system, the planetary embryos resulting from the runaway/oligarchic growth process presumably had masses smaller than the mass of Mars (Wetherill and Stewart 1993; Wetherill 1992; Weidenschilling 1977). This jump of two orders of magnitude in the masses of embryos/cores from the inner to the outer solar system is difficult to understand. In fact, the surface density of solids in the disk should have had a "jump" at the snow line of only a factor of ~2, given the revised solar C/O abundance (Lodders 2003). Moreover, the orbital frequency (that sets the speed of all dynamical processes, including accretion) decreases with increasing distance from the star.

So, what makes the outer solar system so favorable for the formation of massive cores? To answer this question, several investigators searched for mechanisms that can concentrate a large amount of solids (well above the initial surface density) in some localized region of the disk, so to achieve a sweet spot for the formation of a massive object. Some proposed mechanisms still give a pivotal role to the snow line (Morfill and Voelk 1984; Ida and Lin 2008), but others are based on the concentration of boulders in long-lived vortices (Barge and Sommeria 1995; Lyra et al. 2009a, b) or on halting migration of planetary embryos at a given orbital radius (Masset et al. 2006; Morbidelli et al. 2008a; Paardekooper and Papaloizou 2009; Sándor et al. 2011; Horn et al. 2012), which are independent of the snow line location. Obviously, more work is needed to understand which mechanism is relevant and dominant. Depending on future results, it might turn out that the widespread expectation that giant planets form by the core-accretion mechanism only beyond the snow line is naive; instead, some giant planets might have formed in the warmer regions of the disk (Bodenheimer et al. 2000).

2.2 Once Giant Planets are Formed: Type II Migration and Its Consequences

By the action-reaction principle, the torques exerted by the disk onto the planet are symmetrically exerted by the planet onto the disk. Thus, the planet exerts a positive torque onto the outer disk, i.e., it pushes the outer disk outward, while it exerts a negative torque onto the inner disk, i.e., it pushes it inward. For small mass objects, such as the planetary cores considered in the previous section, these torques are overcome by the internal torques of the disk, due to viscosity and pressure, so that the mass distribution in the disk – averaged over the azimuthal coordinate – is not significantly affected by the presence of the planet. However, this is no longer true if the planet is sufficiently massive (several tens of Earth masses for typical values of the disk's parameters); in this case, the internal torques of the disk cannot oppose the torques suffered from the planet. Consequently, the inner and outer parts of the disk are effectively repelled, and a gap in the gas distribution opens around the planet's orbit (Lin and Papaloizou 1986a; Crida et al. 2006).

Once a giant planet has opened a gap in the disk, it is condemned to stay in the middle of the gap. In fact, if it approached the inner edge of the gap, the distance of the planet from the inner disk would decrease and that from the outer disk would increase, so that the torque felt from the inner disk would become stronger than that felt from the outer disk. Thus, the planet would be pushed back toward the center of the gap. The symmetric situation would occur if the

planet approached the outer edge of the gap. Consequently, any radial migration of the planet has to follow the radial migration of the gap. The global evolution of the disk occurs on a viscous timescale $T_\nu = a^2/\nu$, where ν is the disk viscosity (Lynden-Bell and Pringle 1974). So, the radial displacement of the gap and the migration of the planet have to occur on this timescale (Lin and Papaloizou 1986b). Unless the planet is close to the outer edge of the disk (Veras and Armitage 2004), the planet has to migrate toward the star, because this is the natural direction of evolution of an accretion disk (Lynden-Bell and Pringle 1974; Lin and Papaloizou 1986b). The migration of a planet associated with a gap is called "Type II migration."

Type II migration explains in a natural way the presence of giant planets on orbits very close to the parent stars (Lin et al. 1996), which is a very common characteristic in the population of extrasolar planets discovered to date. However, in our solar system, the giant planets have orbital radii of several AUs, comparable to the orbital radii at which said planets are expected to have formed (see Sect. 2.1). Moreover, extrasolar planets have also been discovered on orbits with semimajor axes larger than ~3 AU. What happened in these cases? Why is Type II migration sometimes ineffective?.

For our solar system, the key to answer this question seems to be the coexistence of Jupiter and Saturn, with their specific mass ratio. In fact, hydrodynamical simulations, where Jupiter and Saturn are simultaneously taken into account with fixed masses, show that Saturn migrates the fastest, due to it smaller mass; therefore, it has no trouble in approaching Jupiter until the two planets are caught in resonance (Masset and Snellgrove 2001) (*Fig. 2-1*). In disks with mass comparable to the MMSN, the most likely end state is the capture in the mutual 2/3 mean-motion resonance, where the orbital period of Saturn is 1.5 that of Jupiter. This occurs even if Saturn is initially beyond the 1/2 resonance or locked into the 1/2 resonance (Pierens and Nelson 2008). Stable capture into the 1/2 resonance is possible only for disks with surface density decaying less steeply than 1/r or in low-mass disks (Zhang and Zhou 2010). Once locked in the 2/3 resonance, the inward migration of Jupiter and Saturn stops (Masset and Snellgrove 2001;

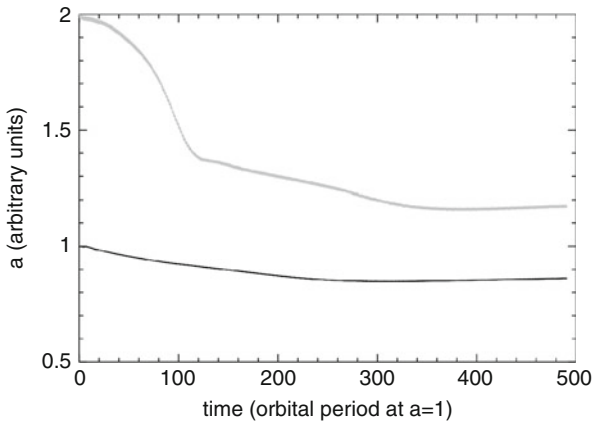


Fig. 2-1

An illustration of the dynamical evolution of Jupiter and Saturn in the gas disk, as in Masset and Snellgrove (2001). The *black* and *grey curves* show the evolutions of the semimajor axes of Jupiter and Saturn, respectively. Capture in the 2/3 mean-motion resonance occurs when the migration of Saturn is reversed

Morbidelli and Crida 2007), which explains why Jupiter did not migrate all the way close to the Sun (*Fig. 2-1*). These results hold both for disks with constant viscosity and for the so-called α-disks (Shakura and Sunyaev 1973) and do not depend critically on the value of the viscosity.

However, things may not be so simple in reality. All the hydrodynamical simulations that show that Saturn captures Jupiter in resonance assume fixed masses for the planets. A natural question arises: Is it still reasonable to expect resonance capture if the migration histories of the planets are coupled with their accretion histories? At first sight, the answer is negative. If the second planet forms later than the first one, its migration history should just replicate that of the first planet but later in time. More simply, the second planet should always lag behind the first one, as it is just repeating the evolution of the first planet, just at a later time. Thus, it appears that Saturn could catch Jupiter in resonance only if the accretion histories of the two planets were different. In particular, if Jupiter grew very rapidly to its current mass, it would have passed very quickly to a Type II migration mode, which is relatively slow. Instead, if Saturn grew more gradually than Jupiter and spent more time near a Saturn mass, it would have undergone fast migration for a longer period and hence could have trapped Jupiter in resonance. It is unclear why Saturn should have grown more slowly than Jupiter. Possibly, the opacity of the disk increased from the time of accretion of Jupiter to that of Saturn, thus slowing down the gas-accretion rate onto the planet.

Interestingly, among the collection of extrasolar planets, there are at least three systems where the Saturn analog did not capture in resonance the Jupiter analog (HD 12661, HD 13498, HIP 14810). But in these cases, the inner, more massive planet is closer than 1 AU to the star, and the outer, lighter planet is more than three times further away. This is obviously very different from the orbital architecture of Jupiter and Saturn or of the system OGLE-06-109L (which is a sort of twin of the Jupiter-Saturn system), which suggests that a different evolution occurred in these cases. It appears that the capture in resonance between a Jupiter analog and a Saturn analog is an event that may or may not happen, according to the accretion histories of these planets; depending on this binary possibility, the systems evolve along clearly different paths.

Assume now that Jupiter and Saturn were captured in their mutual 2/3 resonance. Morbidelli and Crida (2007) showed that the subsequent dynamical evolution of these planets depends on the properties of the disk, particularly the scale height. For thick disks (about 6% in scale height for a typical viscosity), the migration is very slow, and the planets remain at effectively constant distance from the central star. But for disks with decreasing thickness, outward migration becomes increasingly fast. In principle, in thin disks, outward migration can bring the planets up to ten times further than their initial location in a few thousands orbital periods (Crida et al. 2009), which might explain the orbits of some of the planets discovered by direct imaging beyond several tens of AUs from their parent stars. Until now, outward migration of Jupiter and Saturn from inside ~4 AU was considered incompatible with the existence of the asteroid belt – therefore only proto-solar disk models which prohibited outward migration were considered viable (Morbidelli et al. 2007). However, as shown by Walsh et al. (2011) and discussed in Sect. 4.1, this is not true, releasing the constraints against Jupiter's outward migration. This opens a new degree of freedom to model the evolution of the solar system, as it will be shown at the end of this chapter.

For two giant planets to avoid inward migration as discussed above, it is essential that the mass of the outer planet is a fraction of the mass of the inner planet, as in the Jupiter-Saturn case (Masset and Snellgrove 2001; Morbidelli and Crida 2007). Consider a heuristic but intuitive explanation of this statement. When two giant planets are close enough to each other, they evolve inside a common gap of the gas-density distribution. The inner planet, being closer to

the inner edge than to the outer edge of the common gap, feels a positive torque and would tend to migrate outward; the outer planet, being closer to the outer edge of the common gap, feels a negative torque and would tend to migrate inward. If the planets are locked in resonance, their relative orbital separation cannot change (if they are not yet in resonance, they move toward each other until they are captured and locked into a resonance). Thus, the direction of migration of the pair of planets depends on which of the two torques dominates. Each torque is proportional to the surface density of the disk adjacent to the planet and to the square of the mass of the planet itself (Goldreich and Tremaine 1979). Because the planets partially deplete the disk in the region between the star and their innermost orbit (a partial cavity; Crida and Morbidelli 2007), the surface density in the outer disk is typically larger than in the inner disk; so, a necessary condition to avoid inward migration (i.e., to make the torque felt by the inner planet larger than that felt by the outer planet) is that the inner planet is more massive.

Therefore, this mechanism predicts that no pair of resonant giant planets with narrow orbital separation, with the outer planet significantly lighter than the inner one, should ever be found close to the parent star. In fact, planets in this configuration should have avoided inward migration. So far, this prediction is validated by observation. In fact, there are many pairs of planets in resonance (or close to), near their star, but none of these cases exhibits a Jupiter/Saturn mass ratio. The absence of this configuration, which is statistically significant if one assumes that the mass ratio should be random, strongly supports the theoretical result that resonant planets in Jupiter/Saturn mass ratio move outward, and therefore cannot be found within the range of stellar distances that can be probed by radial velocity observations (OGLE-06-109L system was in fact discovered by micro-lensing).

There is an intriguing aspect in this view of the evolution of Jupiter and Saturn. When the two planets do not migrate inward at the nominal Type II migration rate, there is necessarily an inward flow of gas, from the outer disk to the inner disk, through the common gap. Thus, the outer planet should presumably accrete more of the incoming material, narrowing the mass difference with the inner planet. This raises the question of why Saturn remained smaller than Jupiter. The answer may be that the gas disk was rapidly disappearing while Jupiter and Saturn were undergoing the dynamical evolution described above, so that Saturn failed to grow further.

There are several lines of evidence in favor of a formation of Jupiter and Saturn in a dissipating disk. First, Jupiter's atmosphere is enriched in elements heavier than helium by a factor of 3–4 relative to solar composition (Wong et al. 2004), while Saturn is enriched by a factor 11 in Carbon (Fouchet et al. 2009). Guillot and Hueso (2006) have argued that the easiest explanation for this fact is that hydrogen and helium had been already depleted by a factor of 3–4 in the disk by the time Jupiter captured its atmosphere (and, following this logic, by a factor of 11 by the time Saturn captured its atmosphere). Second, the favored model for the accretion of the regular satellites of the giant planets also requires that said satellites formed in gas-poor circumplanetary disks (Canup and Ward 2006). Third, the common explanation for why Uranus and Neptune failed to accrete massive atmospheres is that the gas disappeared before they had a chance to do so (Pollack et al. 1996). All these arguments suggest that the four giant planets formed in a temporal sequence, from Jupiter to Saturn and then to Uranus and Neptune, while the disk was being dispersed. Finally, recall that the solar composition is at the low end of the metallicity range for planet-bearing stars. In the core-accretion model of giant planets, the metallicity of the star is correlated with the speed of accretion of the cores, so that stars that are too poor in metals fail to form giant planets before the disappearance of the disk (Ida and Lin 2004). This suggests that the solar system barely made its giant planets, while the disk was being dispersed.

If this explanation may be satisfying for our solar system, it is nevertheless interesting to discuss what would have happened if Jupiter and Saturn had formed earlier, when the disk was still massive. The issue is not only academic, as it can be pertinent for systems with a higher initial metal content which, as suggested above, should form planets faster. In a gas-rich disk, Saturn would have eventually become as massive, or even more massive, than Jupiter. Thus, the two planets would have resumed an inward migration (Morbidelli and Crida 2007). The orbital eccentricity of two planets migrating in resonance tends to increase monotonically (Ferraz-Mello et al. 2003; Kley et al. 2004). This effect is contrasted by the action of the disk, which tends to damp the planets' eccentricities (Kley and Dirksen 2006). A crucial role is played by the disk inside the orbit of the inner planet. This inner disk tends to be partially depleted due to the presence of the planet(s). The level of depletion depends on several parameters such as the viscosity of the disk, its scale height, its inner radius, and the mass of the planet(s) relative to the disk etc. (Crida and Morbidelli 2007). If the inner disk is depleted substantially, the damping effect on the eccentricity of the inner planet is strongly reduced. In this situation, the eccentricity of the inner planet keeps growing, until the pair of planets becomes dynamically unstable and mutual close encounters are triggered (Kley et al. 2005). This may be the case of many, if not most, of the planetary systems. In fact, mutual scattering seems to be the major mechanism responsible for the eccentricity distribution observed in the extrasolar planets collection (see ❯ Sect. 2.3 for a more complete discussion). However, if the inner disk is not very depleted, the eccentricities of the planets grow until a limit value is achieved (Crida et al. 2008; see ❯ *Fig. 2-2*). This process can leave the planets at the disappearance of the disk on stable resonant orbits with moderate eccentricities, and thus, it can explain the pairs of planets in resonance observed to date.

In conclusion, this section sheds light on the first crucial "bifurcations" in planetary evolution that can account for at least part of the great diversity observed in planetary systems. In fact, assuming that giant planets form in sequence at increasing distances from the central star, most of the observed diversity of planetary systems could stem from the occurrence or avoidance of two events: (i) the capture in resonance of the first, inner planet by the second, initially smaller one, which stops inward migration (often triggering outward migration), and (ii) the growth of the outer planet beyond the mass of the inner one, which causes inward migration of both planets to resume. The solar system structure results from the occurrence of (i) and avoidance of (ii). Systems like HD 12661, with a close-in massive planet and a distant smaller planet, result from the avoidance of (i). Resonant giant planets close to their stars, like those in the GJ876 system, result from the occurrence of both (i) and (ii). Unstable systems, ultimately leaving behind one giant planet on an eccentric orbit, may also result from the occurrence of both (i) and (ii), but in cases where there was not enough eccentricity damping because of a depleted inner disk.

Here, the case with two planets was the only one considered; obviously, the tree of possible evolutions can only become more complicated if more giant planets are involved, and the final outcomes can be even more diverse.

2.3 Planet–Planet Scattering as the Dominant Orbital Excitation Process

One of the greatest surprises that came with the discovery of extrasolar planets is the realization that most planets have orbital eccentricities much larger than those characterizing the planets of

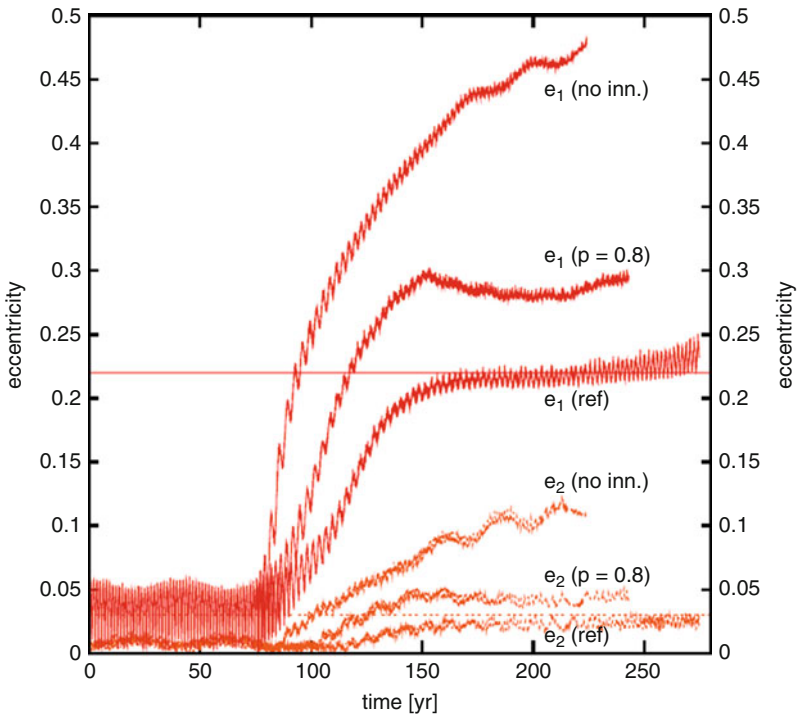


Fig. 2-2

The evolution of the eccentricities of planets *b* and *c* around GJ876 during their putative inward resonant migration. The eccentricity of the interior planet is labeled e_1, that of the exterior planet e_2. The evolutions labeled "no inn." assume that no disk is present inside the orbit of the interior planet; in this case, the eccentricities seem to grow indefinitely. The evolutions labeled "ref." and "p = 0.8" account for an inner disk and differ just for the value of a technical simulation parameter. In these cases the eccentricities attain equilibrium values. In the "ref" case, the final eccentricities reproduce the eccentricities inferred from observations (*horizontal lines*) (from Crida et al. (2008))

our solar system. Eccentricities of about 0.4 are quite common in the extrasolar planets collection; some planets have eccentricities larger than 0.6, values that in our solar system are common only for comets!

Whatever the preferred model of giant planet formation (core-accretion or gravitational instability; see Sect. 2.1), it is expected that planets have originally small orbital eccentricities, because they form from a circumstellar disk whose streamlines are basically circular. There has been a lively debate on whether subsequent planet-disk interactions can raise the planets' eccentricities up to the observed values. Papaloizou et al. (2001) concluded, with numerical experiments and theoretical considerations, that eccentricity growth is not possible for planetary masses below 10–20 Jupiter masses. Instead, Goldreich and Sari (2003) argued with theoretical considerations that, under some conditions, depending mostly on disk's thickness, giant planets of more moderate masses could have their orbital eccentricity excited, although they could not estimate the magnitude of this excitation. More recent hydrodynamical simulations (D'Angelo et al. 2006; Kley and Dirksen 2006) showed that planets with masses larger

than ~2–3 Jupiter masses, under some conditions, can have eccentricities excited by the disk, but only to moderate values (~0.1–0.2), definitely lower than those characterizing many, if not most, of the extrasolar planets. In most cases, the planet-disks interactions rather seem to lead to eccentricity damping. Thus, a more generic orbital excitation mechanism seems to be required to explain the observations.

Soon after the discovery of the first eccentric planets, it was pointed out that mutual encounters between planets can easily provide strong orbital excitation (Rasio and Ford 1996; Weidenschilling and Marzari 1996; Lin and Ida 1997; Levison et al. 1998; Ford et al. 2001; Marzari and Weidenschilling 2002; Adams and Laughlin 2003). More recent studies (Jurić and Tremaine 2008; Chatterjee et al. 2008; Ford and Rasio 2008; Raymond et al. 2009a; Beauge and Nesvorny 2011) show that random systems of giant planets initially on unstable, quasi-circular orbits evolve through close encounters until a dynamical relaxation state is achieved with the ejection or collision of some planets. In these models, the final eccentricity distribution of the surviving planets is remarkably similar to that of known extrasolar planets (see *Fig. 2-3*). Moreover, Jurić and Tremaine (2008) and Raymond et al. (2009a) showed that the final orbital spacing of the surviving planets is also in good agreement with the observations of extrasolar systems of two or more nonresonant planets. These systems typically look "packed," in the sense that the orbital separation (apocenter to pericenter) between neighboring planets is not much larger than what is required by the Hill-stability criterion (Barnes and Greenberg 2006). Finally, the hot Jupiters discovered on orbits strongly inclined relative to the stellar equator could have

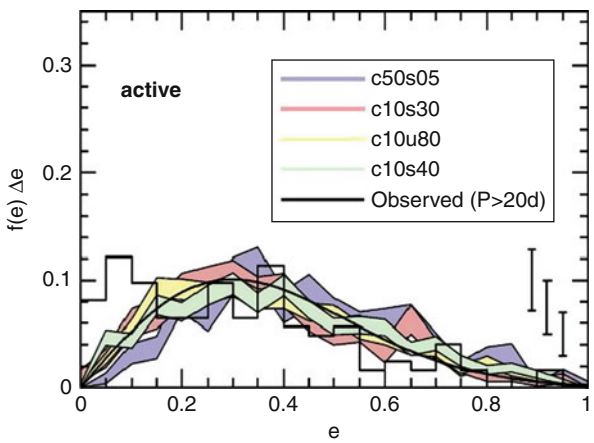


Fig. 2-3

Final eccentricity distribution of simulated ensembles of planetary systems that underwent a dynamical instability sometime during the full simulation time span of 10^8 year. The color bands correspond to different ensembles, characterized by different initial conditions. The histogram shows the observed eccentricity distribution of extrasolar planets with orbital period longer than 20 day, according to Butler et al. (2006). The observed distribution and the final distributions resulting from the simulations agree very well, with the exception of an excess of observed planets with $e < 0.2$. This is probably due to planets that never underwent a significant dynamical instability (from Jurić and Tremaine (2008))

reached their current orbits via planet–planet scattering and tidal circularization (Beauge and Nesvorny 2011).

All these quantitative results give strong support to the idea that planet–planet scattering is the main mechanism sculpting the orbital distribution of extrasolar planets. In addition, Veras et al. (2009) pointed out that planet–planet scattering can propel a planet in the region beyond 100 AU, which can explain some of the extrasolar planets imaged at large distances from their parent stars.[2]

It is important to remember that, assuming giant planets formed beyond an hypothetical snow line at 3–5 AU, the planet–planet scattering mechanism alone cannot explain the observed semimajor axis distribution of extrasolar planets (see, for instance, Marzari and Weidenschilling 2002). In fact, through scattering events, planets can have difficulty reaching orbits with semimajor axis smaller than half of the initial value of the inner planet, with the exception of those objects scattered onto orbits with very large eccentricities and small periastron distances, which may then become circularized by tidal dissipation. Thus, migration is still required to explain the presence of a large number of planets on orbits with small semimajor axes. Adams and Laughlin (2003) and Moorhead and Adams (2005) modeled the interplay between migration and scattering. They used N-body simulations, with fictitious forces to mimic the effect of the disk on the planets, for what concerns both semimajor axis decay and eccentricity damping. After tuning a few parameters (the lifetime of the disk, the timescale of the eccentricity damping, etc.) and accounting for observational biases, Moorhead and Adams obtained a very good reproduction of the two-dimensional (a, e) distribution of the exoplanets detected by the radial velocity technique (*Fig. 2-4*). Although very appealing, this result may be questioned because the numerical recipes used to mimic migration and damping have been based on simple analytic estimates. Reality may be more complicated. For instance, as seen above, a system of two planets in resonance may migrate at a very different speed than a single planet in Type II migration (Morbidelli and Crida 2007); the disk may not always damp the eccentricity of a planet but could also sustain it, depending on the planet mass and initial eccentricity (Kley and Dirksen 2006); migration direction could be reversed for eccentric planets (D'Angelo et al. 2006); and, finally, mass accretion onto the planets is neglected in the Moorhead and Adams model.

For all these reasons, there are two key questions requiring further study in a definitive way: (1) Why do giant planets become unstable in the first place? (2) When do they become unstable, relative to the gas-disk lifetime?

Concerning the first question, there are in principle two answers: either planetary systems become unstable because the planets grow in mass or because the planets are brought too close to each other by migration processes. The first case may be excluded as a dominant mechanism to explain the large eccentricities of extrasolar giant planets for the following reason. Imagine a system of cores close to each other (maybe brought by Type I migration into mutual resonances of type $n/(n+1)$, with quite large n). Given that the time required to trigger the runaway accretion of a massive atmosphere is much longer than the accretion of the atmosphere itself (Pollack et al. 1996), it is unlikely that all the cores would accrete massive atmospheres and become giant planets simultaneously. More realistically, one core would start the accretion of the

[2]In summary, three mechanisms have been proposed to move planets from the snow line region to large distances from the central star: (i) outward Type II migration of planets originally formed in the outer part of a disk in rapid viscous spreading (Veras and Armitage 2004), (ii) outward migration of a pair of resonant planets with a Jupiter-Saturn mass hierarchy (Crida et al. 2009), and (iii) scattering of a planet to a wide elliptic orbit (Veras et al. 2009). These mechanisms are potential alternatives to the possibility that distant planets formed in situ, by the gravitational instability mechanism (Boley 2009)

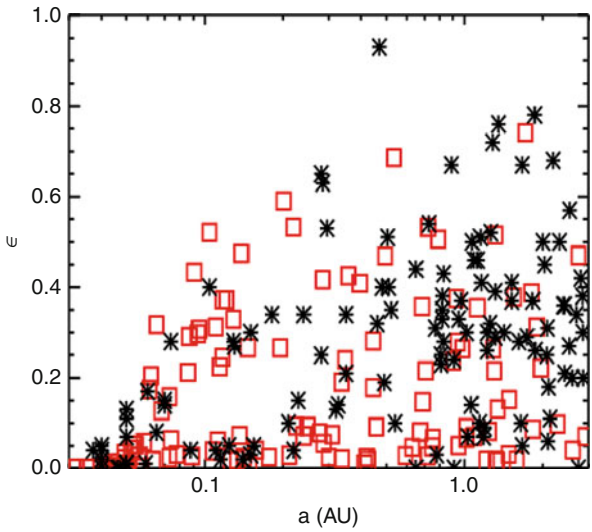


Fig. 2-4

The final semimajor axis versus eccentricity distribution of extrasolar planets (*squares*) in the simulations of Moorhead and Adams (2005) which account for (i) inward migration, (ii) eccentricity damping due to the disk (with an assumed timescale of 0.3 My) (iii) tidal circularization, and (iv) radial velocity detection biases. The stars show the distribution of the extrasolar planets known at the time

atmosphere first, and it would acquire a mass much larger than those of the other planets at that time. The system would become unstable, because it is too closely packed to withstand the newly born giant planet. However, the scattering phase would bring the cores onto orbits with large eccentricities, but it would leave the newly born giant planet onto a quasi-circular orbit: This does not correspond to the measured orbital distribution of the extrasolar planets. Moreover, the orbital eccentricities of the cores would be damped very fast by the disk (Cresswell et al. 2007; Bitsch and Kley 2010), and, by the time these cores grow, in turn, to become giant planets, they would be again on quasi-circular orbits. A system of multiple giant planets probably forms through several repetitions of the events described above but, at each step of the process, the already formed giant planets should be those on the least eccentric orbits.[3]

However, note that the process described above is probably not the one that governs the *final* orbital distribution of giant planet systems. A more generic mechanism to achieve the

[3]The reader should remember that two planets are stable if their orbital separation is a few times their mutual Hill radius $R_H = \bar{a}[(m_1 + m_2)/M_S]^{1/3}$, where m_1 and m_2 are the masses of the two planets and $\bar{a}$ is their mean semimajor axis, while M_S is the mass of the star. Suppose now that planets tend to acquire orbits whose mutual separation is not much larger than this Hill-stability limit. If a planetary system is made of two planetary cores, when one of two objects becomes a giant planet, the system is likely to be destabilized because R_H increases by a factor 3–4 as the mass of one planet grows by a factor 30–60. Instead, a stable system made of one giant planet and one core is not likely to be strongly destabilized when the core grows to the status of a giant planet, because R_H increases only by a factor $\sim 2^{(1/3)} = 1.25$. Similarly, in a system made of one giant planet and two cores, the growth of one of the cores to the status of a giant planet is likely to destabilize the remaining core but not the first planet

ultimate dynamical instability (the one responsible for the final orbits) is that two or more giant planets, once fully formed, are brought in resonance with each other by Type II migration, which causes a subsequent increase of their orbital eccentricities (Ferraz-Mello et al. 2003; Kley et al. 2004, 2005; Crida et al. 2008).

This leads to the second question, on the timing of the instability. In the Adams and Laughlin (2003) and Moorhead and Adams (2005) models, the planets become unstable while they are still migrating in the gas disk. However, limited experience with hydrodynamical simulations of giant planets scattering each other in gas disks shows that the evolution can be quite different from the one modeled in those works. Scattered planets tend to acquire orbits that are more separated in semimajor axis and more eccentric, but then the eccentricities are damped by the disk and the planets migrate back into a new, more stable resonant configuration, with moderate eccentricities (see, for instance, ❯ *Fig. 2-5* or Moekel et al. 2008). It is possible that systems with more planets develop more violent instabilities that extend also longer in time. However, new simulations with three planets, presented in Marzari et al. (2010) and Moeckel

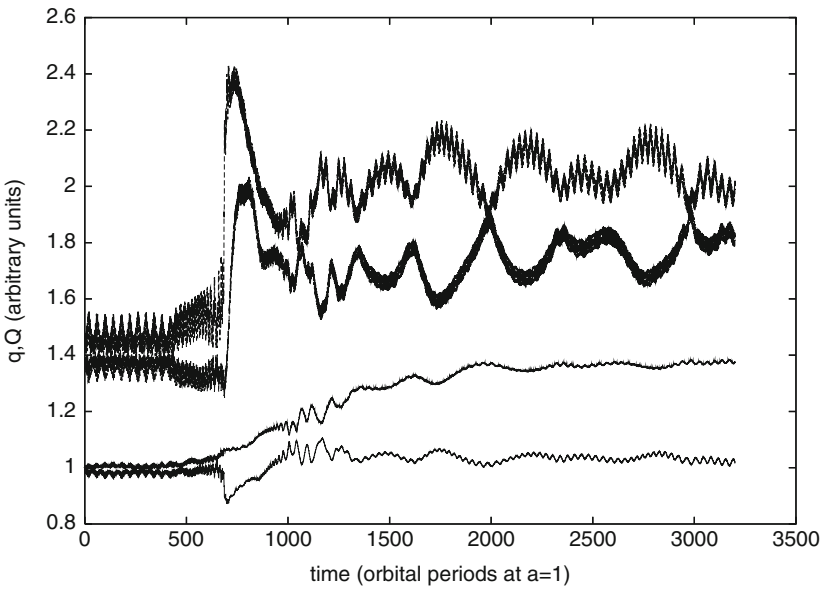


Fig. 2-5

The evolution of a pair of giant planets initially on unstable orbits. For each planet, the pair of curves shows the evolution of the periastron (*q*) and apoastron (*Q*) distances. Thus, when the two curves are close, the orbit is almost circular and $q \sim Q \sim a$, where *a* is the orbital semimajor axis. The *light curves* at the *bottom* are for the inner planet, with a three Jupiter mass; the *thick curves* at the *top* are for the outer planet, with one Jupiter mass. Notice that the orbits become very eccentric and separate from each other at the time of the instability, after about 500–700 orbits of the inner planet. Subsequently, the eccentricities of the planets are damped, and the inward migration of the outer planet brings it in the 1/2 resonance with the inner one ($t = 1{,}000$). Once in resonance, the eccentricity increases again and starts to have long-period oscillations. A relatively stable configuration is achieved (from Morbidelli and Crida (2007))

and Armitage (2012), again show that the planets surviving at the end of the instability phase have low-eccentricity orbits. Until we know more from hydrodynamical simulations about the dynamics of eccentric planets, it is premature to conclude whether instabilities produced during the gas-disk phase would lead to the observed orbital distribution of extrasolar planets. The other possibility is that during the gas-disk phase, planets acquire stable resonant and eccentric orbital configurations, which become unstable when the gas is removed. This is the approach of Lin and Ida (1997), Levison et al. (1998), Jurić and Tremaine (2008), and Chatterjee et al. (2008), to quote just a few works.

The facts that most extrasolar planets have quite large eccentricities and that there are only a few pairs of stable resonant planets suggest that orbital instability in planetary systems is more the rule than the exception. It would be quite surprising if most planet configurations achieved through migration were stable in presence of gas and unstable in absence of gas. In fact, the gas has some stabilizing effect due to the eccentricity damping that it exerts, but it also drives migration, which in turn excites the eccentricity. The two effects cancel out when a limit eccentricity is attained (*Fig. 2-2*). At this point, the gas should not play any longer any crucial role in maintaining stability. One possibility is that most resonant configurations achieved through migration are unstable in both cases (with gas and without gas), but the instability manifests itself on timescales of several millions of years, i.e., well after that the gas has been removed. To have a better appreciation of reality, one needs a more systematic hydrodynamical simulations of the dynamics of sets of giant planets embedded in gas disks, followed by the investigation of their subsequent long-term evolutions after the gas removal. A work of this kind has been done for the planets of our solar system (Morbidelli et al. 2007; see Sect. 2.4), but it is obviously more demanding in general, given the volume of parameter space that needs to be explored.

Naively, planets that develop instabilities and mutual encounters after having migrated to the vicinity of the central star should acquire smaller eccentricities than those that do so further away from the star. This is because the eccentricity acquired in an encounter is proportional to the ratio between the velocity kick received during said encounter and the orbital velocity. The former depends on the escape velocities from the surface of the planets and is independent of the orbital radius, while the latter is larger for the close-in planets. Instead, Jurić and Tremaine (2008) showed numerically that the eccentricity distribution achieved at dynamical relaxation (i.e., after many encounters) is essentially independent of the semimajor axes of the planets. Given that no clear correlation is found between distance and eccentricity in the extrasolar planets collection (apart from the tidal circularization zone), this result gives a quite strong support to the idea that planets first migrate toward the central star and then, after gas removal, become unstable.

2.4 A Plausible Evolution of the Four Giant Planets of the Solar System

Returning to the solar system, what is a plausible scenario for the evolution of the four major planets during the gas-disk phase? As seen above (Sect. 2.2), hydrodynamical simulations strongly suggest that Jupiter and Saturn rapidly reached a 2/3 resonant orbital configuration and that this prevented them from migrating further toward the Sun. Once in resonance, these giant planets either remained on nonmigrating orbits or migrated outward.

What about Uranus and Neptune? According to the latest models on the migration of planetary cores in radiative disks, Uranus and Neptune, or their precursors, should have evolved

on nonmigrating orbits in the intermediate part of the disk (Lyra et al. 2010). If Uranus and Neptune did not migrate while Jupiter and Saturn achieved nonmigrating orbits after a period of inward migration, eventually the giant planets system should have been characterized by a large separation between the Jupiter/Saturn pair and the Uranus/Neptune pair (see left panel of *Fig. 2-6*). How to reconcile this peculiar orbital structure with the current structure of the solar system is unknown. Instead, if Jupiter and Saturn migrated outward, Uranus and Neptune would have been caught in mean-motion resonances with the two major planets (see right panel of *Fig. 2-6*). The same would have occurred also in the case where Jupiter and Saturn remained on nonmigrating orbits but Uranus and Neptune started to migrate inward due to a change of the thermal properties of the disk over time (Lyra et al. 2010). What matters, in fact, is the convergent migration between the Jupiter/Saturn pair and the Uranus/Neptune pair, not which pair of planets is moving.

A search for possible resonant configurations that could have been achieved by the two pairs of planets in convergent migration has been done in Morbidelli et al. (2007), using hydrodynamical simulations, and in Batygin and Brown (2010), using N-body simulations with fictitious forces that mimic the effect of the disk. The search is probably not complete, in the sense that other relative resonant configurations might have been achieved, depending on disk properties, migration speed, and initial Uranus/Neptune configuration. Three results, however, seem robust. First, in case of convergent migration, each planet ends up in resonance with its neighbor. Thus, the planets should have been in a 4-body resonance, the most complex multi-resonant configuration in the solar system.[4] Second, because distant resonances are weak and have a low probability to capture a migrating body, the four planets most likely should have achieved a very compact orbital configuration. Third, the eccentricities of the planets should have remained small (less than 0.005 for Jupiter, 0.02 for Saturn, 0.06 for Uranus, and 0.015 for Neptune in Morbidelli et al. simulations). Thus, Jupiter and Saturn should have been on orbits significantly more circular than now.

Morbidelli et al. also investigated the long-term stability of resonant orbital configurations that they found. This was done by continuing each hydrodynamical simulation for 1,500 Jupiter's orbits, while removing uniformly the gas, exponentially in time, down to a factor of 1/1,000. This procedure was just instrumental for changing adiabatically the potential felt by the planets and was not intended to mimic the real process of evaporation of the disk. The final orbits of the planets were then passed to a symplectic N-body code, and integrated for 1 Gy, without additional perturbations (for instance, from a planetesimal disk). Morbidelli et al. found that only one configuration was stable for 1 Gy. This result was actually affected by an error of the integrator, which was discovered only later. In reality, 4 of the 6 configurations found in Morbidelli et al. are stable for 1 Gy. The only unstable configurations are the two most compact ones, with Uranus in the 3/4 resonance with Saturn and Neptune in either the 4/5 or 5/6 resonances with Uranus. These configurations lead to very violent instabilities, with close encounters of all the planets with each other, including close encounters of Jupiter with Saturn. In these cases, orbital relaxation is achieved when all planets except Jupiter are ejected on hyperbolic or distant eccentric orbits. Obviously, this is not what happened in our solar system. But, in these simulations, the final orbit of Jupiter, with an eccentricity of ~ 0.4, is similar to those of the extrasolar planets discovered to date at a distance of 4–5 AU from their parent stars. This suggests that these extrasolar planets might be the survivors of systems that

[4]Currently, the record for the most complex resonance chain in the solar system is detained by Jupiter's satellites Io, Europa, and Ganymede, which are locked in a 3-body resonance, also known as the Laplace resonance.

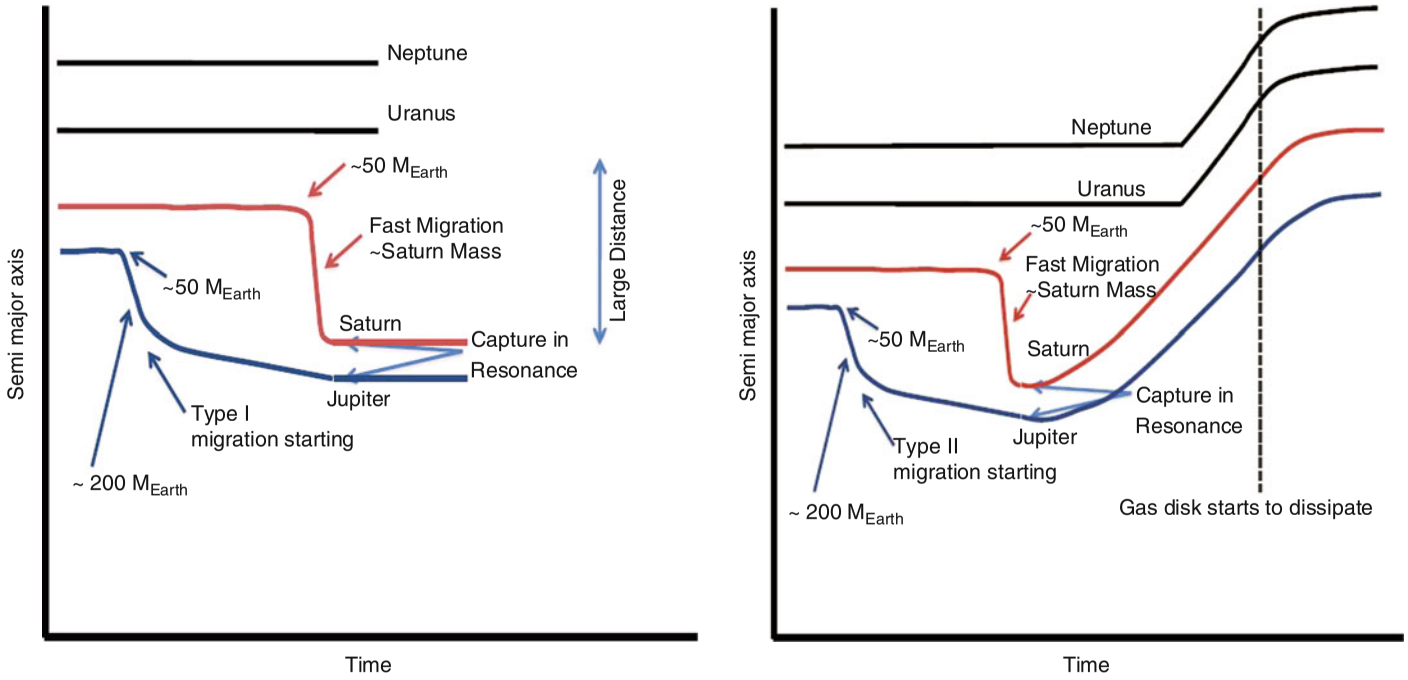


Fig. 2-6

Sketch of the possible migration and accretion histories of the giant planets of the solar system, while they were embedded in the disk of gas. Both scenarios assume that (i) Saturn was eventually caught Jupiter in their mutual 2/3 resonance (see Sect. 2.2 for conditions) and (ii) planets with masses smaller than ~50 Earth masses do not migrate in radiative disks. In the *left panel*, Jupiter and Saturn do not migrate after resonance capture. This leaves a large gap between the Jupiter/Saturn pair and the Uranus/Neptune pair that cannot be reconciled (to our current knowledge) with the current orbital structure of the planets. In the *right panel*, Jupiter and Saturn migrate outward. This leads to the capture of Uranus and Neptune in resonance with Saturn and with themselves. This resonant configuration is consistent with the current orbits of the planet via a phase of dynamical instability after the dispersal of the disk of gas (see Sect. 3.2)

avoided Type II migration, possibly through a mechanism like the Jupiter-Saturn one, but which achieved orbital resonant configurations so compact to undergo a violent instability involving encounters between gas-giant planets. Our solar system was luckier and picked up a less compact multi-resonant configuration, so that encounters between Jupiter and Saturn could be avoided.

The next section shows that the least compact resonant configurations, which are stable when only the four planets (Jupiter to Neptune) are considered, can become unstable when the interactions of the planets with a remnant planetesimal disk are taken into account. The solar system may have passed through such an instability phase, reconciling the current orbits with those that the planets should have had when they emerged from the gas-disk phase (see also Thommes et al. 1999).

3 The Planetesimal-Disk Era

3.1 Brief Tutorial of Planetesimal-Driven Migration

A planet embedded in a planetesimal disk has repeated close encounters with the objects that come close to its orbit. Each of these encounters modifies the trajectory of the incoming planetesimal, and, consequently, the planet has to suffer a small recoil. In this way, if the planetesimal disk is sufficiently massive, significant angular momentum exchange may occur between the planet and the planetesimals, enough to cause a rapid, long-range migration of the planet (Ida et al. 2000). A review of planetesimal-driven migration has been presented in Levison et al. (2007). The sections below review the basic concepts that are relevant for understanding the evolution of our solar system and, possibly, of planetary systems in general.

For a system of giant planets, planetesimal-driven migration is relevant only after the disappearance of the gas. The reason is that the gas contains typically ~100 times more mass than the planetesimals, and therefore it exerts the dominant gravitational forces on the planets. Consequently, planetesimal-driven migration was not mentioned in the previous part of this chapter. It should be remembered, though, that for small planets, gas-driven migration is proportional to the planet's mass (Ward 1997), whereas planetesimal-driven migration is, at first order, independent of the planet's mass (Ida et al. 2000; Kirsh et al. 2009); thus, for small planets, such as planetary embryos of an Earth-mass or smaller, planetesimal-driven migration may rival, under some conditions, Type I migration (Levison et al. 2010; Capobianco et al. 2011).

One may naively think that planetesimal scattering is a purely random process, which consequently cannot force a planet to migrate in a specific direction. This is not true. To zeroth order, during an encounter, the planetesimal is in a Keplerian orbit about the planet. Since the energy of this orbit must be conserved, all the encounter can do is to rotate the relative velocity vector. Thus, the consequences of such an encounter can be effectively computed in most of the cases using an impulse approximation (Öpik 1976; Ida et al. 2000). With this approach, it is easy to compute that on average (that is averaged on all impact parameters and relative orientations), the planetesimals that cause a planet to move outward are those whose z-component of the specific angular momentum, $H = \sqrt{a(1-e^2)}\cos i$, is larger than that of the planet (H_p). In fact, during the encounter, these planetesimals have an azimuthal angular velocity faster than that of the planet: Thus, they tend to be slowed down, propelling in turn the planet along its orbit.

The opposite is true for the planetesimals with $H < H_p$ (Valsecchi and Manara 1997). In these formulæ a, e and i are the semimajor axis, eccentricity, and inclination of the planetesimal.

The direction of migration for a single planet in principle depends on the angular momentum distribution of objects on planet-encountering orbits. Nevertheless, the direction of migration does not simply depend on the sign of $\bar{H} - H_p$, where $\bar{H}$ denotes the mass-weighted value of H of the planet-crossing particles: There is a bias in scattering timescales on either side of the planet's orbit which leads to a very strong tendency for the planet to migrate inward (Kirsh et al. 2009). Consequently, outward migration is found only in systems where $\bar{H} - H_p$ is strongly positive, such as for planetesimal disks with surface density distribution proportional to r^k with $k > 1$ (the value of k for the real disks is expected to be between -2 and -1, definitely giving inward migration).

In order to understand the basic modes of migration, imagine defining a function $\mathcal{F}$ of the H-distribution of the planet-crossing particles, such that the planet's migration rate $\mathrm{d}a_p/dt$ is proportional to $\mathcal{F}$ (which implies that migration is inward if $\mathcal{F}$ is negative, while it is outward otherwise). For a given orbital distribution of the planetesimals in the planet-encountering zone (i.e., for a fixed value of $\mathcal{F}$), the rate of migration in the local units of length and time must be proportional to the total mass M of the planet-encountering planetesimals (Ida et al. 2000; Gomes et al. 2004; Levison et al. 2007). Thus, one can write

$$\frac{\mathrm{d}a_p/a_p}{\mathrm{d}t/T} \sim M\mathcal{F}\,,$$

where T is the planet's orbital period. When a planet migrates, several concomitant processes occur:

(a) The planetesimals that are scattered in the direction opposite to that of planet migration (e.g., onto an orbit with a larger semimajor axis, for an inward migrating planet) can be "left behind," in the sense that they can find themselves on orbits which do not cross the planet's orbit any more (because the orbit of the planet has drifted away; see *Fig. 2-7*). Conversely, other planetesimals, originally situated on stable orbits in the disk through which the planet is moving, start to be scattered. Thus, planetesimals both enter and leave the planet-encountering region as a result of the radial drift of the planet's orbit through the disk. This orbital drift tends to leave $\mathcal{F}$ unchanged but changes M; the latter increases or decreases depending on the gradient of the disk's surface density.

(b) The planet encounters tend to rearrange the angular momentum distribution of the planetesimals to an equilibrium configuration that would induce no migration; this decreases $|\mathcal{F}|$ toward zero, while preserving M.

(c) Some planetesimals may be eliminated from the system through collisions with the planet or ejections onto hyperbolic orbit, which decreases M.

Depending on whether, as a net result of all these processes, $|M\mathcal{F}|$ increases or decreases, radial migration accelerates exponentially or decays to zero. The former case is called *forced* or *self-sustained* migration; the latter is called *damped migration*. Whereas processes (b) and (c) inexorably tend to damp the planet's migration (because they reduce either $|\mathcal{F}|$ or M), process (a) can sustain the migration if it leads to an increase of M. Thus, self-sustained migration occurs if two conditions are met: First, process (a) has to give a positive feedback on migration, which translates into a condition on the gradient of the surface density of the disk; second, the mass M has to be large enough so that the timescale of process (a), which is proportional to $1/M$, is faster than those of (b) and (c), which are independent of M.

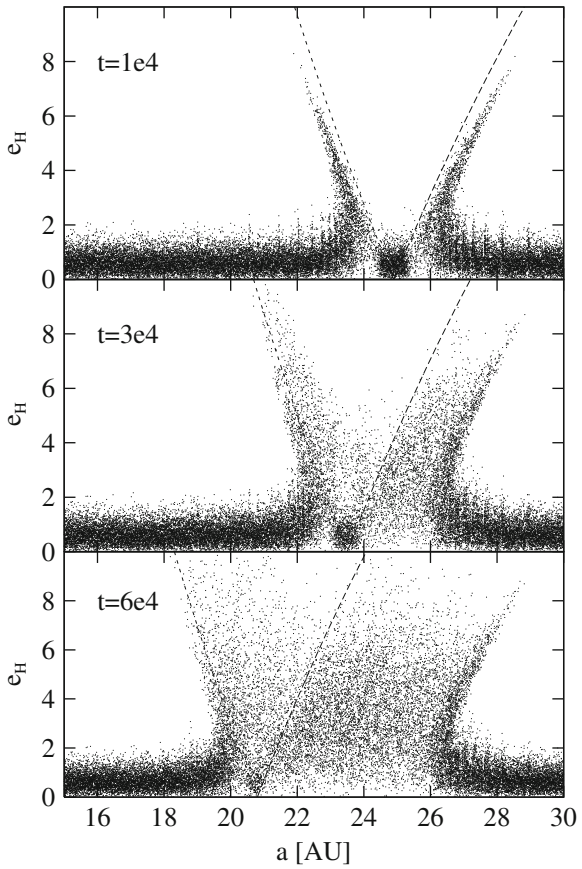


Fig. 2-7

The migration of a 2.3 $M_\oplus$ planet in a planetesimal disk of 230 $M_\oplus$. Each panel shows the eccentricity versus semimajor axis distribution of the planetesimals (*dots*) at the time marked in the *top left corner*. The *dashed curves* delimit the planet-crossing region. The planet is situated at the point of intersection of these curves. Notice how planetesimals are "left behind" on eccentric orbits as the planet migrates inward (adapted from Kirsh et al. (2009))

The dynamical evolution is qualitatively different if there are two planets. In this case, planetesimals can be scattered by one planet onto an orbit that has close encounters with the other planet. This transfer of particles from the "control" of one planet to the other tends to increase the orbital separation between the planets. However, the planets can effectively move away from each other only if they are not locked in a mutual mean-motion resonance (the orbital response of resonant planets is different and will be discussed in Sect. 3.2). Assuming no resonance locking, under some conditions, this orbital divergence can lead the outer planet to migrate outward, even in planetesimal disks in which a single planet would normally migrate inward. In particular, this is the case for a Neptune-mass planet on an orbit exterior to a Jupiter-mass planet. In fact, the planetesimals that the Neptune-mass planet scatters inward onto orbits crossing that of the Jupiter-mass planet, are rapidly ejected onto hyperbolic orbits by the latter;

conversely, the planetesimals that the Neptune-mass planet scatters outward, remain on orbits crossing that same planet and have repeated encounters with it: Sooner or later, most of them will be eventually scattered inward and then removed by an encounter with the Jupiter-mass planet. In conclusions, the net work of the Neptune-mass planet is to transfer planetesimals inward to the control of the Jupiter-mass planet, and consequently the Neptune-mass planet has to move outward.

The case of the giant planets of our solar system, with Jupiter in the innermost orbit, two Neptune-Uranus-mass planets on the outermost orbits, and one intermediate-mass planet (Saturn) in between, is somewhat analog to the simple Jupiter-Neptune system described above (if, again, the planets are not in resonance and are free to migrate relative to each other). With N-body simulations, Fernandez and Ip (1984) showed for the first time that Jupiter migrates inward, while Saturn and, particularly, Uranus and Neptune move outward. Malhotra (1993, 1995) elaborated on this kind of evolution to explain the observed orbital properties of the Kuiper belt (a population of planetesimals, including Pluto, with semimajor axes beyond that of Neptune). The original version of the "Nice model"[5] (Tsiganis et al. 2005; Gomes et al. 2005), that aimed to build a coherent scenario of the late orbital evolution of the outer solar system, was also based on this process of divergent migration of the giant planets. Given our current understanding that the giant planets, at the end of the gas-disk era, had to be in resonance with each other (see previous section), these models are not strictly valid any more and consequently will not be discussed further. A new version of the "Nice model," which starts from a multi-resonant orbital configuration of the giant planets, is instead illustrated in Sects. 3.2 and 3.3.

There are nevertheless some important results of general relevance that these studies brought to light. The concepts of self-sustained migration and damped migration apply to the multi-planet case as well. In the case of self-sustained migration, Neptune tends to migrate to the outer edge of the planetesimal disk or at least up to a large distance (~100 AU) from the Sun (Gomes et al. 2004). In the case of damped migration, the planets attain a final orbital configuration after having removed all planetesimals in between them and leaving a massive disk of objects a few AUs beyond the final orbit of Neptune (Gomes et al. 2004). Given that the Kuiper belt contains very little mass (probably less than 0.01 $M_\oplus$; Fuentes and Holman 2008) and excluding, from considerations based on its orbital and size distributions, that said little mass is the result of collisional grinding (Morbidelli et al. 2008b; see, however, Kenyon et al. 2008 for an opposite view), the original planetesimal disk in our solar system likely had an effective outer edge at ~30 AU, close to the current location of Neptune (Gomes et al. 2004; Morbidelli et al. 2008b).

Another general result is that the eccentricities of the planets are damped during planetesimal-driven migration. This is related to a process called dynamical friction, well known in models of planet formation. In essence, dynamical friction is the mechanism by which gravitating objects of different masses exchange energy so as to evolve toward an equipartition of energy of relative motion (Saslaw 1985): As a general rule, for a system of planets embedded in a massive population of small bodies, the eccentricities and inclinations of the former are damped, while those of the latter are excited (Stewart and Wetherill 1988). Thus, during planetesimal-driven migration, the only possibility for the enhancement of the planets' eccentricities is that the planets pass through mutual resonances as their orbits diverge from each other (Chiang 2003; Tsiganis et al. 2005).

[5]Named for the French city of Nice, where it was developed.

3.2 Multi-resonant Planet Configurations and Planetesimal Scattering: The Solar System Case

When two planets are in a mutual mean-motion resonance, their orbits cannot move freely relative to each other: The resonance holds constant the ratio between the orbital periods, i.e., between the orbital semimajor axes. Thus, both planets have to migrate at the same relative rate. In this case, the differential forces that act on the planets from the planetesimals disk modify instead the orbital eccentricities of the planets. More precisely, if the interaction with the planetesimals is such that, in absence of the mean-motion resonance, the planets would suffer divergent migration (i.e., the ratio between the semimajor axes of the outer and the inner planets would increase), the eccentricities of the planets decrease (Henrard 1993). When the eccentricities are low enough, the planets can eventually exit their resonance. At this point, if the planets are still stable, normal divergent migration, as discussed in the previous section, can start.

Consider now the case of the solar system. Assume, for instance, that the planets were in the least compact of the multi-resonant configurations found in Morbidelli et al. (2007): Saturn is in the 2/3 resonance with Jupiter, Uranus is in the 2/3 resonance with Saturn, and Neptune is in the 3/4 resonance with Uranus. Assume that the system, at the disappearance of the gas, is still embedded in a planetesimal disk, of about 65 $M_\oplus$, with an outer edge at ~30 AU. *Figure 2-8* shows what happens: Due to the gravitational interaction between the planets and the planetesimals, the semimajor axes of the planetary orbits change. Thus, after a short time, the planets are extracted from their multiple resonance. Resonances have a strong stabilizing effect for close orbits (a clear example is that of Pluto which, despite crossing the orbit of Neptune, is stable because it is in the 2/3 mean-motion resonance with it). Once the planets are extracted from their mutual resonances, this stabilizing effect ends. The planets rapidly become unstable, because they are too close to each other. In the simulation illustrated in *Fig. 2-8*, the onset of the instability occurs at $t \sim 2$ My. Uranus and Neptune start to have close encounters with each other and with Saturn. As a result, the orbits of all the planets become more eccentric; Uranus and Neptune are propelled outward by encounters with Saturn, onto very eccentric orbits. Their orbital eccentricities are then strongly damped by dynamical friction, which stabilizes the motion of the planets and prevents further mutual close encounters; the eccentricities of Jupiter and Saturn are also damped but to a lesser extent. At the same time, planetesimal-driven migration, which can operate once the planets are extracted from resonance, increases further the orbital separations between each pair of planets. At the end of the simulation, the planets have acquired orbits very similar to their current ones: The semimajor axes are within 10% of the observed values, and the eccentricities and inclinations are also comparable. Most of the planetesimal disk has been dispersed by then, so that little orbital evolution is expected to occur after 100 My.

This simulation shows that the multi-resonant configuration, which the giant planets should have been driven into during the gas-disk phase, is not incompatible with the current orbital configuration: The interaction with the planetesimals disk and the temporary phase of global instability, which the planets experience after extraction from their original resonances, can very well lead the system to its current dynamical state (Thommes et al. 1999). More examples of this kind of successful evolution, starting also from different multi-resonant orbital configurations, can be found in Batygin and Brown (2010).

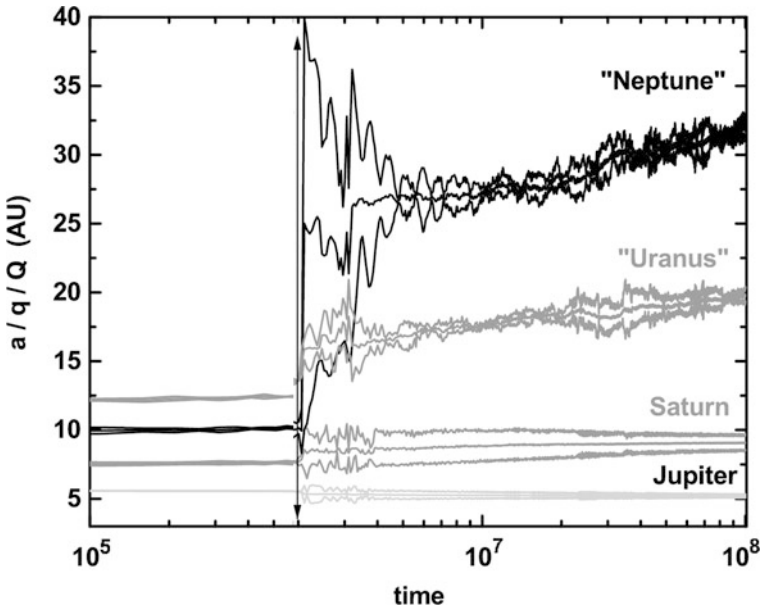


Fig. 2-8

The evolution of the four giant planets of the solar system, starting from a four-body resonance and embedded in a 65 $M_\oplus$ planetesimal disk. Here, each planet is represented by three curves, showing the perihelion distance *q*, the semimajor axis *a*, and the aphelion distance *Q*, respectively. Initially, Saturn is in the 2/3 resonance with Jupiter, Uranus is in the 2/3 resonance with Saturn, and Neptune is in the 3/4 resonance with Uranus. The *vertical arrow* marks the time of the instability, when the orbits of the planets are extracted from the original four-body resonance and become eccentric. In this simulation, Uranus and Neptune are scattered outward by encounters with Saturn and between themselves. The final orbits are quite similar to the current orbits (from Morbidelli et al. (2007))

3.3 The Late Heavy Bombardment as a Smoking Gun for a Late Instability of the Giant Planets

In *Fig. 2-8*, the dynamical instability occurs early, after only 2 My from the beginning of the simulation. However, there is evidence that in our solar system, the onset of the dynamical instability happened much later, approximately 600 My after the disappearance of the disk of gas: This piece of evidence comes from the so-called Late Heavy Bombardment (LHB).

The LHB is a putative cataclysmic period around ~3.9 Gy ago, marked by a collision rate on the Moon that was higher than the one during the immediately preceding period and much higher than the one characterizing the current time. The existence of the LHB was originally proposed by Tera et al. (1974), given the clustering at ~3.9 Gy ago of radiometric impact ages of the lunar samples collected by the Apollo missions (Papanastassiou and Wasserburg 1971a, b; Wasserburg and Papanastassiou 1971; Turner et al. 1973). More recently, the radiometric evidence in favor of the LHB has been reinforced by new laboratory analysis (Cohen et al. 2000) on

lunar meteorites, which should be representative of the whole lunar surface, unlike the Apollo samples.

However, the existence of the LHB has remained controversial since the time when it was first proposed. Some authors (see for instance Neukum and Wilhelms 1982; Neukum and Ivanov 1994; Baldwin 2006; Hartmann et al. 2007) interpret the high bombardment rate ~3.9 Gy ago as the tail of a slowly declining, even-more-intense bombardment occurring since the time of formation of the terrestrial planets. In this view, the paucity of impact ages older that about 4 Gy ago is argued to be the result of a selection effect (see, e.g., Hartmann et al. 2000; Chapman et al. 2007). In essence, the subsequent intense impactor flux would have ground down the oldest impact melts, reducing them to particle sizes smaller than 60 μm. Thus, they would not appear in datable samples.

The analysis of the lunar crater record also gives an ambiguous view of the time evolution of the Lunar bombardment. Neukum and Wilhelms (1982) (see also Neukum 1983; Neukum and Ivanov 1994) studied the crater density over terrains of "known" radiometric age, concluding in favor of a smooth decay of the impactor flux, in opposition to the LHB hypothesis. The problem, however, is that only the youngest units, starting with the Imbrium basin ~3.8–3.9 Gy ago, have well-established radiometric ages, whereas the ages of older basins, like Nectaris, are subjects of debate (e.g., Norman et al. 2010). Neukum and collaborators assumed that age of Nectaris basin is ~4.1 Gy, because this age appears in the samples collected by the Apollo 16 mission in the Descartes region, near Nectaris (Maurer et al. 1978). With this assumption, the density of craters as a function of age of the terrains seems to follow a exponential decline with a decay time of 140 My, which is then extrapolated backward in time, until the Moon formation event (~4.5 Gy ago; see Kleine et al. 2009).

Instead, Stöffler and Ryder (2001) and Ryder (2002) argued that the age of Nectaris is 3.9 Gy, because this age appears more prominently than the 4.1 Gy age among the Descartes region samples of Apollo 16. If one assumes this age, then the *same* crater counts on Nectaris imply a bombardment rate that has a *much steeper* decline over the 3.9–3.5 Gy period than in Neukum and Ivanov (1994). This steeper decline cannot be extrapolated in time back to the lunar formation event, because it would lead to unphysical implications. For instance, the Moon would have accreted more than a Moon mass since its formation (Ryder 2002). Consequently, this implies that the bombardment rate could not have declined smoothly; instead, the bombardment rate should have been smaller before 3.9 Gy ago than in the 3.9–3.5 Gy period, in agreement with the LHB hypothesis. However, as said above, the age of Nectaris is uncertain (Norman et al. 2010). Thus, no definitive conclusion can be derived in favor of the cataclysm or the smooth exponential decline hypothesis.

The following four lines of evidence favor an LHB:

(i) The old upper crustal lithologies of the Moon do not show the enrichment in highly siderophile elements that should be expected if the lunar bombardment had declined exponentially as envisioned in Neukum and Ivanov (1994) (Ryder 2002; Morbidelli et al. 2012).

(ii) Given the fast dynamical and collisional decay of the population of planetesimals that remain in the vicinity of the Earth's orbit at the end of the accretion process of the terrestrial planets, the formation of two huge impact structures such as the Imbrium and Orientale basins (and probably many more) on the Moon 600 My later implies an implausible initial total mass of solids in the inner solar system (Bottke et al. 2007).

(iii) The bombardment rate 3.8–3.9 Gy ago (as deduced from the lunar crater record) was probably not intense enough to vaporize the oceans on Earth (Abramov and Mojzsis 2009). However, if this bombardment rate had been the tail of a more intense bombardment, smoothly decaying over time since lunar formation, the ocean evaporation threshold should have been overcome just a few hundreds of millions of years earlier (~4.2 Gy ago). This contrasts with the oxygen isotopic signature of the oldest known zircons (age: 4.4 Gy), which indicates formation temperatures compatible with the existence of liquid water (Valley et al. 2002).

(iv) The crater size-frequency distributions on the lunar highlands and on the Nectaris basin are different; this difference is well explained by a drastic change in the velocity of the impactors (Marchi et al. 2012). This finding suggests that the dynamics in the inner solar system changed substantially at the time of the formation of the Nectaris basin, as expected in the LHB hypothesis.

The existence of an LHB implies that a massive population of planetesimals must have been stored for ~600 My in a stable reservoir, which then was suddenly destabilized. The only intuitive way to do this is that there was a sudden change in the orbital structure of the planets at that time (Levison et al. 2001). In fact, the planets, in particular the giant ones, control the dynamical evolution of the small bodies and determine which reservoir is stable and which unstable. The planetary evolution illustrated in ❯ *Fig. 2-8* would do a great job in causing an impact spike, by destabilizing the full outer planetesimal disk when Uranus and Neptune are propelled to their current orbital semimajor axes. The asteroid belt would also be partially destabilized when Jupiter and Saturn acquire their current eccentricities and move toward their current orbital separation (Gomes et al. 2005; Minton and Malhotra 2009). The same is true for the putative extension of the asteroid belt toward Mars, named the "E-belt," which probably would have dominated the impact rate on the Moon (Bottke et al. 2010, 2011). However, in the simulation of ❯ *Fig. 2-8*, the spike would occur too early to coincide with the LHB spike. One needs to find a plausible explanation for which the extraction of the giant planets from their original multiple resonance occurred not after only 2 My (as in ❯ *Fig. 2-8*) but approximately 600 My later.

The reason that the instability occurs early in the simulation of ❯ *Fig. 2-8* and in the simulations of Batygin and Brown (2010) is that the planets were assumed to be *embedded* in a planetesimal disk, so that the interaction with said disk was very strong. As pointed out in Gomes et al. (2005), however, this is an unlikely configuration. In fact, the planetesimals that are originally in between the orbits of the giant planets are violently unstable, with a dynamical lifetime well shorter than 1 My. Thus, they should have been removed (by colliding with the planets, being ejected onto distant orbits, etc.) well before the disappearance of the gas, which typically lasts 3–5 My in a protoplanetary disk (Haisch et al. 2001). Then, as said at the beginning of ❯ Sect. 3, the early removal of these planetesimals should not have changed the orbital configuration of the planets, because the dominant forces exerted by the disk of gas forced the planets to stay in their multiple resonance. Therefore, it is more likely that, at the disappearance of the gas, when N-body simulations like that of ❯ *Fig. 2-8* become relevant, the planetesimals were only on those orbits whose dynamical lifetime is of the order of the gas-disk lifetime or longer. This constrains the planetesimals to be in a trans-Neptunian disk, with an inner edge situated at least 1–2 AUs beyond the original semimajor axis of Neptune (for simplicity, the planet that is the most distant from the Sun is designated "Neptune"; notice that in some simulations – that of ❯ *Fig. 2-8* for instance – the two last planets in order of distance from the Sun switch orbits; in these cases, Uranus would have been originally the most distant planet from the Sun).

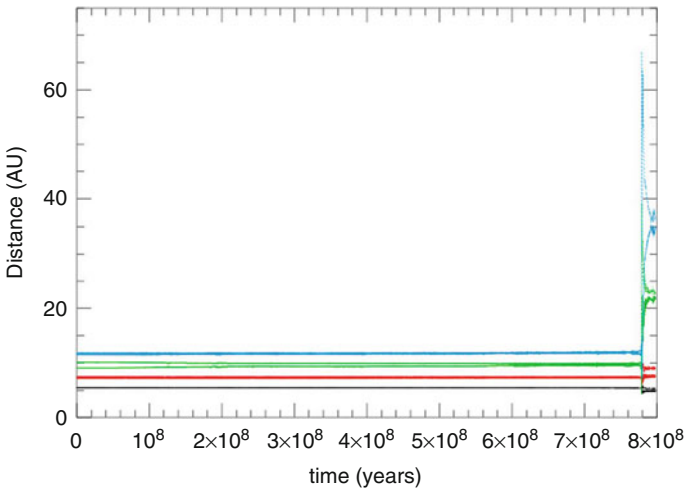


Fig. 2-9

The same as *Fig. 2-8*, but for a planetesimal disk with an inner edge suitably placed beyond the initial orbit of the outermost planet. Here, the instability is delayed to ~700 My, consistent with the timing of the Late Heavy Bombardment of the terrestrial planets

If the planetesimal disk resides beyond the orbit of Neptune, the interactions between the planets and the disk are necessarily much weaker than in the case where the planets are embedded in the disk. In this condition, the instability can occur late, after a time comparable with the LHB chronology, as shown in *Fig. 2-9*. The instability time depends critically on the location of the inner edge of the disk: Disks with inner edges slightly closer to Neptune lead to early instabilities, and disks with edges just a bit further give systems that are stable forever.[6] Such extreme sensitivity looks problematic.

All the simulations presented up to this point, however, were simple, because they assumed that the planetesimals do not to interact dynamically with each other. If self-interactions are taken into account, for instance assuming that there are a few hundreds Pluto-mass objects in the disk perturbing each other and the other particles, then there is a net exchange of energy between the planets and the disk, even if there are no close encounters between planets and planetesimals. This is because the self-stirring of the disk breaks the reversibility of the eccentricity coupling between planet and planetesimals: The planet eccentricities are damped; however, the evolutions of orbital energy and eccentricity are coupled at second order in the masses (Milani et al. 1987): This produces a drift in the planet's energy. In particular, the planets lose energy, i.e., they try to migrate toward the Sun (Levison et al. 2011). The orbits of the planets tend to *approach* each other. This is different from the case where planets scatter planetesimals, and the planetary orbits tend to *separate* from each other. Remember, though, that the planets are in resonances, so the ratios between their semimajor axes cannot change. In response, the planetary

[6]The situation was not nearly as sensitive in Gomes et al. (2005), because the planets were not assumed to be in resonance with each other.

eccentricities slowly *increase*. This eventually drives some planets to pass through secondary or secular resonances, which destabilize the original multi-resonant configuration. The overall evolution is very similar to what is presented in *Fig. 2-9*, but now, the instability time is late in general: In the simulations of Levison et al. (2011), it ranges from 350 My to over 1 Gy for disks with inner edge ranging from 15.5 to 20 AU (Neptune is at ~11.5 AU in these simulations). Unlike the case without self-interactions of disk particles, there is no apparent correlation between instability time and initial location of the inner edge of the disk. This may appear surprising, because the rate of energy exchange between planets and disk decreases with increasing distance of the disk's inner edge. However, this dependence is weak, because the planet-disk interaction is a distant interaction (no close encounters are involved). Then, the expected monotonic dependence of the instability time on the disk's distance can be easily erased by the fact that the evolutions of the disk and of the planets are chaotic, which gives a highly sensitive and nontrivial dependence of the results on the initial conditions. The instability time seems to depend weakly also on the number of Pluto-mass scatterers in the disk, provided that this number exceeds a few hundreds.

Together, the papers by Morbidelli et al. (2007) and Levison et al. (2011) build the new version of the "Nice model." This is much superior than its original version (Tsiganis et al. 2005; Gomes et al. 2005) because (i) it removes the arbitrary character of the initial conditions of the planets by adopting as initial configuration one of the end states of hydrodynamical simulations and (ii) it removes the sensitive dependence of the instability time on the location of the inner edge of the disk; instead, a late instability seems to be a generic outcome.

In the new Nice model, only 10% of the simulations which exhibit a global dynamical instability of the planets lead to a stable four-planets system at the end; in the remaining simulations, one or more planets are lost, ejected onto hyperbolic orbits. The fraction of "successful" simulations in the original work of Tsiganis et al. (2005) was much higher: ~50%. This is because in the new model, the planets are initially in a more compact configuration, and therefore their orbital instability is more violent. This low probability of success may suggests that the model is still missing something important in the reconstruction of the past history of the solar system. For instance, the solar system might have originally had three Neptune-mass planets, one of which was ejected at the time of the instability (Nesvorný 2011). In this case, the fraction of successful simulations (e.g., those ending with four giant planets) increases considerably.

On the positive side, like in Tsiganis et al., when four planets survive in the new Nice model, their final orbits are quite similar to the real ones: The orbital semimajor axes are within 10–15% of the real values, and the final orbital eccentricities and inclinations are also close (within a factor of 2) to the actual ones. This argues that the model, although certainly not perfect, is probably not too far from reality. The match between the final orbits in the simulations and the current real orbits of the giant planets improves in the model assuming that the solar system had originally three Neptune-mass planets and eventually lost one (Nesvorný and Morbidelli 2012).

3.4 The Solar System Debris Disk: Are LHBs Common?

If our understanding of the evolution of the solar system is (even approximately) correct, there should have been a massive belt of planetesimals outside Neptune's orbit during the first ~600 My of history, i.e., up to the LHB time. This disk, through mutual collisions, should have produced a large amount of dust, generating what it is usually referred to as a "debris disk"

(Chapter by Moro-Martin). It is interesting to investigate how our debris disk would have appeared to an extrasolar observer and compare the result to the debris disks that inferred around other stars of various ages.

Booth et al. (2009) addressed this question. The intrinsic collision probabilities and mutual velocities among the planetesimals have been computed from the dynamical simulations of the Nice model (in its old version, but this should not make a big difference in this respect). To compute the outcome of the collisional activity, Booth et al. had to assume an initial size distribution. They adopted the size distribution observed in the current Kuiper belt, with the number of objects in each size bin multiplied by a factor ~1,000. Thus, the initial size distribution contained ~1,000 Pluto-size bodies was relatively steep down to $D \sim 100$ km and then turned over to a shallower slope for sizes smaller than this threshold (Bernstein et al. 2004; Fuentes and Holman 2008). With this assumption, a planetesimal disk with initially 50 $M_\oplus$ of material loses less than 50% of its mass in 600 My because of collisional grinding (see also the supplementary material of Levison et al. 2009). This is an important consistency check for the Nice model: In fact, if collisional grinding rapidly removed most of the mass of the disk no matter the initial size distribution, then there would not be enough mass to affect the evolution of the giant planets at the LHB time.

After having computed the dust production rate as a function of time and the orbital and size distributions of the grains, Booth et al. computed the emission of the disk at wavelengths of 24 and 70 μm, relative to the emission of the Sun. The result is illustrated in ❯ *Fig. 2-10*. It shows that the luminosity of the debris disk should have decayed very slowly, by less than an order of magnitude, during the time preceding the onset of the planetary instability (which occurred at 880 My in the specific simulation of the Nice model used in their calculation). Then, the luminosity should have decayed very rapidly below detectability, as the disk was dynamically dispersed by the planets. The figure also shows, with star symbols, the infrared excess of known stars, as a function of their estimated ages. At 24 μm, only ~15% of the stars younger than 300 My have this kind of excess (Carpenter et al. 2009), and this fraction decreases to a few percent for older stars (Gáspár et al. 2009); at 70 μm, the fraction of stars with detectable infrared excess does not seem to decay with age (Hillenbrand et al. 2008; Trilling et al. 2008).

A first important conclusion, from the comparison of the estimated brightness of our solar system with the IR excess of other stars, is that, before the LHB, our debris disk was fairly typical. A second conclusion, from the fact that there is no general tendency for a sudden disappearance of the IR excess at 70 μm around other stars at about 1 Gy of age, is that a complete dynamical clearing of the planetesimal disk like the one that occurred in our solar system at the LHB time is fairly atypical. From a statistical analysis of the data, Booth et al. estimated that at most 15% of the extrasolar systems undergo such a late dynamical clearing.

The fact that a *late* complete dynamical clearing of the planetesimal disk is a rare event should not be a big surprise. In fact, it is clear from what has been said above that the events described in the Nice model depend on two specific properties that not many planetary systems might have in common with our own. First, the planets in our system did not migrate permanently into the inner solar system; instead, they remained or returned near their birth places, i.e., adjacent to the planetesimal disk that generated them. Thus, when their orbits changed at the time of the instability, they strongly affected the disk. If the planets had migrated close to the Sun and had remained there, they would have lost contact with the distant planetesimal disk. Even if the planets had become unstable in the inner solar system, probably, the distant disk would not have been dynamically cleared. Second, the planetesimal disk of the solar system was small, presumably truncated at ~30 AU (Gomes et al. 2004). If the planetesimal disk had

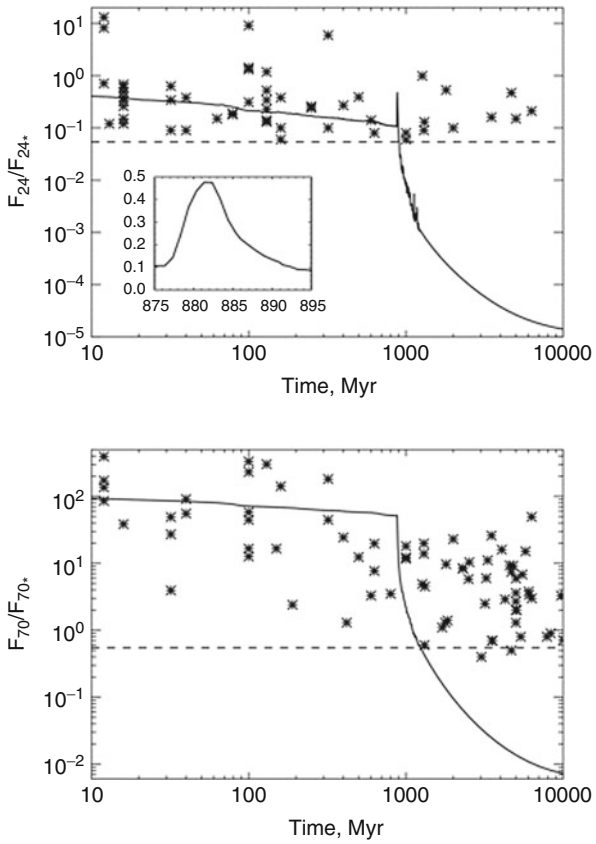


Fig. 2-10

The infrared luminosity of the debris disk of our solar system, according to the Nice model. The *solid line* shows the luminosity of the disk relative to that of the Sun at 24 μm (*top*) and 70 μm (*bottom*). The luminosity decays slowly during the first 880 My, when the planets become unstable (in the adopted simulation). Then, the luminosity of the disk rapidly decays, as the planetesimals are removed from the solar system. The *horizontal dashed line* shows the detection limit for an extrasolar observer with our current measurement capabilities. The *asterisks* represent the observed disks. The window in the *lower left corner* of the *top panel* is a magnification of the evolution around the instability time (from Booth et al. (2009))

been extended to much larger distances, the dynamical instability and the migration of Neptune would have probably cleared the disk up to 50–100 AU; beyond this threshold, the disk would have remained massive and would have continued to produce dust.

Coming back to *Fig. 2-10*, the little spike in the disk's brightness visible at 880 My in the upper plot (magnified in the box in the bottom left corner) is the signature of the LHB event, due to a burst in collisional activity that occurs as the disk starts to be dispersed and its orbital excitation suddenly increases. As one can see, the spike is not prominent enough to make the disk stand out of the natural distribution of brightness of disks of different masses (suggested by the dispersion of the observations reported on the top panel of *Fig. 2-10*).

The Booth et al. calculation, however, does not account for the huge flux of comets into the inner solar system that should have occurred during the disk dispersal: These comets should have liberated a great amount of dust once inside a few AUs from the Sun. Nesvorný et al. (2010) accounted for this effect: They estimated that the inner zodiacal cloud should have been more than 10^4 times brighter during the LHB epoch. As the current infrared excess at 24 μm of the zodiacal cloud is 2×10^{-4} (Kelsall et al. 1998), the excess at the LHB time at this wavelength was probably of order 2–10. This kind of excess is comparable to the upper envelope of the observations. Thus, the conclusion is that luminosity bursts associated to LHB events, totally invisible at 70 μm (cold dust), start to be detectable at 24 μm (hot dust), although they can still be confused with the luminosity of massive disks undergoing gradual collisional grinding. Therefore, the identification of systems that might be undergoing an LHB event at the current time requires a case by case analysis at multiple wavelengths, as done for instance in Wyatt et al. (2007).

4 Terrestrial Planets

Up to this point, this chapter was focussed on the formation and evolution of giant planets. The discussion below considers the evolution of the terrestrial planets.

Terrestrial planets are expected to form in the inner part of a protoplanetary disk, closer to the star than the snow line position. Because of the absence of ices and of the relatively small mass contained in the inner portion of the disk, the largest objects formed by the processes of runaway and oligarchic growth (see ❯ Sect. 2.1) are expected to have masses of about a lunar to a Martian mass (Weidenschilling et al. 1997). These bodies are called *planetary embryos* to distinguish them from the much more massive *planetary cores* that are the precursors of the giant planets. They incorporate about 50% of the mass of solids in the inner disk, the rest remaining in asteroid-sized planetesimals.

The evolution of the system of embryos and planetesimals has been the object of many simulations using N-body integrations (Chambers and Wetherill 1998; Agnor et al. 1999; Chambers 2001; Raymond et al. 2004, 2005, 2006a, 2007, 2009b; O'Brien et al. 2006; Kenyon and Bromley 2006; Kokubo et al. 2006; Thommes et al. 2008; Morishima et al. 2010; Kokubo and Genda 2010). All studies agree on the basic aspects of the terrestrial-planet accretion process, although they may differ in the details. At the disappearance of the gas, the system of embryos becomes violently unstable, due to the mutual interactions among the embryos themselves and to the "external" perturbations from the giant planets. The orbits of the embryos begin to intersect, and accretional collisions between embryos start to occur. This produces a smaller number of more massive objects (see ❯ *Fig. 2-11*), which are eventually identified with the terrestrial planets.

Concerning the planetesimals, a fraction of them contributes to the growth of the planets by colliding with the embryos. The majority of the planetesimals, however, are dispersed onto orbits with large eccentricities and inclinations. In this process (the "dynamical friction" mentioned in ❯ Sect. 3.1), they damp the eccentricities and inclinations of the growing planets, which eventually sets the system into a stable configuration, with the most massive planets on the least eccentric orbits.

This scenario for the formation of the terrestrial planets of our solar system has several strong points:

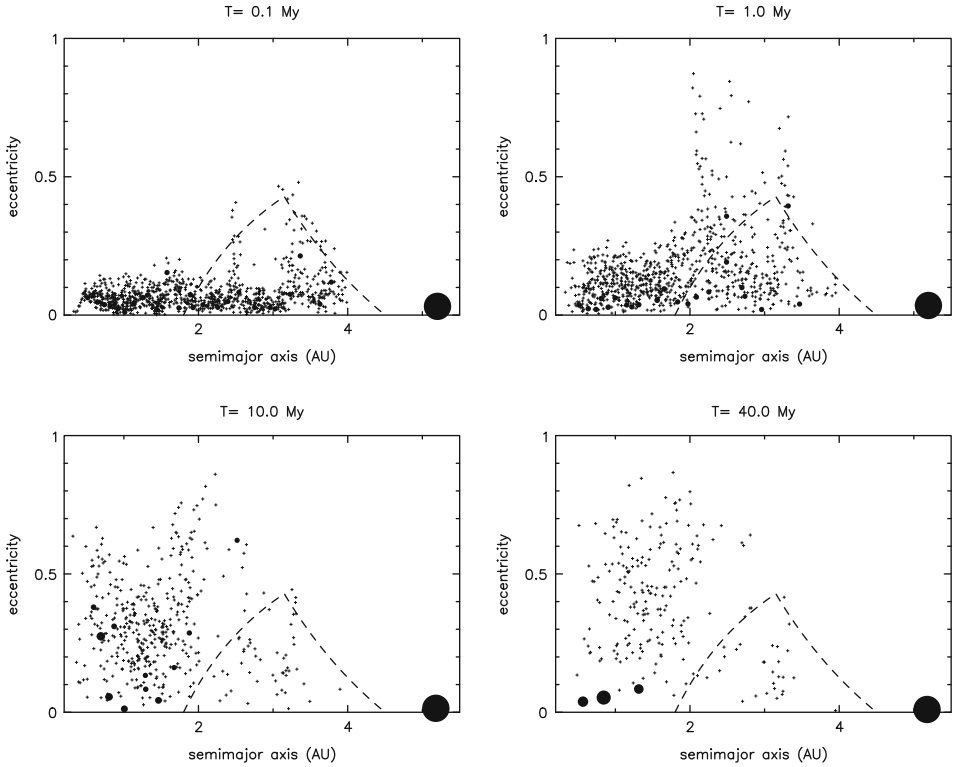


Fig. 2-11

Snapshots of the terrestrial planets formation process, from O'Brien et al. (2006). Each panel shows the eccentricity vs. semimajor axis distribution of planetary embryos (*filled circles*) and planetesimals (*crosses*) at the reported time. The size of each *filled circle* is proportional to the cubic root of the mass of the corresponding embryo. The big filled ball at ~5.2 AU represents Jupiter. The *dashed curves* show the current boundaries of the asteroid belt. Notice the formation of three terrestrial planets in this simulation, the biggest of which is approximately one Earth mass. In the asteroid belt, no terrestrial planet is formed. All embryos have left the asteroid belt region and only a small fraction of the initial planetesimals reside there on excited orbits. In these simulations, all collisions are supposed to be accretional. This approximation has been recently removed by Kokubo and Genda (2010), who considered a database of collisions simulated by the SPH method (like in Agnor and Asphaug 2004; Asphaug et al. 2006) to determine how much mass is accreted or ejected in each collisional event. The final results, though, show very little differences with respect to the simulations that treat all collisions as 100% accretional

- A system of two to five planets is typically formed. The efficiency of the accretion process is about 50%. Thus, starting with $\sim 5 M_{\oplus}$ in embryos and planetesimals typically produces a couple of planets of about an Earth mass each (Chambers 2001). The final orbits of the terrestrial planets produced in the simulations are comparable to those of the real terrestrial planets of our solar system, if the dynamical friction process is properly taken into account (O'Brien et al. 2006).

- Quasi-tangent collisions of Mars-mass embryos onto the protoplanets are quite frequent (Agnor et al. 1999). These collisions are expected to generate a disk of ejecta around the protoplanets (Canup and Asphaug 2001), from which a satellite is likely to accrete (Canup and Esposito 1996). This is the standard, generally accepted scenario for the formation of the Moon.
- The accretion timescale of the Earth analog in the simulations is 30–100 My. This is in the good ballpark with the chronology of Earth accretion as indicated by radioactive chronometers, which still has a comparable uncertainty (Kleine et al. 2009).
- In many/most simulations, terrestrial planets do not form in the asteroid belt. Instead, all the embryos are removed by mutual interactions and perturbations from Jupiter. A small fraction (a few percent) of the planetesimal population is left behind on stable asteroid-belt orbits, with eccentricities and inclinations comparable to those of the real asteroids (Petit et al. 2001; O'Brien et al. 2007).
- A significant fraction (~10–20%) of the mass of the terrestrial planets is accreted from the outer part of the asteroid belt, which provides a formidable mechanism to explain the delivery of water to the Earth (Morbidelli et al. 2000; Raymond et al. 2004, 2007).

On the other hand, this scenario has a major problem: The planet formed in the simulations at the location of Mars is typically too massive (Chambers 2001; Raymond et al. 2009a; Hansen 2009; Morishima et al. 2010). Mars is an oddity not only for its relatively low mass but also for its accretion timescale: In fact, it formed in only a few millions years, like asteroids, and significantly faster than the Earth (Dauphas and Pourmand 2011).

There does not seem to be a simple solution to the Mars problem (Raymond et al. 2009a). Hansen (2009) argued convincingly that a correct mass distribution of the terrestrial planets, with an Earth/Mars mass ratio of ~10, can be achieved only if the initial disk of embryos and planetesimals is assumed to have an outer edge at about 1 AU. The problem is how to justify such an edge and how to reconcile this with the evidence that asteroids exist in the 2–4 AU range. The simulations that assume Jupiter and Saturn initially on orbits with their current separation in semimajor axis but eccentricities two to three times larger do produce very rapidly an effective edge at ~1.5 AU in the distribution of embryos and planetesimals (Raymond et al. 2009a; Morishima et al. 2010) and result in a somewhat small "Mars." However, this initial configuration of the giant planets is inconsistent with our understanding of their orbital evolution through the history of the solar system, described above (see ❯ Sects. 2.4, 3.2 and 3.3). Moreover, none of these simulations are without problems: Mars is often not small enough, the final distribution of bodies in the asteroid belt is not good, water is not delivered to the terrestrial planets, etc.

4.1 Linking Giant Planet Migration to Terrestrial Planet Accretion: The Grand Tack Scenario

The result by Hansen motivated Walsh et al. (2011) to look in more details at the possible orbital history of the giant planets and their ability to sculpt the disk in the inner solar system. For the first time, the giant planets were not assumed to be on static orbits (even if different from the current ones); instead, Walsh et al. studied the coevolution of the orbits of the giant planets and of the precursors of the terrestrial planets, during the era of the disk of gas.

Walsh et al. envisioned the following scenario, based on the considerations reported in ❯ Sect. 2.2: First, Jupiter migrated inward while Saturn was still growing; then, when Saturn

reached a mass close to its current one, it started to migrate inward more rapidly than Jupiter, until it captured the latter in the 2/3 resonance; finally, the two planets migrated outward until the complete disappearance of the disk of gas. The extent of the inward and outward migrations cannot be estimated a priori, because they depend on properties of the disk and of giant planet accretion that are unknown, such as the time lag between Jupiter and Saturn formation, the speed of inward migration (depending on disk's viscosity), the speed of outward migration (depending on disk's scale height), the time lag between the capture in resonance of Jupiter and Saturn, and the photo-evaporation of the gas. However, the extent of the inward and outward migrations of Jupiter can be deduced by looking at the resulting structure of the inner solar system. In particular, Walsh et al. remarked that a reversal of Jupiter's migration at 1.5 AU would provide a natural explanation for the existence of the outer edge at 1 AU of the inner disk of embryos and planetesimals, required to produce a small Mars (see *Fig. 2-12*). Because of the prominent inward-then-outward migration of Jupiter that it assumes, Walsh et al. scenario is nicknamed "Grand Tack."

Several giant extrasolar planets have been discovered orbiting their star at a distance of 1–2 AU, so the idea that Jupiter was sometime in the past at 1.5 AU from the Sun is not shocking by itself. A crucial diagnostic of this scenario, though, is the survival of the asteroid belt. Given that Jupiter should have migrated through the asteroid belt region twice, first inward and, then outward, one could expect that the asteroid belt should now be totally empty. However, the numerical simulations by Walsh et al. show that the asteroid belt is first fully depleted by the passage of the giant planets, but then, while Jupiter leaves the region for the last time, it is repopulated by a small fraction of the planetesimals scattered by the giant planets during their

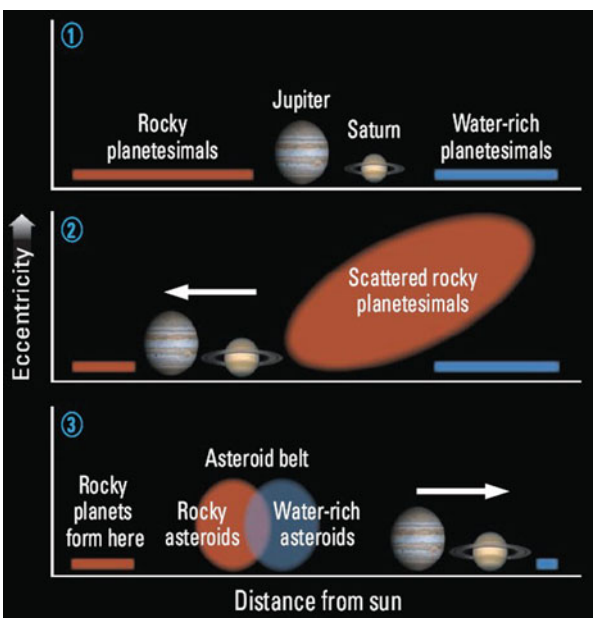


Fig. 2-12

Sketch of the "Grand Tack scenario." The three panels depict subsequent steps in the evolution of the system (form "NEWS&ANALYSIS": Science, 332, 1255 (2011))

migration. In particular, the inner asteroid belt is dominantly repopulated by planetesimals that were originally inside the orbit on which Jupiter formed, while the outer part of the asteroid belt is dominantly repopulated by planetesimals originally in between and beyond the orbits of the giant planets (see *Fig. 2-12*).

Assuming that Jupiter accreted at the location of the snow line, it is then tempting to identify the planetesimals originally closer to the Sun with the un-hydrous asteroids of S type and those originally in between and beyond the orbits of the giant planets with the "primitive" C-type asteroids. With this assumption, the Grand Tack scenario explains the physical structure of the asteroid belt probably better than any other previous model. In fact, the asteroid belt is characterized by a radial gradient in asteroid spectroscopic types (Gradie and Tedesco 1982): The inner belt is dominated S type (usually considered to be the parent bodies of ordinary chondrites; Binzel et al. 1996) and the outer belt by C-type asteroids (usually considered to be the parent bodies of carbonaceous chondrites; (Burbine et al. 2000)), although there is a significant overlapping between the distributions of these different types of asteroids. It is difficult to explain the differences between ordinary chondrite and carbonaceous chondrite parent bodies if they had both formed in the asteroid belt region, given that they are coeval (Villeneuve et al. 2009) and that the radial extent of the asteroid belt is small (~1 AU only). Instead, if ordinary and carbonaceous chondrite parent bodies have been implanted into the asteroid belt from originally well separated reservoirs, as in the Grand Tack scenario, the differences in physical properties are easier to understand in the framework of the classical condensation sequence. The origin of C-type asteroids from the giant planet region would also explain, in a natural way, the similarities with comets that are emerging from recent observational results and sample analysis (see Sect. 7 of the supplementary material of Walsh et al. 2011, for a in-depth discussion). The small mass of the asteroid belt and its eccentricity and inclination distribution are also well reproduced by the Grand Tack scenario.

This scenario also explains why the accretion timescales of Mars and the asteroids are comparable (Dauphas and Pourmand 2011). In fact, the asteroids stopped accreting when they got dispersed and injected onto excited orbits of the main belt; Mars stopped accreting when the inner disk was truncated at 1 AU and the planet was pushed beyond this edge by an encounter with the proto-Earth (Hansen 2009). In the Grand Tack scenario, these two events coincide and mark the time of the passage of Jupiter through the inner solar system.

All these results make the Grand Tack scenario an appealing comprehensive model of terrestrial-planet formation and argue strongly in favor of an evolution of the giant planets of our solar system like that sketched in the right panel of *Fig. 2-6*.

4.2 Terrestrial Planets in Extrasolar Systems

Given that the architecture of the giant planets in our solar system is far from being typical around other stars, it is interesting to investigate the dependence of the terrestrial-planet accretion process on the properties of giant planets, across a wide range of parameters.

From various simulations (Levison and Agnor 2003; Raymond et al. 2004, 2006a), it turns out that the outcome of the terrestrial-planet formation process has a weak dependence on the mass of the giant planets. Obviously, the terrestrial planets cannot form in the vicinity of giants. So, if the giant planets are closer to the star than Jupiter, they leave to the terrestrial-planets a narrower niche to form inside. Instead, the process of terrestrial planets accretion is very sensitive on the eccentricities of the giant planets. Large eccentricities of the giant planets

force large eccentricities on embryos and planetesimals. As a result, the final terrestrial planets will be more eccentric; consequently, they will have a larger separation in semimajor axis and will be less numerous and more massive compared to a simulation where the same giant planets are on circular orbits. Moreover, the planetesimals originally in the vicinity of the giant planets (presumably rich in water and other volatiles) are more likely to be ejected from the system than to collide with the terrestrial planets, if the giant planets are eccentric (Chambers and Cassen 2002; Raymond et al. 2004). So, the resulting terrestrial planets are expected to be water-poor.

The works quoted above assumed giant planets on "fixed" orbits. We know now that the giant planets can have evolutions that lead them to change their orbits, through migration and/or dynamical instabilities. It is interesting to explore how the terrestrial planets, during and after their formation, respond to these changes.

The effect of a Jupiter-mass planet migrating through the disk toward a "hot-Jupiter" orbit has been investigated in Fogg and Nelson (2005, 2007) and Raymond et al. (2006b). These studies showed that a large fraction of the disk's solid mass survives the inward migration of the giant planet in two ways: (i) Planetesimals are captured into mean-motion resonances interior to the orbit of the giant planet and, by mutual collisions, give origin to massive terrestrial planets. (ii) Planetesimals are scattered into external orbits, where gas drag re-circularizes their orbits. The standard terrestrial-planet formation process then resumes. Thus, the widespread expectation that terrestrial-planets could not exist in systems with a hot Jupiter is not correct, and future searches for extrasolar terrestrial planets should not disregard these systems a priori.

4.3 Terrestrial-Planets Evolution During Giant Planets Instabilities

The issue of terrestrial-planet evolution during giant planets instabilities deserves a whole section by itself. As proposed in Sect. 3.3, the giant planets of our solar system passed through a phase of orbital instability ~3.9 Gy ago, i.e., well after the formation of the terrestrial planets (which ended ~4.5 Gy ago; Kleine et al. 2009). During this instability phase, close encounters between the giant planets occurred, the orbits of the giant planets became eccentric, and their separation in semimajor axis increased toward the current values. During this period of chaotic evolution, a wide variety of orbital histories are possible. One may, however, classify the orbital histories in two classes: those in which Jupiter is not involved in close encounters with another planet (only Saturn, Uranus, and Neptune have encounters with each other) and those in which Jupiter has encounters with Uranus and/or Neptune (nicknamed below the *jumping-Jupiter class*). The two classes give a very different evolution of the orbital separation of Jupiter and Saturn.

In the first class of evolutions, the increase in orbital separation between Jupiter and Saturn is due to planetesimal-driven migration. In fact, if Saturn scatters an ice giant (Uranus or Neptune) while Jupiter does not, necessarily, Saturn has to scatter the ice giant outward and recoil toward the Sun. So, planetary encounters in this class of evolutions lead to a reduction of the orbital separation of Jupiter, and Saturn and planetesimal-driven migration is the only mechanism that can increase it. In the jumping-Jupiter class of evolutions, most of the increase in orbital separation between the two gas giants is instead due to planetary encounters. In fact, if Jupiter has an encounter with an ice giant, said ice giant, given that it is beyond the orbit of Saturn both at the beginning and at the end of the evolution, must be first scattered inward by Saturn and then be scattered outward by Jupiter. Thus, Saturn recoils outward and Jupiter inward, which increases the orbital separation between Jupiter and Saturn.

In summary, both classes of evolutions lead to an increase in the orbital separation of Jupiter and Saturn, but the big difference is in the timescale on which this separation occurs. Planetesimal-driven migration is relatively slow: It forces the orbital separation to evolve exponentially as $\Delta a(t) = \Delta a_{\text{current}} - \Delta_0 \exp(-t/\tau)$, with $\tau \sim 5 - 10$ My (the characteristic lifetime of planetesimals crossing the orbits of the giant planets, such as the Centaur objects; Tiscareno and Malhotra 2003; di Sisto and Brunini 2007; Bailey and Malhotra 2009). Conversely, the phase of planetary encounters is short, so that in the jumping-Jupiter class, the orbital separation of Jupiter and Saturn increases in less than 10^5 year (see *Fig. 2-13*).

The increase in orbital separation between Jupiter and Saturn changes the secular frequencies of the orbits of these planets. If the divergent migration of Jupiter and Saturn occurs on a timescale of a few millions of years, as in the case of planetesimal-driven migration, the orbits of the terrestrial planets are quite strongly excited in eccentricity (Brasser et al. 2009). Even if starting from circular orbits, the Earth and Venus acquire orbits whose eccentricity oscillations exceed 0.15, i.e., twice as much as in the real solar system; Mercury is destabilized (*Fig. 2-14*). This happens because the frequency of precession of the perihelion of Jupiter's orbit, denoted by g_5, decreases, and, in sequence, it becomes equal to those of Mars, Earth, Venus, and Mercury ($g_4, \ldots, g_1$), respectively; every time that a frequency of a terrestrial planet g_k is equal to g_5, a secular resonance occurs and the eccentricity of the corresponding planet is strongly affected. Similarly, most of the asteroids in the inner part of the asteroid belt are destabilized while the precession frequency of the perihelion of the orbit of Saturn (g_6) decreases; consequently, the final orbital distribution of the asteroids is incompatible with the one currently observed (Morbidelli et al. 2010). In the jumping-Jupiter class of evolutions, conversely,

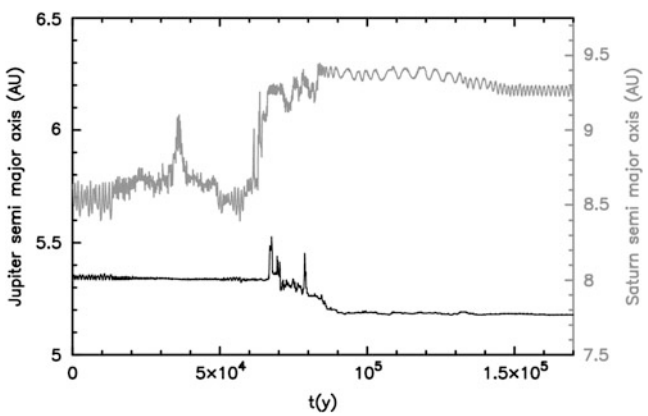


Fig. 2-13

An example of "jumping-Jupiter evolution." The *black and gray curves* show the evolution of the semimajor axes of Jupiter and Saturn, reported in the *left-side* and *right-side vertical scales*, respectively. The stochastic behavior is caused by encounters with a Uranus-Neptune-mass planet, originally placed the third in order of increasing distance from the Sun (not shown here). Time $t = 0$ here is arbitrary and corresponds to the onset of the phase of planetary instability. The full evolution of the planets, which lasts 4.6 My, is illustrated in Fig. 4 of Brasser et al. (2009). All giant planets survived the full 4.6 My simulation on stable orbits, quite similar to those of the real planets of the solar system

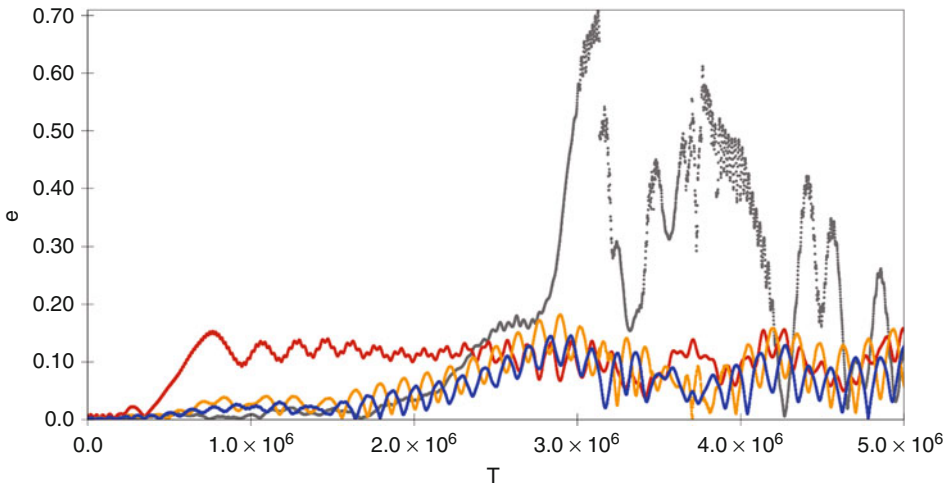


Fig. 2-14

The evolutions of the eccentricities of Mars (*red*), Earth (*blue*), Venus (*orange*) and Mercury (*grey*), during planetesimal-drive migration of Jupiter and Saturn. Here the orbital separation between the two gas giants increases as $\Delta a_{\text{current}} - \Delta_0 \exp(-t/\tau)$, with $\Delta_0 = 1.1$ AU and $\tau = 5$ My

these problems do not exist because the divergent migration of Jupiter and Saturn – and the consequent decrease of g_5 and g_6 – is so fast that the eccentricities of asteroids and terrestrial planets have no time to be seriously affected. Thus, the terrestrial planets can have at the end orbital eccentricities as small (or even smaller) than the current ones, depending on their initial conditions (Brasser et al. 2009), and the asteroid belt preserves roughly the orbital structure acquired during terrestrial-planets formation (Morbidelli et al. 2010).

The conclusion is that the real evolution of the giant planets of our solar system had to be of the jumping-Jupiter class; otherwise, the orbit of the Earth, and the structure of the inner solar system in general, would be substantially different from what they are now.

There is no doubt that many evolutions of the giant planets can be fatal for the formation or the evolution of habitable terrestrial planets. Take the giant extrasolar planets discovered so far: Most of them have very eccentric orbits, which are thought to be the product of a violent instability occurred in the original planetary system, which led to close encounters between giant planets (see Sect. 2.3). Raymond et al. (2011) show that when giant planets acquire similarly large eccentricities, the terrestrial planets in the system are forced to evolve onto orbits with extreme eccentricities: Many of them collide with the central star or start to intersect the orbits of the giant planets and are then rapidly ejected onto hyperbolic orbits. Those terrestrial objects which manage to survive, if any, do so on orbits with eccentricities that can be hardly compatible with habitable worlds. Thus, the existence of an habitable Earth in our system is possible only because our giant planets remained on orbits with exceptionally small eccentricities compared to the orbits of extrasolar planets.

5 Conclusions

This chapter discussed the evolution of planetary systems. The emphasis was directed at reconstructing the history of our system using computer simulations and taking advantage of all possible detailed constraints (the orbits of the planets, the architecture of the populations of small bodies, radioactive chronologies for terrestrial-planet formation, crater records, etc.). Although imperfect, this view of the evolution of the solar system after the completion of giant planets formation has reached a quite satisfactory level of coherence. Conversely, the phase of accretion of the giant planets remains poorly modeled.

According to our understanding, the evolution of the solar system was characterized by three main "eras." In the gas-disk era, Jupiter had a wide-range radial migration. It first migrated inward; then, when it was at about 1.5 AU from the Sun, it got caught in resonance with Saturn, and, given the Jupiter/Saturn mass ratio, it started to migrate outward (Walsh et al. 2011). This inward-then-outward migration explains why the solar system does not have a "hot (or warm) Jupiter." It also left indelible traces in the inner solar system, particularly in the physical structure of the asteroid belt and the small mass of Mars. As a result of the outward migration of Jupiter, the four giant planets acquired a multi-resonant configuration, in which each planet was in a mean-motion resonance with its neighbor. The orbits of the giant planets were at the time much closer to each other than they are now and had significantly smaller eccentricities and inclinations.

At the disappearance of the gas, the system entered in the planetesimal-disk era. The Earth and Venus completed their accretion from a disk of planetary embryos and planetesimals that inherited an outer edge at 1 AU from the earlier incursion of Jupiter into the inner solar system. The accretion of Mars and of the asteroids was frozen. Instead, a massive disk of planetesimals persisted outside the orbit of the outermost giant planet. Its internal collisional activity produced a debris disk comparable to those observed around ~15% of main-sequence stars. Meanwhile, the gravitational interactions between the giant planets and this disk slowly modified the resonant orbits of the former. Eventually, ~600 My later, the giant planets became unstable, as a result of these slow orbital modifications. The chaotic phase that followed reshuffled the structure of the outer solar system: The giant planets acquired their current orbits; most of the distant planetesimal disk was dispersed, and the Kuiper belt is what remains today of that disk (Levison et al. 2008; Batygin et al. 2011). Many asteroids were also released from the asteroid belt. All these destabilized small bodies caused the Late Heavy Bombardment of the terrestrial planets (Bottke et al. 2011).

With this profound reorganization, the solar system entered into the current era, lasting since ~3.8 Gy ago, in which it did not suffer any further significant change.

If this story is true, then the evolution of our solar system was defined by a sequence of specific features. For instance, the mass ratio between Jupiter and Saturn prevented migration toward the Sun; the late formation of the giant planets relative to the gas-disk lifetime prevented Saturn to grow more massive than Jupiter; the instability phase that characterized the giant planets resulted in a jumping-Jupiter evolution, which prevented secular resonances to interfere with the orbits of the terrestrial planets; etc. It was then natural in this chapter to discuss what would have happened if these events had not occurred or if they had occurred differently. This addresses the origin of diversity in extrasolar planetary systems. The possible lines of evolutions described here certainly do not exhaust all possibilities: Nature has much more fantasy than we have. However, they show that the evolution of a planetary system, like the weather on Earth,

is so sensitive on initial and environmental conditions that a huge variety of outcomes is possible, even starting from similar situations.

A frequently asked question, at this point, is whether or not one can predict the probability that a system evolves toward one state or another. The answer at this time, unfortunately, is no. The first stages of planet formation are still poorly understood, including the initial conditions or the properties of protoplanetary disks. Predicting which kinds of planetary systems could be formed and with which probabilities is still out of reach, and without this information, one cannot estimate the probabilities of the subsequent possible evolutions. For instance, the question of why Jupiter and Saturn formed like they are, instead of having different masses or having additional gas-giant companions, is still difficult to answer. In other words, the evolution of the solar system can be reconstructed using all available clues and constraints, but this does not suffice to "predict" our solar system, a priori. This quandary has a parallel with geology. Geologists are able to reconstruct the complicated history of our continents with an amazing precision, but they are not able to say what is the probability that a terrestrial planet develops continents with the properties of our own continents. At this time, the field of "origins" in planetary science is essentially a descriptive discipline. As such, it is led by observations and interpreted by theoretical models, not the other way around.

Acknowledgments

I am grateful to the Helmholtz Alliance's "Planetary Evolution and Life" and the French National Programme for Planetary Science for substantial financial support. I wish to thank all the people with whom I worked over these fabulous years in the quest for a better understanding of the evolution of planetary systems and of our solar system in particular. I am grateful to P. Michel, A. Crida, K. Walsh, and an anonymous reviewer for carefully reading a first draft of this chapter and for providing useful comments and suggestions.

References

Abramov, O., & Mojzsis, S. J. 2009, Microbial habitability of the Hadean Earth during the late heavy bombardment. Nature, 459, 419–422

Adams, F. C., & Laughlin, G. 2003, Migration and dynamical relaxation in crowded systems of giant planets. Icarus, 163, 290–306

Agnor, C., & Asphaug, E. 2004, Accretion efficiency during planetary collisions. ApJ, 613, L157–L160

Agnor, C. B., Canup, R. M., & Levison, H. F. 1999, On the character and consequences of large impacts in the late stage of terrestrial planet formation. Icarus, 142, 219–237

Alibert, Y., Mordasini, C., & Benz, W. 2004, Migration and giant planet formation. A&A, 417, L25–L28

Asphaug, E., Agnor, C. B., & Williams, Q. 2006, Hit-and-run planetary collisions. Nature, 439, 155–160

Bailey, B. L., & Malhotra, R. 2009, Two dynamical classes of Centaurs. Icarus, 203, 155–163

Baldwin, R. B. 2006, Was there ever a Terminal Lunar Cataclysm? With lunar viscosity arguments. Icarus, 184, 308–318

Barge, P., & Sommeria, J. 1995, Did planet formation begin inside persistent gaseous vortices? A&A, 295, L1–L4

Barnes, R., & Greenberg, R. 2006, Stability limits in extrasolar planetary systems. ApJ, 647, L163–L166

Baruteau, C., & Masset, F. 2008, On the corotation torque in a radiatively inefficient disk. ApJ, 672, 1054–1067

Batygin, K., & Brown, M. E. 2010, Early dynamical evolution of the solar system: pinning down the initial condition of the nice model. ArXiv e-prints arXiv:1004.5414

Batygin, K., Brown, M. E., & Fraser, W. C. 2011, In-situ formation of the cold classical Kuiper belt. ApJ, 738, p 13.

Beauge, C., & Nesvorny, D. 2011, Multiple-planet scattering and the origin of hot Jupiters. ArXiv e-prints arXiv:1110.4392

Bernstein, G. M., Trilling, D. E., Allen, R. L., Brown, M. E., Holman, M., & Malhotra, R. 2004, The size distribution of trans-neptunian bodies. AJ, 128, 1364–1390

Binzel, R. P., Bus, S. J., Burbine, T. H., & Sunshine, J. M. 1996, Spectral properties of near-earth asteroids: evidence for sources of ordinary chondrite meteorites. Science, 273, 946–948

Bitsch, B., & Kley, W. 2010, Orbital evolution of eccentric planets in radiative discs. A&A, 523, A30

Bitsch, B., & Kley, W. 2011, Range of outward migration and influence of the disc's mass on the migration of giant planet cores. A&A, 536, A77

Bodenheimer, P., Hubickyj, O., & Lissauer, J. J. 2000, Models of the in situ formation of detected extrasolar giant planets. Icarus, 143, 2–14

Boley, A. C. 2009, The two modes of gas giant planet formation. ApJ, 695, L53–L57

Booth, M., Wyatt, M. C., Morbidelli, A., Moro-Martín, A., & Levison, H. F. 2009, The history of the Solar system's debris disc: observable properties of the Kuiper belt. MNRAS, 399, 385–398

Boss, A. P. 2000, Possible rapid gas giant planet formation in the Solar Nebula and other protoplanetary disks. ApJ, 536, L101–L104

Boss, A. P. 2001, Formation of planetary-mass objects by protostellar collapse and fragmentation. ApJ, 551, L167–L170

Boss, A. P. 2002, Stellar metallicity and the formation of extrasolar gas giant planets. ApJ, 567, L149–L153

Bottke, W. F., Levison, H. F., Nesvorný, D., & Dones, L. 2007, Can planetesimals left over from terrestrial planet formation produce the lunar Late Heavy Bombardment? Icarus, 190, 203–223

Bottke, W. F., Vokrouhlicky, D., Nesvorny, D., Minton, D., Morbidelli, A., & Brasser, R. 2010, The E-belt: a possible missing link in the late heavy bombardment. Lunar Planet. Inst. Sci. Conf. Abstr., 41, 1269

Bottke, W. F., Vokrouhlicky, D., Minton, D., Nesvorny, D., Brasser, R., & Simonson, B. 2011, The great archean bombardment, or the late heavy bombardment. Lunar Planet. Inst. Sci. Conf. Abstr., 42, 2591

Brasser, R., Morbidelli, A., Gomes, R., Tsiganis, K., & Levison, H. F. 2009, Constructing the secular architecture of the solar system II: the terrestrial planets. A&A, 507, 1053–1065

Burbine, T. H., Binzel, R. P., Bus, S. J., Buchanan, P. C., Hinrichs, J. L., Hiroi, T., Meibom, A., & Sunshine, J. M. 2000, Forging asteroid-meteorite relationships through reflectance spectroscopy. Lunar Planet. Inst. Sci. Conf. Abstr., 31, 1844

Butler, R. P., et al. 2006, Catalog of nearby exoplanets. ApJ, 646, 505–522

Cameron, A. G. W. 1978, Physics of the primitive solar accretion disk. Moon Planets, 18, 5–40

Canup, R. M., & Asphaug, E. 2001, Origin of the Moon in a giant impact near the end of the Earth's formation. Nature, 412, 708–712

Canup, R. M., & Esposito, L. W. 1996, Accretion of the moon from an impact-generated disk. Icarus, 119, 427–446

Canup, R. M., & Ward, W. R. 2006, A common mass scaling for satellite systems of gaseous planets. Nature, 441, 834–839

Capobianco, C. C., Duncan, M., & Levison, H. F. 2011, Planetesimal-driven planet migration in the presence of a gas disk. Icarus, 211, 819–831

Carpenter, J. M., et al. 2009, Formation and evolution of planetary systems: properties of debris dust around solar-type stars. ApJSS, 181, 197–226

Cassen, P. M., Smith, B. F., Miller, R. H., & Reynolds, R. T. 1981, Numerical experiments on the stability of preplanetary disks. Icarus, 48, 377–392

Chambers, J. E. 2001, Making more terrestrial planets. Icarus, 152, 205–224

Chambers, J. 2006, A semi-analytic model for oligarchic growth. Icarus, 180, 496–513

Chambers, J. E., & Cassen, P. 2002, The effects of Nebula surface density profile and giant-planet. Meteoritics and Planetary Science 37, 1523–1540

Chambers, J. E., & Wetherill, G. W. 1998, Making the terrestrial planets: N-body integrations of planetary embryos in three dimensions. Icarus, 136, 304–327

Chapman, C. R., Cohen, B. A., & Grinspoon, D. H. 2007, What are the real constraints on the existence and magnitude of the late heavy bombardment? Icarus, 189, 233–245

Chatterjee, S., Ford, E. B., Matsumura, S., & Rasio, F. A. 2008, Dynamical outcomes of planet-planet scattering. ApJ, 686, 580–602

Chiang, E. I. 2003, Excitation of orbital eccentricities by repeated resonance crossings: requirements. ApJ, 584, 465–471

Cohen, B. A., Swindle, T. D., & Kring, D. A. 2000, Support for the lunar cataclysm hypothesis from lunar meteorite impact melt ages. Science, 290, 1754–1756

Cresswell, P., Dirksen, G., Kley, W., & Nelson, R. P. 2007, On the evolution of eccentric and inclined protoplanets embedded in protoplanetary disks. A&A, 473, 329–342

Crida, A., & Morbidelli, A. 2007, Cavity opening by a giant planet in a protoplanetary disc and effects on planetary migration. MNRAS, 377, 1324–1336

Crida, A., Morbidelli, A., & Masset, F. 2006, On the width and shape of gaps in protoplanetary disks. Icarus, 181, 587–604

Crida, A., Sándor, Z., & Kley, W. 2008, Influence of an inner disc on the orbital evolution of massive planets migrating in resonance. A&A, 483, 325–337

Crida, A., Masset, F., & Morbidelli, A. 2009, Long range outward migration of giant planets, with application to fomalhaut b. ApJ, 705, L148–L152

D'Angelo, G., Lubow, S. H., & Bate, M. R. 2006, Evolution of giant planets in eccentric disks. ApJ, 652, 1698–1714

Dauphas, N., & Pourmand, A. 2011, Hf-W-Th evidence for rapid growth of Mars and its status as a planetary embryo. Nature, 473, 489–492

di Sisto, R. P., & Brunini, A. 2007, The origin and distribution of the Centaur population. Icarus, 190, 224–235

Durisen, R. H., Boss, A. P., Mayer, L., Nelson, A. F., Quinn, T., & Rice, W. K. M. 2007, Gravitational instabilities in gaseous protoplanetary disks and implications for giant planet formation, in Protostars and Planets V, ed. B. Reipurth, D. Jewitt, & K. Keil (Tucson: University of Arizona Press), 607–622

Fernandez, J. A., & Ip, W.-H. 1984, Some dynamical aspects of the accretion of Uranus and Neptune – the exchange of orbital angular momentum with planetesimals. Icarus, 58, 109–120

Ferraz-Mello, S., Beaugé, C., & Michtchenko, T. A. 2003, Evolution of migrating planet pairs in resonance. Celest. Mech. Dyn. Astron., 87, 99–112

Fischer, D. A., & Valenti, J. 2005, The planet-metallicity correlation. ApJ, 622, 1102–1117

Fogg, M. J., & Nelson, R. P. 2005, Oligarchic and giant impact growth of terrestrial planets in the presence of gas giant planet migration. A&A, 441, 791–806

Fogg, M. J., & Nelson, R. P. 2007, On the formation of terrestrial planets in hot-Jupiter systems. A&A, 461, 1195–1208

Ford, E. B., Havlickova, M., & Rasio, F. A. 2001, Dynamical instabilities in extrasolar planetary systems containing two giant planets. Icarus, 150, 303–313

Ford, E. B., & Rasio, F. A. 2008, Origins of eccentric extrasolar planets: testing the planet-planet scattering model. ApJ, 686, 621–636

Fouchet, T., Moses, J. I., & Conrath, B. J. 2009, Saturn: composition and chemistry, ed. M. Dougherty, L. Esposito, S. Krimigis et al. (Springer). Saturn from Cassini-Huygens, 83

Fuentes, C. I., & Holman, M. J. 2008, a SUBARU archival search for faint trans-neptunian objects. AJ, 136, 83–97

Gáspár, A., Rieke, G. H., Su, K. Y. L., Balog, Z., Trilling, D., Muzzerole, J., Apai, D., & Kelly, B. C. 2009, The low level of debris disk activity at the time of the late heavy bombardment: a spitzer study of Praesepe. ApJ, 697, 1578–1596

Goldreich, P., & Sari, R. 2003, Eccentricity evolution for planets in gaseous disks. ApJ, 585, 1024–1037

Goldreich, P., & Tremaine, S. 1979, The excitation of density waves at the Lindblad and corotation resonances by an external potential. ApJ, 233, 857–871

Goldreich, P., & Tremaine, S. 1980, Disk-satellite interactions. ApJ, 241, 425–441

Goldreich, P., Lithwick, Y., & Sari, R. 2004, Final stages of planet formation. ApJ, 614, 497–507

Gomes, R. S., Morbidelli, A., & Levison, H. F. 2004, Planetary migration in a planetesimal disk: why did Neptune stop at 30 AU? Icarus, 170, 492–507

Gomes, R., Levison, H. F., Tsiganis, K., & Morbidelli, A. 2005, Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Nature, 435, 466–469

Gradie, J., & Tedesco, E. 1982, Compositional structure of the asteroid belt. Science, 216, 1405–1407

Greenberg, R., Hartmann, W. K., Chapman, C. R., & Wacker, J. F. 1978, Planetesimals to planets – numerical simulation of collisional evolution. Icarus, 35, 1–26

Greenzweig, Y., & Lissauer, J. J. 1992, Accretion rates of protoplanets. II – Gaussian distributions of planetesimal velocities. Icarus, 100, 440–463

Guillot, T. 2005, The interiors of giant planets: models and outstanding questions. Annu. Rev. Earth Planet. Sci., 33, 493–530

Guillot, T., & Hueso, R. 2006, The composition of Jupiter: sign of a (relatively) late formation in a chemically evolved protosolar disc. MNRAS, 367, L47–L51

Guillot, T., Santos, N. C., Pont, F., Iro, N., Melo, C., & Ribas, I. 2006, A correlation between the heavy element content of transiting extrasolar planets and the metallicity of their parent stars. A&A, 453, L21–L24

Haisch, K. E., Jr., Lada, E. A., & Lada, C. J. 2001, Disk frequencies and lifetimes in young clusters. ApJ, 553, L153–L156

Hansen, B. M. S. 2009, Formation of the terrestrial planets from a narrow annulus. ApJ, 703, 1131–1140

Hartmann, W. K., Ryder, G., Dones, L., & Grinspoon, D. 2000, The time-dependent intense bombardment of the primordial earth/moon system, in Origin of the Earth and Moon, ed. R. M. Canup, K. Righter, et al. (Tucson: University of Arizona Press), 493–512

Hartmann, W. K., Quantin, C., & Mangold, N. 2007, Possible long-term decline in impact rates. 2. Lunar impact-melt data regarding impact history. Icarus, 186, 11–23

Hayashi, C. 1981, Structure of the Solar Nebula, growth and decay of magnetic fields and effects of magnetic and turbulent viscosities on the Nebula. Prog. Theor. Phys. Suppl., 70, 35–53

Henrard, J. 1993, The adiabatic invariants in classical mechanics. Dyn. Rep., 2, 117–235

Hillenbrand, L. A., Carpenter, J. M., Kim, J. S., Meyer, M. R., Backman, D. E., Moro-Martín, A., Hollenbach, D. J., Hines, D. C., Pascucci, I., & Bouwman, J. 2008, The complete census of 70 μm-bright Debris Disks within "the Formation and Evolution of Planetary Systems" Spitzer Legacy Survey of Sun-like Stars. ApJ, 677, 630–656

Horn, B., Lyra, W., Mac Low, M.-M., & Sándor, Z. 2012, Orbital migration of interacting low-mass planets in evolutionary radiative turbulent models. ArXiv e-prints arXiv:1202.1868

Ida, S., & Lin, D. N. C. 2004, Toward a deterministic model of planetary formation. II. The formation and retention of gas giant planets around stars with a range of metallicities. ApJ, 616, 567–572

Ida, S., & Lin, D. N. C. 2008, Toward a deterministic model of planetary formation. V. Accumulation near the ice line and super-earths. ApJ, 685, 584–595

Ida, S., & Makino, J. 1993, Scattering of planetesimals by a protoplanet – slowing down of runaway growth. Icarus, 106, 210

Ida, S., Bryden, G., Lin, D. N. C., & Tanaka, H. 2000, Orbital migration of neptune and orbital distribution of trans-neptunian objects. ApJ, 534, 428–445

Johansen, A., Youdin, A., & Mac Low, M.-M. 2009, Particle clumping and planetesimal formation depend strongly on metallicity. ApJ, 704, L75–L79

Jurić, M., & Tremaine, S. 2008, Dynamical origin of extrasolar planet eccentricity distribution. ApJ, 686, 603–620

Kalas, P., Graham, J. R., Chiang, E., Fitzgerald, M. P., Clampin, M., Kite, E. S., Stapelfeldt, K., Marois, C., & Krist, J. 2008, Optical images of an exosolar planet, 25 light-years from Earth. Science, 322, 1345

Kelsall, T., et al. 1998, The COBE diffuse infrared background experiment search for the cosmic infrared background. II. Model of the interplanetary dust cloud. ApJ, 508, 44–73

Kenyon, S. J., & Bromley, B. C. 2006, Terrestrial planet formation. I. The transition from oligarchic growth to chaotic growth. AJ, 131, 1837–1850

Kenyon, S. J., Bromley, B. C., O'Brien, D. P., & Davis, D. R. 2008, Formation and collisional evolution of Kuiper belt objects, in The Solar System Beyond Neptune, ed. M. A. Barucci et al. (Tucson: University of Arizona Press), 293–313

Kirsh, D. R., Duncan, M., Brasser, R., & Levison, H. F. 2009, Simulations of planet migration driven by planetesimal scattering. Icarus, 199, 197–209

Kleine, T., Touboul, M., Bourdon, B., Nimmo, F., Mezger, K., Palme, H., Jacobsen, S. B., Yin, Q.-Z., & Halliday, A. N. 2009, Hf-W chronology of the accretion and early evolution of asteroids and terrestrial planets. Geochim. Cosmochim. Acta, 73, 5150–5188

Kley, W., & Crida, A. 2008, Migration of protoplanets in radiative discs. A&A, 487, L9–L12

Kley, W., & Dirksen, G. 2006, Disk eccentricity and embedded planets. A&A, 447, 369–377

Kley, W., Peitz, J., & Bryden, G. 2004, Evolution of planetary systems in resonance. A&A, 414, 735–747

Kley, W., Lee, M. H., Murray, N., & Peale, S. J. 2005, Modeling the resonant planetary system GJ 876. A&A, 437, 727–742

Kokubo, E., & Genda, H. 2010, Formation of terrestrial planets from protoplanets under a realistic accretion condition. ApJ, 714, L21–L25

Kokubo, E., & Ida, S. 1998, Oligarchic growth of protoplanets. Icarus, 131, 171–178

Kokubo, E., Kominami, J., & Ida, S. 2006, Formation of terrestrial planets from protoplanets. I statistics of basic dynamical properties. ApJ, 642, 1131–1139

Lambrechts, M., & Johansen, A. 2012, Rapid growth of gas-giant cores by pebble accretion. A&A (in press)

Levison, H. F., & Agnor, C. 2003, The role of giant planets in terrestrial planet formation. AJ, 125, 2692–2713

Levison, H. F., & Morbidelli, A. 2007, Models of the collisional damping scenario for ice-giant planets and Kuiper belt formation. Icarus, 189, 196–212

Levison, H. F., Lissauer, J. J., & Duncan, M. J. 1998, Modeling the diversity of outer planetary systems. AJ, 116, 1998–2014

Levison, H. F., Dones, L., Chapman, C. R., Stern, S. A., Duncan, M. J., & Zahnle, K. 2001, Could the Lunar "Late Heavy Bombardment" Have Been

Triggered by the Formation of Uranus and Neptune? Icarus, 151, 286–306

Levison, H. F., Morbidelli, A., Gomes, R., & Backman, D. 2007, Planet migration in planetesimal disks, in Protostars and Planets V, ed. B. Reipurth, D. Jewitt, & K. Keil (Tucson: University of Arizona Press), 669–684

Levison, H. F., Morbidelli, A., Vanlaerhoven, C., Gomes, R., & Tsiganis, K. 2008, Origin of the structure of the Kuiper belt during a dynamical instability in the orbits of Uranus and Neptune. Icarus, 196, 258–273

Levison, H. F., Bottke, W. F., Gounelle, M., Morbidelli, A., Nesvorný, D., & Tsiganis, K. 2009, Contamination of the asteroid belt by primordial trans-Neptunian objects. Nature, 460, 364–366

Levison, H. F., Thommes, E., & Duncan, M. J. 2010, Modeling the formation of giant planet cores. I. evaluating key processes. AJ, 139, 1297–1314

Levison, H. F., Morbidelli, A., Tsiganis, K., Nesvorny, D., & Gomes, R. 2011, Late orbital instabilities in the outer planets induced by interaction with a self-gravitating planetesimal disk. AJ, 142, 152

Lin, D. N. C., & Ida, S. 1997, On the origin of massive eccentric planets. ApJ, 477, 781

Lin, D. N. C., & Papaloizou, J. 1986a. On the tidal interaction between protoplanets and the primordial solar nebula. II – self-consistent nonlinear interaction. ApJ, 307, 395–409

Lin, D. N. C., & Papaloizou, J. 1986b, On the tidal interaction between protoplanets and the protoplanetary disk. III – orbital migration of protoplanets. ApJ, 309, 846–857

Lin, D. N. C., Bodenheimer, P., & Richardson, D. C. 1996, Orbital migration of the planetary companion of 51 Pegasi to its present location. Nature, 380, 606–607

Lodders, K. 2003, Solar system abundances and condensation temperatures of the elements. ApJ, 591, 1220–1247

Lynden-Bell, D., & Pringle, J. E. 1974, The evolution of viscous discs and the origin of the Nebular variables. MNRAS, 168, 603–637

Lyra, W., Johansen, A., Klahr, H., & Piskunov, N. 2009a. Standing on the shoulders of giants. Trojan Earths and vortex trapping in low mass self-gravitating protoplanetary disks of gas and solids. A&A, 493, 1125–1139

Lyra, W., Johansen, A., Zsom, A., Klahr, H., & Piskunov, N. 2009b. Planet formation bursts at the borders of the dead zone in 2D numerical simulations of circumstellar disks. A&A, 497, 869–888

Lyra, W., Paardekooper, S.-J., & Mac Low, M.-M. 2010, Orbital migration of low-mass planets in evolutionary radiative models: avoiding catastrophic infall. ApJ, 715, L68–L73

Malhotra, R. 1993, The origin of Pluto's peculiar orbit. Nature, 365, 819–821

Malhotra, R. 1995, The origin of pluto's orbit: implications for the solar system beyond Neptune. AJ, 110, 420

Marchi, S., Bottke, W. F., Kring, D. A., & Morbidelli, A. 2012, The onset of the lunar cataclysm as recorded in its ancient crater populations. Earth Planet. Sci. Lett., 325, 27–38

Marois, C., Macintosh, B., Barman, T., Zuckerman, B., Song, I., Patience, J., Lafrenière, D., & Doyon, R. 2008, Direct imaging of multiple planets orbiting the star HR 8799. Science, 322, 1348

Marzari, F., & Weidenschilling, S. J. 2002, Eccentric extrasolar planets: the jumping Jupiter model. Icarus, 156, 570–579

Marzari, F., Baruteau, C., & Scholl, H. 2010, Planet-planet scattering in circumstellar gas disks. A&A, 514, L4

Masset, F. S., & Casoli, J. 2010, Saturated torque formula for planetary migration in viscous disks with thermal diffusion: recipe for protoplanet population synthesis. ApJ, 723, 1393–1417

Masset, F., & Snellgrove, M. 2001, Reversing type II migration: resonance trapping of a lighter giant protoplanet. MNRAS, 320, L55–L59

Masset, F. S., Morbidelli, A., Crida, A., & Ferreira, J. 2006, Disk surface density transitions as protoplanet traps. ApJ, 642, 478–487

Maurer, P., Eberhardt, P., Geiss, J., Grogler, N., Stettler, A., Brown, G. M., Peckett, A., & Krahenbuhl, U. 1978, Pre-Imbrian craters and basins – ages, compositions and excavation depths of Apollo 16 breccias. Geochim. Cosmochim. Acta, 42, 1687–1720

Milani, A., Nobili, A. M., & Carpino, M. 1987, Secular variations of the semimajor axes – theory and experiments. A&A, 172, 265–279

Militzer, B., & Hubbard, W. B. 2009, Comparison of Jupiter interior models derived from first-principles simulations. Astrophys. Space Sci., 322, 129–133

Min, M., Dullemond, C. P., Kama, M., & Dominik, C. 2011, The thermal structure and the location of the snow line in the protosolar Nebula: axisymmetric models with full 3-D radiative transfer. Icarus, 212, 416–426

Minton, D. A., & Malhotra, R. 2009, A record of planet migration in the main asteroid belt. Nature, 457, 1109–1111

Moeckel, N., & Armitage, P. J. 2012, Hydrodynamic outcomes of planet scattering in transitional discs. MNRAS, 419, 366–376

Moekel, N., Raymond, S. N., & Armitage, Ph. J. 2008, Extrasolar planets eccentricities from scattering in the presence of residual gas-disks. ApJ, 688, 1361–1367

Moorhead, A. V., & Adams, F. C. 2005, Giant planet migration through the action of disk torques and planet scattering. Icarus, 178, 517–539

Morbidelli, A., & Crida, A. 2007, The dynamics of Jupiter and Saturn in the gaseous protoplanetary disk. Icarus, 191, 158–171

Morbidelli, A., Chambers, J., Lunine, J. I., Petit, J. M., Robert, F., Valsecchi, G. B., & Cyr, K. E. 2000, Source regions and time scales for the delivery of water to Earth. Meteorit. Planet. Sci., 35, 1309–1320

Morbidelli, A., Tsiganis, K., Crida, A., Levison, H. F., & Gomes, R. 2007, Dynamics of the giant planets of the solar system in the gaseous protoplanetary disk and their relationship to the current orbital architecture. AJ, 134, 1790–1798

Morbidelli, A., Crida, A., Masset, F., & Nelson, R. P. 2008a, Building giant-planet cores at a planet trap. A&A, 478, 929–937

Morbidelli, A., Levison, H. F., & Gomes, R. 2008b. The dynamical structure of the Kuiper belt and its primordial origin, in The Solar System Beyond Neptune, ed. M. A. Barucci et al. (Tucson: University of Arizona Press), 275–292

Morbidelli, A., Brasser, R., Gomes, R., Levison H. F., & Tsiganis, K., 2010, Evidence from the asteroid belt for a violent past evolution of Jupiter's orbit. AJ, 140, 1391–1401

Morbidelli, A., Marchi, S., & Bottke, W. F. 2012, The saw timeline of the first billion year of Lunar bombardment. LPI Contrib., 1649, 53–54

Morishima, R., Stadel, J., & Moore, B. 2010, From planetesimals to terrestrial planets: N-body simulations including the effects of Nebular gas and giant planets. Icarus, 207, 517–535

Morfill, G. E., & Voelk, H. J. 1984, Transport of dust and vapor and chemical fractionation in the early protosolar cloud. ApJ, 287, 371–395

Nelson, R. P. 2005, On the orbital evolution of low mass protoplanets in turbulent, magnetised disks. A&A, 443, 1067–1085

Nelson, R. P., & Papaloizou, J. C. B. 2003, The interaction of a giant planet with a disc with MHD turbulence – II. The interaction of the planet with the disc. MNRAS, 339, 993–1005

Nesvorný, D. 2011, Young solar system's fifth giant planet? ApJ, 742, L22

Nesvorný, D., & Morbidelli, A., 2012, Statistical study of the early solar systems instability with 4, 5 and 6 giant planets. AJ (in press)

Nesvorný, D., Jenniskens, P., Levison, H. F., Bottke, W. F., Vokrouhlický, D., & Gounelle, M. 2010, Cometary origin of the zodiacal cloud and carbonaceous micrometeorites. Implications for hot debris disks. ApJ, 713, 816–836

Nettelmann, N., Holst, B., Kietzmann, A., French, M., Redmer, R., & Blaschke, D. 2008, Ab initio equation of state data for hydrogen, helium, and water and the internal structure of Jupiter. ApJ, 683, 1217–1228

Neukum, G. 1983, Habilitation Dissertation for Faculty Membership (Munich: University of Munich)

Neukum, G., & Ivanov, B. A. 1994, Crater size distributions and impact probabilities on Earth from Lunar, terrestrial-planet, and asteroid cratering data, in Hazards Due to Comets and Asteroids, ed. T. Gehrels (Tucson: University of Arizona Press), 359

Neukum, G., & Wilhelms, D. E. 1982, Ancient lunar impact record. Lunar Planet. Inst. Sci. Conf. Abstr., 13, 590–591

Norman, M. D., Duncan, R. A., & Huard, J. J. 2010, Imbrium provenance for the Apollo 16 Descartes terrain: argon ages and geochemistry of lunar breccias 67016 and 67455. Geochim. Cosmochim. Acta, 74, 763–783

O'Brien, D. P., Morbidelli, A., & Levison, H. F. 2006, Terrestrial planet formation with strong dynamical friction. Icarus, 184, 39–58

O'Brien, D. P., Morbidelli, A., & Bottke, W. F. 2007, The primordial excitation and clearing of the asteroid belt – Revisited. Icarus, 191, 434–452

Öpik, E. J., 1976, Interplanetary Encounters: Close Range Gravitational Interactions (Elsevier, New York)

Paardekooper, S.-J., & Mellema, G. 2006, Halting type I planet migration in non-isothermal disks. A&A, 459, L17–L20

Paardekooper, S.-J., & Papaloizou, J. C. B. 2009, On corotation torques, horseshoe drag and the possibility of sustained stalled or outward protoplanetary migration. MNRAS, 394, 2283–2296

Paardekooper, S.-J., Baruteau, C., Crida, A., & Kley, W. 2010, A torque formula for non-isothermal type I planetary migration – I. Unsaturated horseshoe drag. MNRAS, 401, 1950–1964

Papaloizou, J. C. B., Nelson, R. P., & Masset, F. 2001, Orbital eccentricity growth through disc-companion tidal interaction. A&A, 366, 263–275

Papanastassiou, D. A., & Wasserburg, G. J. 1971a. Rb-Sr ages of igneous rocks from the Apollo 14 mission and the age of the Fra Mauro formation. Earth Planet. Sci. Lett., 12, 36

Papanastassiou, D. A., & Wasserburg, G. J. 1971b. Lunar chronology and evolution from Rb-Sr studies of Apollo 11 and 12 samples. Earth Planet. Sci. Lett., 11, 37

Petit, J.-M., Morbidelli, A., & Chambers, J. 2001, The primordial excitation and clearing of the asteroid belt. Icarus, 153, 338–347

Pierens, A., & Nelson, R. P. 2008, Constraints on resonant-trapping for two planets embedded in a protoplanetary disc. A&A, 482, 333–340

Podolak, M., & Zucker, S. 2004, A note on the snow line in protostellar accretion disks. Meteorit. Planet. Sci., 39, 1859–1868

Pollack, J. B., Hubickyj, O., Bodenheimer, P., Lissauer, J. J., Podolak, M., & Greenzweig, Y. 1996, Formation of the giant planets by concurrent accretion of solids and gas. Icarus, 124, 62–85

Rafikov, R. R. 2004, Fast accretion of small planetesimals by protoplanetary cores. AJ, 128, 1348–1363

Rasio, F. A., & Ford, E. B. 1996, Dynamical instabilities and the formation of extrasolar planetary systems. Science, 274, 954–956

Raymond, S. N., Quinn, T., & Lunine, J. I. 2004, Making other earths: dynamical simulations of terrestrial planet formation and water delivery. Icarus, 168, 1–17

Raymond, S. N., Quinn, T., & Lunine, J. I. 2005, Terrestrial planet formation in disks with varying surface density profiles. ApJ, 632, 670–676

Raymond, S. N., Quinn, T., & Lunine, J. I. 2006a, High-resolution simulations of the final assembly of Earth-like planets I. Terrestrial accretion and dynamics. Icarus, 183, 265–282

Raymond, S. N., Mandell, A. M., & Sigurdsson, S. 2006b. Exotic Earths: forming habitable worlds with giant Planet migration. Science, 313, 1413–1416

Raymond, S. N., Quinn, T., & Lunine, J. I. 2007, High-resolution simulations of the final assembly of Earth-like Planets. 2. Water delivery and planetary habitability. Astrobiology, 7, 66–84

Raymond, S. N., Barnes, R., Veras, D., Armitage, P. J., Gorelick, N., & Greenberg, R. 2009a, Planet-Planet scattering leads to tightly packed planetary systems. ApJ, 696, L98–L101

Raymond, S. N., O'Brien, D. P., Morbidelli, A., & Kaib, N. A. 2009b. Building the terrestrial planets: constrained accretion in the inner Solar System. Icarus, 203, 644–662

Raymond, S. N., Armitage, P. J., Moro-Martín, A., Booth, M., Wyatt, M. C., Armstrong, J. C., Mandell, A. M., Selsis, F., & West, A. A. 2011, Debris disks as signposts of terrestrial planet formation. A&A, 530, A62

Ryder, G., 2002, Mass flux in the ancient Earth-Moon system and the benign implications for the origin of life on Earth. J. Geophys. Res.-Planets, 107, 6–14

Sándor, Z., Lyra, W., & Dullemond, C. P. 2011, Formation of planetary cores at type I migration traps. ApJ, 728, L9

Saslaw, W. C. 1985, Thermodynamics and galaxy clustering – relaxation of N-body experiments. ApJ, 297, 49–60

Stamatellos, D., & Whitworth, A. P. 2008, Can giant planets form by gravitational fragmentation of discs? A&A, 480, 879–887

Stewart, G., & Wetherill, G. 1988, Evolution of planetesimal velocities. Icarus, 79, 542–553

Shakura, N. I., & Sunyaev, R. A. 1973, Black holes in binary systems. Observational appearance. A&A, 24, 337–355

Stöffler, D., & Ryder, G. 2001, Stratigraphy and isotope ages of lunar geologic units: chronological standard for the inner solar system. Space Sci. Rev., 96, 9–54

Tanaka, H., Takeuchi, T., & Ward, W. R. 2002, Three-Dimensional Interaction between a Planet and an isothermal gaseous disk. I. Corotation and Lindblad torques and planet migration. ApJ, 565, 1257–1274

Tera, F., Papanastassiou, D. A., & Wasserburg, G. J. 1974, Isotopic evidence for a terminal lunar cataclysm. Earth Planet. Sci. Lett., 22, 1

Thommes, E. W., Duncan, M. J., & Levison, H. F. 1999, The formation of Uranus and Neptune in the Jupiter-Saturn region of the Solar System. Nature, 402, 635–638

Thommes, E. W., Duncan, M. J., & Levison, H. F. 2003, Oligarchic growth of giant planets. Icarus, 161, 431–455

Thommes, E., Nagasawa, M., & Lin, D. N. C. 2008, Dynamical shake-up of planetary systems. II N-body simulations of solar system terrestrial planet formation induced by secular resonance sweeping. ApJ, 676, 728–739

Tiscareno, M. S., & Malhotra, R. 2003, The dynamics of known centaurs. AJ, 126, 3122–3131 Trail, D., Mojzsis, S. J., & Harrison, T. M. 2007, Thermal events documented in Hadean zircons by ion microprobe depth profiles. Geochim. Cosmochim. Acta, 71, 4044–4065

Trilling, D. E., Bryden, G., Beichman, C. A., Rieke, G. H., Su, K. Y. L., Stansberry, J. A., Blaylock, M., Stapelfeldt, K. R., Beeman, J. W., & Haller, E. E. 2008, Debris disks around sun-like stars. ApJ, 674, 1086–1105

Tsiganis, K., Gomes, R., Morbidelli, A., & Levison, H. F. 2005, Origin of the orbital architecture of the giant planets of the Solar System. Nature, 435, 459–461

Turner, G., Cadogan, P. H., & Yonge, C. J. 1973, Apollo 17 age determinations. Nature, 242, 513–515

Valley J. W., Peck W. H., King E. M., & Wilde S. A. 2002, A cool early Earth. Geology, 30, 351–354

Valsecchi, A., & Manara, G. B. 1997, Dynamics of comets in the outer planetary region. II. Enhanced planetary masses and orbital evolutionary paths. A&A, 323, 986–998

Veras, D., & Armitage, P. J. 2004, Outward migration of extrasolar planets to large orbital radii. MNRAS, 347, 613–624

Veras, D., Crepp, J. R., & Ford, E. B. 2009, Formation, survival, and detectability of planets beyond 100 AU. ApJ, 696, 1600–1611

Walsh, K. J., Morbidelli, A., Raymond, S. N. O'Brien, D. P., & Mandell, A. M. 2011, A low mass for Mars from Jupiter's early gas-driven migration. Nature, 475, 206–209

Ward, W. R. 1986, Density waves in the solar Nebula – differential Lindblad torque. Icarus, 67, 164–180

Ward, W. R. 1997, Protoplanet migration by Nebula tides. Icarus, 126, 261–281

Weidenschilling, S. J. 1977, The distribution of mass in the planetary system and solar Nebula. Astrophys. Space Sci., 51, 153–158

Weidenschilling, S. J., & Davis, D. R. 1985, Orbital resonances in the solar Nebula – implications for planetary accretion. Icarus, 62, 16–29

Weidenschilling, S. J., & Marzari, F. 1996, Gravitational scattering as a possible origin for giant planets at small stellar distances. Nature, 384, 619–621

Weidenschilling, S. J., Spaute, D., Davis, D. R., Marzari, F., & Ohtsuki, K. 1997, Accretional evolution of a planetesimal swarm. Icarus, 128, 429–455

Wetherill, G. W. 1992, An alternative model for the formation of the asteroids. Icarus, 100, 307–325

Wetherill, G. W., & Stewart, G. R. 1989, Accumulation of a swarm of small planetesimals. Icarus, 77, 330–357

Wetherill, G. W., & Stewart, G. R. 1993, Formation of planetary embryos – effects of fragmentation, low relative velocity, and independent variation of eccentricity and inclination. Icarus, 106, 190

Villeneuve, J., Chaussidon, M., & Libourel, G. 2009, Homogeneous distribution of 26Al in the solar system from the Mg isotopic composition of chondrules. Science, 325, 985–988

Wasserburg, G. J., & Papanastassiou, D. A. 1971, Age of an Apollo 15 mare basalt: lunar crust and mantle evolution. Earth Planet. Sci. Lett., 13, 97

Wong, M. H., Mahaffy, P. R., Atreya, S. K., Niemann, H. B., & Owen, T. C. 2004, Updated Galileo probe mass spectrometer measurements of carbon, oxygen, nitrogen, and sulfur on Jupiter. Icarus, 171, 153–170

Wyatt, M. C., Smith, R., Greaves, J. S., Beichman, C. A., Bryden, G., & Lisse, C. M. 2007, Transience of hot dust around sun-like stars. ApJ, 658, 569–583

Zhang, H., & Zhou, J.-L. 2010, On the orbital evolution of a giant planet pair embedded in a gaseous disk. I. Jupiter-Saturn configuration. ApJ, 714, 532–548

3 Terrestrial Planets

Nadine G. Barlow
Department of Physics and Astronomy, Northern Arizona University, Flagstaff, AZ, USA

1 Introduction *112*

2 Earth *117*

3 Venus *137*

4 Mars *149*

5 Mercury *173*

6 Moon *178*

7 Summary *189*

References *190*

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_3, © Springer Science+Business Media Dordrecht 2013

Abstract: The four terrestrial planets (Mercury, Venus, Earth, and Mars) and Earth's Moon display similar compositions, interior structures, and geologic histories. The terrestrial planets formed by accretion ~4.5 Ga ago out of the solar nebula, whereas the Moon formed through accretion of material ejected off Earth during a giant impact event shortly after Earth formed. Geophysical investigations (gravity anomalies, seismic analysis, heat flow measurements, and magnetic field studies) reveal that all five bodies have differentiated into a low-density silicate crust, an intermediate density silicate mantle, and an iron-rich core. Seismic and heat flow measurements are only available for Earth and its Moon, and only Earth and Mercury currently exhibit actively produced magnetic fields (although Mars and the Moon retain remanent fields). Surface evolutions of all five bodies have been influenced by impact cratering, volcanism, tectonism, and mass wasting. Aeolian activity only occurs on bodies with a substantial atmosphere (Venus, Earth, and Mars) and only Earth and Mars display evidence of fluvial and glacial processes. Earth's volcanic and tectonic activity is largely driven by plate tectonics, whereas those processes on Venus result from vertical motions associated with hotspots and mantle upwellings. Mercury displays a unique tectonic regime of global contraction caused by gradual solidification of its large iron core. Early large impact events stripped away much of Mercury's crust and mantle, produced Venus' slow retrograde rotation, ejected material off Earth that became the Moon, and may have created the Martian hemispheric dichotomy. The similarities and differences between the interiors and surfaces of these five bodies provide scientists with a better understanding of terrestrial planet evolutionary paths.

Keywords: Aeolian, Asthenosphere, Core, Crust, Earth, Fluvial, Geology, Glacial, Impact craters, Interior structure, Lithosphere, Mantle, Mars, Mass wasting, Mercury, Moon, Plate tectonics, Tectonism, Terrestrial planets, Venus, Volcanism

List of Abbreviations: *ASTER*, Advanced Spaceborne Thermal Emission and Reflection radiometer; *ESA*, European Space Agency; *Ga*, Giga-years (billion years, or 10^9 years); *JHU APL*, Johns Hopkins University Applied Physics Laboratory; *JPL*, Jet Propulsion Laboratory; *KREEP*, Potassium (K), Rare Earth Element, and Phosphorus rich lunar rocks; *KT*, Cretaceous-Tertiary; *LCROSS*, Lunar CRater Observation and Sensing Satellite; *LHB*, Late Heavy Bombardement; *LPI*, Lunar and Planetary Institute; *LRO*, Lunar Reconnaissance Orbiter; *Ma*, Million years (10^6 years); *MESSENGER*, MErcury Surface, Space ENvironment, GEochemistry, and Ranging; *MFF*, Medusae Fossae Formation; *MGS*, Mars Global Surveyor; *MOLA*, Mars Orbiter Laser Altimeter; *MSSS*, Malin Space Science Systems; *NASA*, National Aeronautics and Space Administration; *PLD*, Polar layered deposits; *SELENE*, SELenological and ENgineering Explorer; *SMART-1*, First Small Missions for Advanced Research and Technology; *SNC*, Shergottites, Nakhlites, and Chassignites (Martian meteorites)

1 Introduction

Terrestrial planets are those which resemble Earth: small, rocky, and close to the Sun. The four terrestrial planets are Mercury, Venus, Earth, and Mars, although Earth's Moon is often included in these discussions. Solid surfaces of these five objects reveal the geologic processes which have shaped these bodies. These processes can be divided into endogenic (originating from inside the body) and exogenic (originating outside of the surface). Endogenic processes include volcanism and tectonism, which result from a body's internal heat, and mass wasting, which results from

the planet's gravity. Exogenic processes include aeolian (wind), fluvial (movement of liquids), glacial (movement of ice), and impact cratering processes. Endogenic processes and impact cratering have affected all of the terrestrial planets, but aeolian processes are only seen on bodies with an atmosphere (Venus, Earth, and Mars) and fluvial/glacial processes only occur on bodies with an atmosphere and the right temperature range (Earth and Mars).

Planetary surfaces and interiors are often described using the technique of comparative planetology, which extends our understanding of processes on Earth to other planets. This chapter will first describe the surface and interior of the Earth and then apply this understanding to the other terrestrial planets. The amount of geologic activity on terrestrial planets is dictated by the body's size (*Fig. 3-1*) – large objects retain internal heat for longer periods of time and thus remain geologically active longer than smaller bodies. Earth, being the largest of the terrestrial planets, is the most geologically active, followed by Venus and Mars. Mercury and the Moon are geologically dying or dead worlds at the present time, although their surfaces retain evidence of activity early in their histories.

Geologic processes affecting terrestrial planet surfaces can be directly observed by orbiting and landed spacecraft. Interior structure, however, must be inferred remotely by geophysical techniques: analysis of how mass variations affect orbiting spacecraft through changes in gravitational attraction, seismic studies of how energy travels through the planet's interior, heat flow measurements, and analysis of any present or past magnetic field. Seismic analysis gives the most detailed view of a planetary interior, but such data are lacking except for Earth and the Moon. Heat flow measurements also are currently limited to Earth and the Moon, and only Earth and Mercury have active magnetic fields at present (although Mars and the Moon apparently had magnetic fields early in their histories). Thus most of our information about planetary interiors comes from gravity analysis of trajectory variations of orbiting spacecraft.

Gravity analysis investigates variations in gravitational potential (U) measured by orbiting spacecraft (Hubbard 1984). U is related to the mass of the body (M) and the distance of the spacecraft from the center of the body (r) through

$$U = -\frac{GM}{r}$$

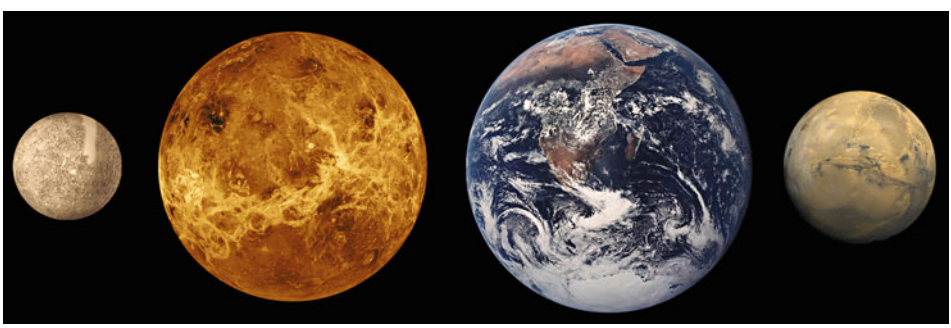


Fig. 3-1

Terrestrial planet sizes. The relative sizes of (*left* to *right*) Mercury, Venus, Earth, and Mars are shown in this diagram. Venus is about 95% of the Earth's diameter (0.95 $D_\oplus$), whereas Mars is 0.53 $D_\oplus$ and Mercury is 0.38 $D_\oplus$. Earth's Moon, not shown here, is 0.27 $D_\oplus$

where G is the universal gravitational constant ($G = 6.672 \times 10^{-11}$ m^3 kg^{-1} s^{-2}). The gravitational potential outside of the planet's surface satisfies Laplace's equation since the planet's entire mass lies inside of the surface:

$$\Delta^2 U_{\text{ext}} = 0$$

The solution to this equation is generally written as

$$U = \left(\frac{GM}{r}\right)\left\{1 + \sum_{l=2}^{n}\sum_{m=0}^{l}\left(\frac{R}{r}\right)^{l} P_l^m(\cos\theta)[C_{lm}\cos(m\lambda) + S_{lm}\sin(m\lambda)]\right\}$$

where R is the radius of the planet, θ is the colatitude ($\theta = 90°$-latitude), λ is longitude, and P_l^m are Legendre polynomials. The index l is called the degree and indicates the rate at which the gravitational potential varies in latitude, whereas m is the order and indicates the variation of U in longitude. Coefficients C_{lm} and S_{lm} are tesseral harmonics and provide information on variations in mass distribution within the planet. C_{lo} values are called zonal harmonics since they provide information about mass distributions parallel to the equator, and S_{lo} values (sectoral harmonics) define mass distributions perpendicular to the equator. Tesseral harmonics can be thought of as a grid pattern subdividing the planet into blocks of mass distributions. A finer grid pattern is obtained with higher values of l and m, resulting in greater detail about the distribution of mass within the planet's interior. Gravity analysis for Earth is completed to degree and order 2,159, meaning the Earth is subdivided into 2,159 zones parallel to the equator and 2,159 sectors perpendicular to the equator. This results in Earth being divided into $(2{,}159)^2 = 4{,}661{,}281$ small blocks over which variations in shape and mass can be determined. The gravity fields for the Moon, Mercury, Venus, and Mars are known to degree and order 165, 25, 180, and 85, respectively.

The C_{lo} nomenclature for zonal harmonics is often replaced by J_l in geophysical applications. J_2 describes the equatorial bulge produced by the planet's rotation. J_2 is related to the body's mass (M), radius (R), and three principal moments of inertia ($A < B < C$):

$$MR^2 J_2 \approx C - A \approx C - B$$

C is the maximum moment of inertia and corresponds to the body's rotation axis (as long as the rotation is not chaotic). A and B are the moments of inertia perpendicular to C and to each other and pass through the body's equator. A and B are approximately equal for planets and moons which have acquired an approximately spherical shape. Precession is the changing direction that a planet's rotation axis points relative to the stars and is caused by gravitational perturbations from other solar system bodies acting on the planet's equatorial bulge. The rate of precession (H) of a planet's rotation axis is given by

$$H = \frac{C - A}{C}$$

Thus, measuring the values of J_2 and H allows determination of the specific values of C and A (and B since $B \approx A$). The value of C/MR^2 is of particular interest in planetary geophysics because it provides insights into the size of a central core. The C/MR^2 value of a homogenous sphere is 0.4. $C/MR^2 < 0.4$ implies central condensation, and a smaller value of C/MR^2 indicates a larger core. Thus gravity analysis obtained by orbiting spacecraft provides information about the planet's shape (from variations in the value of U expected for a perfect sphere), near-surface mass distributions (from U), and interior structure (from C/MR^2).

Radiometric age dating of meteorites indicates that the Solar System formed ~4.5 Ga ago from a collapsing cloud of gas and dust (the solar nebula). The Sun's heat drove volatile-rich

material to the outer solar system, resulting in the terrestrial planets forming from primarily volatile-poor rocky materials. Condensation temperatures through the solar nebula explain the small size, rocky compositions, and lack of extensive atmospheres for the terrestrial planets. Volatiles currently seen on Earth and Mars are from a combination of hydrated minerals incorporated into the planets during formation (Drake and Righter 2002) and water provided by subsequent asteroid and comet impacts (Morbidelli et al. 2000). Terrestrial planets formed by accretion, where small objects stick together during collisions to form larger objects. As these planetismals grew more massive, their gravitational pull increased, pulling in more material until the object became a planetary embryo about the size of Mars. Collisions with other planetismals subsequently altered these planetary embryos, making some planets larger by adding material (Earth and Venus) and making some smaller through ejection of material into space (Mercury) (Canup and Agnor 2000).

Accretion converts the kinetic energy of planetismals into heat, thereby increasing the internal temperature of the growing planet. Decay of short-lived radioactive elements provides additional heat. These two heat sources are sufficient to melt a planet within a few million years after its formation, causing mineral segregation which produces an interior layered structure (differentiation). Minerals are segregated during differentiation by both their density and their chemical affinities. Higher density materials sink to the center, forming the planet's core. Low-density materials float to the surface to produce the crust, which is separated from the core by intermediate density materials comprising the mantle. But elements also can display chemical affinities and follow other elements regardless of density. Siderophile elements follow iron to the core while lithophile elements follow oxygen and silicon to the crust. Thus gold, silver, iridium, and platinum tend to be found deep in planetary interiors because of their siderophile nature whereas aluminum, calcium, and potassium are common in surface rocks because of their lithophilic properties. Gravity data suggest that all four terrestrial planets and the Moon have differentiated interior structures with similar compositions: an iron-rich core, an intermediate-density mantle composed of iron- and magnesium-rich silicates such as olivine, and a crust composed of silicate minerals rich in aluminum, potassium, and sodium (feldspars). The major differences in interior structure among the terrestrial planets are the relative sizes of the core, mantle, and crust and whether planets have solid and/or liquid cores (*Fig. 3-2*).

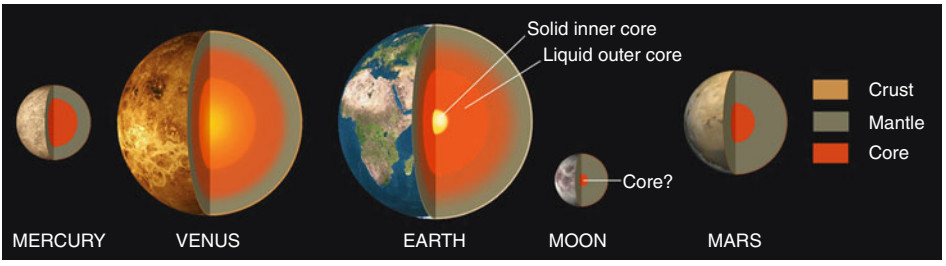


Fig. 3-2

Terrestrial planet interior structure. Geophysical analysis indicates that all four terrestrial planets and the Moon are differentiated into a core, mantle, and crust. Mercury's core is large relative to that of the other terrestrial planets, suggesting that much of Mercury's crust and mantle may have been stripped away by a large impact after differentiation

The terms crust, mantle, and core delineate the chemical structure of a planetary interior. The uppermost portions of the interior (crust and uppermost mantle) also can be divided by their physical characteristics into the lithosphere and asthenosphere. The lithosphere, composed of the crust and uppermost part of the upper mantle, is the rigid outer layer which fractures when stresses are applied. The underlying asthenosphere (lower part of the upper mantle) is warmer and more ductile, causing it to deform under the influence of stress. The lithosphere and asthenosphere are important divisions when considering plate tectonics and the formation of multi-ring impact basins.

Terrestrial planet surfaces cooled and solidified shortly after differentiation. The surfaces began to retain scars of collisions with smaller bodies once they became strong enough. Impact cratering has been a continuous process throughout solar system history and provides a mechanism for determining ages of planetary surface units. Crater density is the number of craters of a certain diameter and larger per area (in km^2). Older surfaces have higher crater densities than younger surfaces (*Fig. 3-3*). However, if the present-day cratering rate is extrapolated back 4.5 Ga, the predicted crater densities underestimate the observed crater densities on surfaces older than 3.8 Ga. This indicates that cratering rates were much higher prior to 3.8 Ga and therefore this period is called the Late Heavy Bombardment (LHB). Traditionally the LHB was proposed to be a gradual decline from extremely high cratering rates at 4.5 Ga resulting from impact of debris remaining from planet formation. However, analysis of lunar samples returned by the Apollo and Luna missions and lunar meteorites recovered on Earth revealed that none of

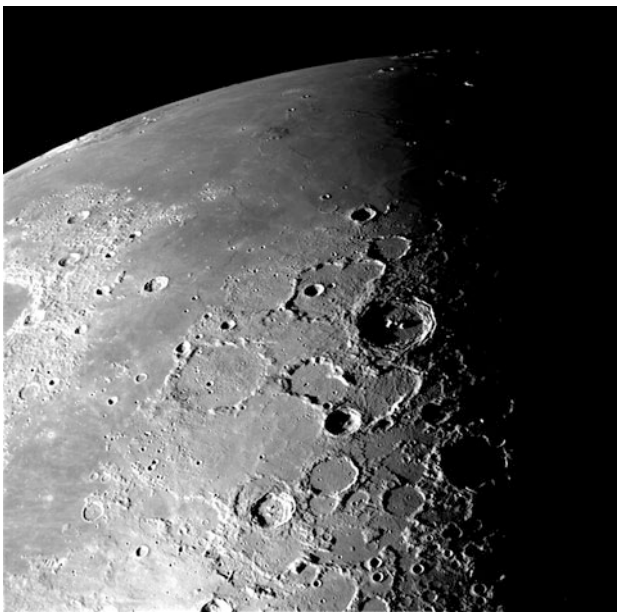


Fig. 3-3
Crater density differences. This image of the Moon, taken by the Galileo spacecraft in December 1992, shows the heavily cratered north polar region at *lower right* and lightly cratered Mare Imbrium at *top*. The higher density of impact craters near the north pole indicates this region is older than Mare Imbrium (NASA/Jet Propulsion Laboratory (JPL))

these samples displayed an age older than ~4.0 Ga (Kring and Cohen 2002; Stöffler and Ryder 2001; Tera et al. 1974). This suggested that the LHB was actually a spike in impact flux between about 3.8 and 4.1 Ga, which was named the Lunar Cataclysm.

Recent dynamical modeling provides a possible mechanism for an inner solar system-wide cataclysm around 3.9 Ga. Numerical modeling suggests the four outer solar system planets (Jupiter, Saturn, Uranus, and Neptune) formed on near-circular orbits closer together than at present. Gravitational interactions with leftover planetismals caused Jupiter to move inward to its present position and Saturn, Uranus, and Neptune to move outward. As Saturn approached the 2:1 resonance with Jupiter (where Saturn's orbital period would be exactly twice that of Jupiter), it rapidly moved outward, jumping over the 2:1 resonance and forcing Uranus and Neptune to also move outward. This rapid migration of the outer planets perturbed the remaining planetismals, causing some to move outward to form the Kuiper Belt, and ejecting others entirely from the solar system (Tsiganis et al. 2005). The migration also perturbed material in the asteroid belt, causing it to move inward and bombard the inner solar system (Gomes et al. 2005; Strom et al. 2005). Lunar geochemical analysis (Kring and Cohen 2002) and dynamical modeling of the original mass of the asteroid belt (Bottke et al. 2007) support the hypothesis that the LHB was produced by this sudden influx of asteroids. Scars of the LHB are seen on ancient surfaces on the Moon, Mercury, and Mars; recent geologic activity on Venus and Earth has erased the record from this time period on these planets.

2 Earth

Earth's large size (12,742 km diameter) has allowed it to retain substantial internal heat (with additional contributions from the decay of long-lived radioactive elements) as well as an atmosphere (*Fig. 3-4*). The average global heat flux for Earth is 75 milliWatts m^{-2}($mW\ m^{-2}$), although this value ranges from near zero in old continental crust to 350 $mW\ m^{-2}$ in volcanically active regions. This heat drives Earth's volcanic and tectonic activity.

Gravity and seismic analyses reveal that Earth is differentiated into a crust, mantle, liquid outer core, and solid inner core. The crust is divided into continental and oceanic components based on composition. Oceanic crust is volcanic in origin, having a basaltic composition with density near 3,000 $kg\ m^{-3}$. Oceanic crust is quite thin, being no thicker than about 12 km. Continental crust is typically about 35 km thick, although its thickness can reach up to 70 km. It has a generally granitic composition with density near 2,700 $kg\ m^{-3}$. Seismic waves show a decrease in velocity at the base of the crust – this region is called the Mohorovicic discontinuity, or Moho. The Moho represents the crust–mantle boundary and corresponds to a change in rock composition from feldspar-rich rocks in the crust to ultrabasic rocks in the mantle.

The mantle extends to a depth of 2,900 km and is composed of silicate rocks rich in iron and magnesium, such as olivine, spinel, and peridotite. The mantle is divided into upper and lower parts, with the upper mantle lying between ~35 and 660 km depth and the lower mantle extending from 660 to 2,900 km depth. Earth's rigid lithosphere is composed of the crust and uppermost part of the upper mantle, extending to an average depth of 60 km. The ductile asthenosphere extends from ~60 to 200 km depth. Seismic analysis indicates the mantle is solid, but material is under high enough pressures and temperatures that it can deform in response

Fig. 3-4
Earth from Space. This image of Earth shows *brown* land of South America, the *white ice* of Antarctica, the *blue* of the oceans, and the *white clouds* in the Earth's atmosphere. Image was taken by the Galileo spacecraft in December 1990 (NASA/JPL)

to stresses over long time periods. The mantle–core boundary is represented by a seismic discontinuity called the Weichert-Gutenberg discontinuity and represents a transition not only in composition but also from solid to liquid phase.

Gravity analysis reveals that $C/MR^2 = 0.3308$, indicating that Earth has a core. Seismic analysis divides the core into outer and inner parts. Seismic S-waves do not travel through the outer core, indicating it is liquid. The outer core extends from 2,900 to 5,150 km depth, is composed of iron mixed with nickel and trace amounts of lighter elements, and has a temperature of ~5,000 K. The inner core was discovered to be solid through seismic analysis by Inge Lehmann (1888–1993) in 1936. It has a radius of ~3,400 km and, like the outer core, is composed primarily of iron with some nickel and traces of light elements. Its temperature is about 6,000 K, but the high pressure produced by overlying material results in this region being solid.

Convective motions within the liquid outer core give rise to Earth's magnetic field. The core's high temperature ionizes iron atoms, resulting in the free flow of electrons. The resulting time-varying electric current induces Earth's magnetic field – this mechanism is called a dynamo. Earth's magnetic field strength is characterized by the dipole magnetic moment, which currently has a value of 7.9×10^{12} Tm^3. The magnetic field axis is offset from the rotation axis by 12°, which means the magnetic poles (identified by a magnetic compass) do not coincide with the geographic poles. A compass needle consists of a dipole magnet and, by definition, the north end of the compass needle's dipole magnet points toward Earth's north magnetic pole. Since opposite poles of a magnet attract each other, this means that Earth's north magnetic pole corresponds

to the south pole of Earth's dipolar magnetic field. Thus, at the present time, the south pole of Earth's magnetic field lies in the planet's northern hemisphere. The orientation of a planet's dipole magnetic field is called polarity.

Iron particles in magma align with the magnetic field lines – once magma has solidified into rock, the orientation of those iron particles is "frozen in," producing a remanent magnetic field that does not vary even if the external magnetic field changes. Analysis of iron-rich basaltic rocks reveals that Earth's polarity has undergone reversals throughout the planet's history, with the last magnetic reversal having occurred about 750,000 years ago. Geophysicists believe that Earth may be heading toward another polarity reversal because the magnetic field strength is decreasing about 1% per decade, the magnetic poles are rapidly shifting positions relative to Earth's surface, and the magnetic field's asymmetry is increasing. The entire reversal is expected to take several thousand years.

In 1915, German geophysicist Alfred Wegener (1880–1930) proposed the idea that continents move ("continental drift") across Earth's surface based on recognition of similar geography, rock types, fossils, and geologic structures along the west coast of Africa and the east coast of South America. However, no driving mechanism for continental drift was identified and the theory languished. In the 1960s, geophysicists recognized that the mirror-image pattern of polarity reversals recorded in basaltic rocks on the floor of the Atlantic Ocean indicated that ocean floors were also moving, a process called sea-floor spreading (McElhinny 1979). The theories of continental drift and sea-floor spreading were combined into the concept of plate tectonics, which states that the entire outermost portion of Earth is moving (see papers in Cox 1973).

Much of Earth's volcanic and tectonic activity results from plate tectonics, driven by the planet's internal heat. Earth's lithosphere is segmented into seven major and dozens of minor plates which move relative to each other because of convection occurring within the asthenosphere. Most of Earth's volcanic and tectonic activity occurs along plate boundaries. New crust is formed and plates move apart from each other at divergent boundaries (*Fig. 3-5*), characterized by extensional faulting, shallow earthquakes, and fluid volcanism. Plates converge at two types of convergent boundaries. A collision convergent boundary results in two plates, each composed of thick continental crust, crumpling to form a mountain range (*Fig. 3-6*). Collision convergent boundaries are characterized by earthquakes but no volcanic activity. Alternately, one ocean crust plate can descend beneath another plate (consisting of either ocean crust or continental crust) at a subduction convergent boundary, characterized by deep earthquakes and explosive volcanism (*Fig. 3-7*). Two plates slide past each other (strike-slip faulting) at transform boundaries, resulting in earthquakes (*Fig. 3-8*).

Although most of Earth's volcanic and tectonic activity occurs along plate boundaries, some earthquakes and volcanoes occur within the plates. Intraplate volcanism, such as Hawaii, results from mantle plumes ("hot spots") which ascend directly to the surface without being incorporated into the convection cells driving plate tectonics. Intraplate earthquakes, such as the series of earthquakes (magnitudes 7.2–8.1) that occurred in New Madrid, Missouri, in 1811–1812, are believed to result from redistribution of unreleased stress from the plate boundaries into the interior of the plate.

Earth displays a wide variety of volcanic landforms, ranging from flat lava flows to steep-sided volcanoes. The type of volcanic landform produced during a volcanic eruption depends on viscosity ("stickiness") of the erupting magma and the eruption rate. Viscosity depends on the magma's temperature (higher temperature corresponds to more fluid magmas) and the amount of silica (SiO_2) in the magma (higher SiO_2 corresponds to stickier (higher viscosity) magmas).

Fig. 3-5
Divergent boundary in Red Sea. This Space Shuttle (STS-41) image shows the Sinai Peninsula and Red Sea. A divergent boundary runs up the Red Sea (*bottom*) into the Gulf of Suez (*left*), splitting Sinai from Egypt (*lower left*). The Gulf of Aqaba (*right*) is formed by a slippage along a transform fault (NASA/Lunar and Planetary Institute (LPI))

High eruption rates and low viscosity magmas produce flat plains of lava flows called flood basalts, which are very common throughout the solar system. Low eruption rates and low-viscosity magmas produce large, low-sloped (typically <3°) volcanic constructs called shield volcanoes (*Fig. 3-9*), which have been observed on Mercury, Venus, and Mars in addition to Earth. Removal of magma during the eruption results in a collapsed depression, called a caldera, at the top of the shield volcano. Gases become trapped by the SiO_2 lattice in higher viscosity gases, resulting in magma fragmentation when the gas is finally able to escape. This fragmented material cools to form cinders and the resulting structures are called cinder cones (*Fig. 3-10*). Higher viscosity magmas produce steep-sided composite or stratovolcanoes (*Fig. 3-11*) which consist of layers of fine-grained ash alternating with lava flows. Composite volcanoes are very explosive because of the high concentration of gas trapped in the SiO_2-rich magma. Composite volcanoes occur in convergent subduction zones on Earth and have thus far not been identified conclusively on any other solar system body.

Fig. 3-6
Mount Everest from the Space Shuttle (STS-66). This image shows the crumpled appearance of the Himalaya Mountains, formed by the collision of the India plate into the Eurasian plate (NASA/LPI)

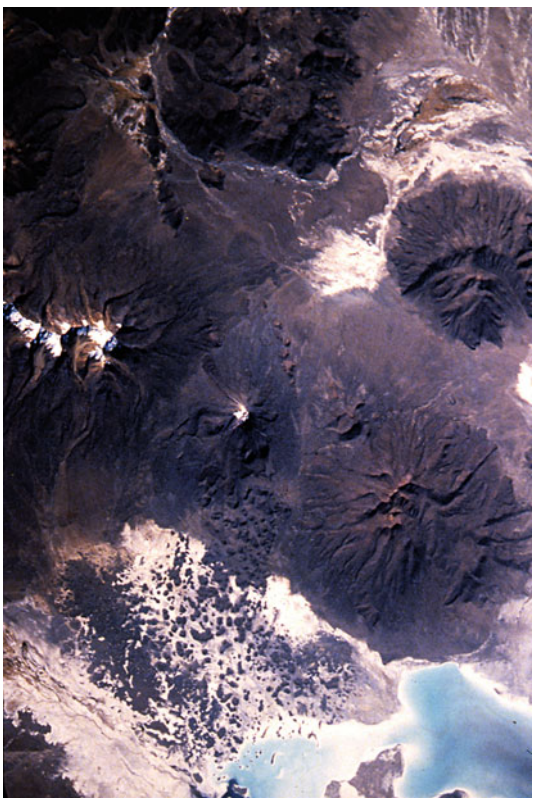


Fig. 3-7
Tata Sabaya Volcano, Bolivia, from the Space Shuttle (STS-41). The 5.4-km-high Tata Sabaya volcano in Bolivia is a composite volcano formed by subduction off the west coast of South America. A debris avalanche emplaced during one of the volcano's eruptions embays Salar de Coipasa lake at the *bottom* of the image (NASA/LPI)

Fig. 3-8
San Andreas fault, California. Landsat 7 imagery is draped over elevation data from the Shuttle Radar Topography Mission to produce this view of Los Angeles and its surroundings. The San Andreas transform fault is clearly seen as the straight line separating the mountains from the desert (NASA/JPL)

Mass wasting is the downward movement of material resulting from Earth's gravity. Mass wasting processes vary in their speed, from very slow deformation of weak (often ice-rich) material called creep to rapid movements of slopes, such as landslides and mudslides (*Fig. 3-12*). The angle of repose is the transition angle between stable soils and sliding materials. Angle of repose depends on the size of the soil particles, with higher angles possible for irregular particles than for fine-grained material. A slope which exceeds the angle of repose for the material is oversteepened and the material will easily slide. Seismic activity can induce slope failure when the slope is close to the angle of repose, and the presence of ice and/or liquid helps lubricate soils so that slope failure can occur even when dry material was stable.

Earth's atmosphere also contributes to its surface geology. The 1 bar atmospheric pressure at Earth's surface combined with warm surface temperatures allow H_2O to exist as gas in the atmosphere (water vapor) and as liquid (water) and solid (ice) on the surface. Liquid water, found in oceans, lakes, and rivers, creates fluvial landforms as it travels from higher to lower elevations.

Fig. 3-9
Mauna Loa, Hawaii. Mauna Loa, with its low slopes and extensive lava flows, is a classic example of a shield volcano (Image by author)

Fig. 3-10
Sunset Crater, Arizona. Sunset Crater is a cinder cone volcano which formed through a series of eruptions between 1040 and 1100 AD. The cinders comprising the volcanic edifice formed by fragmentation of the erupting magma as dissolved gases escaped (Image by author)

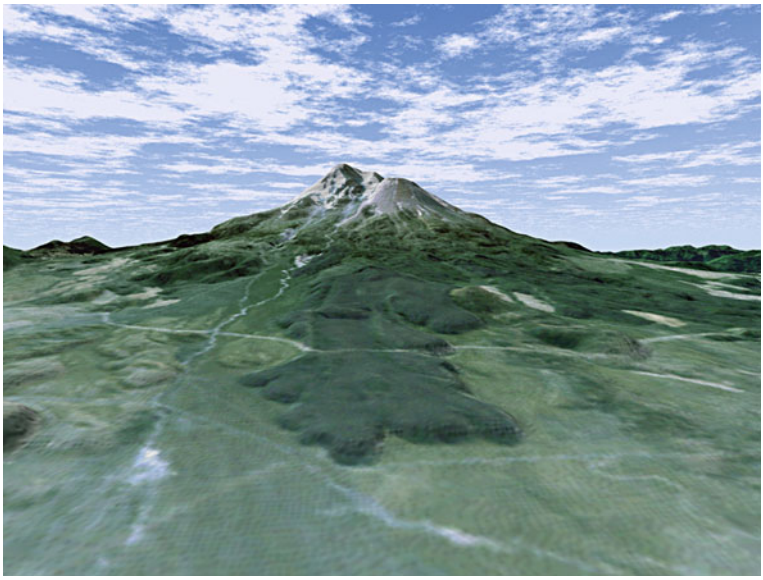


Fig. 3-11
Mount Shasta composite volcano, California. The perspective view of 4.3-km-high Mt. Shasta is obtained by overlaying Landsat 5 imagery onto SRTM topography. The resulting image clearly shows Shasta's caldera and lava flows (NASA/JPL)

River channels are commonly dendritic on Earth, where smaller creeks and streams combine to form larger rivers (*Fig. 3-13*). Most terrestrial channels originate from rainfall or snowmelt, although sapping channels form by collapse following removal of groundwater. Catastrophic floods create much larger channels, such as the Channeled Scablands (*Fig. 3-14*) of eastern Washington which formed between 15,000 and 13,000 years ago following episodic collapses of an ice dam which formed Lake Missoula across western Montana (Baker and Nummedal 1978).

Fluvial activity forms other unique landforms. An alluvial fan forms when a sediment-carrying stream encounters a sudden drop, such as occurs along the edge of a cliff or mountain range. Water and sediments spread out at the base of the cliff, forming the triangular-shaped alluvial fan (*Fig. 3-15*). Deltas are produced when a sediment-laden river encounters a larger body of water, such as a lake or ocean (*Fig. 3-16*). The river drops its sediment load because its water velocity decreases as it enters the larger standing body of water. The influence of liquid water can be inferred even after the water has evaporated. Dry river beds are obvious places where water once flowed, but dry lake beds (playas; *Fig. 3-17*) and shoreline terraces also reveal the prior influence of liquid water on a landscape.

The temperature range on Earth also allows for the existence of water ice. Glaciers are large sheets of ice which build up over many years when summer temperatures are insufficient to melt all the snow which precipitates during winter. Glaciers can be stationary (cold-based glaciers) or can slide on liquid produced at their base (warm-based glaciers). Alpine glaciers (*Fig. 3-18*) are small, localized ice sheets which originate in mountainous environments and, if warm-based, can move downslope under the influence of gravity, carving out U-shaped valleys. Continental glaciers are large, flat ice sheets covering large portions of a surface (*Fig. 3-19*). Continental glaciers were common during the five major ice ages experienced by

Fig. 3-12
Debris Avalanche, Chile. Sometime between 500 and 10,000 years ago, the flank of 6.05-km-high Socompa Volcano in Chile collapsed, producing a debris avalanche which extends up to 40 km from the volcano. This Space Shuttle (STS-41) image shows the volcano at *lower right* and the debris flow near image *center* (NASA/LPI)

Earth within the past 2.1 Ga. The current ice age began about 2.58 Ma ago and has consisted of periods of glacial advance interspersed with warmer interglacial periods. The only continental glaciers remaining during the present interglacial period are ice sheets covering Greenland and Antarctica.

Changes in Earth's orbital eccentricity, precession, and axial tilt (obliquity) affect the planet's climate and may contribute to ice age occurrence (Hays et al. 1976; Imbrie and Imbrie 1980; Muller and MacDonald 1997). These cycles are called Milankovitch cycles after Serbian mathematician Milutin Milanković (1879–1958) who proposed that gravitational perturbations by other planets could affect Earth's orbit and obliquity. Earth completes one entire cycle of precession about every 26,000 years. Gravitational perturbations from the other planets (mainly Jupiter and Saturn) cause Earth's orbital eccentricity to vary between 0.005 and 0.058 (current value is 0.017) over a period of ~100,000 years. These same gravitational perturbations cause the tilt of Earth's rotation axis to vary from 22.1° to 24.5° (current obliquity is 23.5°) over a

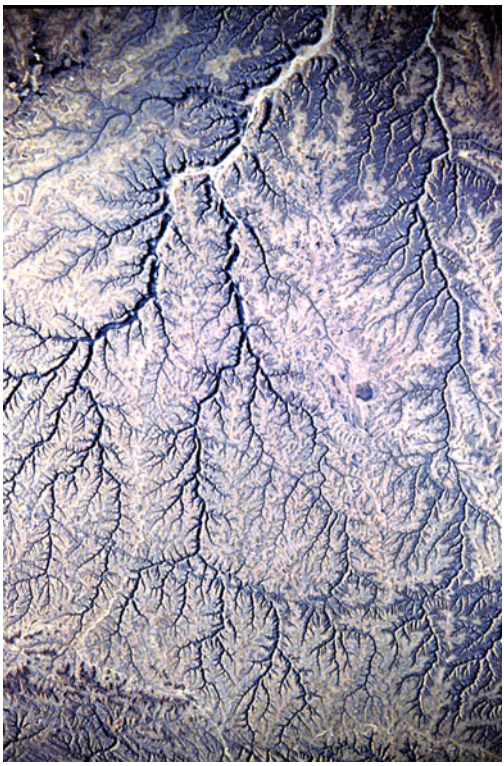


Fig. 3-13
Dendritic channels, Yemen. Yemen is currently a dry desert, but following the last ice age this region was wetter and covered by grasslands. The dendritic channel system shown in this Space Shuttle image (STS-41) was formed during this wetter period. Flow direction is indicated by smaller channels (*bottom*) merging into large channels (*top*) (NASA/LPI)

41,000 year period, cause the perihelion (minimum Earth–Sun distance) position to change on a 23,000 year cycle, and produce a 100,000 year cycle in orbital inclination. Depending on how these different cycles align, they can enhance or reduce climatic effects. Milankovitch cycles are a good but not exact match to the timing of past terrestrial ice ages, indicating that although they are a major contributor, other processes also influence the occurrence of glacial periods on Earth.

Earth's atmosphere moves in response to differences in temperature and pressure, producing wind. Wind, like water, can transport small particles from one location to another. Aeolian features are geologic landforms resulting from the erosional or depositional action of wind. Ventifacts are rocks sandblasted into triangular shapes by small particles carried by the wind. Yardangs are elongated ridges of rock, produced when surrounding soft rock is eroded away by wind coming from a prevailing direction (*Fig. 3-20*). Depositional aeolian features include thick, unsorted deposits of small dust grains called loess and a variety of sand dune shapes (*Fig. 3-21*). Wind streaks also are common aeolian features and can be either bright or dark. Bright wind streaks are dust deposits in the lee of a topographic obstacle which form when bright, fine-grained dust is removed from the surrounding area by wind. Dark wind streaks

Fig. 3-14
Channeled Scablands, Washington. During the last ice age, ice dams alternately formed and collapsed along the Snake River in Idaho. Creation of these ice dams backed up the water to form Lake Missoula in western Montana. Collapse of the ice dams catastrophically released the Lake Missoula water to flood across eastern Washington, carving deep channels to create the Channeled Scablands. This aerial view shows one such channel formed by the catastrophic floods (Image by author)

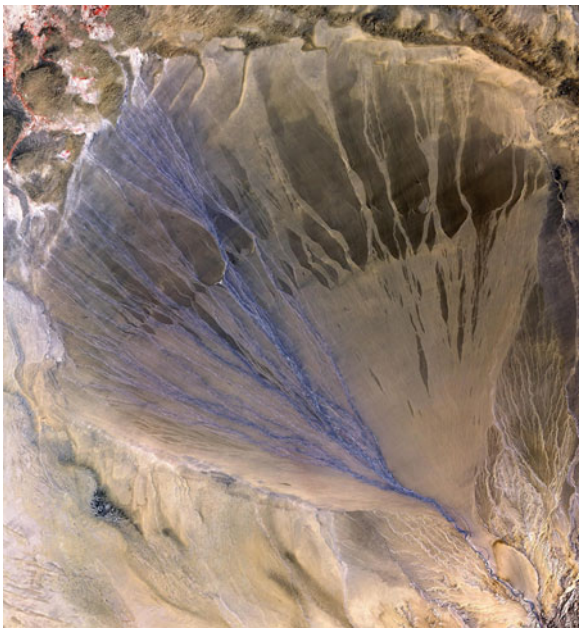


Fig. 3-15
Alluvial fan, China. This 56.6×61.3 km image, taken by the Advanced Spaceborne Thermal Emission and Reflection (ASTER) radiometer on the orbiting Terra spacecraft, shows an alluvial fan near the Kunlun and Altun Mountains in China. Water flows from *lower right* to *upper left*, and the *left side* of the image is still active today (NASA/ASTER)

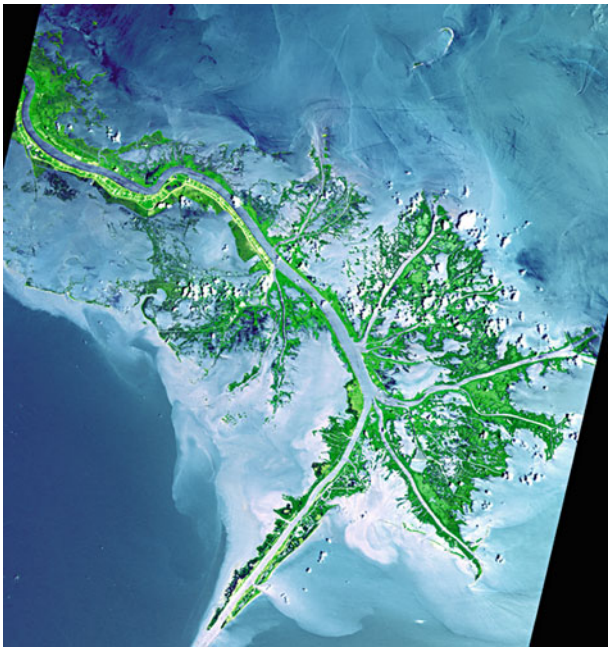


Fig. 3-16
Mississippi River Delta, Louisiana. The Mississippi River drops its sediment load as it enters the waters of the Gulf of Mexico, creating the delta deposit seen in this ASTER image (NASA/ASTER)

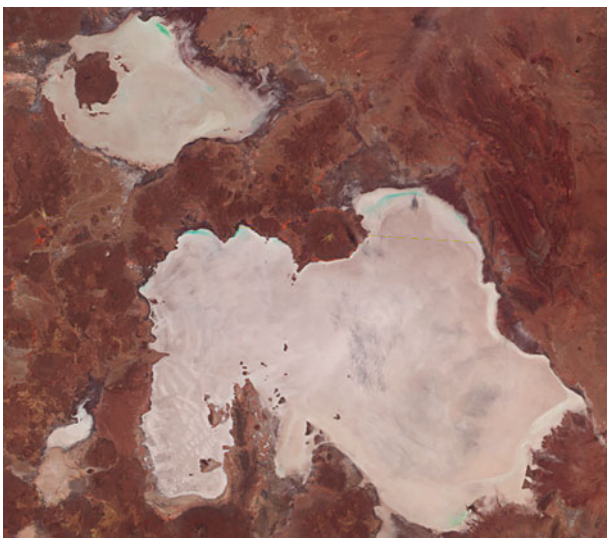


Fig. 3-17
Salar de Uyuni playa, Bolivia. The bright regions in this 199 × 177 km near-infrared image are salt deposits in the Salar de Uyuni playa in Bolivia. The playa is all that remains of a lake which existed during the last ice age when this area was wetter than at present. Image is from the Multi-angle Imaging SpectroRadiometer (MISR) instrument on the Terra spacecraft (NASA/JPL/MISR)

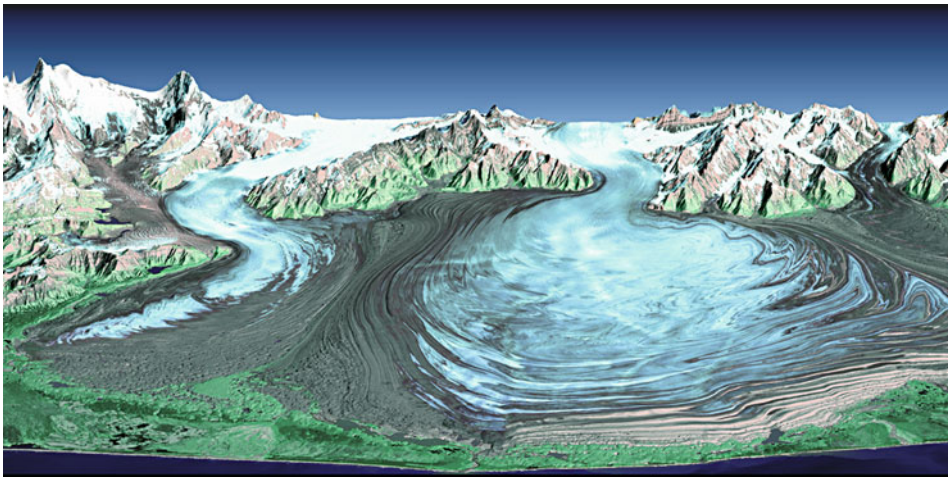


Fig. 3-18
Malaspina Glacier, Alaska. Malaspina Glacier (*bottom*) is formed by the merging of alpine glaciers Agassiz (*top left*) and Seward (*top right*). This perspective image was created by draping Landsat imagery over Shuttle Radar Topography Mission elevation data (NASA/JPL)

Fig. 3-19
Continental Glacier, Greenland. This ASTER image shows part of the continental glacier which covers Greenland. Image covers an area of 47.9 × 42.1 km in northern Greenland (NASA/ASTER)

Fig. 3-20
Yardangs, California. Yardangs form when wind erodes away softer surrounding material, leaving streamlined ridges. These yardangs are on Edwards Air Force Base in the Mojave Desert of California (Image by author)

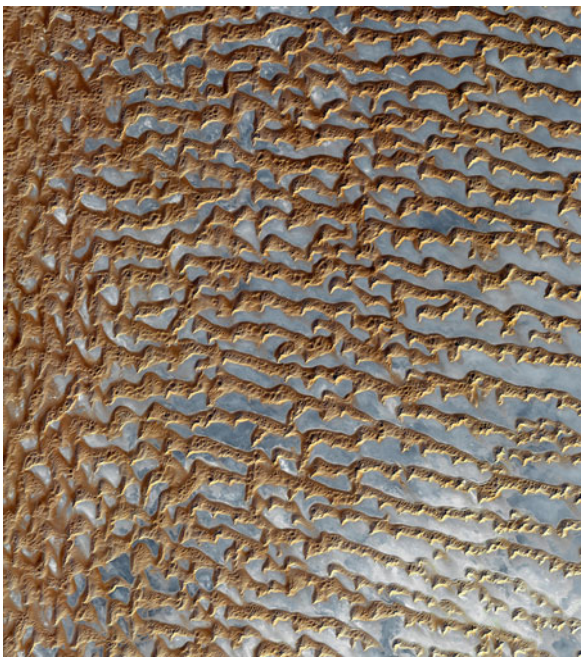


Fig. 3-21
Sand dunes, Rub′ al Khali Desert. The Rub′ al Khali Desert covers 650,000 km^2 on the Arabian Peninsula and is one of the largest sand dune deposits in the world. This image covers an area 54.8 × 61.9 km in size (NASA/ASTER)

result either from deposition of dark material in the lee of a topographic obstacle or from erosion of bright dust deposits by wind turbulence in the downwind area of an elevated structure (◉ *Fig. 3-22*).

Volcanism, earthquakes, mass wasting, fluvial activity, glaciers, and aeolian processes are readily observable on Earth's surface. Impact cratering, however, is a very rare process whose influence on Earth's geology was debated until the mid-twentieth century. Scientists did not accept that rocks fell from space until the early 1800s, and even after the extraterrestrial origin of meteorites was accepted, scientists did not recognize the amount of energy associated with impacts of larger meteorites. Few recognizable impact craters exist on Earth's surface because of our planet's active erosional environment, and the obvious circular depressions on the Moon (first called craters by Galileo in 1610) were interpreted to be volcanic calderas based on terrestrial experience. The circular appearance of lunar craters was used as an argument against an impact origin since everyday experience shows that rocks tossed into a sandbox will produce elongated craters unless the rock is dropped from a vertical angle. Nobody at the time recognized that bolides large enough to produce impact craters are traveling at velocities on the order of tens of kilometer per second (hypervelocities), which are considerably faster than velocities obtained by tossing a rock into a sandbox.

◘ Fig. 3-22

Amboy Crater, California. Amboy Crater (*upper left*) is a 75-m-high cinder cone volcano located in the Mojave Desert in southeastern California. The ~6,000-year-old cone forms a topographic obstacle to the prevailing northwest to southeast winds. Winds deposit brighter sand particles on the surrounding dark lava flows, but diversion of this wind around Amboy Crater keeps the downwind side free of sand, producing the dark wind streak

Laboratory experiments of hypervelocity impacts during the 1920s and 1930s revealed that such impacts are explosion events, destroying the projectile and sending out shock waves in the surface (target) material that excavate a circular crater unless the impact angle is very shallow (<10–15°). The physics of impact crater formation was developed during the 1940s and 1950s by studies of nuclear and large chemical explosion craters (see papers in Roddy et al. 1977). Those studies revealed that it is a combination of downward-propagating compressional shock waves and upward-propagating tensional rarefaction waves which excavate an approximately circular crater that is about ten times larger than the original projectile. Impact crater formation occurs in three stages (Melosh 1989): contact/compression, excavation, and modification. The contact/compression stage occurs when the projectile encounters the target and transfers its energy to the surface material. This first stage lasts only a few seconds and ends with projectile destruction. The crater cavity is produced during the excavation stage, when the shock and rarefaction waves combine to excavate the transient cavity. The excavation stage ends when the transient cavity reaches its maximum size, which takes only a few minutes for even the largest craters. The final stage of crater formation is the modification stage. Initial stages of the modification stage produce central peaks and wall collapse and last only a few minutes. All subsequent processes affecting the crater are included in the modification stage, including erosion, deposition, isostatic uplift, and tectonic modification. Hence, the modification stage continues until the crater is completely destroyed.

Hypervelocity experiments and explosion crater studies provided the theoretical framework for impact crater formation, but no terrestrial impact craters had yet been confirmed by the 1950s. One feature which had been proposed to be of impact origin was a 1.2-km-diameter depression with a raised rim in northern Arizona called Coon Mountain (Hoyt 1987) (*Fig. 3-23*). Small pieces of iron scattered around Coon Mountain had led some to propose it

Fig. 3-23
Meteor Crater, Arizona. The collision of an iron meteorite with Earth ~50,000 years ago created 1.2-km-diameter Meteor Crater in northern Arizona. Meteor Crater is a well-preserved example of a simple crater (US Geological Survey)

was an impact crater formed by destruction of an iron meteorite. However, the famous geologist G. K. Gilbert visited Coon Mountain in 1891 and decided it must be a volcanic landform (formed by a steam explosion as magma encountered groundwater) since he was unable to detect any evidence of a huge chunk of iron meteorite which he expected to be buried under the crater floor. Although Gilbert's reputation convinced most people that Coon Mountain was not an impact crater, mining engineer Daniel M. Barringer disagreed and bought Coon Mountain in 1903. Barringer set up a mining operation on the depression's floor in an attempt to excavate the intact iron-nickel meteorite that he believed was buried under the floor. Barringer was so convinced of Coon Mountain's extraterrestrial origin that he founded the small town Meteor City along nearby Route 66 so he could rename the depression Meteor Crater (impact craters are named after the closest town with a post office). Barringer's mining operations on both the floor and under the crater's south rim only recovered small pieces of iron meteorite and the operation went bankrupt. Scientists were only beginning to realize that impact craters form by explosion events and the projectile explodes during crater formation rather than lying intact at the bottom of the crater.

Barringer's efforts to convince people of the impact origin of Meteor Crater met with limited success since most people were still swayed by Gilbert's volcanic interpretation of the landform. A young geologist named Eugene Shoemaker finally confirmed the impact origin of Meteor Crater through his pioneering studies of the crater's geology (Shoemaker 1963). Shoemaker noted that Meteor Crater displayed numerous differences from volcanic calderas, including the lack of volcanic materials, location on flat ground rather than an elevated mountain, presence of iron meteorite fragments in and around the crater, and inverted stratigraphy in the raised rim (where geologically older layers overlie younger layers; ❯ *Fig. 3-24*). The clincher for an impact origin of Meteor Crater came from the discovery of coesite and stishovite, which are high-pressure phases of quartz (Chao et al. 1960). Stishovite forms when shock pressures reach

Fig. 3-24
Meteor Crater Rim. A characteristic of impact craters is inverted stratigraphy in the rim. This image shows how the older light-colored Kaibab limestone overlies the younger dark-colored Moenkopi formation in the rim of Meteor Crater (Image by author)

12–15 GPa and coesite occurs when pressures exceed 30 GPa. Such high pressures, particularly for coesite, can only occur in impact events.

Recognition of high-pressure mineral phases as well as microscopic and macroscopic shock-metamorphic features (French 1998) has led to the detection of over 200 impact craters on Earth (*Fig. 3-25*). Meteor Crater formed 50,000 years ago, making it one of the youngest and best preserved impact craters on Earth. It is an example of a small, bowl-shaped crater called a simple crater. Complex craters are larger and show more complicated morphologies, including shallower depth-to-diameter ratios, presence of central peaks, and collapse of oversteepened walls to produce wall terraces. The transition from simple to complex craters depends on the strength of the target material and inversely scales with the planet's gravity. The simple-to-complex transition diameter on Earth occurs around 2 km for impacts into sedimentary rocks and 4 km for crystalline rocks like basalt. Most terrestrial impact craters are heavily eroded and their original interior structures and ejecta blankets are missing or highly modified. Observations of the heavily cratered surfaces of the Moon, Mercury, and Mars indicate that Earth has hosted many more impact craters than the 200 currently identified. Most of these craters have been destroyed by Earth's active geologic processes, and the LHB record has been erased since no intact surfaces survive from this early time period. Perusal of *Fig. 3-25* shows that most identified impact craters occur on old, stable continental platforms, such as those found in Canada, northern Eurasia, and Australia. Tectonically active regions such as plate boundaries and the young (<200 Ma) ocean floors are statistically unlikely to display impact scars since the cratering rate over the past 3.8 Ga has been ~100–500 times lower than it was during the LHB.

Earth's restless nature often results in geologic disasters for its life forms. Landslides and mudslides routinely destroy homes built in scenic mountainous regions. Floods can affect anyone on Earth regardless of location. Building collapse triggered by earthquakes kills many people, such as the 73,000 people who died in the 7.6-mag earthquake in Kashmir in 2005. Tsunamis triggered by large earthquakes devastate coastal areas and are accompanied by large

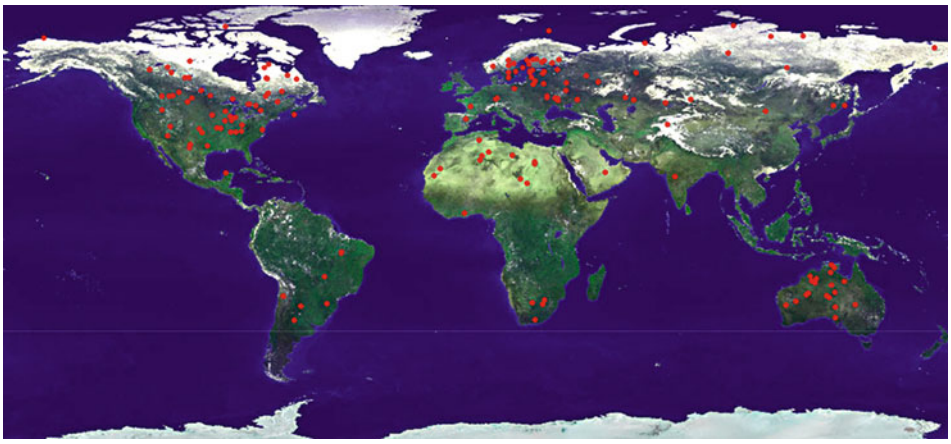


Fig. 3-25

Distribution of Terrestrial Impact Craters. This map shows the distribution of confirmed impact craters on Earth. Earth's active geologic environment has destroyed many of the early craters which formed on its surface (LPI)

death tolls, such as the 2004 tsunami which killed over 200,000 people in Southeast Asia (triggered by a ~9.2 mag earthquake off the coast of Indonesia). Tsunamis also can be generated by large explosive volcanic eruptions, particularly when these volcanoes blow themselves apart, as happened in 1883 with Krakatoa in Indonesia. The ~1500 BC collapse of Thera volcano created the doughnut shape of the Greek island Santorini and destroyed the Minoan civilization. Supervolcanoes affect even larger areas – three eruptions of the Yellowstone supervolcano between 2.1 Ma and 640,000 years ago spread ash over most of the continental USA west of the Mississippi River. But these geologic disasters pale in comparison to the amount of energy released during large impact events (French 1998). Formation of 1.2-km-diameter Meteor Crater released ~6.2×10^{16} J of energy, comparable to the energy released in the 1980 explosive eruption of Mount St. Helens volcano in Washington. A 10-km-diameter impact crater releases 4.6×10^{19} J of energy, more than the 1883 Krakatoa eruption (1.7×10^{15} J) or the largest recorded earthquake (magnitude 9.6 in Chile in 1960, which released 1.5×10^{15} J of energy). Formation of the largest impact craters, such as 2-Ga-old, ~250-km-diameter Vredefort Basin in South Africa, releases almost 300 times more energy than Earth's total annual energy output from heat flow, earthquakes, and volcanism combined.

Consequences of large impacts on terrestrial life were not fully appreciated until 1980 when Luis and Walter Alvarez began investigating possible causes for a layer of iridium found in sediments deposited 65 Ma ago at the boundary between the Cretaceous and Tertiary periods (KT Boundary) (Alvarez et al. 1980). Paleontologists had known for some time that a large number of species, including the dinosaurs, had died out around 65 Ma. The proposed causes of this mass extinction ranged from climate change caused by breakup of the supercontinent Pangaea to radiation from a nearby supernova explosion. Discovery of high iridium concentrations in sediments from this period suggested that the KT mass extinction resulted from the collision of a large asteroid with Earth. Iridium is a siderophile element and is therefore concentrated in Earth's core. Asteroids are small enough that most never underwent differentiation, therefore iridium remains dispersed through the asteroid. Further research revealed that KT sediments across the entire globe showed enhancement of iridium and other impact signatures, such as shocked quartz and impact spherules (Bohor et al. 1987). The impact not only distributed projectile and target debris across the globe but also induced wildfires which injected soot into Earth's stratosphere. This soot was carried by stratospheric winds around the globe, reducing the amount of sunlight reaching the surface and cooling the entire planet. Plants began to die from reduced sunlight levels, leading to a massive die-off of herbivores. Carnivorous animals subsequently starved from the lack of herbivores and the entire food chain collapsed, resulting in the observed mass extinction.

An asteroid about 10 km in diameter would be necessary to produce the observed global effects. An asteroid of this size should produce an impact crater between 100 and 200 km in diameter. No known impact crater was of the correct size and age, leading geologists to begin searching for a buried impact structure to confirm the Alvarez et al. hypothesis. KT sediment layers rich in iridium and other impact debris were thickest in the Americas and thus geologists concentrated their search in the western hemisphere. Tsunami deposits of the correct age were found through central Texas and along the southeastern US coast, suggesting the impact occurred at least partially in the waters of the Gulf of Mexico. The Mexican oil company PEMEX discovered thick sequences of brecciated (fragmented) rocks in their drill cores from the continental shelf off the northern coast of the Yucatan peninsula, which geologists recognized as a signature of an impact. Further inspection revealed that sinkholes (cenotes) in northern Yucatan displayed an arcuate pattern and outlined the edge of a ~175-km-diameter

impact crater (Fig. 3-26), subsequently dated at 65 Ma. The crater, named Chicxulub after the town closest to its center, is now widely accepted as the scar of the impact which initiated the KT mass extinction (Hildebrand et al. 1991; Sharpton et al. 1992). Several other mass extinction events in Earth's history have subsequently been linked to large impact events, revealing that the evolutionary path followed by terrestrial lifeforms has been altered many times by collisions with asteroids and comets. Recognition that impacts have influenced the evolution of life on Earth has led to development of the "impact frustration theory," which states that life forms began and were destroyed by impact events many times during the LHB period (Maher and Stevenson 1988). Widespread distribution of life on Earth did not occur until after the high impact rates of the LHB had declined.

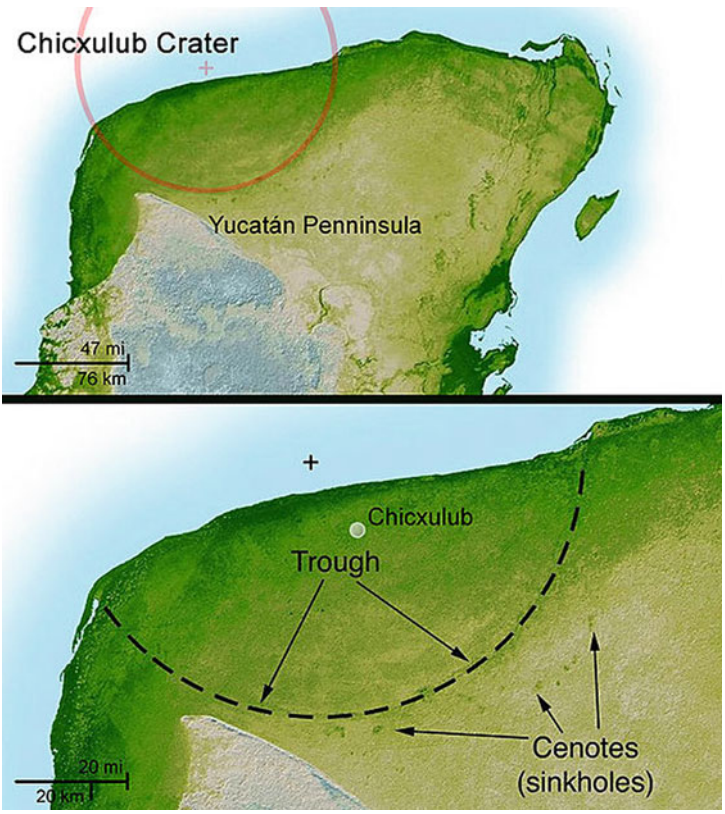


Fig. 3-26

Chicxulub Crater, Mexico. The 180-km-diameter Chicxulub crater on the northern coast of Mexico's Yucatan Peninsula is the scar of an impact which occurred 65 Ma ago. This shaded relief image from the Shuttle Radar Topography Mission shows the trough and cenotes which outline the remaining rim of the crater (NASA/JPL)

3 Venus

Venus is Earth's twin in terms of physical properties. Venus' equatorial diameter of 12,104 km is 95% of Earth's equatorial diameter (12,756), making Venus the planet closest in size to Earth. Venus' mass (4.87×10^{24} kg) is 82% of Earth's mass (5.97×10^{24} kg), meaning the densities of the two worlds are similar (5,200 kg m^{-3} for Venus versus 5,500 kg m^{-3} for Earth). The similar densities indicate that the two planets are composed of similar materials, with crusts and mantles consisting of silicate minerals and cores made primarily of iron. Gravity analysis indicates that C/MR^2 for Venus is 0.33, identical to the value for Earth, suggesting that Venus has a core about the same size as that of Earth. Lack of seismic data for Venus precludes detailed information about the interior structure, but modeling based on gravity data, the planet's mean density, and scaling from Earth's interior structure suggest a core radius of ~2,900 km, a ~3,100-km-thick mantle, and a crust that is <50 km thick (Grimm and Hess 1997).

Venus and Earth, however, are fraternal twins rather than identical. Although Venus is 95% the size of the Earth, it is just small enough that pressures are insufficient to produce a solid inner core. No active magnetic field has been detected around Venus, indicating that the liquid iron core is not convecting sufficiently rapidly to produce a dynamo. This may be at least partially due to Venus' very slow rotation rate of 243 Earth days. Not only is Venus the slowest rotating planet in the solar system, its rotation direction is backwards (retrograde) compared to Earth and most other planets. Most planets rotate in a counterclockwise direction as seen from above the ecliptic plane (direction above Earth's north pole), but Venus's rotation is clockwise from this vantage point. The most likely explanation for Venus' slow retrograde rotation is that it suffered a collision with a large asteroid or comet traveling on a retrograde orbit early in the planet's history. The result is an obliquity value of 177° for Venus' north pole (defined as the pole above which the planet is seen to rotate in a counterclockwise direction). The small 3° tilt of Venus' rotation axis from the normal to its orbital plane means that Venus essentially experiences no seasons.

Early telescopic observations revealed no sign of surface features on Venus, leading astronomers to correctly infer that the planet is shrouded by an opaque cloud layer contained within a planetary atmosphere (*Fig. 3-27*). Venus' thick atmosphere is composed of carbon dioxide (CO_2) which exerts a pressure of 95 bars at the planet's surface (compared to Earth's 1 bar surface pressure). The opaque cloud layer lies between 40 and 60 km above the surface and is composed of sulfuric acid droplets. Visible-wavelength sunlight reaching the planet's surface is absorbed by surface rocks and reemitted at longer thermal infrared (heat) wavelengths. Atmospheric CO_2 is opaque to thermal infrared, thereby trapping this heat next to the surface. This process is called greenhouse warming and has resulted in a 740 K surface temperature for Venus. This temperature is fairly stable across the entire planet, varying by no more than 30° between daytime and nighttime sides in spite of the planet's slow rotation.

Although the cloud layer precludes optical wavelength observations of Venus' surface, longer wavelength radar can penetrate the clouds and provide insights into surface topography and geology. The first radar observations of Venus were made in the 1960s using Earth-based radio telescopes and provided information about the planet's large-scale topography (*Fig. 3-28*). Those radar images revealed that most of Venus' surface consists of flat plains with elevations generally within 2 km of the planet's mean radius ("reference radius"). Three elevated continent-like regions exist, two in the equatorial region (Aphrodite Terra and Beta Region) and one at high northern latitudes (Ishtar Terra). Aphrodite Terra is about the size of Africa and has an elevation about 5 km above the reference radius. Ishtar Terra is about the

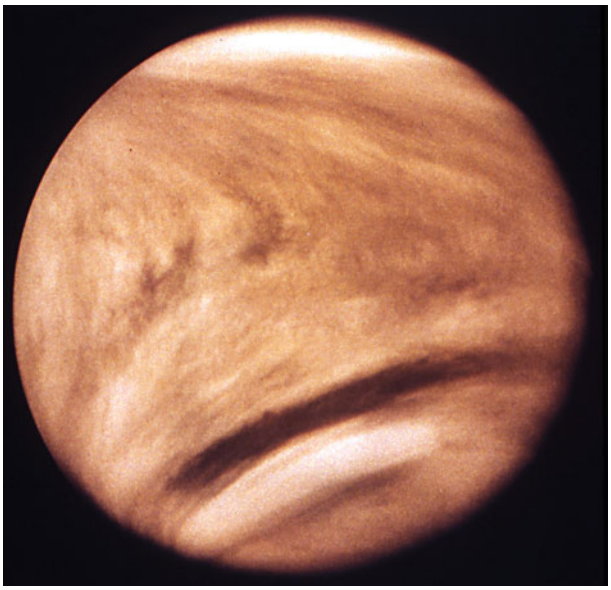


Fig. 3-27
Venus. Venus is covered by a thick layer of sulfuric acid clouds which prevent observations of its surface in visible wavelengths. This Pioneer Venus image shows structure within the clouds as revealed in ultraviolet wavelengths (NASA)

size of Australia and contains a high mountain range (Maxwell Montes) which extends up to 11 km above the reference radius. Beta Regio is much smaller than either Aphrodite or Ishtar and consists of the Theia Mons volcano and a tectonically deformed feature called Rhea Mons.

Higher resolution views of Venus' surface geology were obtained through radar observations from orbiting missions. Four radar-carrying spacecraft have orbited Venus: the Russian Venera 15 and 16 missions in 1983–1984, and the US Pioneer Venus (1978–1992) and Magellan (1990–1994) spacecraft. Pioneer Venus mapped about 90% of Venus' surface at a horizontal resolution of 30 km and vertical resolution of 700 m. It provided scientists with their first detailed views of the planet's surface (Hunten et al. 1983). Venera 15 and 16 improved the areal resolution to 1–2 km and vertical resolution to 30 m, but only mapped 25% of the surface. The most detailed radar mapping to date was obtained by Magellan, which imaged 98% of the surface at horizontal resolutions up to 100 m during the mapping phase from 1990 to 1992 (Bougher et al. 1997). Its synthetic aperture radar had a vertical resolution of 30 m. Later mission cycles were devoted to obtaining high-resolution gravity data, which showed that topographic features on Venus are typically supported by both variations in crustal thickness and mantle plumes.

Soviet/Russian and American probes directly investigated Venus' atmosphere, and a few of these probes were designed to land on the surface. Veneras 8, 9, 10, 13, and 14 and the Vega 1 and 2 landers provided basic analysis of surface soil and rocks at landing sites within the ±30° latitude zone (*Fig. 3-29*). The Venera missions all landed on or east-southeast of Beta Regio whereas the two Vega missions landed along the edge of eastern Aphrodite Terra. Analyses, consisting of gamma ray spectrometry and/or x-ray fluorescence, revealed that Venera 8 and 13 samples were composed of high-potassium (alkaline) basalts whereas the other landing

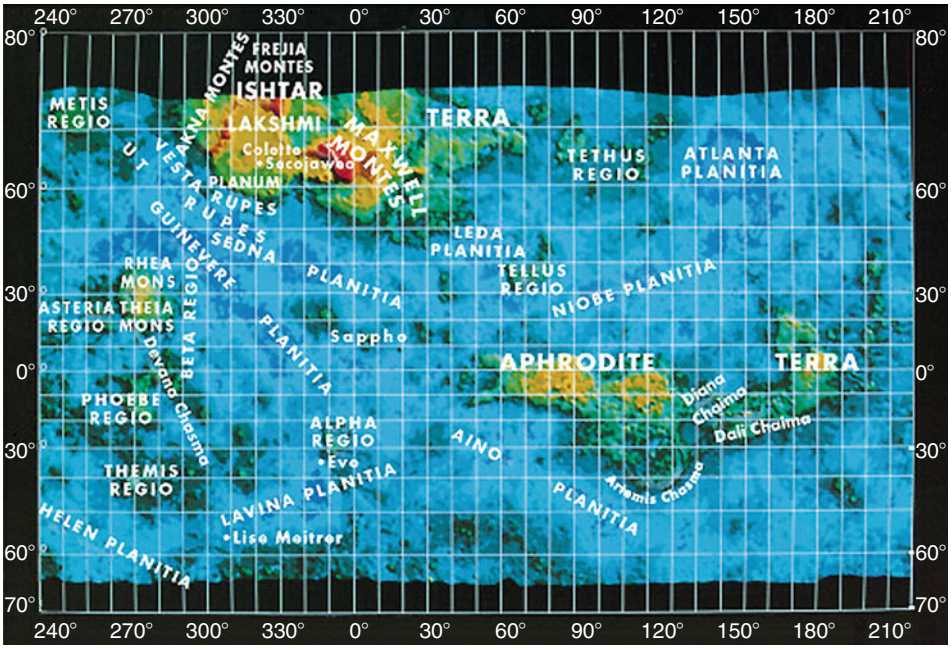


Fig. 3-28

Venus Topography Map. Radar penetrates Venus' thick cloud layer and provides information about the surface topography. This topography map, derived from Pioneer Venus radar measurements, shows that most of Venus is close to the reference radius (*blue*) with only three elevated regions (Beta Regio, Ishtar Terra, and Aphrodite Terra) (NASA)

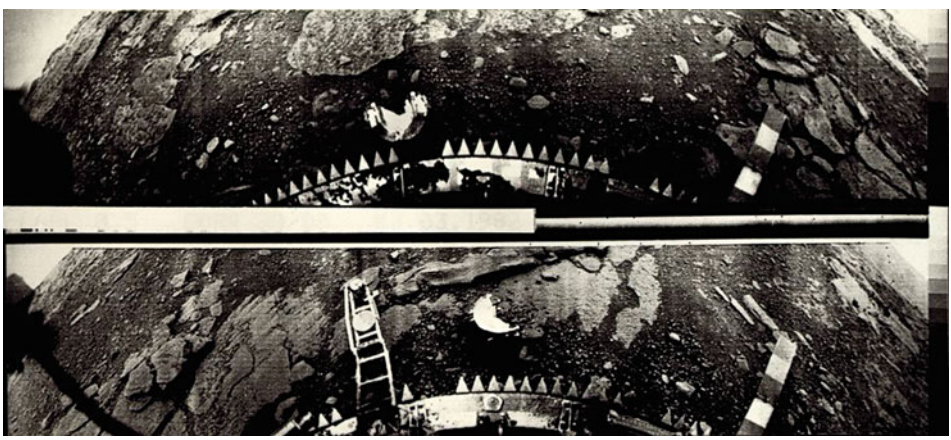


Fig. 3-29

Venus Surface. Several Soviet Venera landers relayed pictures and basic soil information from the Venus surface back to Earth. This pair of images is from the Venera 13 lander, which landed on the east side of Beta Regio (NASA National Space Science Data Center)

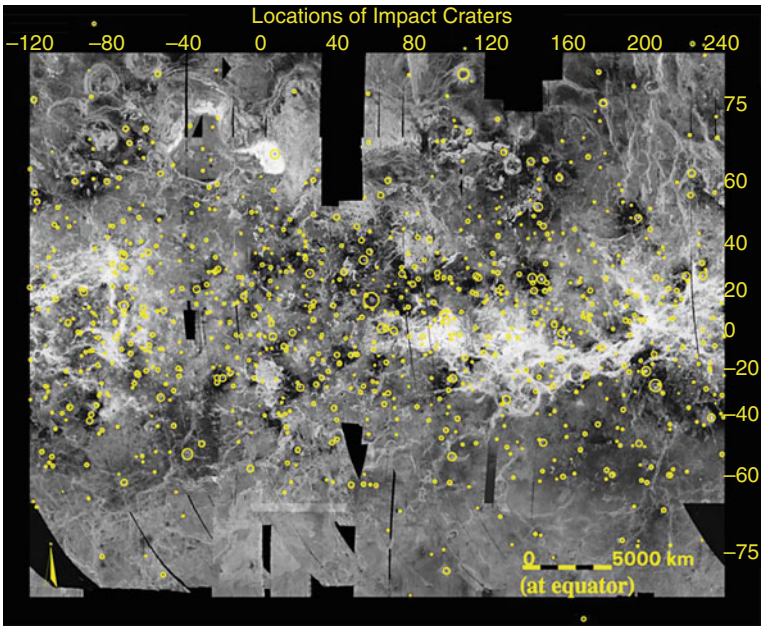


Fig. 3-30

Venus Crater Distribution. Impact craters on Venus show a random distribution, suggesting a global resurfacing event occurred about 0.5 Ga ago (LPI)

site samples consisted of normal (lime-alkaline) to low (tholeiitic) potassium basalts (Barsukov 1992). These compositions are similar to the range of basalts found on Earth, which are produced by melting in the mantle. However, no landers have yet sampled soils or rocks from the interiors of the highlands regions on Venus, so scientists cannot definitively rule out existence of non-basaltic materials on the planet.

Analysis of radar imagery has revealed about 900 impact craters distributed in an approximately random pattern across Venus' surface – there are no regions of extremely high or extremely low crater density (*Fig. 3-30*). This distribution suggests that all surface units on Venus are about the same age, estimated at ~0.5 Ga based on crater density (Phillips et al. 1992; Schaber et al. 1992). Thus, although Venus formed with the rest of the solar system planets about 4.5 Ga, some event occurred around ~0.5 Ga ago which completely resurfaced the planet. No surface units from before this time survived and thus the planet's earlier geologic history is lost. The cause of this massive resurfacing event is not completely understood, but interior models suggest that Venus oscillates between quiescent and active volcanic periods on a timescale of a few hundred million years because of chemical differentiation occurring in the interior. Crustal overturn occurs by vertical resurfacing, where volcanism and tectonism allow underlying magma to erupt onto the surface (Grimm and Hess 1997). Following complete crustal overturn, the planet enters a quiescent period where volcanism and tectonism are regionally limited and erosion rates are low. Craters randomly form on the planet's surface and undergo little modification until the next catastrophic resurfacing episode occurs. This model is consistent with the well-preserved nature of most impact craters on Venus.

Small impact craters typically outnumber larger craters since small asteroids and comets are more common than larger ones. However, Venus displays a distinct lack of small impact craters. Craters smaller than ~3-km-diameter are completely missing and craters between 3 and 15-km-diameter tend to exist as overlapping crater clusters instead of a single crater (*Fig. 3-31*) (Phillips et al. 1992). Both of these observations can be attributed to Venus' thick atmosphere. The smallest projectiles vaporize during passage through the atmosphere and slightly larger materials are fragmented, resulting in almost simultaneous impacts of several pieces of the projectile. Only larger projectiles can survive the pressure and friction of their atmospheric passage and strike the ground intact.

Venus' atmosphere affects its impact craters in additional ways (*Fig. 3-32*, Schultz 1992). Lobate ejecta flows extend from fresh impact craters and were produced by interaction of the curtain of ejected debris with turbulent atmospheric eddies. Winds induced by motion of the ejecta curtain drive a ground-hugging debris flow which extends many crater radii beyond the crater rim. Extensive run-out flows beyond the normal ejecta blanket may result either from debris flows of vaporized material early in the crater formation process or large amounts of impact melt (Asimow and Wood 1992; Schultz 1992). Asymmetric ejecta blankets, indicative of oblique impact events, are enhanced through interactions of the ejecta curtain with atmospheric turbulence. Many impact craters also display radar-bright or radar-dark halos, resulting from scouring and debris deposition by atmospheric blast waves produced as the projectile passes through the atmosphere. Radar-dark parabolas surround some impact craters on Venus and probably result from late-stage fallout of material entrained in the atmosphere.

The high surface temperature precludes the existence of liquid water or ice on present-day Venus, and Magellan's radar images show no evidence that water or ice have affected Venus'

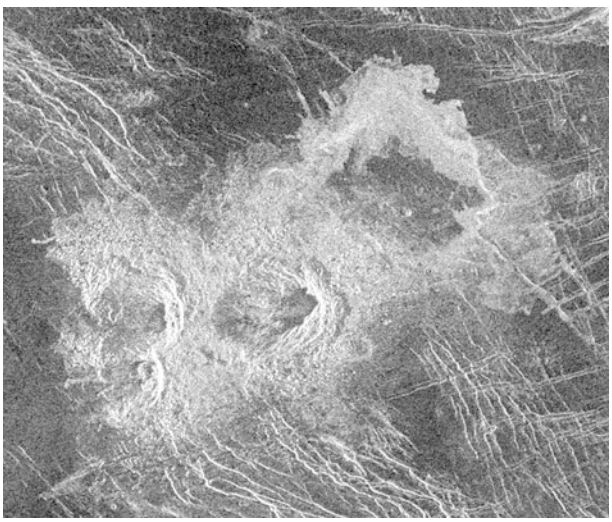


Fig. 3-31

Stein Crater Field, Venus. Small meteorites often fragment as they pass through the thick Venusian atmosphere, resulting in crater clusters rather than a single impact crater once they impact the surface. The three craters comprising the Stein crater field are 14, 11, and 9 km in diameter (NASA/JPL)

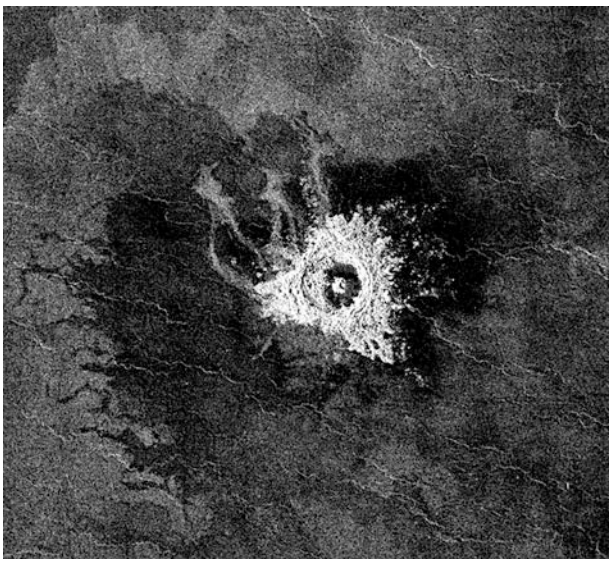


Fig. 3-32
Jeanne Crater, Venus. This Magellan image of 19.5-km-diameter Jeanne crater displays many of the typical features associated with Venusian impact craters. The radar-bright circular crater is surrounded by an asymmetric radar-bright ejecta blanket, indicating an oblique impact. The bright channel-like features (*upper left*) are run-out flows of impact melt or material originally vaporized during impact. The surrounding radar *dark* halo surrounding is a scour zone from an atmospheric blast wave (NASA/JPL)

surface geology within the last 0.5 Ga of recorded history. Venus' atmosphere also shows little evidence of water vapor, indicating the entire planet is currently quite dry. But Venus may have been wetter in the distant past. Venus may have outgassed a water vapor atmosphere early in its history, which caused its surface temperature to increase due to greenhouse warming. Increasing surface temperature resulted in more water vapor accumulating in the atmosphere, which led to a runaway greenhouse effect. Eventually all of the planet's water accumulated in the atmosphere. Atmospheric mixing gradually brought the water molecules to the upper part of the atmosphere where they could be photodissociated by solar ultraviolet radiation, allowing hydrogen atoms to escape to space. Oxygen would combine with carbon atoms to form the CO_2 atmosphere seen today. Alternatively, the moist Greenhouse model proposes that Venus' surface temperature was initially cool enough for a liquid water ocean on the planet's surface. Air above the ocean was saturated with water vapor and convection transported this water vapor to the top of the atmosphere where water molecules were photodissociated. Atmospheric water vapor produced Greenhouse warming near the surface, but the higher temperature resulted in increasing rates of evaporation from the ocean, leading to more water vapor in the atmosphere and high enough atmospheric pressure to keep the ocean from boiling away. Molecular oxygen (O_2) and CO_2 were removed from the atmosphere through interactions with ocean water, resulting in production of carbonate minerals. Water continued to evaporate from the ocean as photodissociation removed water vapor from the top of the atmosphere. Eventually all liquid water was removed from the surface through evaporation, and atmospheric convection led to

removal of water vapor through photodissociation. CO_2 began to accumulate in the atmosphere since lack of liquid water on the surface kept it from being removed through formation of carbonates. The planet's sulfuric acid clouds may have resulted from volcanism associated with the crustal overturn 0.5 Ga (Bullock and Grinspoon 2001) and are likely maintained through recent regional volcanism.

More than 6,000 aeolian landforms have been detected in Magellan images of Venus' surface. Most of these are wind streaks, produced by interaction of prevailing winds with topographic obstacles. Depositional landforms such as dune fields (*Fig. 3-33*) and erosional features including yardangs have been identified. Orientation of wind streaks, dune fields, and yardangs have been used to determine near-surface wind directions, which are found to be consistent with those predicted by Hadley Cell circulation produced from solar heating (Greeley et al. 1997).

Mass wasting features resulting from Venus' gravity are observed. Complex craters on Venus show evidence of the downward pull of gravity through the presence of wall terraces. Landslides are seen in association with mountainous regions on the planet (*Fig. 3-34*).

The most prevalent geologic landforms on Venus are volcanic and tectonic features. Volcanism primarily consists of low-viscosity lavas, as indicated by widespread presence of flood basalts and shield volcanoes (*Fig. 3-35*) and Venera/Vega lander detection of basalt. Volcanic features are divided into large individual volcanic centers consisting of edifices and calderas produced by high eruption rates, intermediate volcanoes, and fields of small shield volcanoes formed by lower eruption rates (Crumpler et al. 1997; Guest et al. 1992; Head et al. 1992). Slightly more than half of all volcanoes on Venus are found in the Beta-Atla-Themis region. The low-viscosity nature of Venus' lava also is indicated by the presence of more than 200 lava channels (Baker et al. 1997). Some channels are morphologically similar to sinuous rilles seen on other solar system bodies and typically originate from collapse regions. Extremely long channels, called canali, display constant width along their entire length (>500 km) and are commonly

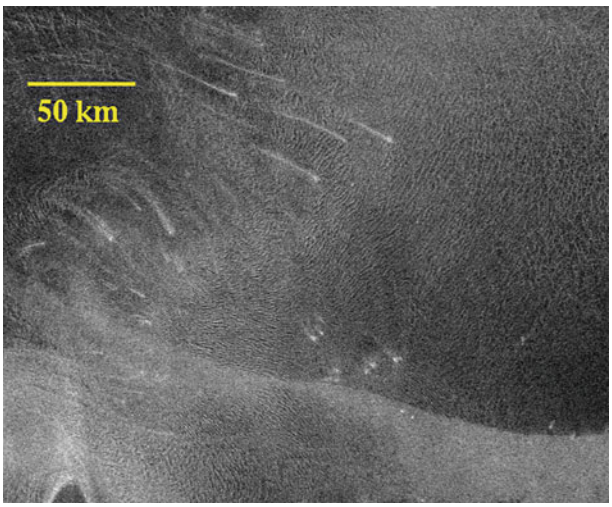


Fig. 3-33
Aeolian Features, Venus. Sand dunes and wind streaks are evidence of wind activity in this Magellan image of the region near 68°N 90°E (NASA/JPL/LPI)

Fig. 3-34
Landslides, Venus. This Magellan image shows landslides on the northeast and northwest slopes of a 17.4-km-diameter volcano on Venus. These landslides formed when the volcano flanks collapsed under the planet's gravitational influence (NASA/JPL)

found on the Venusian plains (*Fig. 3-36*). Morphologic characteristics of the canali indicate they form by a highly erosive fluid, such as high iron-titanium basalts, komatiites, alkali carbonatite, or sulfur lavas (Kargel et al. 1994; Williams-Jones et al. 1998).

Intermediate volcanoes are 20 to 100 km in diameter and display features such as radial lava flows, fracture patterns, and/or a volcanic edifice (Head et al. 1992). Intermediate volcanoes include the anemone and tick subclasses. Anemones have a petal-like lava flow pattern surrounding a central elongated caldera. Ticks have a circular flat or depressed interior surrounded by a rim and display a series of ridges radiating outward from the rim. Several different models for tick formation have been proposed, including extrusion and rifting, localized dike intrusion, erosion of ash flows, or post-volcanism slumping.

Although most volcanic features on Venus form from fluid magmas, some examples of high-viscosity lavas are observed. Pancake domes (*Fig. 3-37*) are steep-sided, flat-topped edifices which show morphologic similarities to high-viscosity silicic lava domes on Earth (Pavri et al. 1992). Head et al. (1992) classify them in the intermediate volcano category. The higher viscosity lavas forming pancake domes likely originate from differentiation and evolution of basaltic magmas within subsurface magma chambers.

Large volcanoes include coronae, arachnoids, and novae. Coronae are large (>200 km diameter) circular volcanic-tectonic structures common on Venus but with no terrestrial analogs (*Fig. 3-38*). Coronae are slightly elevated regions (~1–1.5 km high) bounded by concentric fractures. Volcanic features such as calderas, domes, and lava flows are common in their interiors, which also display radial fractures. The entire feature is surrounded by a topographic moat (Stofan et al. 1997). Coronae tend to occur as rises within the plains and along large extensional tectonic features called chasmata. They form by upwelling plateau-shaped mantle diapirs,

Fig. 3-35
Atla Regio Volcanoes, Venus. Magellan images of Atla Region display a wide variety of volcanic and tectonic fractures. The bright linear features in this image (350 km across) are fractures. Most other features are various volcanic edifices and lava flows (NASA/JPL)

Fig. 3-36
Sinuous Channel, Venus. The 200-km-long sinuous channel (canali) in this Magellan image was formed by highly erosive lava (NASA/JPL)

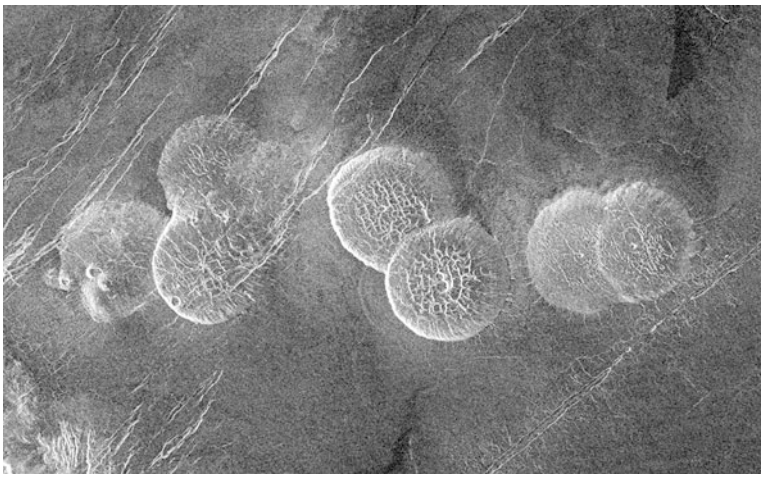


Fig. 3-37
Pancake Domes, Venus. The morphologic appearance of pancake domes suggests they consist of viscous lava. This series of pancake domes, each about 25-km-diameter and 750 m high, is located on the eastern edge of Alpha Regio (NASA/JPL)

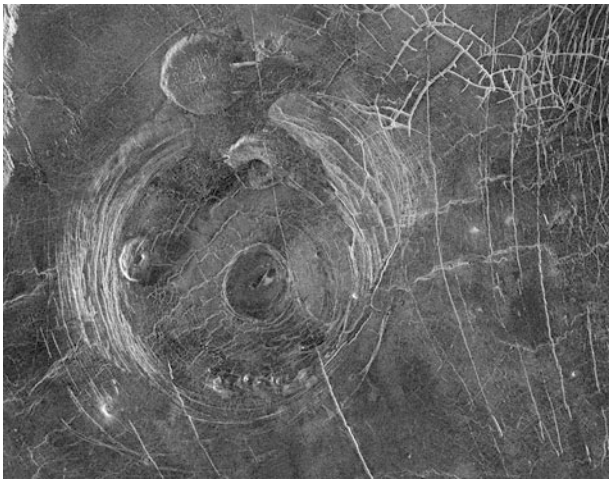


Fig. 3-38
Aine Corona, Venus. Coronae are large volcanic features bordered by circumferential graben. This Magellan image shows 200-km-diameter Aine Corona and its associated volcanic and tectonic features (NASA/JPL)

which fracture the surface and produce the associated volcanic landforms. Following the eruption, the diapir cools and loses its buoyancy which causes the corona region to relax, producing additional fractures and the surrounding moat.

Arachnoids display a concentric pattern of fractures or ridges surrounded by extensive radial fractures or ridges. Their name is derived from their resemblance to a multilegged spider. Arachnoids are smaller than coronae, with diameters <200 km. They display fewer signs of

extrusive flows than coronae but otherwise are morphologically similar to their larger counterparts and therefore likely share a similar origin. Magmatism associated with arachnoids may be largely intrusive (underground), explaining the lack of obvious flow structures, whereas that for coronae is predominantly extrusive (aboveground).

Unlike coronae and arachnoids, novae display no concentric structures, appearing as outward radiating fractures similar to a starburst. The fracture pattern often is centered on a broad elevated dome and fractures narrow in width as they extend outward from the center. Novae typically occur in association with other large volcanic features and often are found in the interiors of coronae and arachnoids. Their characteristics suggest that novae form in response to uplift and fracturing of the surface by underlying magma intrusions (Head et al. 1992).

Both extensional and compressional tectonic features are observed in Magellan data and much of the tectonic activity is associated with volcanic landforms (Solomon et al. 1992). The most intensely fractured regions are tessera terrain (*Fig. 3-39*), which are elevated above the surrounding plains and characterized by at least two intersecting sets of tectonic features (Hansen et al. 1997). Tessera were first detected in Venera 15 and 16 radar images and have been subdivided into different classes based on morphology (folds, extensional fractures and graben, compressional ridges, etc.). Tessera are among the oldest surface units on Venus, although not all tessera units were formed simultaneously. The variety of landforms within tessera units indicates a complex tectonic history and variations in stress regimes in different portions of the crust.

Extensional fractures are common across Venus' surface, typically forming flat-floored valleys called graben (*Fig. 3-40*). Long linear graben are commonly associated with large volcanoes and may represent underlying dike swarms. Several large extensional rift zones (chasmata) exist on Venus, including Devana Chasma in Beta Regio and Ganis Chasma in the Atla Regio region of western Aphrodite Terra, both of which are associated with large volcanic features. A different type of extensional faulting is observed in central Aphrodite Terra with the west-southwest to east-northeast trending Diana Chasma and east-west trending Dali Chasma. These chasmata have been variously interpreted as a fold-and-thrust belt, a transform zone, a

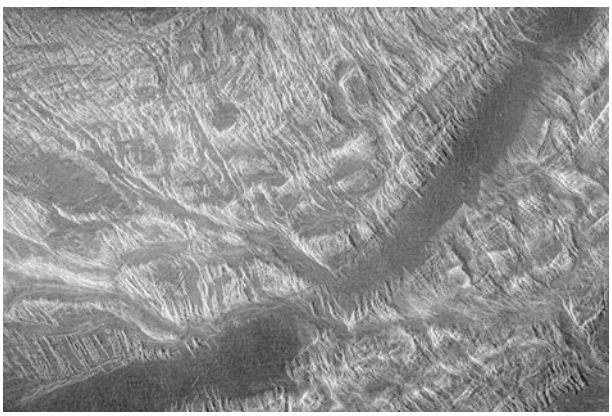


Fig. 3-39

Tessera, Venus. Tessera terrain, such as that shown in this 225 × 150 km Magellan image of Ovda Regio, consists of a complex series of tectonic deformation (NASA/JPL)

Fig. 3-40
Graben System, Venus. The Venusian crust in the Lavinia region west of Alpha Region has undergone considerable extension as indicated by the large number of graben in this region. This Magellan image is about 70 km across (NASA/JPL)

crustal spreading zone, and even an area of lithospheric subduction (Solomon et al. 1992). The prevalence of extensional fractures across the planet indicates that a considerable amount of lithospheric stretching occurs on Venus.

Ishtar Terra displays evidence of both extensional and compressional tectonics (Hansen et al. 1997; Solomon et al. 1992). Ishtar rises more than 2 km above the planetary reference radius and is composed of a variety of landforms (*Fig. 3-41*). Most of Ishtar Terra is dominated by Lakshmi Planum, a broad plateau (4 km above reference) bordered by mountain belts (Danu, Akna, Freyja, and Maxwell Montes, extending from 3.5–11 km in elevation) and steep smooth scarps on the north and south sides. Tessera terrains extend beyond the mountain belts whereas the scarps descend to plains-filled basins. The surface of Lakshmi Planum is characterized by wrinkle ridges formed in response to compressional stresses in lava flows, smooth plains formed by lava flows, and marginal troughs formed by extension. The center of Lakshmi is dominated by volcanic calderas, including 100-km-diameter Colette and 200-km-diameter Sacajewea. The four mountain belts surrounding Lakshmi Planum are compressional folds and thrust faults which formed by crustal shortening of the surrounding plains after Lakshmi formed. Some modification of the mountain belts by volcanism and extension are seen on the non-Lakshmi sides. The range of landforms indicates that Ishtar Terra has experienced a complex history which cannot be described by simple mantle upwelling or downwelling (Hansen et al. 1997). Magellan gravity data suggest that Ishtar Terra formed in an area of mantle downwelling, which

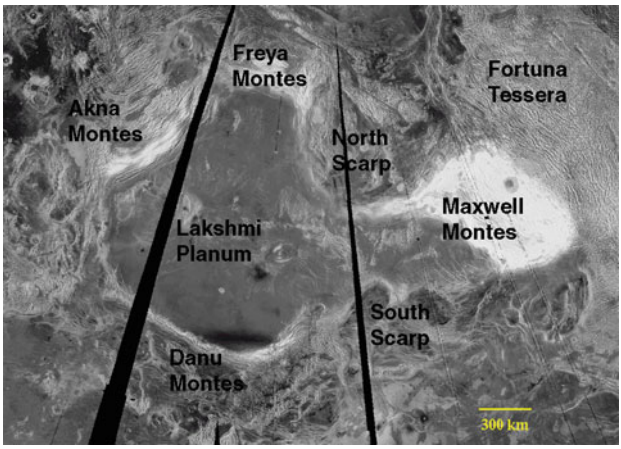


Fig. 3-41
Ishtar Terra, Venus. Ishtar Terra is a highlands region near the north pole. It contains the highest elevation on Venus (Maxwell Montes) and the complexity of terrains indicates a unique evolution (NASA/JPL/LPI)

resulted in the ponding and thickening of a residuum layer formed by partial melting of the mantle. Lakshmi Planum formed in this region, overlying normal plains crust. Shear displacement of the lower part of the crust (which was decoupled from the upper crust) toward the center of Lakshmi produced the marginal mountain belts and their associated tessera.

The variety of volcanic and tectonic features on Venus suggests a different mechanism for removal of interior heat than the plate tectonics regime operating on Earth (Hansen et al. 1997; Solomon et al. 1992). Venus' lithosphere undergoes predominantly vertical motions but the lithosphere itself remains fixed in position, in contrast to the horizontal movement of Earth's tectonic plates. Convection within Venus' asthenosphere produces zones of mantle upwelling (producing extensional faulting, coronae, and chasmata) and downwelling (resulting in plains and compressional ridge belts), and these convection cells migrate with time. Deep thermal plumes (hot spots) produce volcanic rises and may have contributed to formation of crustal plateaus within the highlands regions through partial melting. Tessera may represent original crust which has been highly deformed by subsequent coronae-chasmata formation and/or by collapse of original high-standing plateaus. Isthar Terra is a unique region produced by a combination of the residuum layer formed by partial melting over a mantle downwelling zone, a thickened lower crust, and stress-induced flow in the upper crust. Interior convection and thermal plumes stretch Venus's lithosphere, causing it to fracture and allow access of magma to the surface. If this process occurs at a slow pace, heat builds up in the interior for a few hundred million years until a catastrophic volcanic resurfacing event occurs. Regardless of whether the volcanic resurfacing is episodic or continuous, Venus is in a constant state of reinventing herself.

4 Mars

Venus may be Earth's twin in terms of physical properties, but aspects of Mars' geologic history appear to have more closely mirrored that of Earth (*Fig. 3-42*). Mars' smaller size (53%

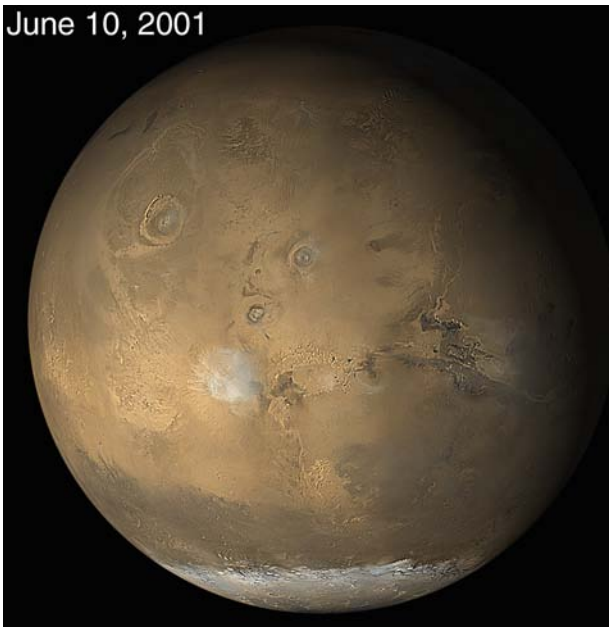


Fig. 3-42
Mars. The Mars Orbiter Camera on Mars Global Surveyor (MGS) obtained this global view of the western hemisphere of Mars on June 10, 2001. Image is centered on the Tharsis bulge. Olympus Mons, the largest volcano in the solar system, is at *upper left* and its companion shield volcanoes are in image *center*. The Valles Marineris canyon system extends eastward at the *right side* of the image. Also visible is the south polar cap and *white clouds* in the Martian atmosphere (NASA/JPL/Malin Space Science Systems (MSSS))

of Earth's diameter) means that it has lost more of its internal heat over the past 4.5 Ga and therefore is not as geologically active as Earth or Venus at the present time. However, Mars has been, and continues to be, more geologically active than smaller Mercury or the Moon. In fact, Mars is the only terrestrial planet whose surface retains terrains which have formed throughout the planet's history. The active geologic environments of Earth and Venus have destroyed the earliest surfaces of these planets whereas Mercury and the Moon display no surfaces younger than ~1–2 Ga. Mars displays ancient surfaces dating back to the LHB period, intermediate aged surfaces, and very young terrain which have formed within the past few million years.

Numerous flyby, orbiter, lander, and rover missions from the US, USSR/Russia, and Europe have successfully explored Mars since 1965 and both orbiters (Mars Odyssey (NASA), Mars Express (European Space Agency), and Mars Reconnaissance Orbiter (NASA)) and NASA's Mars Exploration Rover Opportunity continue to operate as of mid-2011. These missions have provided detailed information about Mars' interior structure, topography, atmosphere, geologic landforms, and surface mineralogy which have been used to reconstruct the planet's history and investigate the role played by H_2O on the planet's climate, geology, and possibility of life through time (Barlow 2005; Kieffer et al. 1992).

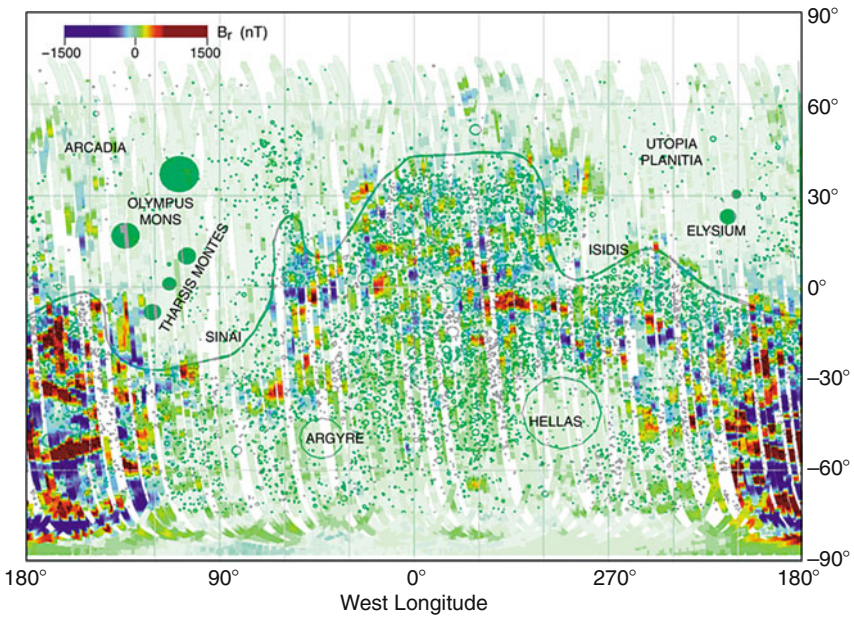


Fig. 3-43

Crustal Remanent Magnetization, Mars. The magnetometer/electron reflectometer experiment on MGS detected zones of remanent magnetization in the ancient Martian crust. The lack of magnetization around the Hellas and Argyre impact basins indicate the planet's magnetic dynamo had ceased by the time these basins formed ~3.8–4.0 Ga (NASA/JPL/Goddard)

Mars, like the other terrestrial planets, formed by accretion in the solar nebula 4.5 Ga ago and differentiated within a few million years as a consequence of heating by accretion and radioactive decay. The planet's two small heavily cratered moons, Phobos and Deimos, are likely asteroids which were captured by Mars' gravity after the planet formed. Mars has a mean density of 3,933 kg m^{-3}, indicating it is a rocky body with a small iron-rich core. Gravity and topography analyses suggest that the mean thickness of the crust ranges between 38 and 62 km, being thicker in the southern hemisphere than in the north (Wieczorek and Zuber 2004). Geochemical analysis from both orbital and surface missions indicate the crust is primarily composed of basaltic materials, ranging from very fresh to heavily weathered. The mantle extends to a depth of ~1,700–2,100 km and is primarily composed of olivine and spinel (Zuber 2001). Gravity measurements from orbiting spacecraft combined with information about the precessional constant obtained from landed spacecraft indicate a C/MR^2 value of 0.3662, indicating that Mars has a small core. The core is proposed to consist of iron mixed with some sulfur and has a radius of ~1,300–1,700 km. The core may be either liquid or solid, but most recent geophysical results favor a liquid state. Mars has no active magnetic field, indicating that the core either is solid or is not undergoing vigorous convection if liquid. However, portions of the ancient crust display remanent magnetization (*Fig. 3-43*) and thus Mars had an active magnetic dynamo for about the first 0.5 Ga of its history (Acuña et al. 1999).

Detailed topographic analysis by the Mars Orbiter Laser Altimeter (MOLA) instrument on Mars Global Surveyor (MGS) revealed that surface elevations range from 21.287 km (top of

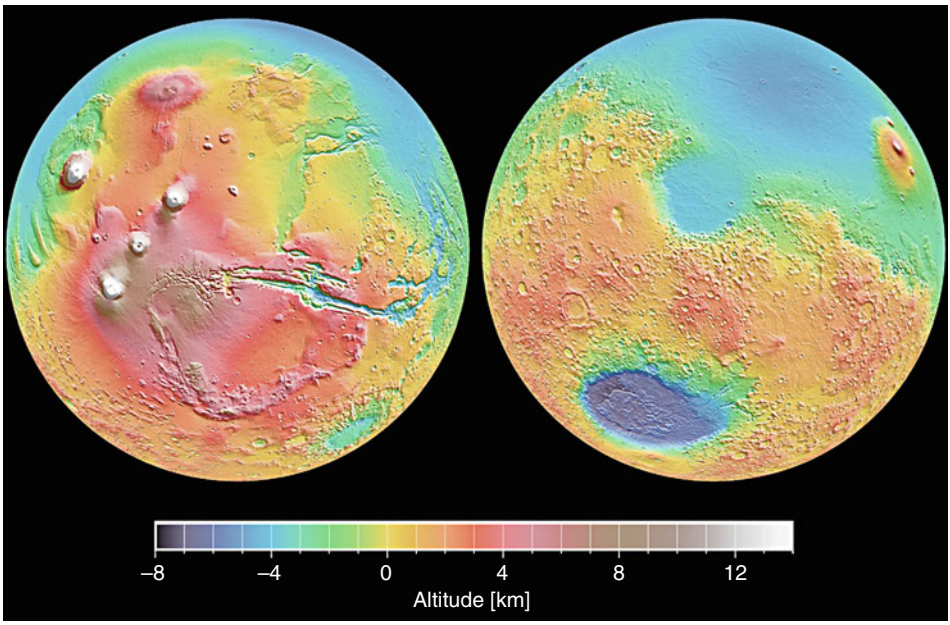


Fig. 3-44

Martian Topography. MGS's MOLA experiment produced the first detailed topography map for Mars. The highest topography is found in the western hemisphere (*left*) associated with the Tharsis bulge. The lowest elevation on the planet is in the Hellas impact basin in the eastern hemisphere (*right*). The hemispheric dichotomy is clearly visible in the topography (NASA/JPL/Goddard)

the Olympus Mons volcano) to −8.180 km (inside the 2,100-km-diameter Hellas impact basin), measured relative to the planet's equipotential surface, or geoid (3,396 km from the planet's center of mass) (*Fig. 3-44*). The lithospheric thickness necessary to support the observed surface topography ranges from <12 km to as much as 200 km under the large volcanoes. Topography and gravity analyses reveal that the Tharsis volcanic region is a bulge of excess mass sitting along the Martian equator. This bulge results in Mars' center of mass being offset from the planet's center of figure by about 2 km in the direction of Tharsis.

Figure 3-44 shows the distinct elevation difference between the northern and southern hemispheres of Mars. Early Mariner 9 (1971–1972) and Viking Orbiter (1976–1980) observations revealed this dichotomy in both elevation and surface ages, with the area south of the hemispheric dichotomy having higher elevation and a more heavily cratered surface, indicating an older age, than the region north of the boundary. The dichotomy is an ancient feature, likely forming shortly after the planet differentiated. One model for dichotomy formation invokes enhanced convection under the northern plains, which would have thinned the crust and allowed for a longer period of volcanism (Breuer et al. 1993; Wise et al. 1979). Another model proposes that a large number of overlapping impact basins thinned the crust north of the dichotomy boundary and later volcanism and sedimentary deposits covered this crater record (Frey and Schultz 1988). MOLA topography and ground-penetrating radars on Mars Express and Mars Reconnaissance Orbiter reveal that a heavily cratered crust of similar age to the highlands lies buried beneath the northern plains (Frey et al. 2002; Watters et al. 2006), but the

distribution of these buried basins does not correlate with crustal thickness or topography of the northern plains and dichotomy boundary (Zuber et al. 2000). A third model invokes a single gigantic impact to carve out the northern plains and produce the dichotomy boundary (Wilhelms and Squyres 1984). Although there is no correlation between crustal thickness and topography of the dichotomy boundary, as would be expected from a giant impact crater, recent reconstruction of the boundary under the younger Tharsis volcanics suggests that an oblique impact could explain the 10,600 × 8,500 km elliptical shape of the dichotomy boundary (Andrews-Hanna et al. 2008). At present, the two leading hypotheses for formation of the dichotomy boundary are the convection and giant impact models.

Geochemical data about Mars come from spectroscopic measurements obtained from ground-based, orbiting, and surface observations, and analysis of Martian meteorites. Most meteorites have formation ages of 4.5 Ga, but by 1979 three meteorites with different mineralogies and isotopic compositions were known. These three meteorites were Shergotty (named for Shergahti, India, where the meteorite fell in 1865), Nakhla (fell in 1941 near El-Nakhla, Egypt), and Chassigny (fell in 1815 near Chassigny, France), and the group of meteorites having characteristics similar to these three representatives became known as the shergottites, nakhlites, and chassignites (SNC meteorites). SNC meteorites, unlike regular meteorites which originate from asteroids, are volcanic rocks with formation ages <1.35 Ga. The youngest SNC has a formation age of 0.16 Ga, indicating these rocks were derived from a planet large enough to have been volcanically active up to a few million years ago. Composition and formation ages of SNCs ruled out all solar system bodies except for Earth, Venus, and Mars. Dynamical considerations make it difficult to eject a rock from Venus and have it move away from the Sun to land on Earth. Venus' large gravitational pull and thick atmosphere also would be detrimental to getting material off the planet's surface, so Venus was eliminated as the possible source of the SNCs. Analysis of oxygen isotopic ratios ($^{17}O/^{16}O$ versus $^{18}O/^{16}O$) in rocks provides information about the solar system location where rocks formed. Such an analysis revealed that SNC meteorites did not form from the same reservoir as terrestrial and lunar rocks, eliminating Earth and the Moon as possible source of these meteorites. Thus, by a process of elimination, Mars became the most likely parent body for the SNCs. Trapped noble gases were discovered in the shergottite EETA79001 (❯ *Fig. 3-45*) in 1982 which were found to be statistically identical to the isotopic ratios of these gases in the Martian atmosphere (Bogard and Johnson 1983). This discovery clinched the case for SNC meteorites to be pieces of the Martian crust.

Both volcanic eruptions and impact crater formation were considered as possible mechanisms for ejecting the SNCs off the Martian surface. The rocks would have to reach or exceed a velocity of 5.03 km s^{-1} to escape the planet, and even the most explosive volcanic eruptions (which are not observed on Mars) do not provide the necessary velocity. Thus impact cratering is the only natural process which can accelerate rocks off the Martian surface. Over 30 Martian meteorites are now cataloged and variations in their ages and compositions indicate they came from five to eight different impact events (Nyquist et al. 2001). Many different craters have been proposed as sources of these meteorites based on geologic and mineralogic analysis, but scientists still cannot definitively link a particular meteorite with its source crater. So even though scientists have samples of the Martian crust, they do not know the exact locations from where those samples were derived. However, one perplexing issue is that all but one of the Martian meteorites have formation ages <1.35 Ga, yet <40% of Mars' surface area displays such a young age. The answer to this riddle likely lies in the fact that older surfaces are covered by thick layers of fragmented material (regolith) – only younger volcanic surfaces are coherent enough to produce rocks which can survive ejection off the planet's surface.

Fig. 3-45
EETA79001. Trapped gases from the Martian atmosphere were first detected in the EETA79001 Martian meteorite (NASA/Johnson Space Center)

Martian meteorites are classified into four main groups based on composition. Shergottites are subdivided into basaltic, lherzolitic, olivine-orthopyroxene, and olivine-phyric, depending on their mineralogy. Nakhlites are clinopyroxene cumulates and chassignites are olivine-rich rocks called dunites. Of the 37 Martian meteorites currently cataloged, 27 are shergotties, seven are nakhlites, and two are chassignites. One additional Martian meteorite (ALH84001) is different from the SNC meteorites in age (formation age 4.1 Ga) and composition (orthopyroxene) and appears to represent a piece of the planet's original cumulate.

Analysis of Martian meteorites combined with spectroscopic analysis of the surface from ground-based, orbiting, and surface landers reveals that Mars has experienced three distinct mineralogic periods throughout its history (Bibring et al. 2006). Mars currently has a thin CO_2 atmosphere (0.7 mbar surface pressure) and cold surface temperatures (mean surface temperature is 240 K), which preclude the existence of liquid water on its surface – any liquid water will either evaporate into the atmosphere or freeze into ice. Martian surfaces which have formed in the past ~3.5 Ga support the concept of dry environmental conditions through the presence of ferric oxides. However, surfaces which formed prior to ~3.5 Ga show evidence of interactions with liquid water, indicating a change in climatic conditions from a wetter period before 3.5 Ga to drier conditions after that time. The period from ~3.8 to 3.5 Ga is characterized by acidic aqueous alteration processes, including high concentrations of sulfur compounds in some areas of Mars. The period prior to ~3.8 Ga was water-rich but less acidic, as indicated by large concentrations of hydrated minerals such as phyllosilicate clays in rocks formed during this period. Mineralogy results suggest that Mars had a thicker atmosphere early in its history which allowed liquid water to exist on the surface. The transition to a more acidic environment at ~3.8 Ga correlates with geologic evidence of a period of increased volcanic activity. Geologic evidence also supports the mineralogy results that the Martian atmosphere thinned around 3.5 Ga and transitioned to the drier environment seen today. Three mechanisms combined around 3.5 Ga to

thin the planet's atmosphere. Mars, being about half the size of the Earth, has a weaker gravitational pull (1/3 Earth's gravity) and thus lighter gases could escape to space by the process of Jean's escape. Geologic evidence from the ages of rocks with and without remanent magnetization indicates that the magnetic dynamo ceased operation around 4.0 Ga, allowing solar wind particles to erode the outer portions of Mars' atmosphere. In addition, formation of large impact basins (such as Hellas and Argyre) around 3.8–4.0 Ga would heat up and expel some of the Martian atmosphere as bolides made their way to the surface. The processes of Jean's escape, solar wind erosion, and impact erosion combined between 3.5 and 4.0 Ga to produce the thin atmosphere and drier climatic conditions seen today (Jakosky and Phillips 2001).

Mars' surface landforms support the notion that the planet has experienced changing climatic and geologic conditions throughout its history. The planet's history is divided into three periods, based on stratigraphic relationships, crater densities, and occurrence of specific geologic processes. The earliest period, the Noachian, extended from the time when the planet's surface solidified until ~3.7 Ga. The Noachian is characterized by high impact rates during the LHB, formation of large impact basins, creation of the hemispheric dichotomy, high degradation rates, and formation of small channels and hydrated minerals by liquid water. The Hesperian Period ran from ~3.7–3.0 Ga and was a period of declining impact rates, increased volcanic activity, transition from an acidic aqueous environment to dry conditions, deposition of much of the northern plains sedimentary deposits, and formation of the large outflow channels by catastrophic floods. The most recent period, the Amazonian, covers the period since 3.0 Ga and is characterized by low impact rates, localized volcanic activity (mainly in the Tharsis and Elysium regions), localized channel formation, low degradation rates, deposition and sublimation of ice in the polar regions, aeolian activity, and localized ice deposits in nonpolar regions.

The Martian surface retains a larger number of impact craters than the surfaces of Earth and Venus. Over 42,000 craters ⩾5-km-diameter have been cataloged, but the crater density varies considerably across the surface. The region south of the dichotomy boundary is heavily cratered and dates to the LHB during the Noachian period. Noachian-aged craters are typically heavily degraded, with missing rims, infilled floors, and missing ejecta blankets and interior structures. Modeling of crater degradation by different geologic processes suggests that Noachian-aged craters were modified mainly by fluvial, mass wasting, and aeolian processes (Craddock and Howard 2002). Craters formed during the Hesperian and Amazonian periods are less heavily modified and retain evidence of their original morphologies. Many of these craters are surrounded by a layered (fluidized) ejecta blanket which displays one, two, or multiple ejecta layers (*Fig. 3-46*). Characteristics of these layered ejecta blankets suggest that vaporization of subsurface volatiles is primarily responsible for formation of this ejecta morphology, although interaction of the ejecta curtain with the Martian atmosphere also may contribute (Barlow 2005). The diameter of the smallest crater displaying a layered ejecta morphology can be used with crater depth/diameter relationships to estimate the depth to subsurface volatile (ice and maybe liquid) layers. Such analyses indicate that ice lies within a few centimeters of the surface at latitudes poleward of ~45° and within 10s to 100s of meters depth in the equatorial region. Martian complex craters (>6–7 km in diameter) display shallower depths and flatter floors than simple craters and often exhibit wall terraces and central peaks. However, some complex craters display a central depression (central pit) which can occur either directly on the crater floor or atop a central peak (*Fig. 3-47*). Central pit craters are only observed on Mars and Jupiter's two largest moons (Ganymede and Callisto), suggesting that the pit forms by vaporization or melting of target ice during crater formation.

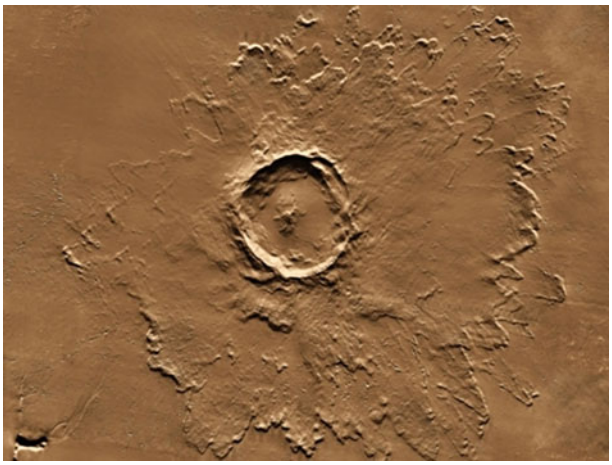


Fig. 3-46
Tooting Crater, Mars. Tooting Crater is a ~28-km-diameter pristine multiple-layer ejecta crater west of Olympus Mons (NASA)

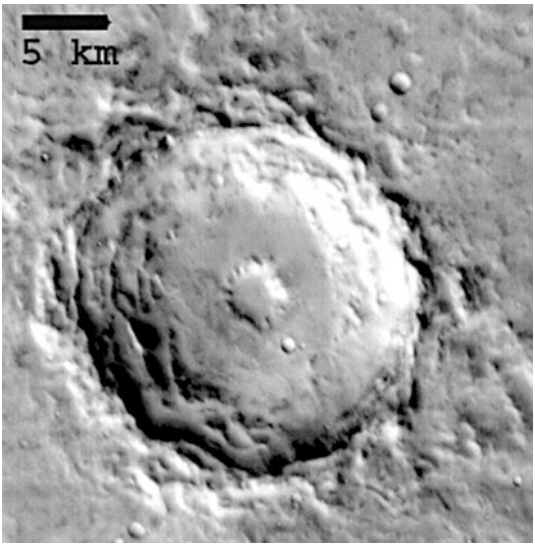


Fig. 3-47
Central Pit Crater, Mars. Central pit craters are common on Mars and likely result from removal of subsurface ice during crater formation. This 21-km-diameter crater (22.5°N 340.4°E) is an example of a floor pit crater (NASA/Arizona State University)

Orbiting spacecraft have been continuously observing the Martian surface since 1996, which has allowed long-term monitoring for changing events. This continuous surface monitoring has revealed formation of several new small impact craters (Malin et al. 2006), which has helped to constrain the impact flux at Mars during recent times. Near-surface ice is

sometimes exposed by these new impact craters, providing additional constraints on the depth to subsurface ice layers at different latitudes (Byrne et al. 2009).

Information on the depth to subsurface ice-rich layers obtained by crater analysis can be compared with orbiter and lander experiments. Neutron spectrometers aboard Mars Odyssey provide insights into the present-day distribution of hydrogen, and thus water, within the upper meter of Martian regolith. High-energy cosmic rays strike the Martian surface, dislodging neutrons from atoms within surface minerals. Energies of these neutrons determine if they are thermal (energy <0.4 eV), epithermal (0.4 eV–0.7 MeV), or fast (0.7–1.6 MeV) neutrons. Various types of neutrons interact in different ways with atoms and molecules within surface materials. Hydrogen, which typically occurs as H_2O within the inner solar system, is a good absorber of epithermal and fast neutrons, whereas CO_2 is a poor absorber of epithermal and thermal neutrons. Distribution of H_2O within the upper meter of the surface can be determined by comparing the fluxes of different neutrons given off by the Martian surface. Results of this study (*Fig. 3-48*) reveal that the polar regions have high concentrations of near-surface H_2O ice, consistent with the proposed distribution of subsurface ice based on models of solar insolation and the expected geothermal heat flux (Feldman et al. 2004). The high northern latitude landing site of the 2008 Phoenix mission was selected based on neutron spectrometer results,

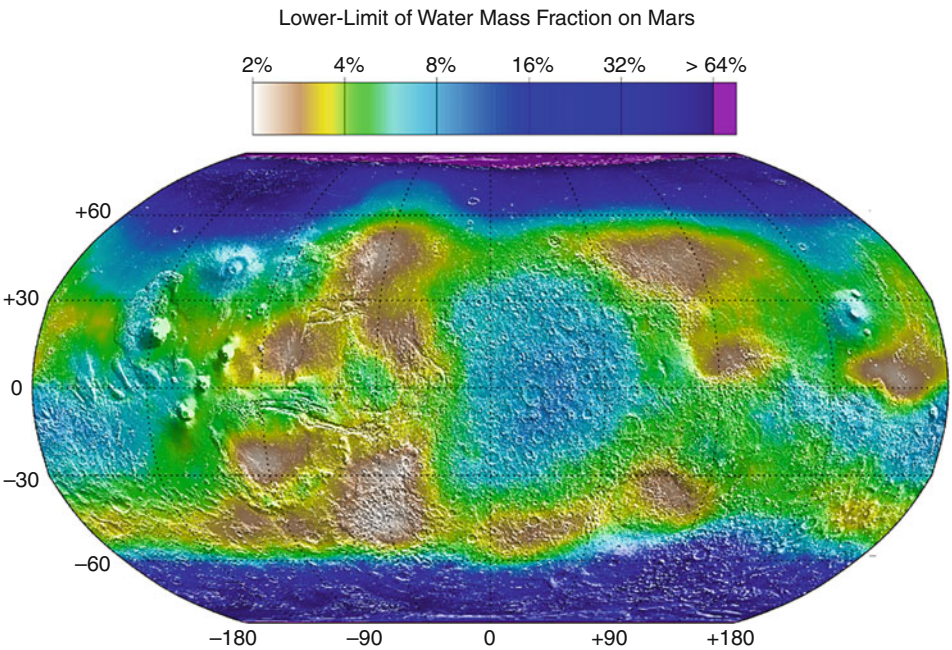


Fig. 3-48

Distribution of Water on Mars. Mars Odyssey's neutron spectrometers measured the flux of epithermal, thermal, and fast neutrons emitted by the planet's surface. Low epithermal and fast neutron fluxes are consistent with high concentrations of hydrogen, and thus water, in the upper meter of the Martian regolith. This map provides estimates of water content based on neutron analysis (NASA/JPL/Los Alamos National Lab)

and Phoenix succeeded in detecting ice within a few centimeters depth. Equatorial regions of Mars are expected to be dry based on ice stability models, but H_2O was detected over broad areas within this region by the neutron spectrometers. This H_2O signature could result from either buried ice or the presence of hydrated minerals.

H_2O and CO_2 ices are exposed on the Martian surface in the polar regions. Earth-based telescopic observations revealed that the Martian polar caps expand and retreat on seasonal cycles (*Fig. 3-49*). The polar cap is largest in winter and a small remnant cap persists through the summer. Temperature and compositional analyses reveal that the seasonally varying cap is composed of CO_2 ice, which condenses directly from atmospheric clouds ("the polar hood") during Martian fall and winter. CO_2 ice sublimates during spring into summer, leaving behind the remnant cap which is composed of H_2O ice. The north polar remnant cap is about 1,100 km in diameter and is centered near the north rotational pole. The ~400-km-diameter south polar remnant cap is offset from the south pole, with its center near 87°S 315°E. Deposition and sublimation of CO_2 ice produces a variety of intriguing features around the polar caps (*Fig. 3-50*). The polar ice caps lie atop a thick sequence of layered ice and dust (the polar layered deposits, PLD), which likely preserves a record of climate cycles similar to terrestrial tree rings.

Landforms suggestive of subsurface and surface ice deposits are seen in many regions of Mars, including within the equatorial zone. Irregular depressions, similar to thermokarst pits on Earth which form from removal of subsurface ice, are seen in the Martian mid-northern latitudes, and polygonal terrain, formed by freeze-thaw cycles in terrestrial permafrost regions, are

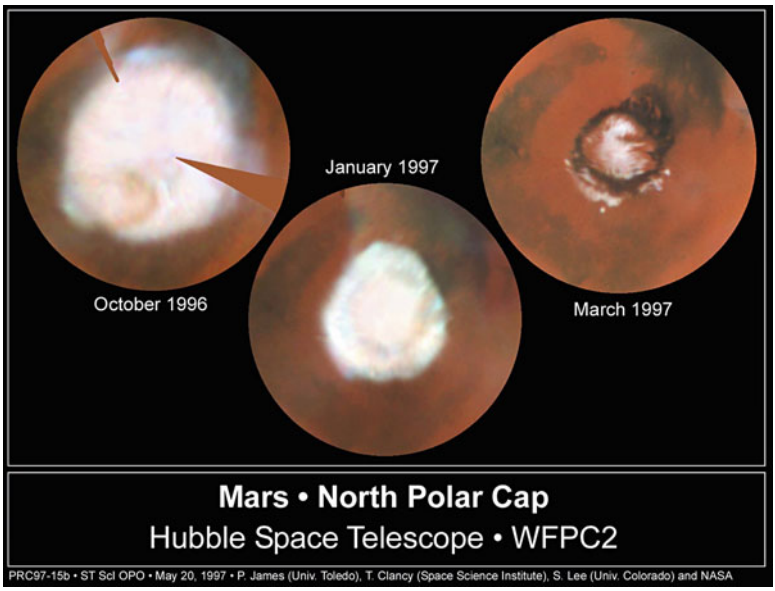


Fig. 3-49
Martian Polar Cap. These composite images of the Martian north polar region were assembled from multiple Hubble Space Telescope images. The three views show the transition from winter (*top left*) to summer (*top right*). The seasonal CO_2 cap which covers the pole during winter disappears in summer, leaving behind the H_2O remnant cap (NASA/JPL/Space Telescope Science Institute)

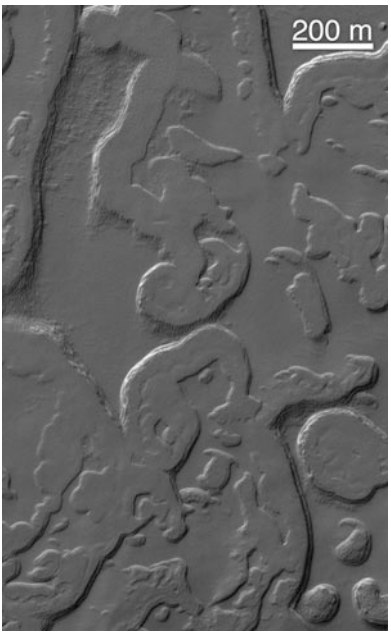


Fig. 3-50
South Polar Terrain, Mars. The deposition and sublimation of CO_2 throughout the Martian year leads to formation of unusual landforms. This series of mesas, pits, and grooves (nicknamed "Swiss Cheese Terrain") is one example of the icy landforms associated with the south polar cap (NASA/JPL/MSSS)

common near both poles (*Fig. 3-51*). Sinuous ridges near the south pole have been proposed to be glacial eskers and parallel lines of small mounds ("thumbprint terrain") may be blocks deposited in moraines of retreating glaciers. Numerous deposits showing evidence of flow lines have been revealed to be ice-rich by radar investigations and are now generally interpreted to be dust-covered glacial deposits (*Fig. 3-52*).

Ice is only deposited over the Martian poles under current climatic conditions, but many of the putative glacial deposits are found at lower latitudes. These features are typically no more than a few million years old, based on counts of superposed craters, indicating that the Martian climate must have undergone recent short-lived changes if these deposits formed from precipitation of snow. Dynamical modeling suggests that gravitational perturbations from Jupiter affect Mars' orbital eccentricity, inclination, and obliquity on cycles of a few million years, similar to Earth's Milankovitch cycles. These models find that Mars' obliquity can vary from almost 0° to over 80° with a most probable value of 41.8° (compared to its present value of 25°) over a period of ~2.5 Ma. Eccentricity can vary from ~0 to 0.12 with a most probable value of 0.068 (current value is 0.093) on a ~1.7 Ma period (Laskar et al. 2004). Mars is currently moving from a lower to higher eccentricity period and from higher to lower obliquity. Polar regions receive more solar insolation during higher obliquity periods, causing CO_2 and H_2O ices in the polar caps and PLDs to sublimate. These gases increase the density of the Martian atmosphere and surface temperatures may rise from greenhouse warming. Precipitation becomes more common under

Fig. 3-51
Polygons, Mars. Polygonal terrain is common in terrestrial permafrost environments where H_2O-rich soils change volume as they alternately freeze and thaw during the seasons. Polygonal terrain also is common in the Martian polar regions, such as this outcrop near Lyot Crater, and is suggestive of ice-rich materials. Image is 3 km across (NASA/JPL/MSSS)

denser atmospheric conditions, but whether this occurs as snow or rain depends on the combination of obliquity and orbital values. The last high obliquity period occurred about 1–2 Ma ago when obliquity reached a moderate value of ~35°, but obliquities near 45° likely occurred around 6 and 9 Ma ago. The putative equatorial cold-based glaciers and a mid-latitude ice-rich mantling layer likely formed by snowfall during the moderately high obliquity period around 1–2 Ma ago.

Glacial processes may have affected the Martian surface in recent times, but liquid water produced numerous fluvial landforms earlier in Martian history (Baker 1982). Martian channels formed by flowing liquid are divided into valley networks and outflow channels. A third type of channel (fretted channels) occurs along the dichotomy boundary and may have been carved by warm-based glaciers. Valley networks are found in the ancient southern highlands and on flanks of younger volcanoes (*Fig. 3-53*). Many display dendritic patterns, and interconnections between valley networks and drainage basins suggest that rainfall was a major source for the water carving these channels. However, some valley networks show evidence of a sapping origin, suggesting both rainfall and groundwater contributed to formation of these features. Valley network formation in Noachian terrains rapidly declines near the end of the LHB, with younger systems only being found in association with volcanoes. Thinning of the Martian atmosphere between 3.5 and 4.0 Ga and the accompanying change to a more arid climate probably eliminated valley network formation by rainfall and concentrated subsequent

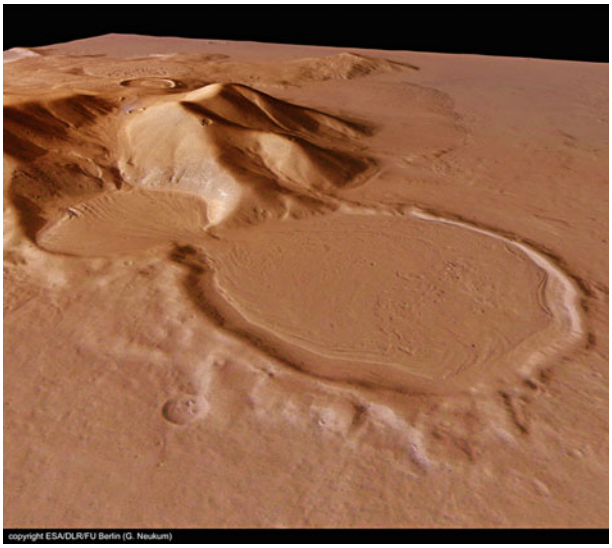


Fig. 3-52
Possible Glacier, Mars. Many features displaying evidence of downslope flow lines have been identified on Mars. These features may be dust-covered glaciers from snow deposited during the last high obliquity period. This Mars Express High Resolution Stereo Camera image is a perspective view of such a deposit flow from the upper 9-km-diameter crater into the lower 16-km-diameter crater (ESA/DLR/Frei University, Berlin)

channel formation in regions where either magma could maintain near-surface groundwater systems or volcanic activity produced localized precipitation.

Outflow channels are much larger than valley networks, with widths on the order of tens of kilometers and lengths of thousands of kilometers (*Fig. 3-54*). Outflow channels formed during the Hesperian and Amazonian periods under climatic conditions similar to today. Outflow channels originate from large depressions, often filled with a jumble of rocks called chaotic terrain. Characteristics of outflow channels suggest that they form from catastrophic release of groundwater, similar to the catastrophic flooding that formed the Channeled Scablands in eastern Washington. The source of this groundwater is likely ground ice melted by volcanic or impact processes.

Other evidences of liquid water, particularly during the Noachian, include impact craters with rims breached by channels and sediment-covered floors. These are interpreted to be ancient lakebeds (paleolakes). Distributary fans are sometimes observed where the channel enters the crater floor and may be analogous to terrestrial deltas (*Fig. 3-55*). Alluvial fans also have been recognized in several impact craters. Opportunity rover has found mineralogic evidence that an acidic, briny sea existed in the Meridiani Planum region for an extended period of time (*Fig. 3-56*), and Spirit rover has found evidence of lakebed and hydrothermal deposits near its landing site in Gusev crater. Recent evidence of water flow can be seen in small gullies formed along crater and canyon walls (*Fig. 3-57*). Although CO_2 and methane (CH_4) have been proposed as the eroding fluids, models show that current conditions on Mars favor gully formation by H_2O. Seepage of groundwater (Malin and Edgett 2000) and/or melting of

Fig. 3-53
Warrego Valles, Mars. Warrego Valles displays the typical dendritic morphology of valley network systems. This Thermal Emission Imaging System image is 17 × 62 km (NASA/JPL/Arizona State University)

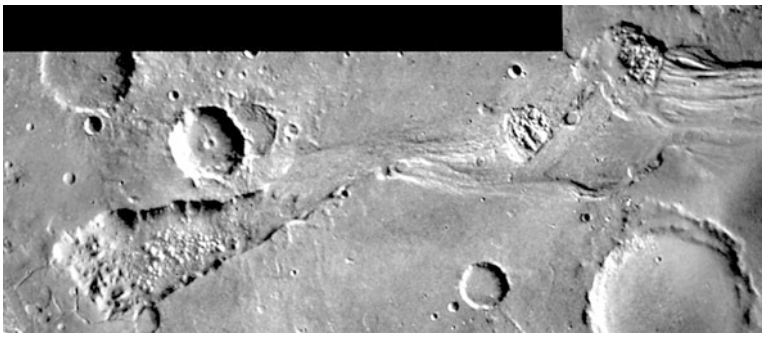


Fig. 3-54
Ravi Vallis, Mars. Ravi Vallis is an outflow channel which originated in a depression filled with chaotic terrain. The channel extends about 300 km in length in this image (NASA/LPI)

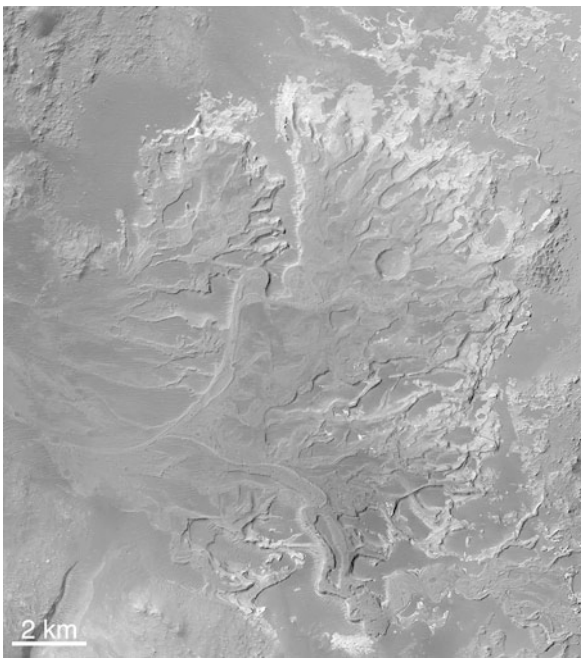


Fig. 3-55
Eberswalde Fan, Mars. The distributary fan in Eberswalde Crater (24.0°S 326.3°E) occurs where a channel has cut the crater rim and enters the crater floor. The fan shows characteristics similar to delta deposits on Earth and likely represents sediment deposition into a paleolake environment (NASA/JPL/MSSS)

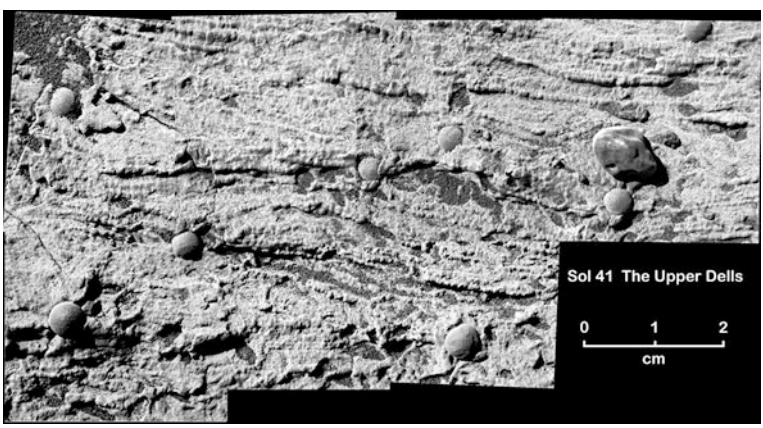


Fig. 3-56
Upper Dells, Mars. Images by Opportunity rover reveal a series of layers in the rocks near its landing site. The non-horizontal layers are suggestive of emplacement by small water ripples. The small spheres are hematite concretions formed by iron-rich water percolating through the rocks (NASA/JPL/Cornell/US Geological Survey)

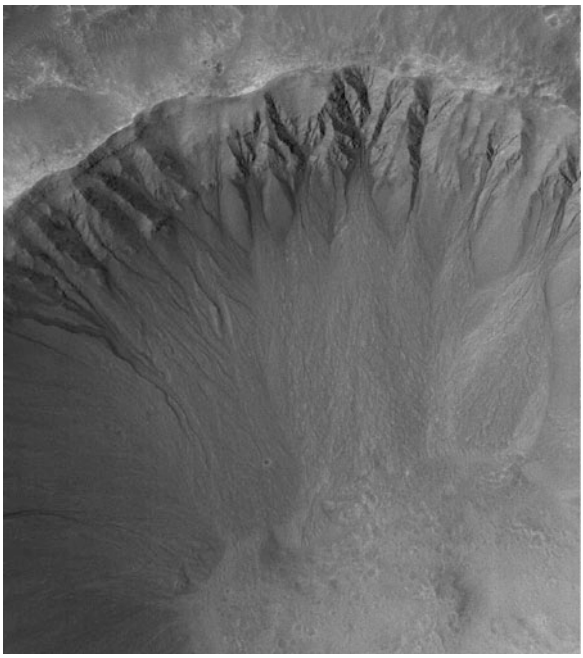


Fig. 3-57
Martian Gullies. Many slopes, such as the wall of this impact crater, display small gullies. Gullies formed recently, as indicated by the lack of superposed dunes or craters. They may form by groundwater seepage, mass wasting, or snowmelt. This MGS Mars Orbiter Camera image is 3 km wide (NASA/JPL/MSSS)

ice deposited during the last high obliquity cycle (Christensen 2003) are the most likely sources of the fluids carving these gullies.

The thin Martian atmosphere undergoes large pressure and temperature variations on daily, seasonal, and annual cycles which produce surface winds. Aeolian features are therefore common on the Martian surface and are among the landforms seen to change in appearance during the decades of spacecraft observation. Dust transport is obvious on Mars, both in terms of everyday dust in the atmosphere (landers often see a pink sky because of atmospheric dust) and the frequent dust storms which can range in size from localized to global. Dust devils (*Fig. 3-58*) help to transfer dust particles from the surface into the atmosphere (and sometimes help clean dust off the solar panels on Spirit and Opportunity) and can help initiate the saltation (hopping motion) of larger sand particles. Aeolian depositional features on Mars range from large sand seas called ergs to numerous sand dune deposits and small ripples (*Fig. 3-59*). Both bright and dark wind streaks are seen on Mars and their changing orientations provide insights into wind shifts on both seasonal and longer timescales. Yardangs, ventifacts, and the presence of pits and grooves in surface rocks (*Fig. 3-60*) are indicators of the erosive power of sediment-carrying winds. Dust devils remove dust from the surface, leaving behind dark tracks where the underlying darker bedrock is revealed.

Mass wasting processes also are evident on the surface of Mars. Large complex craters often show wall terraces, formed when their oversteepened walls collapsed under the influence

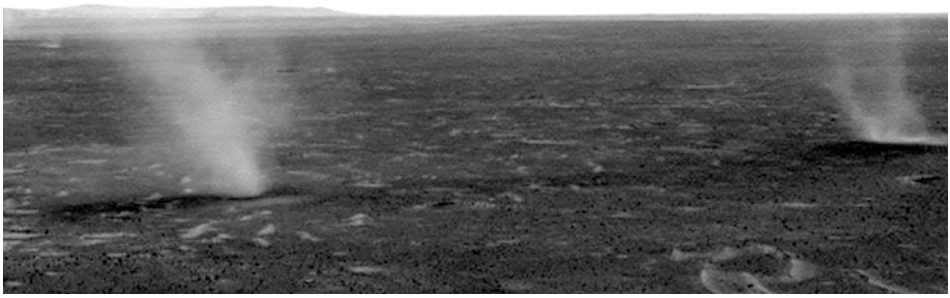


Fig. 3-58
Martian Dust Devils. Spirit rover has captured many images of dust devils crossing the floor of Gusev Crater. Dust devils are thought to help lift dust from the surface into the Martian atmosphere (NASA/JPL/Texas A&M University)

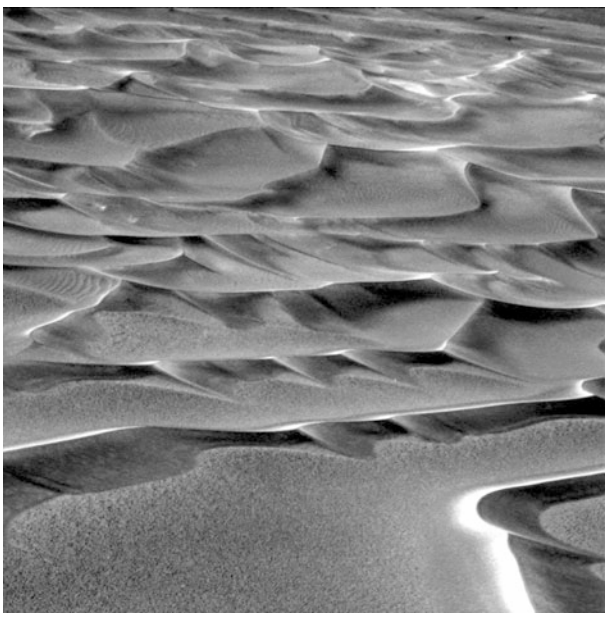


Fig. 3-59
Martian Sand Dunes. Opportunity rover imaged these sand dunes on the floor of Endurance Crater. Small ripples are superposed on *top* of the larger dunes (NASA/JPL/Cornell)

of Martian gravity. Landslides are common along steep slopes, particularly within the Valles Marineris canyon system, which likely has increased in width by the action of landslides over the canyon's long history (*Fig. 3-61*). Crater analysis indicates that these landslides have been occurring since at least 3.5 Ga ago and the youngest slides are <50 Ma. Both dry granular flows and wet flows match the morphologies of these landslides and both mechanisms probably have operated at different times.

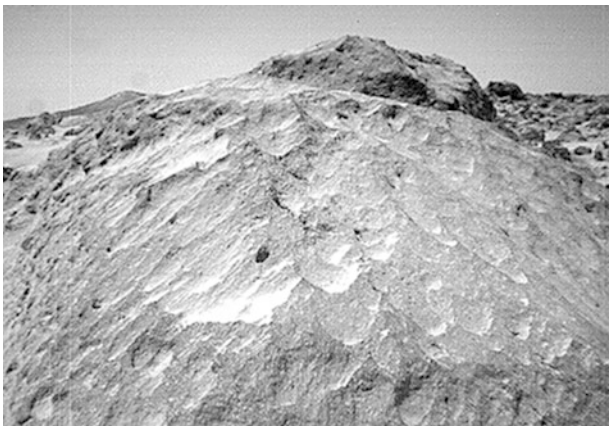


Fig. 3-60
Moe Rock, Mars. Moe rock displays numerous pits formed by wind abrasion. This image was taken by Mars Pathfinder's Sojourner rover (NASA/JPL)

Martian volcanic features were first detected by Mariner 9 and range from lava plains to a variety of volcanic edifices. Spectroscopic observations from orbiting spacecraft suggest that the surface composition is largely basalt, and this is consistent with the types of volcanic features observed across the planet. Flood basalts appear to have erupted throughout Martian history, being exposed as Noachian-aged intercrater plains, Hesperian-aged ridged plains, and Amazonian-aged lava flows in the Tharsis and Elysium volcanic provinces. Both large and small shield volcanoes are scattered across the planet, with the highest concentrations occurring in Tharsis and Elysium.

The largest volcano known in the solar system is 21.3-km-high Olympus Mons, which lies on the western edge of the Tharsis bulge (*Fig. 3-62*). Three other large shield volcanoes (Ascraeus, Pavonis, and Arsia Montes) are aligned in a northeast-to-southwest orientation along the crest of the Tharsis bulge. The multiple calderas at the summits of these large shield volcanoes together with numerous lava flows along their flanks indicate these volcanoes underwent multiple eruptions. Crater counts suggest that the most recent eruptions occurred about 40 Ma ago, suggesting that these volcanoes, particularly Olympus Mons, might still be active. The other large volcano in Tharsis is 1,600-km-diameter Alba Mons, which reaches a height of 6.8 km and is surrounded by a system of circumferential graben (*Fig. 3-63*). Smaller volcanic domes (tholii) in the Tharsis region are small but older shield volcanoes, many of which have their bases covered by more recent lava flows (Hodges and Moore 1994). Crater counts indicate that volcanism in the Tharsis region has been occurring since the Noachian and successive lava flows over Mars' history have gradually constructed the bulge seen today.

The Tharsis volcanic province dominates the western hemisphere of Mars, whereas volcanic activity in the eastern hemisphere is concentrated in Elysium. Three volcanoes constitute the Elysium volcanic region: Albor Tholus, Hecates Tholus, and Elysim Mons. Albor Tholus is a 160-km-diameter Noachian-to-Hesperian-aged shield volcano. Hecates Tholus primarily formed in the Hesperian, although its activity may have continued into the Amazonian. It is 200 km in diameter with a small caldera, numerous channels on its flanks, and a possible pyroclastic (ash-rich) deposit on its western flank. Four-hundred-km-diameter Elysium Mons is the youngest

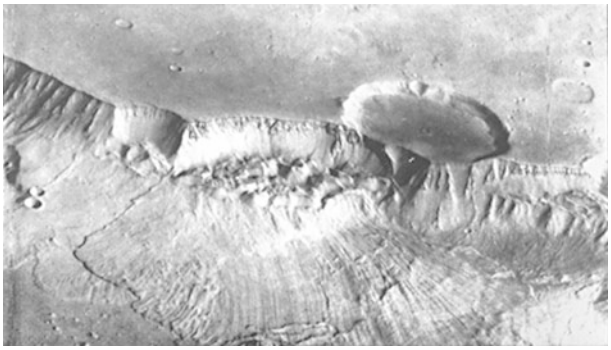


Fig. 3-61
Landslide in Valles Marineris, Mars. Landslides occur when an oversteepened slope collapses under the influence of gravity. Landslides in Valles Marineris have helped to widen the canyon system. Image is 60 km across (NASA/LPI)

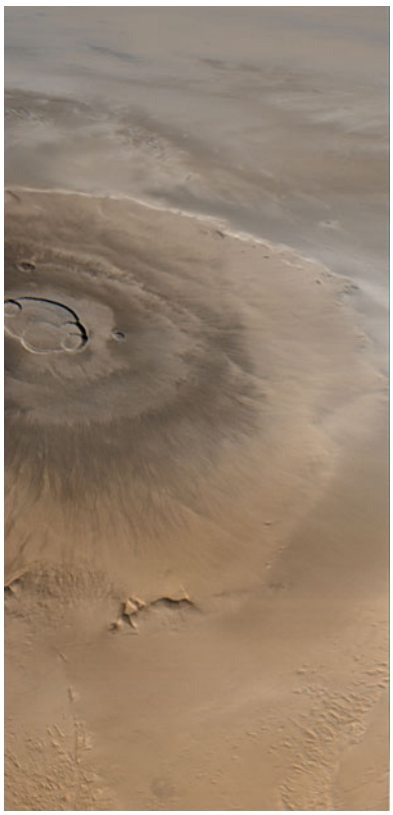


Fig. 3-62
Olympus Mons, Mars. The 550-km-diameter, 27-km-high Olympus Mons volcano is the largest volcano known in the solar system. It displays the classic morphologic characteristics of a shield volcano (NASA/JPL/MSSS)

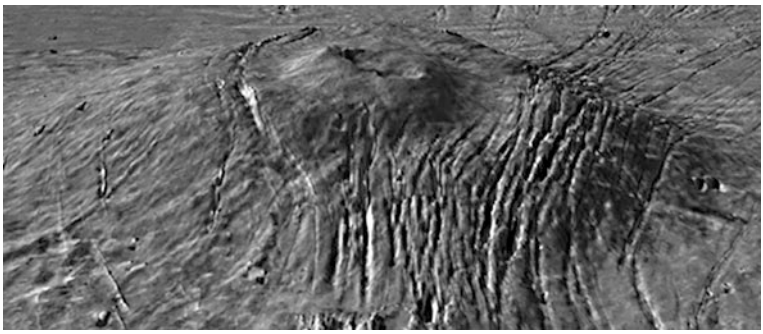


Fig. 3-63
Alba Mons, Mars. This three-dimensional view of 1,600-km-diameter Alba Mons volcano was obtained by draping Viking imagery over MOLA topography. The circumferential graben system surrounding the volcano is clearly visible (NASA/Goddard)

of the Elysium volcanoes with a Late Hesperian to Early Amazonian age. It is the tallest of the Elysium volcanoes with a height of 14 km.

Recent volcanic activity has been concentrated in the Tharsis and Elysium volcanic provinces, but earlier activity was more extensive. The Hesperian period witnessed not only the formation of volcanic edifices in Tharsis and Elysium but also creation of the low-relief Syrtis Major volcano and extrusion of low-viscosity lavas which formed the ridged plains. The oldest volcanic edifices on Mars are the paterae, which are large but very low-relief (slopes <3°) volcanoes. Paterae have basaltic compositions, but their flanks are covered by easily erodible ash layers which are highly dissected by wind and/or fluvial activity (*Fig. 3-64*). The ash layers are produced either by a steam explosion caused when magma encounters water, or by rapid ascent of deep-sourced magmas. The explosive phase of paterea formation occurred in Late Noachian to Early Hesperian, but later effusive lava flows continued into the Amazonian.

The youngest volcanic unit on Mars lies in the Cerberus Planitia area, which is west of Tharsis and east of Elysium. Crater analysis of flood basalts in this area suggest formation ages from 3 to 100 Ma ago (Berman and Hartmann 2002; Werner et al. 2003). These low-viscosity lavas were extruded from fractures and/or small shield volcanoes on the western side of Cerberus Planitia.

Most volcanic activity on Mars consists of low-viscosity basaltic eruptions, but a possible Amazonian-aged ash deposit from explosive volcanism lies southwest of the Tharsis shields in the Medusae Fossae Formation (MFF). Radar observations suggest a maximum thickness of 580 m for this deposit (Carter et al. 2009). The material is fine-grained, as indicated by the presence of aeolian erosional features such as yardangs, is porous, and shows no evidence of fine-scale internal layering similar to the PLDs, although larger-scale layers are observed (Carter et al. 2009; Mandt et al. 2008). Mars Odyssey's Gamma Ray Spectrometer detected a high concentration of chlorine coincident with the deposit (Karunatillake et al. 2009). All of MFF's characteristics are consistent with it being a type of pyroclastic deposit known as an ignimbrite (Mandt et al. 2008), which originated from explosive activity in the corridor between the Tharsis and Elysium volcanic provinces. Explosive eruptions in Tharsis are suggested to have produced layered deposits in other regions around the planet.

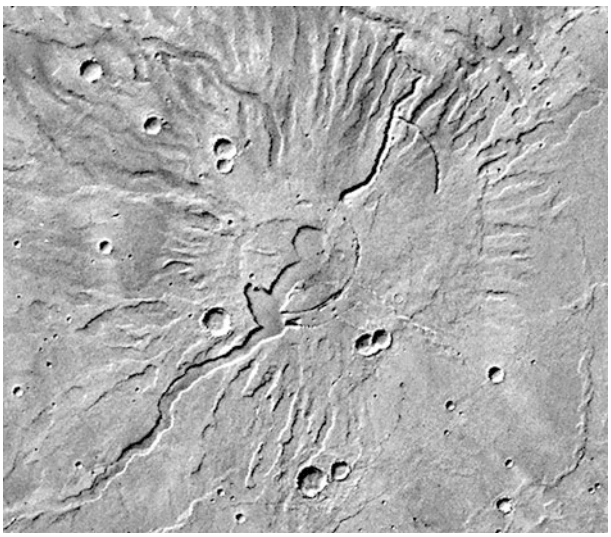


Fig. 3-64
Tyrrhena Mons, Mars. Paterae are low-relief, heavily dissected volcanic edifices which may have formed by magma interactions with water early in Martian history. A classic example of a patera is 300-km-diameter, 2-km-high Tyrrhena Mons, located northeast of the Hellas impact basin (NASA/JPL)

Northwest of Olympus Mons and extending 300–700 km from the edge of the volcano is a region of blocks, arcuate ridges, and deep extensional troughs called the Olympus Mons aureole (*Fig. 3-65*). Possible origins for the aureole deposit include both volcanic (eroded pre-Olympus Mons shield volcano, eroded lava flows, or pyroclastic deposits) and tectonic (failure of the volcano flanks) mechanisms. Gravity, topography, and image data support a tectonic origin for the aureole, suggesting that the northwestern flank of Olympus Mons collapsed and slid away from the volcano (McGovern et al. 2004). A scarp up to 10 km in height defines the edge of Olympus Mons on the side facing the aureole and its origin is consistent with the flank collapse model.

Like Venus, Mars is a single-plate planet and shows no sign of plate tectonic activity at the present time. Plate tectonics early in Martian history have been proposed, particularly to explain the striped nature of the remanent magnetization recorded by rocks in the highlands (*Fig. 3-43*). However, other mechanisms can explain the magnetization pattern and models of early plate tectonics cannot explain observations of present day crustal thickness, timing of the magnetic dynamo, declining rate of volcanism, and formation of the crustal dichotomy (Breuer and Spohn 2003). Thus tectonism is localized in extent and is primarily associated with mantle upwelling and thermal plume activity responsible for Martian volcanism. Wrinkle ridges are the dominant form of compressional tectonics observed on Mars and form when layered lava plains sag under their weight (*Fig. 3-66*). Wrinkle ridges are broad linear arches with superposed ridges which are surface expressions of subsurface thrust faults. Extensional fractures are commonly associated with the volcanic centers, particularly the Tharsis bulge, and indicate that the lithosphere underwent substantial stretching in response to volcanic loads.

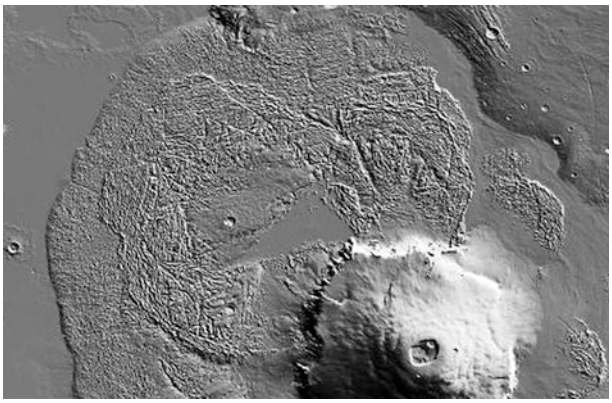


Fig. 3-65
Olympus Mons Aureole, Mars. The Olympus Mons aureole extends up to 700 km from the volcano's edge and consists of a series of blocks, ridges, and troughs. The aureole probably formed by the collapse of the northwestern flank of Olympus Mons (MOLA shaded relief image, NASA/Goddard)

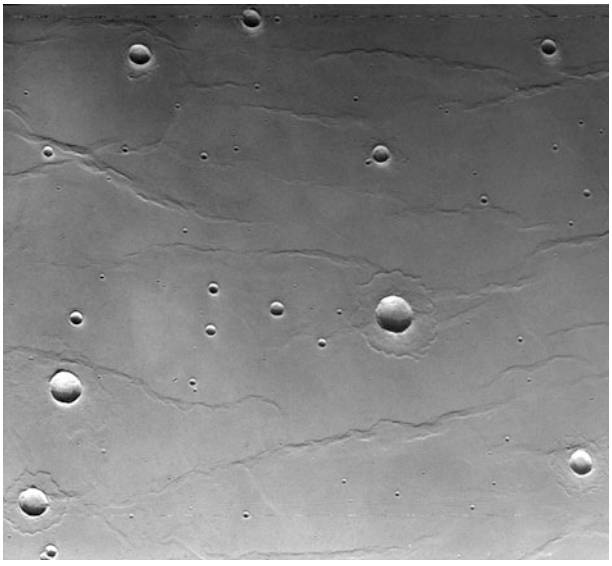


Fig. 3-66
Lunae Planum, Mars. This Viking image shows a portion of the Lunar Planum ridged plains. The ridges are called wrinkle ridges and are surface expressions of subsurface thrust faults (NASA/JPL)

The largest extensional feature on Mars is the Valles Marineris canyon system (*Fig. 3-67*), first discovered in Mariner 9 images. Valles Marineris consists of a series of smaller canyons, stretching ~4,000 km along the equator between 250° E and 330° E longitude. Parts of the canyon reach depths of 6 km below the reference radius and 11 km below the surrounding plains. Dike emplacement from Syria Planum to the south or stresses associated with the uplift of the Tharsis

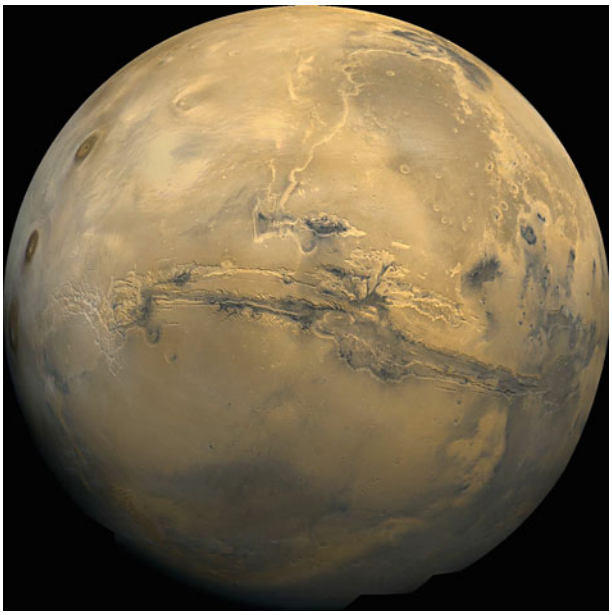


Fig. 3-67
Valles Marineris, Mars. The Valles Marineris canyon system stretches across the equatorial region in this global mosaic of Mars (NASA/JPL)

region to the west initiated the fracturing which formed Valles Marineris, beginning in the Late Noachian to Early Hesperian. The canyon has subsequently expanded through mass wasting processes and possibly some fluvial erosion.

The possibility of current (extant) or past (extinct) life on Mars has intrigued scientists and the public for centuries (see discussion in Barlow 2008). Italian astronomer Giovanni Schiaparelli (1835–1910) reported telescopic observations of thin dark lines crossing the planet's surface in 1877. Schiaparelli termed these lines canali, Italian for channel. Channels are formed by natural processes such as aeolian or fluvial erosion or by tectonic fracturing. However, the term was incorrectly translated into English as canal, which implies a waterway engineered by intelligent lifeforms. American astronomer Percival Lowell (1855–1916) promulgated the idea that the canals were artificial waterways, constructed by a dying race of Martians to bring water from the melting polar caps to the thirsty populace living in cities near the equator. Astronomers argued for many decades about whether the canals were real or simply optical illusions caused by poor seeing conditions. The debate was not resolved until spacecraft began visiting Mars in 1965. Early images from the Mariners 4, 6, and 7 flyby missions revealed no signs of canals, Martian cities, or even vegetation, and these missions were the first to reveal the thinness of the atmosphere and inability of Mars to retain liquid water on its surface. However, the Mariner 9 discovery of channels formed by past episodes of flowing water reignited the possibility that lifeforms may have arisen on Mars. This led to the two 1976 Viking missions, each of which consisted of an orbiter and a lander. The landers carried a series of experiments designed to test Martian soil for evidence of microbial life. Viking Lander 1 landed in Chryse Planitia (22.48°N 312.03°E) on July 20, 1976, followed by Viking Lander 2 in Utopia Planitia (47.97°N 134.26°E)

on September 9, 1976. The biology experiments were designed to detect soil organics and waste products produced by microorganisms under a variety of humidity, temperature, and radiation conditions. The experiments indicated that Martian soil is highly reactive but that it is primarily chemical reactions resulting from the oxidized nature of the soil rather than biologic activity which produced the observed results.

Scientists generally considered the question of Martian life to be resolved until 1996, when analysis of the ancient Martian meteorite ALH84001 suggested it contained signatures of Martian life (McKay et al. 1996). Carbonates within the meteorite contain mixtures of polycyclic aromatic hydrocarbons, magnetite, and iron sulfides which can be produced by bacteria. Electron microscope images also revealed tiny elongated features interpreted as possible fossilized bacteria (*Fig. 3-68*). Subsequent investigations of the ALH84001 signatures have questioned the original biologic interpretations. The putative fossilized bacteria are either an artifact from sample preparation or terrestrial contamination. Chemical signatures are generally attributed to inorganic processes or terrestrial contamination. Thus the ALH84001 evidence is considered too weak to support the claim that it contains evidence of ancient Martian life.

Low temperature, lack of abundant liquid water, oxidizing soil conditions, and penetration of radiation to the surface produce surface conditions which are inhospitable to terrestrial biology. Any extant Martian life has likely migrated underground where conditions may be more conducive for biologic activity. No signs of extinct or extant life have been detected by surface landers or rovers, but future missions, such as NASA's 2011 Mars Science Laboratory (recently named Curiosity) and ESA's 2018 ExoMars, are designed to more completely explore Mars' exobiological potential.

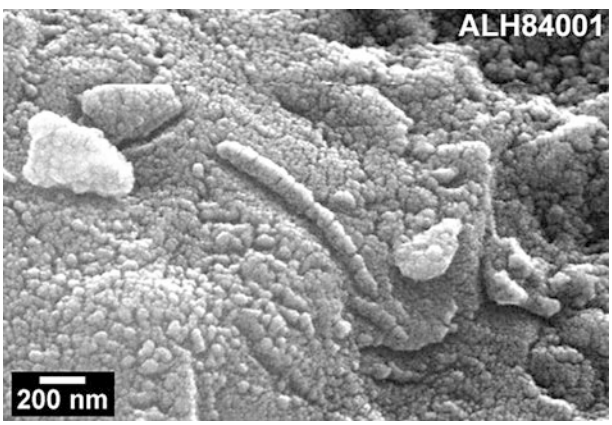


Fig. 3-68
Putative Martian Life. Electron microscopic images of the ALH84001 meteorite revealed the presence of intriguing features within carbonate globules. McKay et al. (1996) interpreted these features, such as the elongated, segmented feature near image center, as fossilized nanobacteria (NASA/Johnson Space Center)

5 Mercury

Mercury is the smallest of the terrestrial planets, with a mean diameter of 4,879 km (38% of Earth). The size, combined with a mass of 3.30×10^{23} kg, gives the planet a high density of 5,400 kg m^{-3}, indicating that Mercury contains a large amount of iron in its interior. The planet has the most elliptical orbit among the eight major planets (e = 0.2058), resulting in the intensity of sunlight varying by a factor of 2.5 between the planet's closest and furthest distances from the Sun. Tidal forces from the Sun combined with Mercury's elliptical orbit have placed Mercury in a 3:2 spin-orbit coupling, meaning that the planet rotates exactly three times for every two orbits around the Sun. Tidal forces also have affected Mercury's obliquity, placing the rotation axis almost perpendicular to the planet's orbital plane.

Mercury is difficult to observe with Earth-based telescopes because of its small size and proximity to the Sun. The first detailed information about the planet's surface and interior came from the Mariner 10 mission, which passed by Mercury three times in 1974–1975 (Strom and Sprague 2003; Vilas et al. 1988). The mission's orbital characteristics combined with Mercury's 3:2 spin-orbit coupling resulted in image acquisition of only 45% of the planet's surface (*Fig. 3-69*). Mariner 10 was actually in orbit around the Sun, not Mercury, limiting

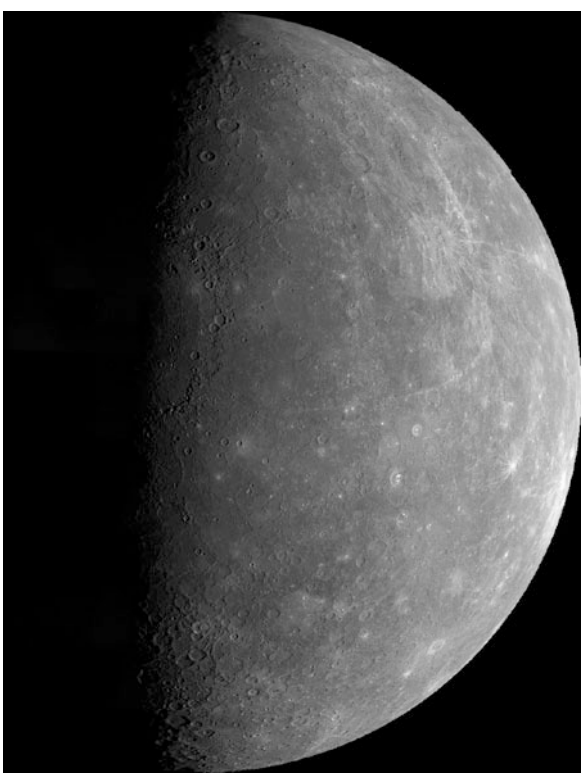


Fig. 3-69
Mercury. Mariner 10 revealed the first images of Mercury's surface during its flybys in 1974–1975. This image shows Mercury's heavily cratered surface (NASA/JPL/Northwestern University)

the amount of gravity information which could be obtained about the planet. Improvements in both surface coverage and internal structure will come from the MErcury Surface, Space ENvironment, GEochemistry, and Ranging (MESSENGER) mission, which made three flybys of Mercury between January 2008 and September 2009 and entered Mercury orbit in March 2011.

Mariner 10 revealed the high density of the planet and also detected a weak magnetic field (0.07% Earth's field strength) around Mercury. These two observations indicate that Mercury's iron core is larger relative to the planet's volume than that of any other terrestrial planet core. Lack of orbital gravity data precludes a precise determination of the C/MR^2 value until MESSENGER has been in orbit for a longer period, but models suggest that 42% of Mercury's volume is taken up by its iron core. In comparison, Earth's core is only 16% of our planet's volume. Mercury's core is estimated to have a radius of ~1,800 km while the mantle is ~600 km thick and the crustal thickness ranges from 100–300 km. The large volume occupied by the core may be the result of an early catastrophic impact on Mercury which stripped away a substantial portion of the planet's original crust and mantle.

Scientists have debated the origin of Mercury's magnetic field since its detection by Mariner 10. One possibility is that Mercury's core has completely solidified, shutting off the dynamo process and leaving the magnetic field as a remnant field preserved in Mercury's crust. The other model is that Mercury's core remains partially molten and the field is produced by an active dynamo at the present time. Results from MESSENGER's flybys do not support the remnant magnetic field model and thus Mercury's dynamo appears to still be operating (Solomon et al. 2008).

Mariner 10's cameras imaged about 45% of Mercury's surface and MESSENGER has largely completed image mapping of the planet. Mercury's geologic history is divided into five periods based on stratigraphic and crater density analysis. The oldest terrains have formation ages >4.0 and are placed in the pre-Tolstojan period. The Tolstojan period covers materials formed between ~4.0–3.9 Ga and is followed by the Calorian period, which extended from ~3.9–3.5 Ga. The youngest geologic periods are the Mansurian (3.5–3.0 Ga) and the Kuiperian (~1.0 Ga).

Mercury has only a very low-density haze of atoms and molecules above its surface, producing an exosphere. The particles in the exosphere include those captured from the solar wind (hydrogen, helium, and oxygen) and atoms sputtered from the surface by impact of solar wind particles (calcium, sodium, and potassium). This exosphere is too thin to produce any fluvial, glacial, or aeolian features on Mercury.

Impact craters dominate in all surface images, indicating that Mercury's crust cooled and solidified early in the planet's history. Most of the surface dates from the LHB, although the smooth plains are slightly younger (Strom et al. 2008). Several large impact basins occur on Mercury, including the 1,550-km-diameter Caloris basin (*Fig. 3-70*) and the recently discovered 715-km-diameter Rembrandt basin. Basin formation strongly influenced the subsequent volcanic and tectonic evolution of regions in and around these large impact structures.

Earth-based radar observations suggest that some impact craters near Mercury's poles may contain ice deposits. Floors of these high-latitude impact craters are in perpetual shadow because of Mercury's small obliquity. The cold temperatures in these perpetually shadowed regions trap volatiles, which will migrate to these cold traps regardless of where the volatiles originated. MESSENGER will investigate the polar regions and its neutron spectrometer will determine if the radar signature seen from Earth is indicative of water ice deposits on crater floors.

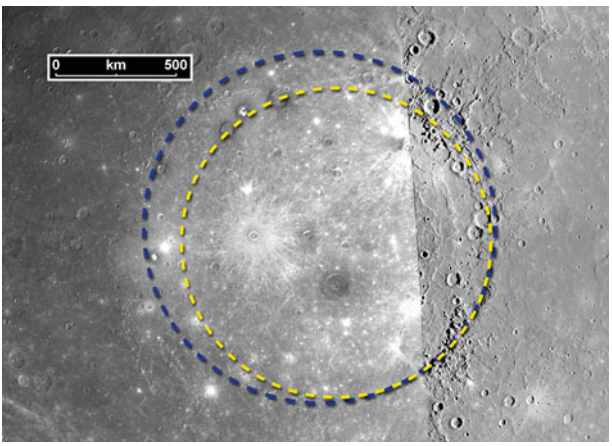


Fig. 3-70
Caloris Basin, Mercury. Mariner 10 observations were limited to the eastern rim of the Caloris Basin. Based on those images, Caloris was estimated to be ~1,300 km in diameter (*inner yellow dashed circle*). MESSENGER has imaged the entire basin and finds the diameter to be closer to 1,550 km (*outer blue dashed circle*) (NASA/Johns Hopkins Univ. Applied Physics Lab (JHU APL)/Carnegie Institution of Washington/Brown University)

Mariner 10 imagery revealed two types of plains units, the intercrater plains and smooth plains. Intercrater plains, which exist between craters on heavily cratered terrain, were generally considered to be volcanic in origin, but controversy ensued regarding whether the smooth plains were volcanic or impact ejecta. MESSENGER albedo and color data suggest that most of the planet's crust is of volcanic origin (Denevi et al. 2009) and crater analysis indicates that the smooth plains often display statistically different ages than the large basins (Strom et al. 2008). These observations indicate that both intercrater and smooth plains are likely volcanic in origin.

MESSENGER data have revealed an enhanced role of volcanism in Mercury's history (Head et al. 2008). Mercury's albedo is lower than the Moon's, suggesting widespread distribution of opaque components with high iron and/or titanium components which originated in Mercury's mantle (Denevi et al. 2009). High-resolution images reveal volcanic vents which are the likely source regions of plains material (Head et al. 2008). Low-viscosity magmas dominated in Mercury's volcanic history, as indicated by the prevalence of lava flows and lack of large volcanic edifices. The floors of most large impact basins appear to be flooded by postimpact volcanism. The first possible shield volcano detected on Mercury is a small domed feature with irregular depressions located along the southern margin of the Caloris basin (*Fig. 3-71*).

Mercury's tectonic environment is unique among the terrestrial planets. Extensional tectonics are limited to the interiors of large impact basin, typically as circumferential graben near the basin rim but occasionally as radial graben near the basin center (*Fig. 3-72*) (Watters et al. 2009). The circumferential graben are similar to those seen in mare-filled lunar basins and result from sagging of lava-filled basin floors, creating extensional stresses near the basin rim. Radial graben patterns may result from updoming of the surface by subsurface magma intrusions.

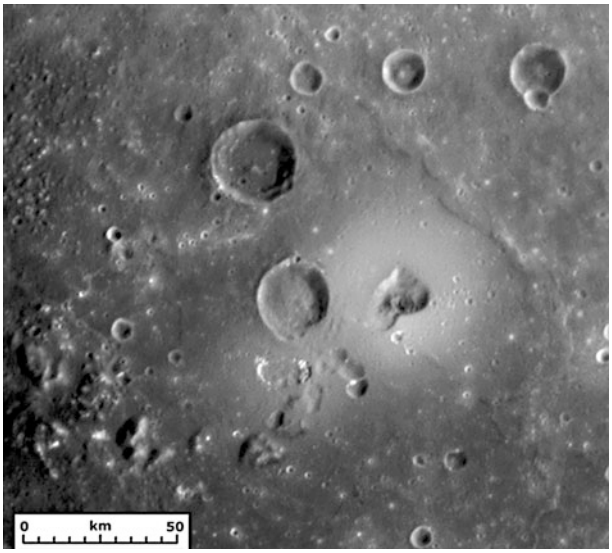


Fig. 3-71
Shield Volcano, Mercury. Mariner 10 imagery revealed widespread lava flows on Mercury, but the first evidence of a possible shield volcano comes from MESSENGER images. The feature at center is characterized by irregular depressions which could be calderas. The brighter deposit surrounding the caldera is possibly an ash deposit (NASA/JHU APL/Arizona State Univ./Carnegie Institution of Washington)

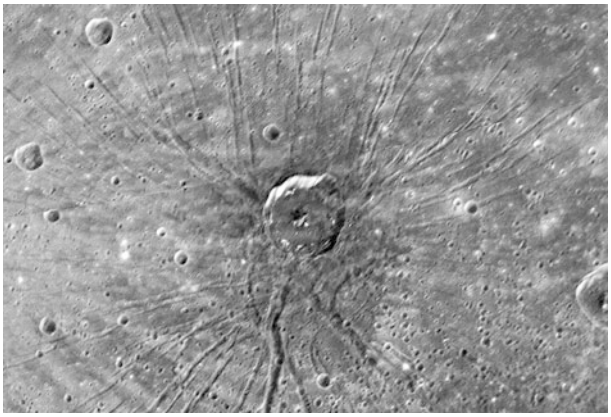


Fig. 3-72
Pantheon Fossae, Mercury. Pantheon Fossae is an unusual feature located near the center of the Caloris Basin. It consists of a series of radial faults extending outward from a 40-km-diameter impact crater. Magma upwelling after the crater impact is proposed to have stretched the crust and formed the fracture pattern (NASA/JHU APL/Carnegie Institution of Washington)

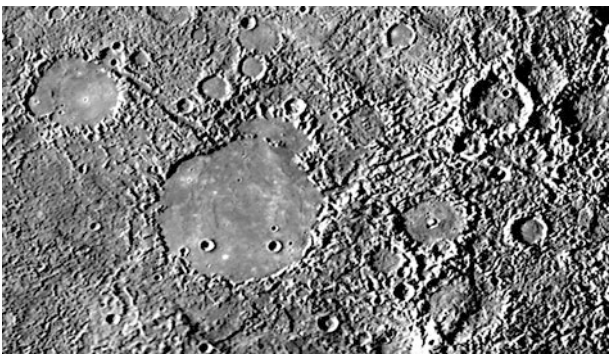


Fig. 3-73
Hilly and Lineated Terrain, Mercury. Antipodal to the center of the Caloris Basin is a region of disrupted terrain called the Hilly and Lineated Terrain. Seismic waves created by the Caloris impact and focused on the antipodal point by the liquid iron core provided the energy to cause vertical displacement of the surface (NASA/JHU APL/Carnegie Institution of Washington)

The Hilly and Lineated Terrain is an unusual area antipodal to the Caloris Basin. It consists of severely disrupted terrain and covers an area of at least 360,000 km^2. The hills, depressions, and valleys constituting the Hilly and Lineated Terrain disrupt preexisting landforms such as impact craters (*Fig. 3-73*). These unusual features resulted from the impact forming the Caloris Basin. Seismic waves generated by the Caloris impact passed through the interior, including the core which was probably entirely molten at the time. Passage through the liquid core focused the seismic waves on the point antipodal to the center of Caloris. The convergence of the seismic energy on this point resulted in very strong shaking and vertical displacement of the surface, creating the highly disrupted terrain.

Most of the tectonic features seen on Mercury are compressional features, including wrinkle ridges within lava plains. Mariner 10 and MESSENGER imagery reveal a global distribution of linear to arcuate lobate scarps (*Fig. 3-74*), which are geologic expressions of surface-breaking thrust faults (Watters et al. 2009). Lobate scarps appear to be Calorian or younger in age, showing little embayment by lava flows in smooth plains units that they cross. They often transect preexisting impact craters, resulting in apparent rim offsets which can be used to determine the angle and height of the thrust fault. Distribution of the lobate scarps indicates global contraction, and the amount of crustal displacement produced by the scarps suggests that Mercury's radius has decreased by 1–2 km since the planet's formation. The cause of this planetary shrinking is the gradual solidification of Mercury's large iron core. Iron undergoes a decrease in volume as it transitions from liquid to solid. Thus solidification results in a decrease in the core's size, which in turn causes the entire planet to shrink. Complete solidification of Mercury's core would reduce the planet's radius by 17 km, supporting magnetic field results that the core is still partially molten.

Mariner 10 observations suggested that volcanism on Mercury had ceased shortly after the LHB as planetary contraction shut off magma access to the surface. Crater analysis from MESSENGER imagery indicate that volcanism has continued into more recent times, perhaps as recently as the last 1 Ga in localized areas (Strom et al. 2008). The post-Mariner 10 view was that Mercury was a geologically dead world which currently only experienced the occasional

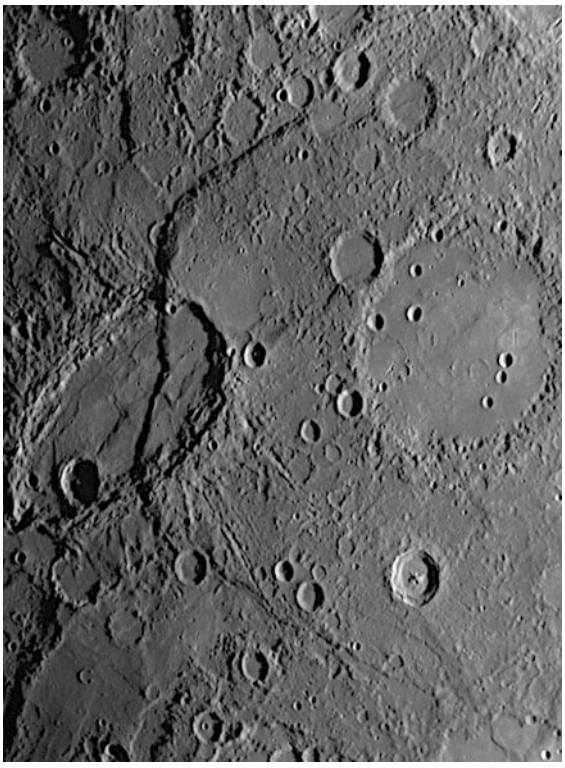


Fig. 3-74
Beagle Rupes, Mercury. The lobate scarp Beagle Rupes is more than 600 km long and is the surface expression of a thrust fault formed as Mercury contracts in response to the solidification of its large iron core (NASA/JHU APL/Carnegie Institution of Washington)

formation of impact craters and perhaps compressional tectonics as the planet continued to contract. Initial results from the MESSENGER flybys suggest that this paradigm requires some revision. A better understanding of Mercury's geologic history and internal structure will be obtained now that MESSENGER has achieved orbit and has begun its full investigation of the innermost planet.

6 Moon

Earth's Moon is not officially a terrestrial planet since it orbits Earth rather than the Sun, but its rocky composition, interior structure, and geologic history make it similar enough to the terrestrial planets that it is usually included in discussions of the inner solar system. The Moon is Earth's closest celestial neighbor and was therefore the first objective of our planet's space exploration program. It is also the only body other than Earth which has been investigated by human explorers. Earth-based telescopes revealed that the Moon's two terrain units, the bright highlands and the dark maria, were different in terms of crater density (and thus age), roughness,

Fig. 3-75
Moon. The brightness differences between the brighter highlands and dark maria are evident in this view of the lunar nearside obtained by Clementine (NASA/JPL/US Geological Survey)

and topography (*Fig. 3-75*). The heavily cratered highlands are higher, older, and rougher than the sparsely cratered maria. However, Earth-based telescopic observations only provide information about ~50% of the lunar surface. Gravitational interactions (tidal forces) between Earth and the Moon have slowed the Moon's rotation so it is synchronous with its orbital period of 27.3 Earth days. This synchronous rotation (or 1:1 spin-orbit coupling) means that one side of the Moon, the nearside, always faces Earth. The side facing away from Earth is the lunar farside and can only be explored by spacecraft.

Early lunar landers, including the Soviet Luna 2, 9, and 13 missions and the US Ranger and Surveyor series, proved that impact cratering had not produced an exceptionally thick regolith which would swallow up landed spacecraft and their human occupants. Orbiting missions, including the Soviet Luna 10, 11, 12, 14, 19, and 22 missions and the US Lunar Orbiters, provided detailed views of the surface geology, including that of the farside which is never visible from Earth. These missions paved the way for human visitation of the Moon in the Apollo missions, including surface exploration and sample return by the Apollo 11, 12, 14, 15, 16, and 17 crews (*Fig. 3-76*). Although the Soviet Union never landed cosmonauts on the lunar surface, they did achieve successful teleoperation of surface rovers (Lunakhod 1 and 2) and robotic sample return (Lunas 16, 20, and 24). This first phase of lunar exploration lasted from 1959 to 1976 and significantly expanded our understanding of the Moon's interior structure, surface geology, topography, and geologic history (Heiken et al. 1991).

A second phase of lunar exploration occurred between 1994 and 1999 with the US Clementine and Lunar Prospector missions. Clementine orbited the Moon between February 19 and May 3, 1994, providing detailed maps of crustal composition and topography. Lunar Prospector

Fig. 3-76
Lunar Surface. Apollo 17 astronaut Harrison Schmidt is seen standing next to large boulder in the Taurus-Littrow valley of the Moon (NASA)

orbited the Moon between January 11, 1998 and July 31, 1999. That mission produced global maps of elemental variations within the crust, investigated evidence of ice deposits near the lunar poles, provided low-altitude mapping of the lunar gravity field, and identified regions of crustal remnant magnetization.

The Moon is currently undergoing a third period of intense exploration with a much more international flavor than the previous exploration periods. ESA's first Small Missions for Advanced Research and Technology (SMART-1) orbiter provided images and compositional information between 2004 and 2006. Japan's Selenological and Engineering Explorer (SELENE; also called Kaguya) orbited the Moon in 2007–2009 and included instruments to provide surface images, topography data, composition maps, crustal remanent magnetization data, gravity field detail, and information about the space environment. China's Chang'e 1 orbiter (2007–2009) imaged the Moon's surface in preparation for future landers, provided topography data, produced composition maps, and investigated the interaction of the solar wind with the Earth–Moon environment. India also participated in this international investigation of the Moon with the Chandrayaan-1 orbiter (2008–2009), which carried a camera, spectrometers, a laser altimeter, and a radar system to investigate the lunar surface. The most recent addition to this armada of lunar missions is the US Lunar Reconnaissance Orbiter (LRO) and Lunar Crater Observation and Sensing Satellite (LCROSS), which reached lunar orbit on June 23, 2009. Like the other lunar missions, LRO carries a camera, laser altimeter, radar, and a variety of spectrometers operating in different wavelengths to thoroughly explore the lunar surface. LCROSS was deliberately crashed into Cabeus crater near the lunar south pole on October 9, 2009, to investigate the possibility of ice deposits within this crater.

The Moon is the only solar system body other than Earth for which scientists have acquired a full complement of geophysical measurements. The numerous orbiting missions have produced detailed gravity maps up to degree and order 165 for the nearside and degree and order 90 for the farside. Such maps reveal that nearside impact basins are typically associated with gravity highs, indicating the presence of excess mass (mascons) under these basins. However, farside basins usually exhibit negative gravity anomaly rings (indicating mass loss), although some of these features also exhibit a central gravity high due to mantle uplift under the basins (Namiki et al. 2009). Gravity measurements give a lunar C/MR^2 value of 0.39, indicating that the Moon has only a very small core. These gravity results are supported by seismic data recorded by passive seismometers deployed by the Apollo missions, which suggest that the Moon has a solid inner core of approximately 240 km radius surrounded by a liquid core extending out to 480 km radius from the Moon's center. The Moon exhibits little seismic activity, with most moonquakes being initiated either by tidal forces when the Moon's elliptical orbit brings it closest to Earth or by small impact events. Surface heat flow was measured at the Apollo 15 and 17 landing sites and was found to vary from 22 to 31 mW m^{-2} (Heiken et al. 1991). These values are much less than Earth's average heat flow (75 mW m^{-2}), supporting geologic analyses which indicate that endogenic geologic processes are not actively occurring at the present time on the Moon.

Magnetometers aboard orbiting lunar spacecraft have found no indication of an actively produced magnetic field, but some rocks on the lunar surface retain a remnant magnetization suggesting a lunar magnetic dynamo operated in the past. About 5% of the maria retain a significant remnant magnetization whereas that of the highlands is more heterogenous and diverse. Localized strong magnetic anomalies have been detected (*Fig. 3-77*) and are proposed to result from cometary impacts, solar magnetic storms, or by ejecta interactions causing magnetization of regions antipodal to major impacts.

Global topography maps have now been produced for the Moon from laser altimeter measurements on many of the orbiting spacecraft (*Fig. 3-78*). These maps reveal that topography varies by 16 km across the Moon (Zuber et al. 1994). Maria are located in topographic depressions, typically ranging in depth from 2 to 4 km below the reference radius. The lowest region on the Moon is the 2,500-km-diameter South Pole-Aitken Basin, located in the southern hemisphere of the farside and reaching a maximum depth of 8.2 km below the reference radius. The highlands are topographically higher, ranging up to 6 km or more above the reference radius. Farside highlands are higher than their counterparts on the nearside and have been suggested to include ejecta deposits from the South Pole-Aitken Basin. Combined topography and gravity data indicate the lunar crust is thicker on the farside (typically >60 km) than on the nearside (generally <30 km) (Wieczorek and Phillips 1998). The Moon's center of mass is displaced ~1.68 km from its center of figure in the direction of the Earth.

Lunar rock and soil samples returned by the Apollo and Luna missions revealed that albedo (brightness) differences between the highlands and maria result from variations in composition. The dark maria are volcanic lava flows composed of basalt. Mare basalts often are subdivided into three groups based on titanium concentrations: high-titanium, low-titanium, and very-low-titanium. Titanium concentrations can reach up to 15% in the high-titanium basalts, leading to speculation about the economic value of eventually mining these deposits. The bright highlands are primarily composed of ferroan anorthosite, a plagioclase-rich rock with minor mafic (iron and magnesium) components. The highlands display strong uniformity in composition (SELENE observations report plagioclase concentrations near 100% (Ohtake et al. 2009)), suggesting this portion of the crust formed from cooling of a single well-mixed

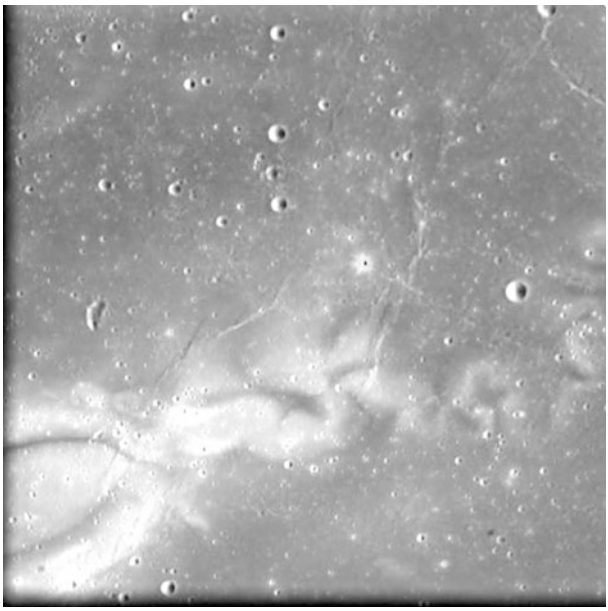


Fig. 3-77
Reiner Gamma, Moon. Reiner Gamma is an enigmatic feature displaying a strong magnetic anomaly. This SMART-1 image of Reiner Gamma shows the feature's bright swirl pattern (ESA/Space Exploration Institute)

magmatic reservoir. In addition to anorthosite, the lunar highlands also show localized outcrops of magnesium-rich rocks and KREEP rocks. KREEP rocks have high concentrations of potassium (K), the rare-earth elements (REE), and phosphorous (P).

Mineralogical information from lunar rock analysis provides insights into the Moon's formation and early history. Early models for the Moon's formation included co-accretion with Earth from the solar nebula, formation elsewhere in the solar system and subsequent capture by Earth's gravity, and fission of the Moon from an early rapidly rotating Earth. Analysis of lunar samples returned by Apollo and Luna revealed that the Moon is very similar in overall composition to Earth except for a lower percentage of iron and volatile elements. Lunar and terrestrial rocks have identical oxygen isotope ratios, indicating both bodies formed in the same region of space out of the same reservoir of material. Geochemical analysis combined with dynamical models revealed major problems with all three existing formation models, leading scientists to propose a new model of lunar formation through a giant impact (Hartmann et al. 1986). The giant impact model proposes that Earth formed alone out of the solar nebula and underwent differentiation within a few million years. Following differentiation, Earth collided with a planetismal about half its size. The collision was off-center, resulting in terrestrial crust and mantle material being ejected along with destroyed remnants of the impacting planetismal (Canup and Esposito 1996). Ejected material formed a ring of debris around Earth which eventually accreted into the Moon. This model explains the similar compositions between Earth and Moon since the Moon is largely derived from Earth's crust and mantle. Most of Earth's iron is located in the core, which was not included in the ejected material, hence the Moon ended up with a lower

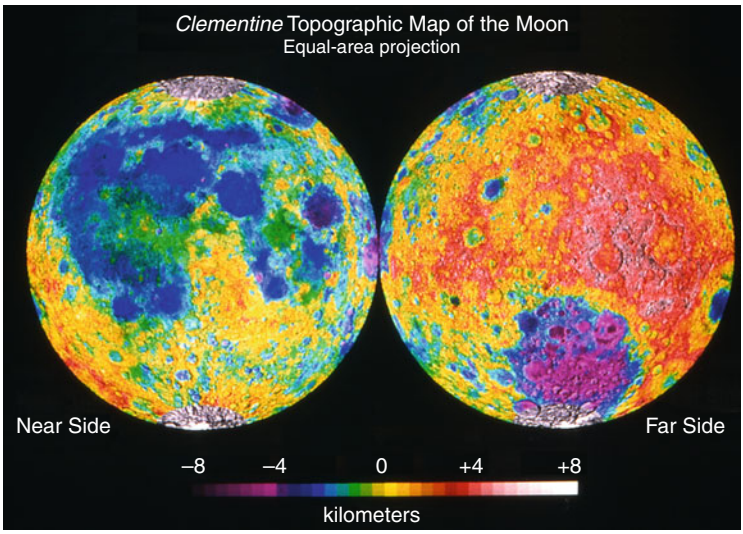


Fig. 3-78

Lunar Topography. The first complete topographic map of the Moon was provided by the Clementine mission. The farside highlands are 6 km or higher in elevation, perhaps in part due to ejecta from the South Pole-Aitken Basin seen near the south pole on the farside (NASA/LPI)

concentration of iron because only the impacting planetismal contributed iron to the material forming the Moon. The volatile-poor nature of lunar rocks is a result of impacts being energetic events which easily vaporize volatile elements. The Moon thus accreted from terrestrial-derived crust and mantle material which was depleted in iron and volatile elements.

Accretion and the decay of radioactive elements provided enough heat to melt at least the outer portion of the Moon. The compositional variation of the present-day lunar crust can be explained by this magma ocean stage. The outer ~500 km of the Moon was partially to completely molten, allowing differentiation to occur. Higher density basalt sunk to the base of the ocean while lower density plagioclase floated to the top. The plagioclase-rich reservoir was at least 100 km thick and well-mixed, whereas the basalt formed isolated pods which underwent subsequent fractional crystallization to alter the magma compositions. The uppermost part of the plagioclase reservoir solidified, forming the homogenous anorthositic lunar highlands crust. Magnesium-rich and KREEP materials formed by localized eruptions/intrusions from magma lying between the anorthositic crust and basaltic reservoirs. Large impact basins forming during the LHB excavated and fractured the crust deeply enough to allow the basaltic magmas access to the surface.

Observations by the Moon Mineralogy Mapper instrument on Chandrayaan-1 recently have revealed the presence of hydroxyl (OH) and H_2O in the lunar regolith, with highest concentrations near the poles and around some fresh impact craters (Pieters et al. 2009) (*Fig. 3-79*). The OH/H_2O is detected as an absorption line near 3-μm wavelength and is indicative of water and hydroxyl adsorbed by other minerals. The OH/H_2O likely originates through implantation of oxygen and hydrogen from the solar wind, although delivery by cometary impacts also may contribute. These findings do not contradict Apollo and Luna

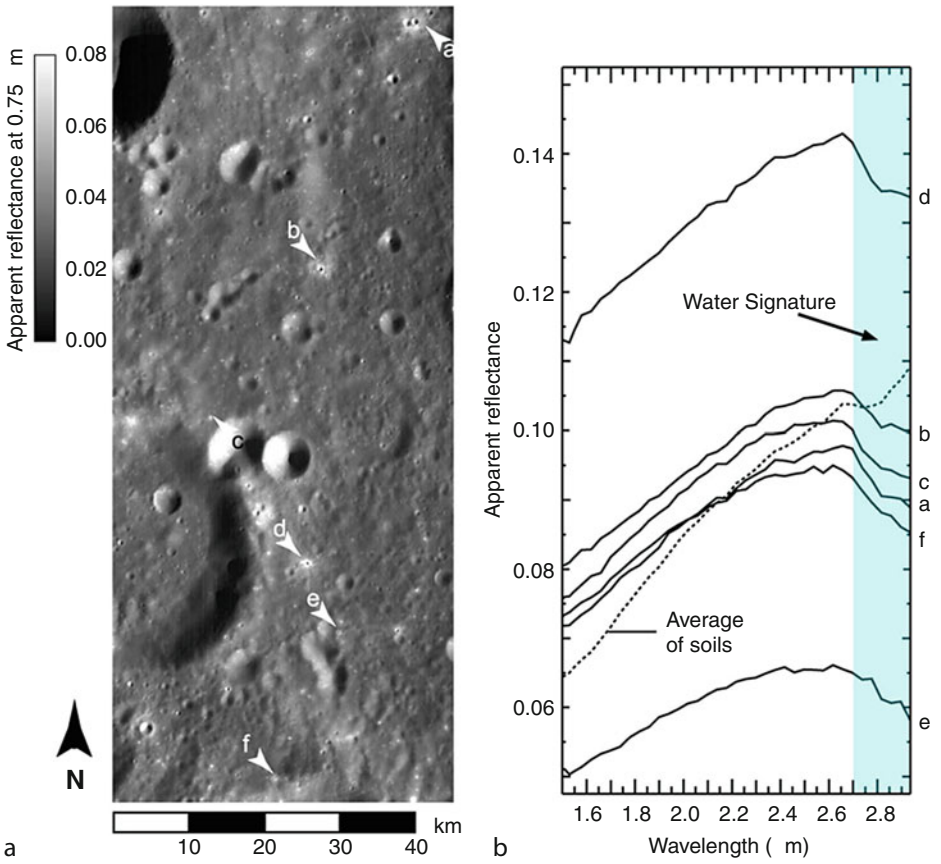


Fig. 3-79

Water Signature on the Moon. Spectral results from the Moon Mineralogy Mapper on Chandrayaan-1 detected absorption lines indicative of OH/H_2O (India Space Research Organization/NASA/JPL/Brown University)

sample results that the Moon is largely anhydrous because the OH/H_2O adsorptions are only a surficial process occurring in lunar regolith and are not representative of the bulk lunar composition.

The Moon, like Mercury, retains only a thin exosphere of particles derived from solar wind capture and surface sputtering. Lack of any aeolian, fluvial, or glacial landforms indicates that the Moon has never possessed an extensive atmosphere. However, like Mercury, radar observations suggest that ice deposits have accumulated in the floors of high-latitude impact craters whose floors are in permanent shadow. Lunar Prospector's neutron spectrometer revealed significant declines in epithermal and fast neutron fluxes near the polar regions, consistent with the presence of H_2O ice in these regions (*Fig. 3-80*). Direct evidence of H_2O deposits on floors of polar craters was provided by the LCROSS impact into Cabeus crater (*Fig. 3-81*). Spectroscopic analysis of the plume produced by the impact revealed the presence of H_2O and OH at concentrations of about ten parts per million.

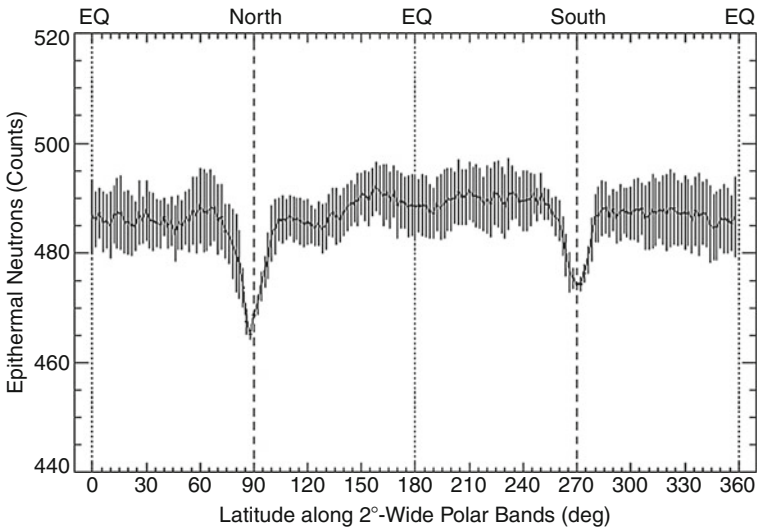


Fig. 3-80

Lunar Polar Ice. The neutron spectrometer on Lunar Prospector investigated the neutron fluxes over the lunar surface. Strong dips in the flux of epithermal neutrons over the polar regions are consistent with the presence of H_2O ice in permanently shadowed crater floors (NASA/JHU APL/Los Alamos National Lab)

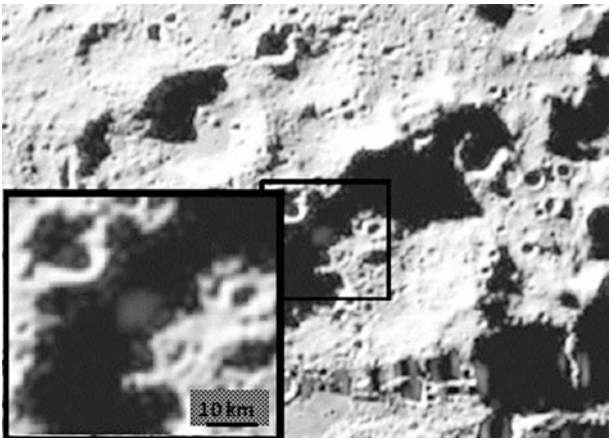


Fig. 3-81

LCROSS Plume. LCROSS impacted in two parts into Cabeus crater on October 9, 2009. This image from the shepherding satellite records the plume created from the Centaur rocket impact in Cabeus crater. Further analysis revealed the presence of OH/H_2O in the debris plume (NASA/LCROSS)

Impact craters are the dominant geologic landform on the Moon, ranging from tiny (micrometeorite) pits on lunar rocks to multi-ring basins thousands of kilometers in diameter. The volcanic versus impact origin of lunar craters was the source of considerable debate until return of lunar samples led to widespread acceptance of the impact origin of these features.

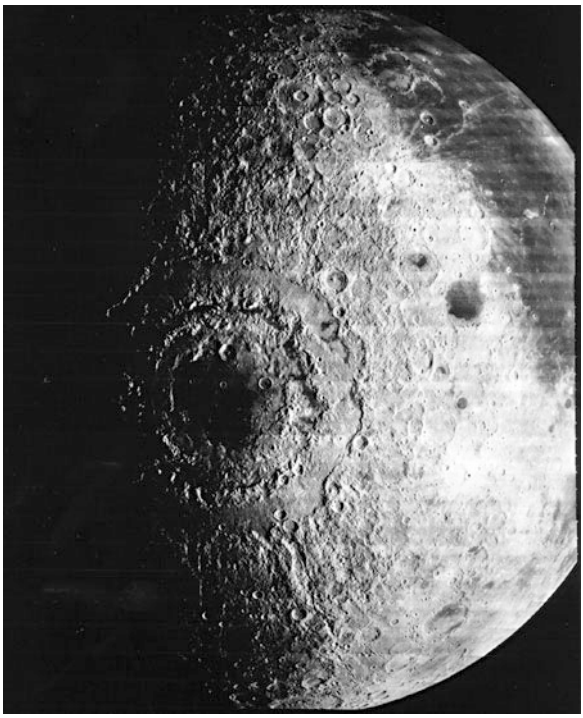


Fig. 3-82
Orientale Basin, Moon. Lunar Orbiter 4 obtained this image of the Orientale Basin, located along the nearside-farside border. Orientale is an excellent example of a multi-ring impact basin, formed when the impact excavates through most of the body's lithosphere (NASA)

The transition from simple bowl-shaped craters to complex craters occurs near a diameter of 17 km, but larger craters display peak-ring and multi-ring morphologies. Multi-ring basins (*Fig. 3-82*) occur when an impact excavates through most or all of the lithosphere, setting up localized convection in the asthenosphere which fractures the lithosphere to produce the ring structure (Melosh and McKinnon 1978). The number of rings decreases as lithospheric strength increases, thus providing insights into the lithospheric thickness at the time of impact. Most of the large lunar basins are multi-ring structures, although few are obvious on the nearside because of postimpact intrusion by mare lava flows.

Crater density analysis indicates that the maria are younger than the highlands. The ability to radiometrically date lunar samples combined with knowledge of their original location on the lunar surface has allowed scientists to relate crater density to surface unit formation age. The resulting Lunar Crater Chronology allows determination of a lunar surface unit's age based entirely on its crater density. The combination of radiometric ages for actual samples and crater density-derived ages suggest that the lunar highlands date to ~4.0–4.2 Ga and the maria mainly formed between 3.0 and 3.5 Ga, although localized lava flows may have continued up to 2.0 Ga. The Lunar Crater Chronology has been extrapolated to other solar system bodies to estimate absolute surface ages on those worlds and is the basis for non-terrestrial surface ages cited throughout this chapter.

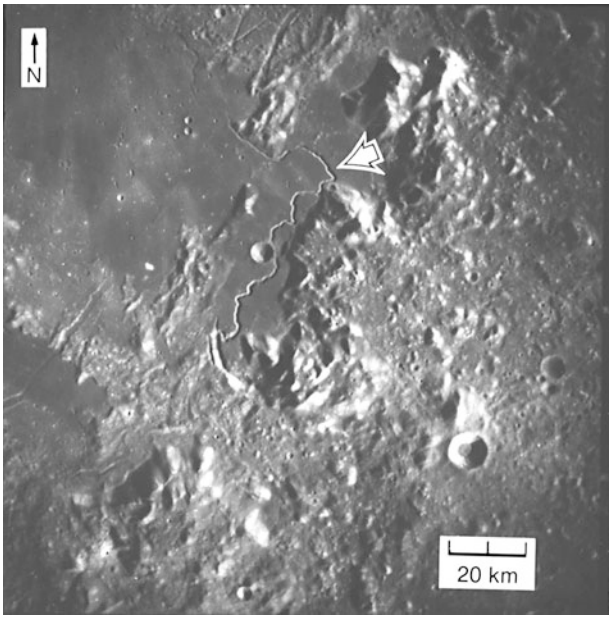


Fig. 3-83
Hadley Rille, Moon. Sinuous rilles are common in the lunar mare and were formed by low-viscosity lava flows. This image shows the Apollo 15 landing site adjacent to Hadley Rille. The nearby mountains are the Apennine Mountains which form part of the Imbrium Basin's rim (NASA/LPI)

Lunar volcanism was concentrated in the maria and primarily consisted of low-viscosity magmas with high eruption rates. These characteristics produced flood basalts that comprise the maria regions. Small, low-sloped domes are the only volcanic edifices identified on the Moon, but lava channels (sinuous rilles; *Fig. 3-83*) and collapse pits suggesting underground lava tunnels are common. Superposition relationships indicate that the mare deposits were emplaced through a series of eruptions spaced over the ~1 Ga of time during which volcanism was occurring. Individual flows can be less than 50 m thick, but the superposition of these flows results in mare deposits having total thickness of 100s to 1,000s of meters.

One surprising result from first images of the lunar farside was that the maria are not evenly distributed across the Moon's surface (*Fig. 3-84*). Maria constitute about 30% of the nearside surface area but only about 2% of the farside area. The approximately circular shape and lower topography of the maria revealed that they are lava flows concentrated in the floors of large impact basins. The concentration of maria on the lunar nearside is not because of an asymmetric distribution of large impact basins (the distribution of basins is statistically identical on the nearside versus farside), but rather can be attributed to the thinner nearside crust. Large projectiles excavate to greater depths than smaller projectiles and the large basins would excavate to depths comparable to the thickness of the nearside crust. This would allow basaltic magma retained in reservoirs near the base of the original anorthositic lunar crust to ascend to the floor of the basin via impact-induced fractures, producing the mare flood basalts. The farside crust is thick enough that basin excavation is insufficient to reach the deeper magma reservoirs, reducing the amount of mare fill in farside basins.

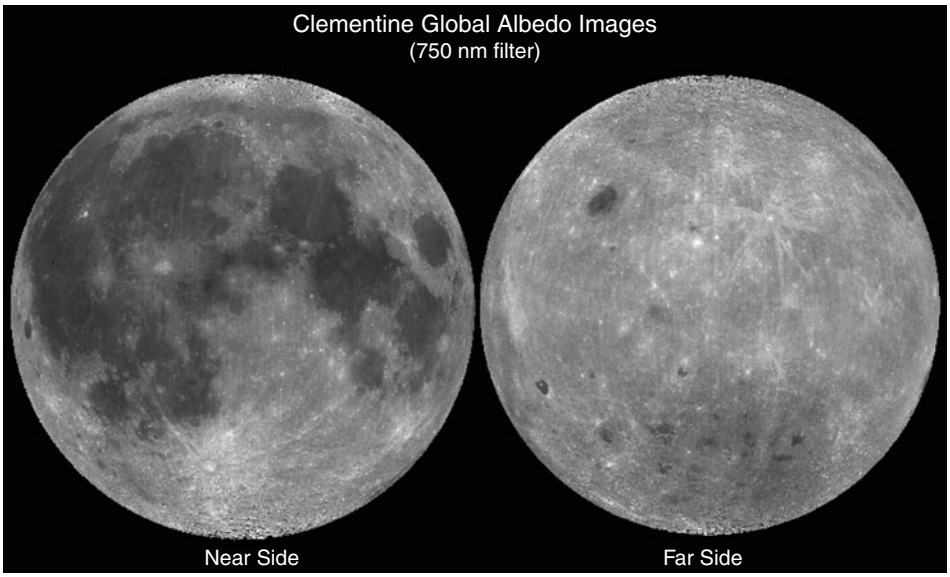


Fig. 3-84
Lunar Albedo. Albedo maps derived from Clementine data reveal the difference in dark maria distribution between the lunar nearside and farside. Approximately 30% of the nearside is covered by maria but only the dark lava flows cover only ~2% of the farside (NASA)

Although low-viscosity volcanism dominates on the Moon, localized regions of more explosive activity have been identified. Apollo 17 astronauts returned samples of orange soil from dark-haloed Shorty crater. Orange and black glass in this sample was formed by a gas-rich eruption at ~3.6–3.7 Ga. This deposit was exposed much later by the impact which created Shorty crater. Multispectral mapping of the lunar surface by Clementine revealed additional sites covered by pyroclastic glasses, such as the Aristarchus region (McEwen et al. 1994). However, explosive volcanism is very limited in extent across the Moon and can be explained by eruption of gas-rich lavas from localized magma reservoirs.

The Moon primarily exhibits localized tectonic landforms, although recent Lunar Reconnaissance Orbiter images suggest it may have undergone global compression of ~0.9 km. Most tectonic features are associated with the maria, which often display extensional graben along the edges and compressional wrinkle ridges toward their centers (*Fig. 3-85*). Both of these tectonic features can be attributed to infilling of impact basins by mare flood basalts. The impact basins formed millions of years before the volcanic eruptions producing the maria, giving the basins time to isostatically adjust to a stable state. Addition of high-density basaltic lava flows caused the basin floor to sag under the increased weight, causing extensional stresses along the edges as these regions pulled away from the basin rim and compressional stresses near the basin center where downward-directed forces merged to produce thrust faults. These stresses and the production of their resulting landforms continued until the mare-filled basins reached a new isostatically stable configuration.

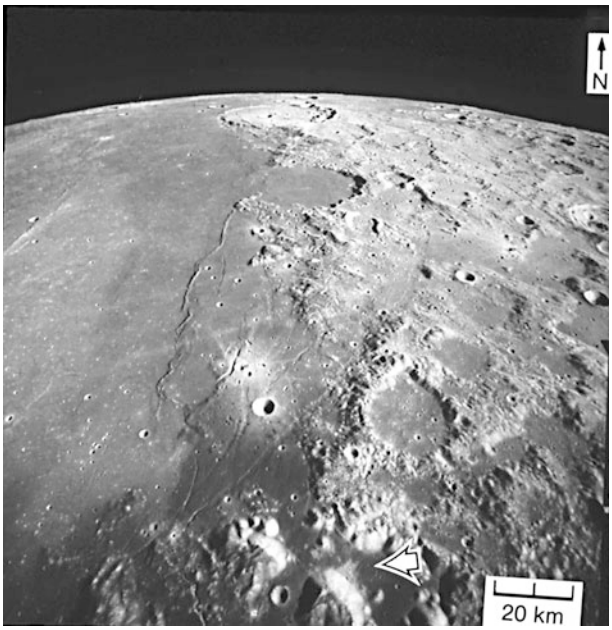


Fig. 3-85
Lunar Tectonics. Extensional graben and compressional wrinkle ridges are seen along the eastern edge of Mare Serenitatis. *Arrow* points to the Apollo 17 landing site in the Taurus-Littrow valley (NASA/LPI)

7 Summary

Surfaces and interiors of the four terrestrial planets (Mercury, Venus, Earth, and Mars) and Earth's Moon display a number of similarities which indicate common evolutionary paths but also a number of differences which make each body unique. All five bodies accreted from smaller debris and underwent differentiation to produce a core, mantle, and crust. All have likely had dynamos producing magnetic fields at some time in their histories, but only Earth and Mercury have active dynamos today. Geologic histories of each body have been shaped by impact cratering, volcanism, tectonic activity, and mass wasting processes. However, aeolian processes have only affected the three bodies with a substantial atmosphere (Venus, Earth, and Mars) and fluvial/glacial activity has been restricted to Earth and Mars. Large impacts early in their histories can explain the large size of Mercury's core, the slow retrograde rotation of Venus, the presence of Earth's Moon, and the hemispheric dichotomy on Mars. Earth and Venus are large enough that they retain enough internal heat to drive active volcanism and tectonism at the present time, but the two bodies lose heat by very different processes. Earth's volcanic and tectonic activity is dominated by horizontal movement of lithospheric plates whereas Venus' volcanism and tectonism results from vertical motion produced by mantle upwelling and hot spot volcanism. Mars is about half the size of Earth and Venus, and the level of its volcanic activity has declined through time, becoming concentrated in the Tharsis and Elysium regions at the present time. Martian landforms have been heavily affected by fluvial and glacial activity both

in the distant past when the planet had a thicker atmosphere and in more recent times as obliquity variations produce dramatic examples of climate change. Volcanic activity has declined on Mercury since its formation and may no longer occur, but the planet displays a unique tectonic regime of global contraction as its large iron core gradually solidifies and shrinks. The Moon is a geologically dead world, where volcanism ceased about 2 Ga ago, but the recent discovery of adsorbed hydrogen and oxygen combined with confirmation of H_2O ice in polar craters have reignited interest in our closest celestial neighbor and the resources it may contain for future human exploration and settlement. Continued exploration of the surfaces and interiors of the worlds constituting the inner solar system will not only provide a better understanding of our own Earth but also will give humans the knowledge to make us better caretakers of our home planet.

References

Acuña, M. H., Connerney, J. E. P., Ness, N. F., et al. 1999, Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. Science, 284, 790–793

Alvarez, L. W., Alvarez, W., Asaro, F., & Michel, H. V. 1980, Extraterrestrial cause for the Cretaceous-Tertiary extinction. Science, 208, 1095–1108

Andrews-Hanna, J. C., Zuber, M. T., & Banerdt, W. B. 2008, The Borealis basin and the origin of the Martian crustal dichotomy. Nature, 453, 1212–1215

Asimow, P. D., & Wood, J. A. 1992, Fluid outflows from Venus impact craters: analysis from Magellan data. J Geophys Res, 97, 13643–13665

Baker, V. R. 1982, The channels of Mars (Austin: University of Texas Press), 198

Baker, V. R., & Nummedal, D. 1978, The channeled scabland: a guide to the geomorphology of the Columbia basin, Washington (Washington, DC: NASA Planetary Geology Program), 186

Baker, V. R., Komatsu, G., Gulick, V. C., & Parker, T. J. 1997, Channels and valleys, in Venus II, eds. S. W. Bougher, D. M. Hunten, & R. J. Phillips (Tucson: University of Arizona Press), 757–793

Barlow, N. G. 2005, A review of Martian impact crater ejecta structures and their implications for target properties, in Large meteorite impacts III, eds. T. Kenkmann, F. Hörz, & A. Deutsch, Geological Society of America Special Paper 384, Boulder, CO (Washington, DC: The Society), 433–442

Barlow, N. G. 2008, Mars: an introduction to its interior, surface, and atmosphere (Cambridge: Cambridge University Press), 264

Barsukov, V. L. 1992, Venusian igneous rocks, in Venus geology, geochemistry, and geophysics, eds. V. L. Barsukov, A. T. Basilevsky, V. P. Volkov, & V. N. Zharkov (Tucson: University of Arizona Press), 165–176

Berman, D. C., & Hartmann, W. K. 2002, Recent fluvial, volcanic, and tectonic activity on the Cerberus Plains of Mars. Icarus, 159, 1–17

Bibring, J.-P., Langevin, Y., Mustard, J. F., et al. 2006, Global mineraological and aqueous Mars history derived from OMEGA/Mars Express data. Science, 312, 400–404

Bogard, D. D., & Johnson, P. 1983, Martian gases in an Antarctic meteorite? Science, 221, 651–654

Bohor, B. F., Modreski, P. J., & Foord, E. E. 1987, Shocked quartz in the Cretaceous-Tertiary boundary clays: evidence for a global distribution. Science, 236, 703–709

Bottke, W. F., Levison, H. F., Nesvorný, D., & Dones, L. 2007, Can planetismals left over from terrestrial planet formation produce the lunar Late Heavy Bombardment? Icarus, 190, 203–223

Bougher, S. W., Hunten, D. M., & Phillips, R. J. 1997, Venus II (Tucson: University of Arizona Press), 1362

Breuer, D., & Spohn, T. 2003, Early plate tectonics versus single-plate tectonics on Mars: evidence from manetic field history and crust evolution. J Geophys Res, 108, E75072, doi: 10.1029/2002JE001999

Breuer, D., Spohn, T., & Wüllner, U. 1993, Mantle differentiation and the crustal dichotomy of Mars: evidence from magnetic field history and crust evolution. J Geophys Res, 108, E75072, doi: 10.1029/2002JE001999

Bullock, M. A., & Grinspoon, D. H. 2001, The recent evolution of climate on Venus. Icarus, 150, 19–37

Byrne, S., Dundas, C. M., Kennedy, M. R., et al. 2009, Distribution of mid-latitude ground ice on Mars from new impact craters. Science, 325, 1674–1676

Canup, R. M., & Agnor, C. B. 2000, Accretion of the terrestrial planets and the Earth-Moon system, in Origin of the Earth and Moon, eds. R. M. Canup & K. Righter (Tucson: University of Arizona Press), 113–129

Canup, R. M., & Esposito, L. W. 1996, Accretion of the Moon from an impact-generated disk. Icarus, 119, 427–446

Carter, L. M., Campbell, B. A., Watters, T. R., et al. 2009, Shallow radar (SHARAD) sounding observations of the Medusae Fossae Formation, Mars. Icarus, 199, 295–302

Chao, E. C. T., Shoemaker, E. M., & Madsen, B. M. 1960, First natural occurrence of coesite. Science, 132, 220–222

Christensen, P. R. 2003, Formation of recent Martian gullies through melting of extensive water-rich snow deposits. Nature, 422, 45–47

Cox, A. 1973, Plate tectonics and geomagnetic reversals (San Francisco: W.H. Freeman), 702

Craddock, R .A., & Howard, A. D. 2002, The case for rainfall on a warm, wet early Mars. J Geophys Res, 107, 5111, doi: 10.1029/2001JE001505

Crumpler, L. S., Aubele, J. C., Senske, D. A., et al. 1997, Volcanoes and centers of volcanism on Venus, in Venus II, eds. S. W. Bougher, D. M. Hunten, & R. J. Phillips (Tucson: University of Arizona Press), 697–756

Denevi, B. W., Robinson, M. S., Solomon, S. C., et al. 2009, The evolution of Mercury's crust: a global perspective from MESSENGER. Science, 324, 613–618

Drake, M. J., & Righter, K. 2002, Determining the composition of the Earth. Nature, 416, 39–44

Feldman, W. C., Prettyman, T. H., Maurice, S., et al. 2004, Global distribution of near-surface hydrogen on Mars. J Geophys Res, 109, E09006, doi: 10.1029/2003JE002160

French, B. M. 1998, Traces of catastrophe: a handbook of shock-metamorphic effects in terrestrial meteorite impact structures (Houston, TX: Lunar and Planetary Institute, Contribution No. 954), 120

Frey, H., & Schultz, R. A. 1988, Large impact basins and the mega-impact origin for the crustal dichotomy on Mars. Geophys Res Lett, 15, 229–232

Frey, H. V., Roark, J. H., Shockey, K. M., Frey, E. L., & Sakimoto, S. E. H. 2002, Ancient lowlands on Mars. Geophys Res Lett, 29, 1384, doi: 10.1029/2001GL013832

Gomes, R., Levison, H. F., Tsiganis, K., & Morbidelli, A. 2005, Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Nature, 435, 466–469

Greeley, R., Bender, K. C., Saunders, R. S., Schubert, G., & Weitz, C. M. 1997, Aeolian processes and features on Venus, in Venus II, eds. S. W. Bougher, D. M. Hunten, & R. J. Phillips (Tucson: University of Arizona Press), 547–589

Grimm, R. E., & Hess, P. C. 1997, The crust of Venus, in Venus II, eds. S. W. Bougher, D. M. Hunten, & R. J. Phillips (Tucson: University of Arizona Press), 1205–1244

Guest, J. E., Bulmer, M. H., Aubele, J., et al. 1992, Small volcanic edifices and volcanism in the plains of Venus. J Geophys Res, 97, 15949–15966

Hansen, V. L., Willis, J. J., & Banerdt, W. B. 1997, Tectonic overview and synthesis, in Venus II, eds. S. W. Bougher, D. M. Hunten, & R. J. Phillips (Tucson: University of Arizona Press), 797–844

Hartmann, W. K., Phillips, R. J., & Taylor, G. J. 1986, Origin of the Moon (Houston: Lunar and Planetary Institute), 797

Hays, J. D., Imbrie, J., & Shackleton, N. J., 1976, Variations in the Earth's orbit: pacemaker of the ice ages. Science, 194, 1121–1132

Head, J. W., Crumpler, L. S., Aubele, J. C., et al. 1992, Venus volcanism: classification of volcanic features and structures, associations and global distribution from Magellan data. J Geophys Res, 97, 13153–13197

Head, J. W., Murchie, S. L., Prockter, L. M., et al. 2008, Volcanism on Mercury: evidence from the first MESSENGER flyby. Science, 321, 69–72

Heiken, G. H., Vaniman, D. T., & French, B. M. 1991, Lunar sourcebook: a user's guide to the Moon (Cambridge: Cambridge University Press), 736

Hildebrand, A. R., Penfield, G. T., Kring, D. A., Pilkington, M., Camargo-Zanoguera, A., Jacobsen, S. B., & Boynton, W. V. 1991, Chicxulub crater: a possible Cretaceous/Tertiary boundary impact crater on the Yucatan Peninsula, Mexico. Geology, 19, 867–871

Hodges, C. A., & Moore, H. J. 1994, Atlas of volcanic landforms on Mars, Professional Paper No. 1534 (Washington, DC: US Geological Survey), 194

Hoyt, W. G. 1987, Coon mountain controversies: meteor crater and the development of the impact theory (Tucson: University of Arizona Press), 442

Hubbard, W. B. 1984, Planetary interiors. (New York: Van Nostrand Reinhold Press), 334

Hunten, D. M., Colin, L., Donahue, T. M., & Moroz, V. I. 1983, Venus (Tucson: University of Arizona Press), 1143

Imbrie, J., & Imbrie, J. Z. 1980, Modeling the climatic response to orbital variations. Science, 207, 943–953

Jakosky, B. M., & Phillips, R. J. 2001, Mars' volatile and climate history. Nature, 412, 237–244

Kargel, J. S., Fegley, B., Treiman, A., & Kirk, R. L. 1994, Carbonatite-sulfate volcanism on Venus. Icarus, 112, 219–252

Karunatillake, S., Wray, J. J., Squyres, S. W., et al. 2009, Chemically striking regions on Mars and Stealth revisited. J Geophys Res, 114, E12001, doi: 10.1029/2008JE003303

Kieffer, H. H., Jakosky, B. M., Snyder, C. W., & Matthews, M. S. 1992, Mars (Tucson: University of Arizona Press), 1498

Kring, D. A., & Cohen, B. A. 2002, Cataclysmic bombardment throughout the inner solar system 3.9–4.0 Ga. J Geophys Res, 107, 5009, doi: 10.1029/2001JE001529

Laskar, J., Correia, A. C. M., Gastineau, M., et al. 2004, Long term evolution and chaotic diffusion of the insolation quantities of Mars. Icarus, 170, 343–364

Maher, K. A., & Stevenson, D. J. 1988, Impact frustration of the origin of life. Nature, 331, 612–614

Malin, M. C., & Edgett, K. S. 2000, Evidence for recent groundwater seepage and surface runoff on Mars. Science, 288, 2330–2335

Malin, M. C., Edgett, K. S., Poslolova, L. V., McColley, S. M., & Noe Dobrea, E. Z. 2006, Present-day impact cratering rate and contemporary gully activity on Mars. Science, 314, 1573–1577

Mandt, K. E., DeSilva, S., Zimbelman, J. R., & Crown, D. A. 2008, Origin of the Medusae Fossae Formation, Mars: insights from a synoptic approach. J Geophys Res E, 113, E12011, doi: 10.1029/2008JE003076

McElhinny, M. W. 1979, Palaeomagnetism and plate tectonics (Cambridge: Cambridge University Press), 385

McEwen, A. S., Robinson, M. S., Eliason, E. M., et al. 1994, Clementine observations of the Aristarchus region of the Moon. Science, 266, 1858–1862

McGovern, P. J., Smith, J. R., Morgan, J. K., & Bulmer, M. H. 2004, Olympus Mons aureole deposits: new evidence for a flank failure origin. J Geophys Res, 109, E08008, doi: 10.1029/2004JE002258

McKay, D. S., Gibson, E. K., Thomas-Keprta, K. L., et al. 1996, Search for past life on Mars: Possible relic biogenic activity in Martian meteorite ALH84001. Science, 273, 924–930

Melosh, H. J. 1989, Impact cratering: a geological process. (New York: Oxford University Press), 245

Melosh, H. J., & McKinnon, W. B. 1978, The mechanics of ringed basin formation. Geophys Res Lett, 5, 985–988

Morbidelli, A., Chambers, J., Lunine, J. L., et al. 2000, Source regions and time scales for the delivery of water to Earth. Meteoritics, 35, 1309–1320

Muller, R. A., & MacDonald, G. J. 1997, Glacial cycles and astronomical forcing. Science, 277, 215–218

Namiki, N., Iwata, T., Matsumoto, K., et al. 2009, Farside gravity field of the Moon from four-way Doppler measurements of SELENE (Kaguya). Science, 323, 900–905

Nyquist, L. E., Bogard, D. D., Shih, C.-Y., et al. 2001, Ages and geologic histories of Martian meteorites. Space Sci Rev, 96, 105–164

Ohtake, M., Matsunaga, T., Haruyama, J., et al. 2009, The global distribution of pure anorthosite on the Moon. Nature, 461, 236–240

Pavri, B., Head, J. W., Klose, K. B., Wilson, L. 1992, Steep-sided domes on Venus: characteristics, geologic setting, and eruption conditions from Magellan data. J Geophys Res, 97, 13445–13478

Phillips, R. J., Raubertas, R. F., Arvidson, R. E., et al. 1992, Impact craters and Venus resurfacing history. J Geophys Res, 97, 15923–15948

Pieters, C. M., Goswami, J. N., Clark, R. N., et al. 2009, Character and spatial distribution of OH/H_2O on the surface of the Moon seen by M^3 on Chandrayaan-1. Science, 326, 568–572

Roddy, D. J., Pepin, R. O, & Merrill, R. B. 1977, Impact and explosion cratering: planetary and terrestrial implications (New York: Pergamon Press), 1301

Schaber, G. G., Strom, R. G., Moore, H. J., et al. 1992, Geology and distribution of impact craters on Venus: What are they telling us? J Geophys Res, 97, 13257–13301

Schultz, P. H. 1992, Atmospheric effects on ejecta emplacement and crater formation on Venus from Magellan. J Geophys Res, 97, 16183–16248

Sharpton, V. L., Dalrymple, G. B., Marín, L. E., et al. 1992, New links between the Chicxulub impact structure and the Cretaceous/Tertiary boundary. Nature, 359, 819–821

Shoemaker, E. M. 1963, Impact mechanics at Meteor Crater, Arizona, in The Moon, Meteorites, and Comets, eds. B. M. Middlehurst & G. P. Kuiper (Chicago, IL: University of Chicago Press), 301–336

Solomon, S. C., Smrekar, S. E., Bindschadler, D. L., et al. 1992, Venus tectonics: an overview of Magellan observations. J Geophys Res, 97, 13199–13255

Solomon, S. C., McNutt, R. L., Watters, T. R., et al. 2008, Return to Mercury: a global perspective on MESSENGER's first Mercury flyby. Science, 321, 59–62

Stofan, E. R., Hamilton, V. E., Janes, D. M., & Smrekar, S. E. 1997, Coronae on Venus: morphology and origin, in Venus II, ed. S. W. Bougher,

D. M. Hunten, & R. J. Phillips (Tucson: University of Arizona Press), 931–965

Stöffler, D., & Ryder, G. 2001, Stratigraphy and isotope ages of lunar geologic units: chronological standard for the inner solar system. Space Sci Rev, 96, 9–54

Strom, R. G., & Sprague, A. L. 2003, Exploring Mercury: the iron planet (Chichester, UK: Springer Praxis Publishing), 216

Strom, R. G., Malhotra, R., Ito, T., Yoshida, F., & Kring, D. A. 2005, The origin of planetary impactors in the inner solar system. Science, 309, 1847–1850

Strom, R. G., Chapman, C. R., Merline, W. J., Solomon, S. C., & Head, J. W. 2008, Mercury cratering record viewed from MESSENGER's first flyby. Science, 321, 79–81

Tera, F., Papanastassiou, D. A., & Wasserburg, G. J. 1974, Isotopic evidence for a terminal lunar cataclysm. Earth Planet Sci Lett, 22, 1–21

Tsiganis, K., Gomes, R., Morbidelli, A., & Levison, H. F. 2005, Origin of the orbital architecture of the giant planets of the Solar System. Nature, 435, 459–461

Vilas, F., Chapman, C. R., & Matthews, M. S., 1988, Mercury (Tucson: University of Arizona Press), 794

Watters, T. R., Leuschen, C. J., Plaut, J. J., et al. 2006, MARSIS radar sounder evidence of buried basins in the northern lowlands on Mars. Nature, 444, 905–908

Watters, T. R., Solomon, S. C., Robinson, M. S., et al. 2009, The tectonics of Mercury: the view after MESSENGER's first flyby. Earth Planet Sci Lett, 285, 283–296

Werner, S. C., van Gasselt, S., & Neukum, G. 2003, Continual geological activity in Athabasca Valles, Mars. J Geophys Res, 108, 8081, doi: 10.1029/2002JE002020

Wieczorek, M. A., & Phillips, R. J. 1998, Potential anomalies on a sphere: applications to the thickness of the lunar crust. J Geophys Res, 103, 1715–1724

Wieczorek, M. A., & Zuber, M. T. 2004, Thickness of the Martian crust: improved constraints from geoid-to-topography ratios. J Geophys Res, 109, E01009, doi: 10.1029/2003JE002153

Wilhelms, D. E., & Squyres, S. W. 1984, The Martian hemispheric dichotomy may be due to a giant impact. Nature, 309, 138–140

Williams-Jones, G., Willaims-Jones, A. E., & Stix, J. 1998, The nature and origin of Venusian canali. J Geophys Res, 103, 8545–8556

Wise, D. U., Golombek, M. P., & McGill, G. E. 1979, Tectonic evolution of Mars. J Geophys Res, 84, 7934–7939

Zuber, M. T. 2001, The crust and mantle of Mars. Nature, 412, 220–227

Zuber, M. T., Smith, D. E., Lemoine, F. G., & Neumann, G. A. 1994, The shape and internal structure of the Moon from the Clementine mission. Science, 266, 1839–1843

Zuber, M. T., Solomon, S. C., Phillips, R. J., et al. 2000, Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. Science, 287, 1788–1793

4 Gas and Ice Giant Interiors

David J. Stevenson
Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

1 Introduction *196*

2 What Are These Planets Made Of? *197*

3 What Kinds of Materials Exist in Planets? *198*

4 What Are the Temperatures in Planets? *205*

5 How Does One Explain the Heat Flows? *207*
5.1 Radioactivity 210
5.2 Secular Cooling 211
5.3 Differentiation 211

6 The Gravity Field *214*

7 Magnetic Fields *218*

8 Detailed Models *219*

9 The Challenges *221*

References *221*

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_4, © Springer Science+Business Media Dordrecht 2013

Abstract: The interiors of the gas giants (Jupiter and Saturn) and ice giants (Uranus and Neptune) are discussed. Emphasis is on the basic physical principles rather than the details of specific models. This article covers the nature and properties of the materials that comprise these planets and the basic principles of their behavior at high temperature and pressure, leading to the predominance of metallic hydrogen in the interiors of Jupiter and Saturn but no first order phase transition between molecular and metallic states. The consequences of hydrostatic equilibrium are assessed, and the resulting dependence of radius on mass and dependence of internal temperature on pressure are described. Thermal evolution of planets is described through a complete explanation of the consequences of the virial theorem, and from this, cooling times for these planets are estimated, leading to the implication that Jupiter is primarily cooling throughout whereas Saturn depends in part on an additional energy source (helium rainout) while Uranus and Neptune are storing part of their energy of formation. The method for using gravity data to constrain models, in particular moments of inertia, is outlined but not developed in detail. A summary is provided of current interior modeling and how it relates to the observed magnetic fields. The general discussion also allows an appreciation of the nature of the many giant exoplanets being found, but these bodies are not explicitly described.

1 Introduction

Planets are diverse. Within our solar system, the four "major" planets (Jupiter, Saturn, Uranus, Neptune) have some superficial similarities: They all have hydrogen-dominated atmospheres, large radii and low densities, magnetic fields, and satellite and ring systems. It makes some sense to discuss them as a group. They are nonetheless different in important ways. My goal here is to explain the main principles that apply to these bodies as a group but also convey the ways in which they express their individuality. Uranus and Neptune still appear to be similar because our vision is still blurred (dedicated missions not having taken place yet), but previous experience suggests that we should not be surprised if important differences exist. The explosion of exoplanet discoveries tells us that these kinds of planets are very common in the universe.

Why should one care what is deep inside a planet? First, knowledge of the interior structure of a planet provides the basis for explaining many phenomena that arise within but manifest themselves externally. One can not understand why some planets have magnetic fields (and others do not) and why Jupiter emits as much heat as it does without an understanding of the insides of these bodies. Second, this knowledge provides an essential part of the understanding of external phenomena and processes that are indirectly linked to the interior. For example, one cannot claim to understand why Jupiter's atmosphere is depleted in light noble gases (helium and neon) except by appealing to internal processes. Third, it provides a unifying framework. If one wants to build a story of how the planet formed and evolved to the present state, this will depend in large part on understanding the interior. Fourth, it provides a testing ground for fundamental physics – theory and experiment – under "extreme" conditions. For example, the most common metal in the universe is probably metallic hydrogen. This material is of fundamental interest to condensed matter physicists. It also happens to dominate the mass of Jupiter. It is not yet well understood, either theoretically or experimentally. This last reason may be of less interest to planetary scientists, but it is of much interest to condensed matter physicists.

The goal in this chapter will be to identify and explain the main ideas. This is not in the form of a conventional review paper because it goes back to fundamentals and relies to some

extent on toy models. It should thus be complementary to review papers such as Guillot (2005) or Fortney and Nettelmann (2010) and be suitable as a basis for teaching. Readers seeking the details of models or the input data should seek these in the review papers; they are merely summarized here. A reader seeking an understanding of basic assumptions and their limitations or an elementary exposition of the fundamentals can find them here.

If one wants to construct a model of a planet, then the most obvious requirement is that it agrees with observations and be consistent with the cosmic environment in which it formed. It is expected that the body will be close to hydrostatic equilibrium almost everywhere. However, observables and basic physics are never sufficient to remove ambiguities, especially since we do not have reliable seismology for these bodies. The model constructed also needs to be consistent with the age of the body. There is a great deal of uncertainty about how planets form, but there are some general principles about how they cool after forming, and these constrain likely thermal structures. Planetary models thus rest on four foundations: cosmic abundances, planetary observations, material properties (at both high and low pressure), and (to a lesser extent) transport properties, especially our ideas of convection. The deep interiors of these planets have material properties unlike those of everyday experience. If the atmosphere is defined to be that part where sunlight is scattered or absorbed or where IR outgoing radiation is emitted, then this is a negligible part of the planet mass, but it plays the role as a boundary condition on the interior and so its material and transport properties matter. The text is structured to explain each of these requirements and ideas for satisfactory models.

2 What Are These Planets Made Of?

The best way to answer this question is to "ask the planet," which means deducing planetary composition from its observed properties. In the special case of Jupiter, we shall find that the planet is mostly hydrogen. This is unambiguous because hydrogen is an end-member: There is nothing less dense. However, most planets have ambiguous compositions. This is especially true of the so-called ice giants (Uranus and Neptune), but even true of the planet we know best: Earth. To make the argument that Earth's core is mostly iron, one must appeal, at least in part, to cosmochemical arguments like "What is abundant as a planet forming material?" These arguments are necessarily plausible rather than rigorous, but they still make a strong case due to the large differences in cosmic abundance among materials of similar chemistry.

Cosmic abundances of elements are determined by nuclear physics. *Hydrogen* overwhelmingly dominates because it is an elementary particle; *helium* is also abundant because it is stable and can be formed in the Big Bang era. Heavier elements (what astronomers call "metals") are formed in stars and then ejected into the interstellar medium, where the material then becomes available to form solar systems. Combinations of alpha particles, multiples of 4 mass units with proton number equal to neutron number, are very favorable at low mass; oxygen, formed by combining four alphas, is especially favorable because of its nuclear shell structure. So *oxygen* is the next most abundant element followed closely by *carbon*. *Neon* (five alphas) and *nitrogen* (not a combination of alphas) follow somewhat behind. It is more complicated as one goes to larger masses, but *magnesium*, *silicon*, and *iron* are particularly favored by nuclear physics. Iron is the "endpoint" of equilibrium nucleosynthesis in the sense that all more massive nuclei are less stable.

Table 4-1
Solar system abundances

Element	Number fraction	Mass fraction
H	0.92	0.71
He	0.08	0.27
O	7×10^{-4}	0.011
C	4×10^{-4}	0.005
Ne	1.2×10^{-4}	0.002
N	1×10^{-4}	0.0015
Mg	4×10^{-5}	0.001
Si	4×10^{-5}	0.0011
Fe	3×10^{-5}	0.0016
S	1.65×10^{-5}	0.0004
Ar	5×10^{-6}	1.5×10^{-4}
Al	4×10^{-6}	1.2×10^{-4}
Ca	3×10^{-6}	1.2×10^{-4}

Cosmic abundances can be estimated by observations of the interstellar medium and other stars. However, they are not spatially uniform because of differences in stellar activity from place to place. These abundances are continuously evolving in response to both ongoing thermonuclear synthesis and the recycling of material back into the interstellar medium as a star dies; since both of these processes occur at different rates depending on the local environment, abundances themselves evolve in both space and time. This means that planets forming around other stars could have significantly different properties, even though the fundamental classes of materials (discussed below) are surely universal.

Solar system abundances are determined from the solar photosphere and correlate well with the *relative* abundances measured in meteorites, except for the most volatile elements. Accordingly, the ratio of hydrogen to silicon is enormously different in the Sun from the value in a meteorite or in Earth, but the ratio of Ca to Al, say, is very similar between the Sun and a meteorite. This comparison is not easy to do at high precision, and determination of solar abundances from spectroscopy is not straightforward (and still debated for some important ratios, e.g., C/O, Ne/H). A useful summary of abundance and other data can be found in Lodders and Fegley (1998).

Here is a summary list, useful for understanding the most abundant elements (*Table 4-1*).

3 What Kinds of Materials Exist in Planets?

Of course, elemental abundances do not tell all one wants to know; one also needs to know the chemical form that the materials take. The abundances of elements are determined by *nuclear physics*. However, most elements are not stable in elemental form but are found as molecules or in compounds. Additionally, it is important to understand the difference between the chemistry of matter and its phase. Chemistry speaks only about the elements that make up the matter

(think of the chemical formula). Phase, on the other hand, refers to the physical properties of the matter. For example, water has the chemistry of H_2O but has three different accessible phases on the Earth's surface: vapor, liquid, and ice. In fact, water has many high-pressure ice phases, which are all unique in their crystal structure. At sufficiently high pressure, water will be a metal. The ways in which the atoms are bonded together into a solid network differ between different solid phases. The phase of matter depends on the conditions of the environment, most importantly the temperature and pressure. The properties of planet-forming materials are the subject of physical chemistry and condensed matter physics. The cosmically most abundant materials can be divided into three general groups:

1. "Gases": those that do not condense (i.e., form solids or liquids) under conditions plausibly reached in the medium from which planets form
2. "Ices": those that form volatile compounds and condense but only at low temperatures (usually beyond the asteroid belt)
3. "Rocks": those that condense at high temperatures and provide the building blocks for the terrestrial planets

It is important to understand that these are labels of convenience; nothing is ever so simple that one could so easily subdivide materials. The quotation marks are there to remind the reader that what one calls an ice is sometimes in the gas or liquid phase, etc. But the subdivision proves useful nonetheless because of the large differences in behavior and abundances among these groups. Remarkably, there is considerable correspondence between abundances and classes of materials even though they involve completely unrelated physics: Generally speaking, the gases are most abundant, the ices are next most abundant, and the "rock" is least abundant (though not by much).

The "gases" are hydrogen, helium, and (to a much lesser extent) the heavier noble gases. The gases overwhelmingly dominate the "observable," (i.e., baryonic part of the) universe, the Sun, and giant planets such as Jupiter.

The hydrogen molecule H_2 is the low-pressure thermodynamic ground state of H and it interacts with other hydrogen molecules and with helium by a van der Waals interaction, which is a very weak attractive force except when the molecules are pushed close together – this is why condensation of hydrogen requires very low temperatures. It is also why it is easy to squeeze hydrogen (as a liquid, solid, or, of course, as a gas) until one approaches densities where the distance between molecules is about the size of a molecule.

The "ices" are mostly hydrides of the next set of light elements: O, C, and N (e.g., H_2O water, CH_4 methane, and NH_3 ammonia). But they also include other combinations among themselves (e.g., N_2, CO, CO_2, HCN). Hydrides do not necessarily dominate – they do not seem to in the interstellar medium – but they are thermodynamically favored when the partial pressure of hydrogen is high and will thus form if temperature or pressure permits reactions to occur. Water is the least volatile of this set because of hydrogen bonding between water molecules, which one can think of loosely as a weak form of ionic bonding arising from the very nonuniform charge distribution around the water molecule. Ammonia also has some hydrogen bonding. Methane and molecular nitrogen rely on van der Waals bonding in the liquid and solid state. CO has a small dipole moment but also interacts mostly by van der Waals forces.

"Rocks" include both metallic materials (iron and iron-nickel alloys) and as well as what we might usually call rock (oxides and silicates). The latter both have strong ionic and covalent bonding. Metallic bonding can be thought of as a special case of ionic bonding with electrons providing a spatially distributed charge, rather than the discrete charges of an ionic material

such as NaCl. These materials are much more tightly bound than "gases" or "ices" and are therefore stiffer and less volatile.

As one considers the material classifications in order of decreasing volatility,

$$\text{“GAS”} \rightarrow \text{“ICE”} \rightarrow \text{“ROCK”}$$

low-pressure density increases, and material stiffness increases. The densities remain in this order even at very high pressure (e.g., ice is always less dense than rock even though it is more compressible). The parameter that describes stiffness is the adiabatic bulk modulus:

$$K_s = \rho \left(\frac{\partial p}{\partial \rho} \right)_s$$

where ρ is mass density, P is pressure, and S is entropy . Bulk modulus has units of pressure and is a measure of the strength of interaction among the molecules in the material – soft (weakly bound/volatile) materials compress easily, while tightly bound materials are stiff. It is also a guide as to how much density change might arise in a planet due to internal pressure.

$$\Delta\rho/\rho \sim p/K$$

Thus, the fractional change in density between surface and deep interior is roughly the pressure in the deep interior divided by the bulk modulus. This assumes negligible surface pressure and that $P \lesssim K$. Therefore, if the interior pressure is comparable to the bulk modulus, then you expect the density inside the planet to be considerably larger than at the surface. In the giant planets, this increase in density occurs even in the interior well below the region that behaves like an ideal gas.

To summarize (Table 4-2):

The type of bonding refers to *low-pressure* behavior. However, everything becomes a metal at high-enough pressure.

As an important first step to understanding the conditions inside planets, the pressure can be estimated by *assuming* that the density is roughly constant. Obviously, this is only a very rough guide, but it helps the reader appreciate what must be known about material properties. From hydrostatic equilibrium,

$$\frac{dp}{dr} = -\rho(r)g(r) \tag{4.1}$$

Table 4-2
Types of bonding

Type of bonding	Examples	Solid densities at low P	Bulk modulus of solid	Locations found
Van der Waals	"Gases," hydrogen, helium, methane, N_2	E.g., hydrogen is ~0.07 g/cc	E.g., hydrogen is a few kilobars	Giant planets (also CH_4 and N_2 on icy satellites)
Hydrogen bonding	"Ices," water, ammonia	Around unity	10 kilobar (roughly)	Giant planets, icy satellites
Ionic and covalent (including metallic)	"Rocks," metallic iron	Rocks are around 3 g/cc; iron is near 8 g/cc	Typically of order 1 Megabar	Terrestrial planets, cores of giant planets(?), icy satellites

where r is radius and g is gravitational acceleration. But also,

$$g(r) = \frac{GM(r)}{r^2} \approx \frac{4}{3}\pi G\overline{\rho}r \tag{4.2}$$

where M(r) is the mass enclosed within radius r and $\overline{\rho}$ is the mean density. By substituting for g and integrating,

$$p(r) \approx \int_r^R \frac{4}{3}\pi G\overline{\rho}^2 x\,dx = \frac{2}{3}\pi G\overline{\rho}^2(R^2 - r^2) \tag{4.3}$$

where the fact has been used that surface pressure is negligible, $p(R) \approx 0$. Evaluating at r = 0,

$$P_{\text{center}} \approx (1.4 \text{ kilobars})\left(\frac{\overline{\rho}}{1\text{ g/cm}^3}\right)^2\left(\frac{R}{1{,}000\text{ km}}\right)^2 \tag{4.4}$$

For the Moon, this predicts about the true central pressure (not surprisingly because it is a small body). For Earth, it predicts around 2 Megabars (the true value is about 3.6). For Uranus and Neptune, it predicts similar values (because the lower mean density is offset by the larger radius). For Jupiter, it predicts around 10 Megabars (the actual is 40 or more). This crude formula under predicts the central pressure of differentiated bodies (which is to say, all planets), especially when the core has a density much larger than the mean density.

If one wants to figure out what goes on in a planet, then the behavior of the materials listed above at planetary pressures must be known. One approach is *experiment.* The experimental techniques are of two kinds: shock waves and static compression. Shock waves generate very high pressures for tiny fractions of a second by accelerating samples into a stationary target. Static compression is more ideal for studies of what happens in equilibrium, but attaining extremely high pressures is more difficult. The highest static pressures are obtained by squeezing tiny (~tens of microns) samples between the tips of two diamonds in what are called diamond anvil cells (remember that pressure is force over area). But experimental data alone are not enough for two reasons: (1) Experiments do not usually get to the full range of conditions encountered in planets. For example, pressures deep within Jupiter are unattainable by conventional techniques. (2) Even when experiments reach relevant conditions, they seldom map out enough of the thermodynamic phase space – T, P, and composition – to be sufficient for planetary modeling, where one needs a fine grid of parameter values. So it helps greatly to have a *theoretical framework* to incorporate experimental results and to extrapolate and interpolate. Many of these frameworks seek to find one of the most basic and important properties of a material, which is its equation of state. This relates how the density of a material depends on the environmental conditions; strictly, this includes both pressure and temperature, but in many cases, the thermal effects can be safely neglected. The reason is that the electronic energies from the interactions at high pressure much exceed thermal energies. Typical electronic energies are many electron volts, whereas the thermal energy per ion of molecule is of order $k_B T$ where k_B is Boltzmann's constant, and this equals 1 eV at 12,000 K.

It is a good idea to get some physical understanding of why materials resist compression, and one way to do this is to construct a theory that works at very high pressures. This turns out to be an excellent approximation for the deep interior of Jupiter, as well as being pedagogically valuable. Consider, first, the energy of an electron that is confined to a sphere of radius r. By the Heisenberg uncertainty principle, it has a momentum $\sim\hbar/r$, and therefore an energy $\sim(\hbar/r)^2/2$ m where m is the electron mass and t_v is planck's constant divided be 2π. Notice that since density scales as $1/r^3$, this means the energy scales as $\rho^{2/3}$. Since Coulomb energies scale as 1/r,

it follows that the kinetic energy implied by quantum mechanics always win at sufficiently high density and pressure. At very high pressures, the electrons will form a Fermi gas, and the typical momentum of an electron will be ~ħ/r even though the electrons are not spatially localized but form a degenerate "sea." This arises because of the Pauli exclusion principle, which limits each quantum state occupancy to just two electrons (corresponding to the two-spin states). The resulting energy per electron thus scales as $\rho^{2/3}$, and the resulting pressure (minus the derivative of this energy with respect to volume) scales as $\rho^{5/3}$. This is called the Fermi pressure, and it is the asymptotic (that is to say, extreme high pressure) limit of all nonrelativistic materials. Of course, the electrons are nonuniformly distributed in planets and the materials need not even be metals, but even so, this kind of equation of state forms a guide for what to expect.

In Jupiter, this limit is approached but not reached. To see this, consider the density one would obtain if one simply filled space with spheres, each of which had the mass and first Bohr radius of a hydrogen atom. Neglecting interstitial space (i.e., squashing the atoms into dodecahedra of the same volume as the spheres), this would yield

$$\rho = \frac{m_p}{\frac{4}{3}\pi a_o^3} \sim 2.69\ \mathrm{g/cc} \tag{4.5}$$

where m_p is the proton mass, and a_o is the first Bohr radius (about 0.53 Å). The mean density of Jupiter is only 1.33 g/cc, but a substantial part of the interior is at densities exceeding this value. In the hydrogen atom, there is balance between Coulomb and the kinetic energy required by the uncertainty principle. As a consequence, the actual behavior of hydrogen deep within Jupiter has not reached the Fermi gas limit. Still, it is beginning to approach the limit in which the density is determined by the electron density. Since hydrogen uniquely has only one heavy particle (the proton) per electron, whereas all other materials have two or more heavy particles per electron, it turns out that hydrogen is distinctly less dense than any other material at *any* pressure, usually by more than a factor of 2. This is what enables us to be so confident about the mean composition of Jupiter. However, there remains one more step in the logic: Can we be sure that temperature does not reduce the density significantly? That is, could we imagine a Jupiter made of mostly helium (say) but very hot? To answer this, it is necessary to appreciate the typical electronic energies within planets. In the case above of hydrogen atoms stuffed together, the energy of the electrons is of order 10 eV (the binding energy of the hydrogen atom). To disturb the energy and significantly change the density, one would have to heat the material to ~(10 eV)/k_B ~100,000 K. Aside from the likely lack of a means of heating to that level, the problem with such a high T is that the body would then be wildly unstable to convection. Imagine displacing upward a blob of material at such a high T. It will rise and expand adiabatically and thus find itself surrounded by material more dense than itself. It therefore accelerates upward, at high velocity carrying its excess heat outward and relaxing back toward an isentropic state. The story is actually somewhat more complicated than this since there may be compositional gradients (see, e.g., Leconte and Chabrier 2012), and one must also consider the limited amount of energy available in formation. But the conclusion is that planets are close to being *degenerate*, that is, their thermodynamic properties are dominated by the electrons and not by the thermal motions. Indeed, this is the central difference between planets and main sequence stars. Planets can be compared to white dwarf stars, which have a mass similar to our Sun but a radius similar to Earth.

It is useful to think about the behavior of a planet that has an equation of state of polytropic form. Astrophysicists write such an equation in the form $P \propto \rho^{1+1/n}$ where n is the

polytropic index. We shall here adopt for convenience the form $P \propto \rho^n$. This will have relevance for materials that expand without limit as $P \to 0$ or bodies that are so massive that the mean density is much larger than the zero pressure density. Since the central pressure according to hydrostatic equilibrium must scale as $\langle\rho\rangle gR \sim (M/R^3)(GM/R^2)R \sim GM^2/R^4$ and the density scales as M/R^3, we immediately have

$$\frac{M^2}{R^4} \propto \left[\frac{M}{R^3}\right]^n \Longrightarrow R \propto M^{\frac{2-n}{4-3n}} \tag{4.6}$$

Notice four things:

1. In the limit $n \to \infty$, the result $R \propto M^{1/3}$ is recovered; this makes sense since that limit is incompressible material (infinitesimal density change gives large pressure change).
2. When $n = 5/3$, $R \propto M^{-1/3}$. Recall that this is the case for *an ideal Fermi gas.* It is therefore applicable to super-Jupiters, brown dwarfs, and white dwarfs. At sufficiently high mass (but still nonrelativistic), it applies to all materials irrespective of atomic mass. *At sufficiently high mass, all degenerate bodies become smaller as mass is added to them.* It should be noted, however, that bodies more massive than Jupiter are also hotter; more precisely, they usually have higher entropy. This means that the thermal effects become increasingly important. For this reason, bodies of ten or so Jupiter masses are often slightly bigger than Jupiter itself.
3. R is independent of mass when $n = 2$; this is approximately relevant to Jupiter and Saturn.
4. For $n = 4$ (crudely relevant for Superearths and parts of Uranus and Neptune) $R \propto M^{1/4}$. There is a large literature on polytropes; the reader can find out all about them in Chandrasekhar (1939).

In general, the equation of hydrostatic equilibrium does not have an analytical solution, but $n = 2$ is a special case since it leads to a linear differential equation. Let us derive this, since it is of practical use. We assume $P = K\rho^2$:

$$\begin{aligned} \frac{dP}{dr} &= 2K\rho(r)\frac{d\rho(r)}{dr} = -\rho(r)g(r) \\ 2K\frac{d\rho(r)}{dr} &= -g(r) = -\frac{G}{r^2}\int_0^r 4\pi\rho(x)x^2dx \end{aligned} \tag{4.7}$$

Multiplying by r^2 and performing another derivative eliminates the integral:

$$\frac{d}{dr}\left(r^2\frac{d\rho}{dr}\right) = -k^2r^2\rho; \quad k^2 = \frac{2\pi G}{K} \tag{4.8}$$

This can be conveniently be written as a standard differential equation

$$\frac{d^2}{dr^2}(r\rho) = -k^2r\rho \tag{4.9}$$

for which the solutions for $r\rho(r)$ are sin(kr) and cos(kr), and it is convenient to write the solution this way:

$$\rho(r) = A\frac{\sin kr}{kr} + B\frac{\cos kr}{kr} \tag{4.10}$$

If this solution applies for all radii including the center ($r = 0$), then finiteness requires that $B = 0$. In that case, A is identified as the central density ρ_c because $\sin(x)/x$ is unity as $x \to 0$. In this coreless case, the outer surface of the planet r=R must be the first zero of sin(kr)/kr:

$$kR = \pi \Longrightarrow R = \sqrt{\frac{\pi K}{2G}} \tag{4.11}$$

Thus, an explicit formula for the radius is derived, and it is independent of mass, as promised. For the realistic choice of $K = 2.1 \times 10^{11}$ cgs units (which approximates a cosmic hydrogen/helium mixture), this formula gives a radius of 70,300 km. The mean radius of Jupiter is 69,800 km. We can also compute the mean density in terms of the central density:

$$\bar{\rho} = \frac{M}{\frac{4}{3}\pi R^3} = 3\int_0^1 \rho_c \frac{\sin \pi x}{\pi x} x^2 dx = \frac{3}{\pi^2}\rho_c \tag{4.12}$$

The inferred central density is $\pi^2/3$ times the mean density, corresponding to 4.38 g/cc and a pressure of 40 Megabars for Jupiter. The radius thus obtained should apply equally well to Saturn, but the observed radius of Saturn is only ~58,000 km. *The fact that Saturn is smaller than Jupiter must be because it has heavier constituents, not because it has lower mass.* Detailed models confirm this elementary observation.

A common way of estimating the effect of a uniform admixture of heavier elements is to assume *volume additivity.* This is typically accurate to about a percent or two, sufficiently precise for the purpose here. If the mass fraction of "heavies" is y then one obtains

$$\frac{1}{\rho(P)} = \frac{1-y}{\sqrt{P/K}} + \frac{y}{\rho_{\text{heavy}}} \tag{4.13}$$

where ρ(P) is the density at pressure P; the LHS is just the specific volume, and the RHS terms represent the specific volumes of the hydrogen–helium mixture and the heavy component respectively. In the limit where the heavy component is very much more dense than the hydrogen–helium mix and y is not too large, this equation is simply $P = K_{\text{eff}}\rho^2$ where $K_{\text{eff}} = K(1-y)^2$. This is still of the right functional form for application of the derived solution, and the radius is accordingly reduced by 1 – y. Saturn would require that y ~ 0.2, implying ~15 Earth masses of heavier stuff. The *same* amount of heavy stuff in Jupiter is a smaller fraction of the total mass ($y_{\text{Jupiter}} \sim 0.05$) and permitted by the observed radius! Detailed models allow this, though the uncertainties remain large. The presence of a core (as distinct from just heavy element enrichment) is unresolved, though likely, especially for Saturn. Determination of the presence of a core cannot be done from radius alone; it requires careful consideration of the gravitational moments (discussed below).

The above solution can still be used in the region external to a core, but A is no longer the central density, and B is no longer zero. The coefficient B is then determined by making sure that the equation of hydrostatic equilibrium is correctly satisfied at the core surface (i.e., one must allow for the gravitational acceleration of the core). The relative magnitudes of A and B then determine the shift in the radius R. This exercise can be carried through to show that the effect of a core is roughly the same as the effect of a uniform enrichment of heavy elements. *A Jupiter-like planet will have the same radius (for a given mass) irrespective of whether the heavy elements within it are in a core or uniformly mixed!*

In the simple discussion provided here, a detailed description of the current state of theory and experiments has been avoided; more details on this can be found in Guillot (2005) and in French et al. (2012). In respect of hydrogen, it is now generally agreed that at the temperatures of interest, there is no first order phase transition between the low-pressure molecular form and the high-pressure metallic (monatomic) form. Instead, there is a gradual breakdown of the molecules and populating of a conduction band, much as in a semiconductor where the band gap decreases and then vanishes with increasing pressure. The "transition" occurs at around 1 Megabar, out at 0.8–0.9 of Jupiter's radius and at ~0.6 of Saturn's radius. It probably does not

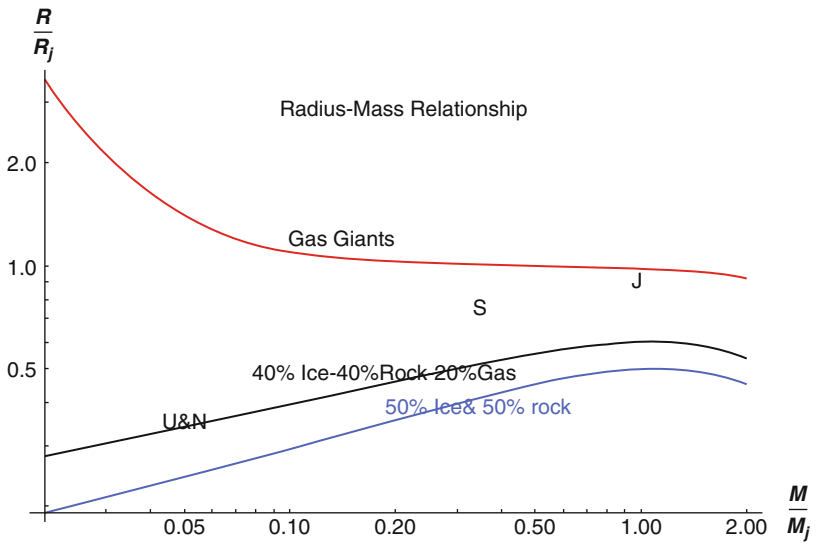


Fig. 4-1

Radius vs. mass for gas giants and ice giants, both in Jupiter units. The gas giants are assumed to have a Jupiter-like internal entropy. This is the reason why the radius does not decrease with decreasing mass. The ice/rock curve assumes 50% ice and 50% rock by mass. There is also an ice/rock + 20% hydrogen/helium by mass similar to Uranus and Neptune. The positions of the planets are indicated

occur in Uranus and Neptune because the material at that pressure is not predominantly hydrogen. The helium component does not significantly metalize (i.e., it does not contribute to the electron gas even at the higher pressures encountered in Jupiter). The interaction between the helium atoms and the Fermi gas is repulsive, leading to the possibility of limited solubility of helium, just as all noble gases tend to be insoluble in metals (and oil is insoluble in water). In Uranus and Neptune, a combination of theory and experiment constrains the behavior of rock and ice components. However, their interpretation is ambiguous: A mixture of gas and rock can have similar density and bulk modulus as ice.

In *Fig. 4-1*, one sees that hydrogen–helium *adiabatic* bodies have roughly constant radius (as promised) but actually expanding as they approach low-mass ideal gas adiabatic behavior ($P \propto \rho^{1.45}$ for a cosmic H_2–He mixture). The radius actually declines as one goes to still higher masses (the brown dwarf regime) though the effect is modest if (as is usually the case) these bodies are also hotter and thus less close to the degenerate limit. This figure assumes isentropic bodies, and we turn now to the basis for that assumption. This figure also shows us that Uranus and Neptune do not have a simple interpretation.

4 What Are the Temperatures in Planets?

The temperature inside a planet is determined by the heat flow from the interior. In the giant planets, this heat flow is sufficiently large that it can be observed as excess luminosity. The observed values are shown in *Table 4-3*.

Table 4-3
Planetary heat flows

Body	Heat flux (erg/cm^2.sec)	Luminosity (erg/sec)	Luminosity per unit mass (erg/g.sec)
Sun	6.3×10^{10}	4×10^{33}	2
240 K earth (black body)	1.9×10^{5}	1×10^{24}	1.7×10^{-4}
Earth	80	4.3×10^{20}	7×10^{-8}
Jupiter	5,400	3×10^{24}	1.7×10^{-6}
Saturn	2,000	8×10^{23}	1.5×10^{-6}
Uranus	<180	$<1.5 \times 10^{22}$	$<1.7 \times 10^{-7}$
Neptune	285	2.2×10^{22}	2.2×10^{-7}

(Earth is included for purposes of comparison). Knowledge of plausible planetary materials can be used to pose the following question: *What would the thermal state be inside a planet, were the heat to be carried by microscopic (i.e., conductive or radiative) processes alone?*

First, consider the deep atmosphere. These bodies are dominated by molecular hydrogen atmospheres, which also usually provides the dominant opacity source through pressure-induced absorption. The property of pressure-induced absorption is that the opacity is proportional to pressure. Now the heat flux carried by radiation can be written in the form

$$F_{rad} \sim d(\sigma T^4)/d\tau \tag{4.14}$$

where the increment in optical depth is

$$d\tau = -\rho\kappa dr \tag{4.15}$$

and κ is the opacity. (This equation should be intuitively obvious to order of magnitude since the photons go one optical depth before being absorbed.) Optical depth is dimensionless and here defined to decrease as one goes outward, so an optical depth of one corresponds to the place from which photons escape to space. This predicts that if radiative transport dominates, then $T^4 \propto \tau$. However, $T = T_e$ (the effective temperature) at $\tau = 1$, by definition, whence $T \sim T_e \tau^{1/4}$.

From hydrostatic equilibrium, $dP/dr = -g$. Dividing by (4.15) gives

$$dP/d\tau = g/\kappa \tag{4.16}$$

But in molecular hydrogen, the opacity is roughly proportional to pressure

$$\kappa_{mol\ H} \propto P \tag{4.17}$$

because it is the result of pressure induced opacity. (Hydrogen molecules have no dipole moment.) Integrating (4.16),

$$p \propto \tau^{1/2} \tag{4.18}$$

and constancy of F_{rad} ($T^4 \propto \tau$) then implies $T \propto p^{1/2}$. This is a stronger dependence than the increase of temperature along an adiabat. In a cosmic H_2–He mixture, $T \sim p^{0.3}$ along an adiabat. Radiative transport in molecular hydrogen would therefore require that the entropy decrease with height. This leads to a convective instability. This argument could break down around $T \sim 1{,}000$ K where molecular hydrogen becomes somewhat transparent. The opacity

does not just depend on pressure, it also depends on the wavelengths of the thermal photons involved or equivalently on temperature. Guillot (2005) has studied this in detail and finds that there are minor constituents, for example, alkali metals, that will provide sufficient opacity to guarantee that the heat flow will be carried by radiation.

Deep within the planet, the highest thermal conductivity you can find is that attributable to metallic hydrogen. This is certainly less than 10^9 erg/cm.s.K (see French et al. 2012). Even for this upper bound, the observed heat flow would lead to a temperature gradient of 0.3 K/km, implying a temperature increase of around 20,000 K for a radial range of 60,000 km. As we shall see, this is convectively unstable. The situation in dense (insulating or semiconducting) molecular hydrogen is much worse …a thermal conductivity of perhaps 10^7 erg/cm.s.K and a conductive gradient of around 30 K/km, implying ridiculous temperature increase with depth, were the heat flow carried conductively.

In all planets, and in giant gas and ice planets in particular, the conductive profiles are usually convectively unstable (at least if the material is not too viscous). This leads to *the adiabatic hypothesis*. The planet convects, and because it is a fluid, it resides very close to the isentrope (often referred to as the adiabat). Of course, this assumes that the body is chemically homogenous or is in discreet layers. But which adiabat? The answer lies in the atmosphere; the place where heat can finally escape to space. *The specific entropy at optical depth unity is the specific entropy in the very deepest parts of the planet (assuming uniform composition).* This is very different from terrestrial bodies where the choice of adiabat is dictated by the rheological law (i.e., how hot must the mantle be before the material can flow so as to eliminate the heat?).

It follows from the adiabatic hypothesis that the temperatures within these planets are above the melting curves of the major constituents. They are also supercritical, meaning that the adiabat never encounters a first-order phase transition. For example, hydrogen may have a first order phase transition between molecular and metallic forms below a few thousand degrees, but this is well below the actual temperature (see French et al. 2012). In ❯ *Fig. 4-2*, we see typical temperatures for these planets as a function of pressure, assuming adiabatic structures.

5 How Does One Explain the Heat Flows?

To answer this question, it is first useful to understand the *virial theorem.*

Hydrostatic equilibrium together with the first law of thermodynamics and other thermodynamic relations can enable one to construct an evolution of a planet or star. However, there is an extremely valuable result called the *virial theorem* that enables one to see some of the general behavior of an evolving body. There is no new physics in this theorem, merely a clever use of the existing physics. It assumes slow evolution relative to dynamical timescales, and in the current application, this is the same as the hydrostatic assumption. The virial theorem also shows up in other areas of physics (e.g., the expected velocity dispersion of a self-gravitating star cluster).

The gravitational energy of a planet can be written

$$E_G = -\int_0^M \frac{Gmdm}{r(m)} \tag{4.19}$$

where we have chosen to use m (the mass inside radius r) as the independent variable. Thus, r is the radius that contains mass m. But we also have the hydrostatic equation, which together

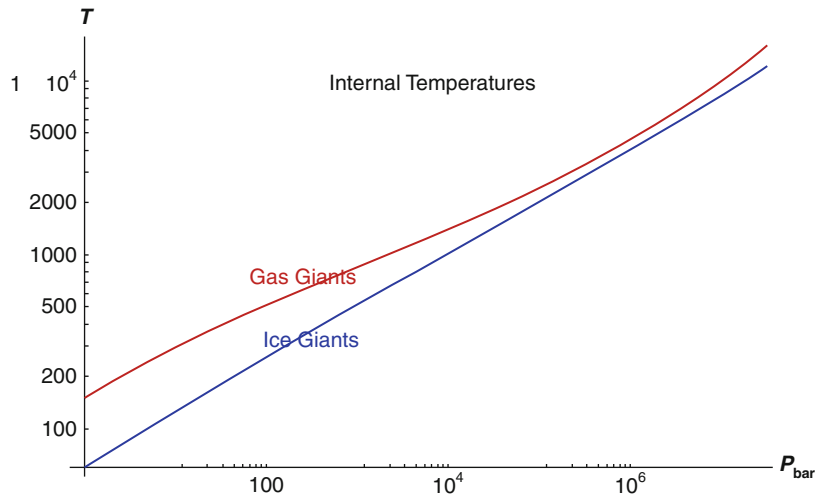


Fig. 4-2

Temperature vs. pressure for Jupiter/Saturn and Uranus/Neptune. Although Saturn is cooler than Jupiter, the difference is not large enough to be evident on this scale. Similarly Uranus and Neptune

with the definition of m(r) leads to the following:

$$\begin{aligned} \frac{dP}{dr} &= -\frac{Gm\rho}{r^2} \\ dm &= 4\pi r^2 \rho dr \\ \Longrightarrow \frac{dP}{dm} &= -\frac{Gm}{4\pi r^4} \end{aligned} \tag{4.20}$$

Looking back at the gravitational energy, we see that

$$E_G = \int_0^M 4\pi r^3 dP = 3\int_0^M \left\{ d\left[\frac{4}{3}\pi r^3 P\right] - PdV \right\} \tag{4.21}$$

However, $P.(4\pi r^3/3)$ is zero at both the center and outer surface of the planet, since r = 0 and P = 0 respectively, at those locations. Consequently, only the PdV terms remain:

$$E_G = -3\int_0^M PdV = -3\int_0^M \frac{P}{\rho} dm \tag{4.22}$$

This last result is one of several ways of stating the *virial theorem*.

In the conventional discussion of stellar structure (*not planets!*), ideal gas formulas apply to a good approximation. The pressure is then nk_BT, where n is the number density of atoms or molecules. Moreover, the thermal energy can be written in the form nk_BT/γ where $\gamma = C_p/C_v - 1$ and is known as the Gruneisen parameter. (Warning: Some texts use γ to represent the ratio of specific heats. Remember also that this formula is only correct for an ideal gas.) It follows that

$$E_{\text{total}} = E_G + E_{\text{thermal}} = \left(3 - \frac{1}{\gamma}\right)\int_0^M PdV \tag{4.23}$$

whence it follows immediately that provided $\gamma > 1/3$, the total energy is negative. Atomic hydrogen has $\gamma = 2/3$; molecular hydrogen has $\gamma = 2/5$. Radiation dominated systems can approach

$\gamma = 1/3$, and the specialness of this limit is related to the undetermined character of a polytrope when n = 4/3 and the resulting instability (collapse or explosion), as discussed at the beginning of this chapter.

The following discussion is only for the usual case where $\gamma > 1/3$. For a body in hydrostatic equilibrium and supported by ideal gas pressure, the integral above increases as the temperature goes up since P = nkT and the integral of ndV is constant. But this corresponds to a decrease in total energy and therefore corresponds to contraction as heat is lost to space. In other words, a star with no other energy sources has "negative specific heat": As it loses energy by radiation to space, it contracts and heats up internally. This was a major puzzle in the early work on stellar evolution. In reality, the contraction and heating is truncated either by the introduction of a new energy source: thermonuclear ignition (arrival at the main sequence) or the introduction of a new pressure: the onset of degeneracy. Provided ideal gas dominates, $T \sim GM\mu/k_B R$, either during slow contraction or on the main sequence. Here, T is the typical internal temperature, and μ is the molecular weight (normally ~ proton mass for a main sequence star).

In a planet, one cannot use the virial theorem to determine temperature directly, but it can still be used to set up equations for the thermal history, as explained below, and it leads to a very different answer from stars.

It is most useful to consider a perturbation to the planet in which hydrostatic equilibrium is preserved, but there are infinitesimal changes in density and pressure at each mass element. In the following formulas, the reader should think of δ as labeling the change that comes about because of some evolutionary change at that location. For example, $\delta(1/r)$ means the change in 1/r that comes from the change in the radius that contains mass m. Thus,

$$\delta E_G = -\int_0^M Gm\delta\left(\frac{1}{r}\right)dm = \int_0^M \frac{Gm}{r^2}\delta(r)dm \tag{4.24}$$

But as before we apply hydrostatic equilibrium (9.12) to get

$$\begin{aligned}\delta E_G &= -\int_0^M 4\pi r^2 \frac{dP}{dm}\delta(r)dm = -\int_0^M \delta\left(\frac{4}{3}\pi r^3\right)\frac{dP}{dm}dm \\ &= -\int_0^M \left\{d\left[\delta\left(\frac{4}{3}\pi r^3\right)P\right] - \delta\left[\frac{d\left(\frac{4}{3}\pi r^3\right)}{dm}\right]Pdm\right\}\end{aligned}$$

However, the first term is zero just as it was previously, and $dm = \rho.d(4\pi r^3/3)$, whence

$$\delta E_G = \int_0^M P.\delta\left(\frac{1}{\rho}\right)dm \tag{4.25}$$

So *the change in gravitational energy equals the work done on the sample.* (This interpretation only works if you think about constant composition. Obviously one can also lower the energy by moving the denser stuff to higher pressures and the less dense stuff to lower pressures, that is, $\delta(1/\rho)$ is negative where P is high and positive where P is low. The formula is always correct but the interpretation of it depends on the circumstances.)

We can now use this result to derive a very important result for the energy output of planets. The intrinsic luminosity L of a planet comes from explicit energy sources (e.g., radioactive decay, tidal heating), here labeled Q (per unit mass), but can also come from changes in internal and gravitational energy. When we refer to "intrinsic" luminosity, we are excluding the (sometimes dominant) energy arising from external sources, especially the absorbed sunlight. Thus,

$$L = \int_0^M Q dm - \frac{d(E_G + E_{\text{int}})}{dt} \tag{4.26}$$

Consider a planet that is not differentiating (i.e., not changing the distribution of constituents). Now, planets are degenerate bodies, and we can thus conveniently subdivide the internal energy and pressure into zero temperature pieces and finite temperature corrections in the form

$$\begin{aligned} E_{\text{int}} &= E_0 + C_V T \\ P &= P_0 + \gamma \rho C_V T \\ P_0 &= -\frac{dE_0}{d\left(\frac{1}{\rho}\right)} \end{aligned} \tag{4.27}$$

(This approximation is not essential to the result but aids the understanding of the result). So

$$\frac{d(E_{\text{int}} + E_G)}{dt} = \int_0^M \left[\frac{dE_0}{dt} + C_V \frac{dT}{dt} + P_0 \frac{d(1/\rho)}{dt} + \gamma \rho C_V T \frac{d(1/\rho)}{dt} \right] dm \tag{4.28}$$

where the expression for E_G was derived above. But the first and third terms in the integral cancel. The last term is small because we can estimate that $d(1/\rho)/dt = (\alpha . dT/dt)/\rho$, where α is the coefficient of thermal expansion, and $\alpha T \ll 1$, but can also be combined with the earlier thermal term using the thermodynamic identity (Maxwell relation) $C_p = C_V(1 + \alpha\gamma T)$. So we finally get, to an excellent approximation,

$$L = \int_0^M \left[-C_P \frac{dT}{dt} + Q \right] dm \tag{4.29}$$

(The extent to which this is approximate rather than exact is not straightforward to estimate analytically. Numerical evaluation does however indicate that it is a very good approximation.)

What this result means is that as a planet cools and contracts, the gravitational energy becomes more negative, and the dominant part of the internal energy (the part due to compression) becomes more positive, but that these cancel! The consequence is that the dominant energy output associated with the planet evolution (aside from radioactivity and external sources of heating) is the change in the thermal energy at constant pressure. This seems intuitively obvious, but it is nonetheless widely misunderstood (e.g., there are books which say that Jupiter's energy output is derived from contraction when in fact the dominant effect is simply cooling. Of course, cooling *implies* contraction, but the energy available is nonetheless thermal).

The result is sometimes stated with C_V replacing C_P, but this is often a minor point since these differ typically by less than 10%, at least for Jupiter and Saturn (unlike "hot Jupiters" or brown dwarfs). *Very importantly*, the result does not include gravitational energy release due to compositional changes (e.g., core formation, differentiation in general). This can be immediately recognized by noting that changes in density (determining changes in gravitational energy) do not only come about from work done by pressure …they can also arise from moving constituents around.

Let us now consider the various possible heat sources.

5.1 Radioactivity

By *mass,* the cosmic abundances of K, U, and Th are in the ratios 60,000:4:1. For those ratios, (close to those actually observed in CI carbonaceous chondrites) the expected present-day heat production of carbonaceous chondritic material is around 6 or 7×10^{-8} erg/g.sec. This comes roughly one half from ^{40}K and one quarter each from ^{238}U and ^{232}Th (for more reliable numbers look at the various tables in Lodders and Fegley (1998)). The CI value for heat production looks roughly like the Earth value of L/M in ❯ *Table 4-3* but that is a fallacy because Earth's mantle

is depleted in K by about an order of magnitude relative to CI. There is no reason to suppose that Earth's core contains such large amounts of potassium; the depletion is more reasonably attributable to volatility. (Other elements of similar volatility to K are also depleted in Earth, and there is no expectation that they would also go into the core.) Consequently, Earth's energy output is believed to be roughly a factor of 2 larger than that due to radioactivity alone. Moon's heat flow is too poorly determined to yield to any useful analysis. It may be anomalously high in the places measured (or just very poorly measured).

Note two important things about radioactive heating: *First*, it scales as the mass of the planet (which means that the surface heat flux will scale roughly as mass over area or, equivalently, radius). So Mars should have a factor of two-lower heat flux than Earth (all else being equal, which it isn't!) *Second*, radioactive heat decays with time so the heat flow should be much larger in the early history of the planet. For Earth, radioactive heating should be about four times larger in early history (with ^{40}K and ^{235}U playing major roles).

Clearly, none of the heat flows measured in the outer solar system can be attributed to radioactivity, especially when one considers that the mass of these bodies (except Io) is dominated by material (H, He, C, N, O) that contains no radioactive elements.

Short-lived radioactive heating (^{26}Al, ^{60}Fe) may also be important in setting up the initial conditions for planet evolution and allowing even small bodies to melt and differentiate. The half-lives of these elements are less than a million years. There is enough energy release from the decay of ^{26}Al to raise the temperature by ~1,500 K (assuming no heat loss). Of course, it is only available if a substantial-sized body forms very quickly after the initial collapse to form the solar system.

5.2 Secular Cooling

Suppose a planet has cooled by an amount T over 4.5 billion years. If it has a mean-specific heat C_P, then the expected luminosity per unit mass, L/M, is obviously $C_P T/(4.5 \times 10^9$ years) which is

$$L/M = (7 \times 10^{-8} \text{ erg/g.sec}).(C_P/1 \times 10^7 \text{cgs}).(T/10^3 \text{K}) \tag{4.30}$$

For Earth, this will be significant, or even dominant for cooling rates of ~200 K/billion years. For Jupiter and Saturn where the specific heat is $\sim 2 \times 10^8$ cgs (because of the low molecular weight), the observed luminosity can be explained for total cooling of a few thousand degrees (though, as we shall see, the cooling may be fast early on). Neptune could also be explained with ~1,000–2,000 K cooling and an ice-dominated-specific heat ($C_P \sim 2 \times 10^7$).

5.3 Differentiation

If some fraction x of a planet settled to the bottom and was ~twice as dense as mean density, then you might expect the total energy release to be $\sim xGM^2/2R$. Averaged over 4.5×10^9 years, one then has

$$\begin{aligned} &L/M \sim 5 \times 10^{-5} x \text{ for Jupiter} \\ &\text{and} \sim 1.5 \times 10^{-5} x \text{ for Saturn,} \end{aligned} \tag{4.31}$$

so this could easily be important.

Let us now consider what happens if we attribute all the heat flow from the giant planets to secular cooling. Suppose the atmosphere has opacity κ. Optical depth unity is by definition the level at which $\rho\kappa H \sim 1$, where H is the atmospheric pressure scale height. But $P \sim \rho g H \sim g/\kappa$. If g does not change much through the planet's history, and the opacity is likewise constant, then the outer boundary condition for the planet's adiabatic interior is a fixed pressure P_e and it is associated temperature T_e (which will certainly vary). For Jupiter and Saturn, this "effective pressure" for the outer boundary condition is about 1 bar.

Now let us attribute to Jupiter a "mean adiabatic index" Γ defined so that $T \sim P^{\Gamma}$ along an adiabat. It follows that

$$\frac{T_{\rm int}}{T_e} = \left(\frac{P_{\rm int}}{P_e}\right)^{\Gamma} = A(\text{some constant, independent of time}) \tag{4.32}$$

where T_{int} and P_{int} are typical internal or central temperatures and pressures. This is our basis for estimating the temperature inside Jupiter and how it varies over geologic time. In fact, it turns out that since $T_e \sim 170$ K (Jupiter, now), $P_e \sim 1$ bar, $P_{int} \sim 10$ Megabars, and $\Gamma \sim 0.3$ (all the way from ideal gas to deep interior), we get $T_{int} \sim 20,000$ K. Thus, A is around 100 *not just now but throughout the history of the planet.* This ignores the role of any rock or ice core (but note that the heat stored in such a core is small because the specific heat of such materials is low compared to hydrogen.)

Since we know that the radioactive decay of long-lived isotopes cannot contribute significantly to the heat output of giant planets, let us try a Kelvin model in which the heat output is due to the leakage of primordial heat (heat trapped when the planet formed). Let us first estimate the maximum temperature rise during formation:

$$\Delta T \sim \frac{GM}{RC_p} \sim 10^5 \text{ K} \tag{4.33}$$

for Jupiter. This is much larger than current estimates for Jupiter's mean temperature so the idea is not unreasonable. From the virial theorem,

$$L = 4\pi R^2 \sigma\left(T_e^4 - T_0^4\right) = -\frac{d}{dt}\int_0^M C_p T dm = -A\overline{C}_p M \frac{dT_e}{dt} \tag{4.34}$$

where T_0 is the effective temperature the planet would have in the absence of internal heat sources. If we define a *Kelvin time* τ as the *current* heat content of Jupiter divided by the *current* excess luminosity, then we can nondimensionalize the equation in this form

$$\begin{aligned} &\frac{\left(x^4 - x_0^4\right)}{\left(1 - x_0^4\right)} = -\tau_K \frac{dx}{dt} \\ &x \equiv \frac{T_e(t)}{T_e(t = \text{now})};\; x_0 \equiv \frac{T_0}{T_e(t = \text{now})} \end{aligned} \tag{4.35}$$

which can be solved straightforwardly to find the time that must elapse to get from an initial condition ($x = x_i$) to the current state ($x = 1$; note that for Jupiter $x_0^4 = 0.5$ approx). If one uses the approximation $x \gg 1$ (which is not really true except at early times), then clearly $x \sim t^{-1/3}$. The time it takes to cool from a much hotter state to the present state is thus found to be about 0.25 Kelvin times, where the numerical factor arises because the initial cooling is fast. Applied to the data, one finds that:

Table 4-4
Cooling times

Planet	Cooling time from hot initial state to present observed state
Jupiter	5 Ga
Saturn	2.5–3.5 Ga
Uranus	~15 Ga (or more)
Neptune	~10 Ga

Table 4-5
Observed helium abundances

Body	Helium abundance (expressed as a mass fraction of the total)
Post–Big Bang universe[a]	~ 0.26
Primordial solar[b]	0.272
Jupiter atmosphere[c]	0.24
Saturn atmosphere[d]	0.15?
Uranus, Neptune[d]	~ 0.27

[a]Determined in part by the need to get the correct deuterium and He-3 (this is a theoretical calculation but well constrained)
[b]Determined from solar models and solar seismology in particular. Very well constrained
[c]Determined by the Galileo probe (a dedicated instrument that measured refractive index of the gas). Accurate to ±0.005
[d]The value determined indirectly by the far IR opacity of the atmosphere was originally estimated to be 0–0.1. But this method gave about 0.18 for Jupiter and we know this to be wrong. The true value is more plausibly ~0.15. Cassini is improving this, and the rumors (2007–2008) suggest 0.15. But this indirect method is notoriously unreliable

The interpretation for *Jupiter* is that Kelvin cooling could explain the luminosity (but see below). The interpretation for *Saturn* is that we need an extra energy source. The interpretation for *Uranus and Neptune* is that either the planets did not start out hot or they do not convect from top to bottom. The latter seems more plausible, given their energy of formation.

In Jupiter and Saturn, helium is expected to be insoluble below some critical temperature. This is because they contain metallic hydrogen, and noble gases are highly insoluble in metallic hydrogen. Droplets would then form and settle toward the center of the planet. The energy release from this process in Jupiter is potentially enormous such that only 10% of the helium needs to rainout to supply all of Jupiter's luminosity for a billion years or more. In Saturn, where the gravity is smaller, the planet is smaller and the rainout distance probably smaller (even expressed as a fraction of the radius), the energy release is much less …about half of the helium would need to rain out. The observations are these (*Table 4-5*).

The likely interpretation is that *Jupiter* is experiencing rainout and the size of the helium change is small because the planet is hotter than Saturn and also somewhat buffered …. Rainout releases heat which reduces further rainout. *Saturn* has experienced more rainout because it is a smaller planet and it is also somewhat colder.

The lower heat flows (longer "ages") for Uranus and Neptune are best explained by supposing that these planets are not well mixed: Much of the heat emplaced during accretion is unable to escape because of a compositional gradient. About one-half of the heat content must be trapped and not contributing to the current heat flow. Uranus is more affected than Neptune and it is also closer to the Sun, which buffers the escape of heat.

6 The Gravity Field

External to a planet (in a region where there is no or negligible mass), the gravitational potential V satisfies Laplace's equation

$$\nabla^2 V = 0 \tag{4.36}$$

It makes sense to use a planet-centered spherical coordinate system, and in that case, the general solution to Laplace's equation can be written in the form

$$V = \frac{1}{a}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell}\left[\frac{a}{r}\right]^{\ell+1}(C_{\ell m}\cos m\varphi + S_{\ell m}\sin m\varphi)P_\ell^m(\cos\theta) \tag{4.37}$$

Angle θ is colatitude and φ is longitude. The reference radius a is conventionally taken to be the equatorial radius. The solution assumes no external sources of mass (i.e., all terms decay as r goes to infinity). The Ps are associated Legendre functions, and along with the sines and cosines of longitude, they define spherical harmonics usually written $Y_{\ell m}$. The C's and S's are called spherical harmonic coefficients.

Obviously, C_{00} is nothing other than GM. (By the way, GM can be measured to twelve-figure accuracy for Earth, but that does not mean we know M that well! G is the least well-known fundamental constant and very hard to measure.) Precise tracking of an orbiting spacecraft can give you these coefficients. As in any mathematical representation, one always truncate the representation and thus fail to characterize very high harmonics. This is a potential problem even in quite low-planetary orbits.

In the special case where the planet is a rotating hydrostatic fluid, symmetry arguments alone dictate that the potential is axisymmetric (only m = 0 terms allowed) provided we choose our polar axis to be coincident with the rotation axis. Moreover, there is no physical distinction between northern and southern hemispheres, so odd λ values are excluded. (This assumes we have chosen the origin of coordinates to be the center of mass.) We can then write the potential in the form

$$V = \frac{GM}{r}\left[1 - \sum_{\ell=1}^{\infty} J_{2\ell}\left[\frac{a}{r}\right]^{2\ell} P_{2\ell}(\cos\theta)\right] \tag{4.38}$$

In this simple case, the Ps are now the simple Legendre polynomials, and the Js are called gravitational moments. In rapidly rotating planets, J_2 is generally far larger than any of the other harmonics (except of course C_{00}).

The fundamental definition of the gravitational potential is, of course

$$V(\vec{r}) = G\int_{\text{all space}} \frac{\rho(\vec{r}')d^3r'}{|\vec{r}-\vec{r}'|} \tag{4.39}$$

obtained by adding up the contributions of all masses and appealing to the superposition principle (the linearity of Newtonian gravity). Outside the planet, one can appeal to the fundamental theorem (also known in mathematics as the generating function)

$$\frac{1}{|\vec{r}-\vec{r}'|} = \sum_{\ell=0}^{\infty}\frac{r'^{\ell}}{r^{\ell+1}}P_\ell(\cos\gamma) \tag{4.40}$$

where γ is the angle between vectors r and r′, and r is outside the planet (r′ is inside the planet). Moreover,

$$P_\ell(\cos\gamma) = \sum_{m=0}^{\ell}\frac{(\ell-m)!}{(l+m)!}P_\ell^m(\theta)P_\ell^m(\theta')\cos[m(\phi-\phi')] \tag{4.41}$$

But we only need to keep track of m = 0, for evaluating $J_{2\lambda}$ (because these are the only terms that contribute to the integral):

$$P_\ell(\cos\gamma) = P_\ell(\cos\theta)P_\ell(\cos\theta') + m \neq 0 \text{ terms} \tag{4.42}$$

so it follows immediately that

$$J_{2\ell} = -\frac{1}{Ma^{2\ell}} \int r'^{2\ell} P_{2\ell}(\cos\theta')\rho(r')d^3r' \tag{4.43}$$

One can see already why the Js are called gravitational moments since they are integrals of the internal density distribution, weighted by progressively higher powers of the radius.

Recalling that $P_2(\cos\theta) = (3\cos^2\theta - 1)/2$, one can see immediately that

$$Ma^2J_2 = -\int \left[\frac{3}{2}z^2 - \frac{1}{2}(x^2 + y^2 + z^2)\right]\rho(r)d^3r \tag{4.44}$$

where x, y, z is a Cartesian coordinate system in which z is along the rotation axis. Now if we define the axial moment of inertia as C and the other two-principle moments of inertia as A and B, then we have

$$\begin{aligned} C &\equiv \int (x^2 + y^2)\rho(r)d^3r \\ A &\equiv \int (z^2 + y^2)\rho(r)d^3r; \quad B \equiv \int (z^2 + x^2)\rho(r)d^3r \end{aligned} \tag{4.45}$$

and it is easy to see that

$$Ma^2J_2 = C - \frac{1}{2}[A + B] \tag{4.46}$$

In the special (and highly relevant case) where A = B (i.e., the planet is a body of rotation rather than triaxial), we have

$$Ma^2J_2 = C - A \tag{4.47}$$

So this moment is related to the *difference* between axial and equatorial moments of inertia. In a similar manner you can show that $C_{22} \sim (B–A)/Ma^2$. (With appropriate choice of zero longitude, S_{22} will be zero.) *But you cannot get the actual values of A, B, and C from the gravity field; there is insufficient information. You can only get their differences (e.g., C–A).*

We saw that J_2 can be related to the difference in the moments of inertia about equatorial and polar axes. An independent piece of information is often available from the precession rate of a planet. Since the torque acting on a planet depends on C–A and the rate of change of angular momentum of a planet is proportional to C, precession rate gives you (C–A)/C. One can then solve for the individual moments of inertia C and A. This is how we know Earth, Mars, and lunar moments of inertia. (Determination of the Mars' moment of inertia using Mars Pathfinder tracking was one of the major accomplishments of that mission.) A somewhat more complicated version of this approach can work for Mercury. But we ought to be able to figure out something more from J_2 alone because its value depends on how the planet responds to its own rotation, and this response depends on its density distribution. It should be obvious, for example, (just by looking at the definition as an integral over the interior weighted by radius squared) that J_2 will be small for a body that is centrally concentrated.

Although this is intuitively reasonable, the appropriate theory is quite nasty, and rather little insight emerges from wallowing in the nastiness. (The full theory involves integro-differential equations that must be solved on a computer and even then converge slowly.) I will only give

a feeling for the theory. *This will only work if J_2 is dominated by hydrostatic effects. The theory explicitly assumes hydrostaticity.*

Consider, first, the constant density body. The external potential is obviously completely determined by the shape of the free surface. Approximate the free surface of this body by the lowest order non-spherical shape permitted, that is,

$$r_s = r_0(1 + \varepsilon.P_2) \tag{4.48}$$

where r_0 is some mean radius, and ε is a dimensionless constant. Inserting in the fundamental equation for the gravitational potential and using the generating function, one immediately finds that the only part that depends on P_2 is of the form

$$V_2 = \frac{GP_2}{r^3} \int_{-1}^{1} P_2(\cos\theta')d(\cos\theta') \int_{0}^{r_0(1+\varepsilon P_2)} x^2.2\pi x^2.\rho_0.dx \tag{4.49}$$

(V_2 is the amplitude of the coefficient of the P_2 term in the external potential. Here and below, the explicit angular dependence is sometimes omitted since it should be obvious and it makes the equations less cluttered.) Since the Ps are orthogonal to each other (and remember that P_0 is unity), the only part of the integral over x that contributes is the part proportional to P_2. Therefore,

$$\begin{aligned} V_2 &= \frac{GP_2}{r^3} \int_{-1}^{1} P_2^2(\cos\theta')d(\cos\theta').\varepsilon.2\pi\rho_0 a^5 \\ &= \frac{3G}{5r^3} Ma^2.\varepsilon \end{aligned} \tag{4.50}$$

(where the fact has been used that the normalization integral for Legendre functions is $2/(2\lambda + 1)$). We have set $r_o = a$ (equatorial radius), which is correct to this order of approximation. We must compare with the expression that defines J_2 in terms of the external expansion of the field:

$$\begin{aligned} &V_2P_2 = -\frac{GMa^2}{r^3} J_2P_2 \\ &\Rightarrow J_2 = -\frac{3}{5}\varepsilon \end{aligned} \tag{4.51}$$

Consider, now, the gravitational potential evaluated at the actual surface of the body. We must of course include the effect of rotation (the "centrifugal" effect). Recall that the acceleration is $\omega^2 s$ where s is the perpendicular distance from the rotation axis. The potential that yields this acceleration is (by integration) obviously $\omega^2 s^2/2 = \omega^2 r^2 \sin^2\theta/2 = \omega^2 r^2(1 - P_2)/3$. Now the total potential must be constant at the surface. This is where the assumption of hydrostaticity enters. To lowest nonvanishing order in P_2, this implies that

$$\frac{GM}{r_0(1 + \varepsilon P_2)} - \frac{GM}{a} J_2P_2 + \frac{1}{3}\omega^2 a^2[1 - P_2] \tag{4.52}$$

must have no dependence on P_2. If it had a dependence on P_2, then it would not be a constant on that surface! Notice that we can ignore the differences between r and a, etc., in terms that are already small (J_2 and ε are small parameters). Also, $\varepsilon \ll 1$ means that $1/(1 + \varepsilon P_2) = 1 - \varepsilon P_2$ to an excellent approximation. In other words,

$$\begin{aligned} &\left\{\frac{GM}{a}[-\varepsilon - J_2] - \frac{1}{3}\omega^2 a^2\right\} P_2 \equiv 0 \\ &\Rightarrow -\varepsilon - J_2 - \frac{q}{3} = 0 \end{aligned} \tag{4.53}$$

where q is a dimensionless measure of planetary rotation:

$$q \equiv \frac{\omega^2 a^3}{GM} \tag{4.54}$$

But we already have $J_2 = -3\varepsilon/5$. Substituting, we get

$$J_2 = \frac{q}{2}; \quad \varepsilon = -\frac{5q}{6} \tag{4.55}$$

This is what we mean by the response of the planet to rotation: The gravitational moment is related to a dimensionless measure of the strength of rotation. In general, we expect that

$$J_{2\ell} = \sum_{n=0}^{\infty} \Lambda_{2\ell,n} q^{n+\ell} \tag{4.56}$$

with the n = 0 term dominating. For example, $\Lambda_{2,0} = 0.5$ for a Maclaurin spheroid (the technical name for the uniform density case we studied here) and $\Lambda_{4,0} = -0.536$.

Of course, all that we have done here is find these coefficients in the special case of a uniform density body (a case where we already know the moment of inertia). But it should be plausible, and turns out to be actually true, that these coefficients are diagnostic of the density structure. For example,

$$\Lambda_{2,0} = \left(\frac{5}{\pi^2} - \frac{1}{3}\right) = 0.173 \tag{4.57}$$

for the case of the model we studied for Jupiter where $P = K\rho^2$. This is very different from 0.5.

From J_2/q, it is possible to estimate the moment of inertia. In practice, use is also made of the higher harmonics (J_4, J_6, …) to better constrain the internal structure. More details on this theory of figures are found in Guillot (2005) (Table 4-6). Table 4-6 shows the observed values of $J_2/q \equiv \Lambda_2$.

Recall the $P = K\rho^2$ model (for giant planets). For this we have $\rho(r) = \rho_c \sin(kr)/kr$ and

$$\frac{1}{MR^2} = \frac{2}{3}.\frac{\int_0^1 x^3 \sin(\pi x)dx}{\int_0^1 x \sin(\pi x)dx} = \frac{2}{3}\left[1 - \frac{6}{\pi^2}\right] \approx 0.26 \tag{4.58}$$

Where I is the mean moment of inertia. Notice that this result is not changed much if you add heavy elements *uniformly* since (as discussed previously) that will not change the form of

Table 4-6
Rotational response of planets

Body	Measured J_2	Measured q	J_2/q	Inferred C/Ma^2
Earth	1.0826×10^{-3}	3.5×10^{-3}	0.31	0.33
Jupiter	1.4733×10^{-2}	0.089	0.166	~0.25[a]
Saturn	1.646×10^{-2}	0.153	0.107	~0.23[a]
Uranus	3.352×10^{-3}	0.035	0.096	~0.20[a]
Neptune	3.538×10^{-3}	0.028	0.125	~0.22[a]

[a] Although these bodies are hydrostatic, the exact theory does not provide a highly precise or unique relationship between density structure and measured J_2, so the results are expressed here only approximately. Note that the result for Jupiter does agree quite well with the exact prediction of $\Lambda_{2,0} = 0.173$ for $P \propto \rho^2$. J_4 and even J_6 are used for these planets to improve the estimates of internal structure

the density profile (even though the radius of the body may change substantially!) However, if you add a dense core then $I/MR^2 \sim 0.26(1 - M_{core}/M_{total})$. This comes from including the cos(kr)/kr term in the density profile (see earlier chapter). Since I/MR^2 is somewhat smaller for Saturn than for Jupiter, it is likely that Saturn has a core.

7 Magnetic Fields

Giant planet magnetic fields are interesting because all these planets have large fields (greater or comparable to earth's field). When a planet has a large field, this requires a dynamic process (called a dynamo) deep within the planet. Thus, the observed field provides us with insight into the state of matter and the dynamics deep within a planet; there is no other way to do this. Except for the special case of Jupiter, which is a synchrotron source of radio waves, we learn about planetary magnetic fields by the direct detection of the field (the magnetosphere) in a flyby or orbiter spacecraft. This is usually done with a *magnetometer*, which is an instrument that detects change in flux through a coil.

Table 4-7 shows the observations (including earth for comparison), with likely interpretations (explained more fully below).

For *Earth, Jupiter, and Saturn* (and probably *Ganymede*), the field is predominantly dipolar but with detected higher harmonics (except for Ganymede). The tilt of the dipole relative to the rotation axis is of order 10° (Jupiter and Earth) and zero (Saturn).

For *Uranus and Neptune*, the field is about equally dipole and quadrupole, and the tilt of the dipole is 40°–60°.

There is a detailed discussion of planetary magnetic fields in Stevenson (2003) and the Chap. 6. In Jupiter and Saturn, the magnetic fields are thought to be generated by convection in the metallic hydrogen region. Jupiter's field is larger because the dynamo region extends to larger radius and perhaps also because the energy source is larger. However, this does not seem to fully explain the relative size of Jupiter's field compared to Saturn. The remarkable axisymmetry of Saturn's field may be explained by the effect of differential rotation at depth, but even in this model, it is difficult to get such a small dipole tilt as that observed. It is hoped that this will be better understood at the end of the Cassini mission. The fields of Uranus and Neptune may be explainable by convection in a shell, bounded above by insulating fluid and below by electrically conducting but stably stratified fluid. This is compatible with the unexpectedly low heat flow explained in Sect. 5 above.

Table 4-7
Magnetic fields

Planet or satellite	Observed surface field (in gauss, approx)	Comments and interpretation
Earth	0.5	Dynamo; dipolar
Jupiter	4.2	Dynamo; dipolar (extends to near surface)
Saturn	0.2	Dynamo; dipolar (deeper than Jupiter, spin-axisymmetric)
Uranus	0.2	Dynamo; quadrupolar
Neptune	0.2	Dynamo; quadrupolar

8 Detailed Models

There is no doubt that Jupiter and Saturn are mostly hydrogen and helium. No other material of sufficiently low density can explain the global properties of these bodies. However, Jupiter is much closer to cosmic (or primordial solar) composition than Saturn. Roughly speaking, Jupiter and Saturn have similar total amounts of heavy elements (all elements heavier than hydrogen or helium), but Jupiter is over three times the mass of Saturn. We have constraints on the interior structure of these bodies that arise primarily from the observed gravity field (including the gravitational moments caused by planetary rotation) but also from deep atmosphere composition, magnetic field, heat flow, and laboratory and theoretical equations of state. Jupiter may have a dense core; Saturn almost certainly has a dense core. It may seem surprising that we can be more precise about Saturn than about Jupiter, since better data exist for Jupiter. However, the putative core of Jupiter is a tiny fraction of the total mass, perhaps at most a few percent (i.e., three to ten or so Earth masses), and it accordingly has a very small effect on planetary structure. The common practice of placing a separate core of heavy elements at the centers of these planets is governed by simplicity, rather than by observation. To varying degrees, the "core" could have a fuzzy boundary with the overlying hydrogen-rich envelope. Uncertainties in the hydrogen equation of state continue to be a major source of uncertainty in the interior models.

Both of these planets are enriched in heavy elements throughout (and separate from the presence or absence of a core). In Jupiter, this enrichment is probably about a factor of 3 relative to cosmic abundance and is readily observed in those volatile components that do not condense out in the observable atmosphere. We cannot be sure that this is the enrichment for water, which condenses out deep and was not observed to be enriched in the presumably dry region that Galileo probe sampled. Water is the most abundant carrier of oxygen and therefore presumably the most abundant heavy material in Jupiter. However, interior models support about ten Earth masses of heavy elements mixed throughout the hydrogen, consistent with this factor of 3. It is particularly interesting that this factor of 3 is even seen in the heavy noble gases, including argon. The threefold enrichment of argon suggests delivery to Jupiter of material that condensed at very low temperatures, probably around 40 K, since there is no known way of incorporating argon into solid bodies in large amounts at higher temperature. These planets also supported in situ formation of a satellite system. The Galilean satellite system is particularly impressive and may contain important clues to the last stages of giant planet formation.

Uranus and Neptune are far less well understood than Jupiter or Saturn. However, there is no doubt that they are mostly ice and rock, yet also possess two or so Earth masses of gas each. The atmospheres have solar hydrogen to helium ratios (though with large uncertainty because this determination is based on the pressure induced absorption features of hydrogen, a method that has been unreliable for Jupiter and Saturn). The amount of hydrogen extractable from the ices is in principle about 0.2 of the total mass (assuming the hydrogen was delivered as water, methane, and ammonia), and this is marginally close to the hydrogen mass required by interior models. Moreover, there is the possibility that methane would decompose into carbon and hydrogen at extreme pressures. However, the atmospheres of Uranus and Neptune are highly enriched in methane (thus limiting massive decomposition of this compound to very deep regions, if any), and there is no experimental or theoretical evidence for extensive decomposition of water or ammonia under the conditions encountered inside these bodies. Consequently, it is not plausible to derive even one Earth mass of predominantly hydrogen gas from the breakdown of hydrogen-bearing ice or rock, even leaving aside the dubious proposition that such decomposed

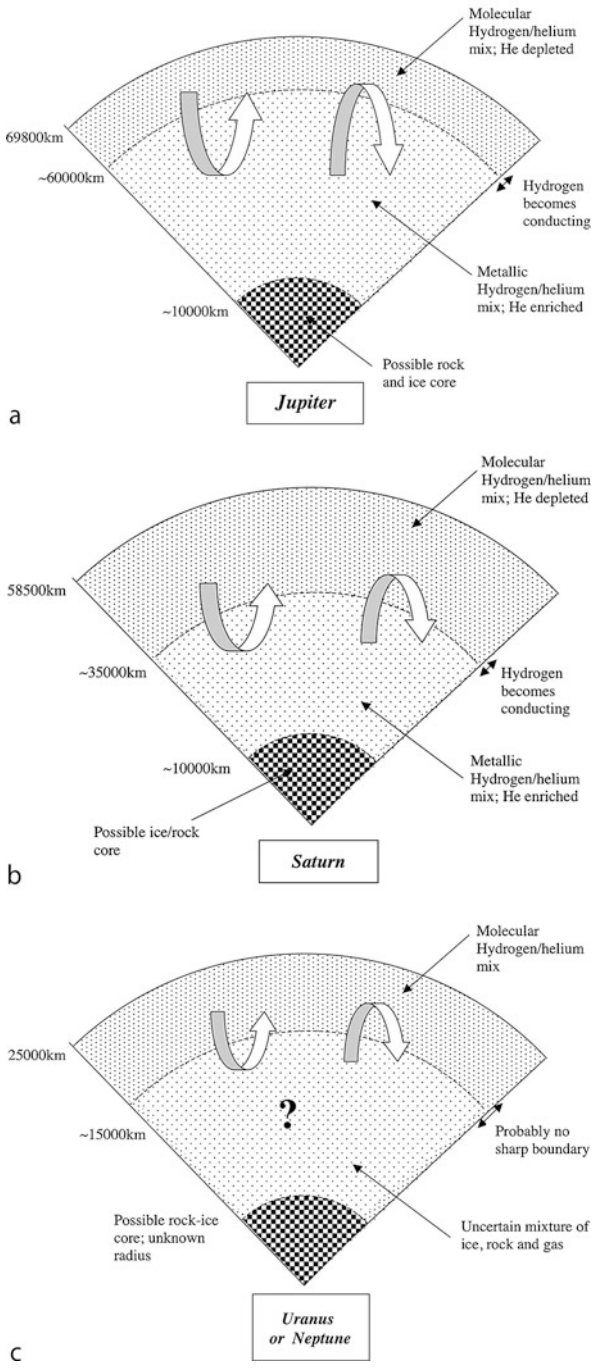


Fig. 4-3

Cartoon cross sections for the interiors of (a) Jupiter, (b) Saturn, and (c) Uranus/Neptune (the ice giants being too similar to merit separate figures)

hydrogen would rise to the outer regions of the planet. This gas appears to have come from the solar nebula. Uranus and Neptune must have formed largely in the presence of the solar nebula, a very stringent constraint on the formation of solid bodies.

It is often supposed that the presence or absence of a core in Jupiter (say) can be placed in one-to-one correspondence with the presence or absence of a nucleating body that caused the inflow of gas to form the much more massive envelope. However, there is no neat correspondence between mode of giant planet formation and current presence of a core. The nucleation model is still the favored model for making these planets, but a discussion of this is beyond the intent of this chapter (for more details, see Lissauer and Stevenson (2006)) (*Fig. 4-3*).

9 The Challenges

There are many possibilities that have not been discussed here. Gravity can involve more than the response of the planet to uniform rotation; perhaps one can see the signal of differential rotation or even convection. Certainly, the tides raised by large satellites may be detectable, and the gravitational signal of the Lense-Thirring effect (a general relativistic effect) may be seen by the Juno spacecraft. Perhaps the most intriguing and most important possibility lies in seismology, a technique that has been immensely important for understanding earth structure, there are tentative detections of normal modes for Jupiter that provide information on the internal structure (Gaulme et al. 2011).

Future developments in this field are likely to be in three areas: (1) improved observations of our planets by spacecraft, (2) improved thermodynamic data from theory and experiment, and (3) exoplanet measurements.

Cross-References

Planetary Magnetospheres

References

Chandrasekhar, S. 1939, An Introduction to the Study of Stellar Structure. Reprinted by Dover Publications, New York. ISBN: 0486604136

Fortney, J. J., Nettelmann, N. 2010, The interior structure, composition and evolution of giant planets. Space Sci. Rev., 152, 423–447

French, M., Becker, A., Lorenzen, W., et al. 2012, Ab initio simulations for material properties along the Jupiter adiabat. Astrophys. J., in press

Gaulme P., Schmider F. -X., Gay J., et al. 2011, Detection of Jovian seismic waves: a new probe of its interior structure. Astron. Astrophys., 531, Article number: A104. doi: 10.1051/0004-6361/201116903

Guillot, T. 2005, The interiors of giant planets. Annu. Rev. Earth Planet. Sci., 33, 493–530

Leconte, J., Chabrier, G. 2012, A new vision of giant planet interiors: impact of double-diffusive convection. Astron. Astrophys., A20. doi: http://dx.doi.org/10.1051/0004-6361/201117595

Lissauer, J. J., Stevenson, D. J. 2006, Formation of giant planets, in Protostars and Planets V, ed. B. Reipurth, D. Jewitt, K. Keil (Tucson: University Arizona Press), 591–606

Lodders, K, Fegley, B. 1998, The Planetary Scientist's Companion (New York: Oxford University Press), QB601 .L84

Stevenson, D. J. 2003, Planetary magnetic fields. Earth Planet. Sci. Lett., 208, 1–11

5 Atmospheres of Jovian Planets

Nancy Chanover
Astronomy Department, New Mexico State University, Las Cruces, NM, USA

1 Introduction ***224***

2 Atmospheric Composition and Structure ***227***
2.1 Cloud Locations 228
2.1.1 Limitations of Remote Sensing 231
2.2 In Situ Measurements 233

3 Atmospheric Dynamics ***234***
3.1 Winds 235
3.1.1 Observational Evidence for Seasonal Changes on Uranus and Neptune 236
3.2 Storm Features 239
3.2.1 Jupiter 239
3.2.2 Saturn 241
3.2.3 Uranus and Neptune 243

4 Atmospheric Chemistry ***244***
4.1 Energy Balance 245
4.2 Case Study: Shoemaker-Levy 9 Impacts on Jupiter 245

5 Future Directions ***246***
5.1 Unanswered Questions 246
5.2 Future Missions to the Outer Solar System 246
5.3 Links to Exoplanets 248

Acknowledgments ***248***

References ***248***

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_5, © Springer Science+Business Media Dordrecht 2013

Abstract: The giant planets of the solar system have been studied for centuries using a wide range of remote sensing and in situ techniques. An understanding of the atmospheres of Jupiter, Saturn, Uranus, and Neptune has dramatically improved since the dawn of spacecraft exploration of the outer solar system in the 1970s. Cloud decks that were predicted to exist from thermochemical equilibrium arguments have been observationally confirmed, although the exact vertical distribution of condensible species in these atmospheres remains an active area of study. All four of the giant planets have fast zonal (east-west) winds with prograde and retrograde jets, which dominate their atmospheric circulations. Each planet also contains long-lived cyclonic features or convective cloud features that appear and disappear on short timescales. These features suggest a link between the energy transport in the deep atmosphere and the visible cloud tops; the exact nature of this connection remains an outstanding question in giant planet atmosphere studies. The chemistry of the giant planet atmospheres is driven by both the convective processes that loft disequilibrium species from the deep atmosphere into the stratosphere and the interaction between stratospheric materials and ultraviolet sunlight. A unique opportunity to study these interactions was presented to planetary scientists in 1994, when the 22 fragments of Comet Shoemaker-Levy 9 impacted Jupiter. The future of giant planet atmospheric studies is promising. Several mission concepts that will answer fundamental questions regarding giant planet atmospheres are in various stages of development, and the James Webb Space Telescope will also contribute especially to our understanding of Uranus and Neptune. As an understanding of giant planet formation and evolution expands and deepens, these knowledge gains must be examined against the backdrop of the numerous exoplanet systems recently discovered, very few of which resemble our own.

1 Introduction

The gas giant planets differ in many ways from the rocky planets of the inner solar system. Historically, Jupiter, Saturn, Uranus, and Neptune have been considered to be a class of objects within the solar system unto themselves – they are large, made primarily of gas and ices, they each possess ring systems, and they are orbited by a plethora of moons. More recently, planetary scientists have made the distinction between the larger *gas giants*, Jupiter and Saturn, and the smaller *ice giants,* Uranus and Neptune. *Figure 5-1* contains an image montage of the four giant planets of our solar system, and *Table 5-1* lists some of the orbital and physical parameters for all four planets.

An examination of *Table 5-1* reveals obvious distinctions between the gas giants and ice giants in terms of size, mass, and composition. The masses of Uranus and Neptune are similar to each other, whereas Jupiter and Saturn have widely different masses that are both much larger than those of Uranus and Neptune. Given that the radii of Jupiter and Saturn are not vastly different, this suggests that Jupiter is much more internally compressed than Saturn. The gas giants, Jupiter and Saturn, are composed mostly of hydrogen and helium, while Uranus and Neptune are primarily composed of ices such as methane and water. All four of the giant planets rotate rapidly, resulting in oblateness values that are much larger than those of the terrestrial planets (e.g., the oblateness of Earth is 0.00034).

Two competing models of solar system formation, the *gravitational instability* and the *core accretion* models, are invoked to explain many of the bulk properties of the solar system as

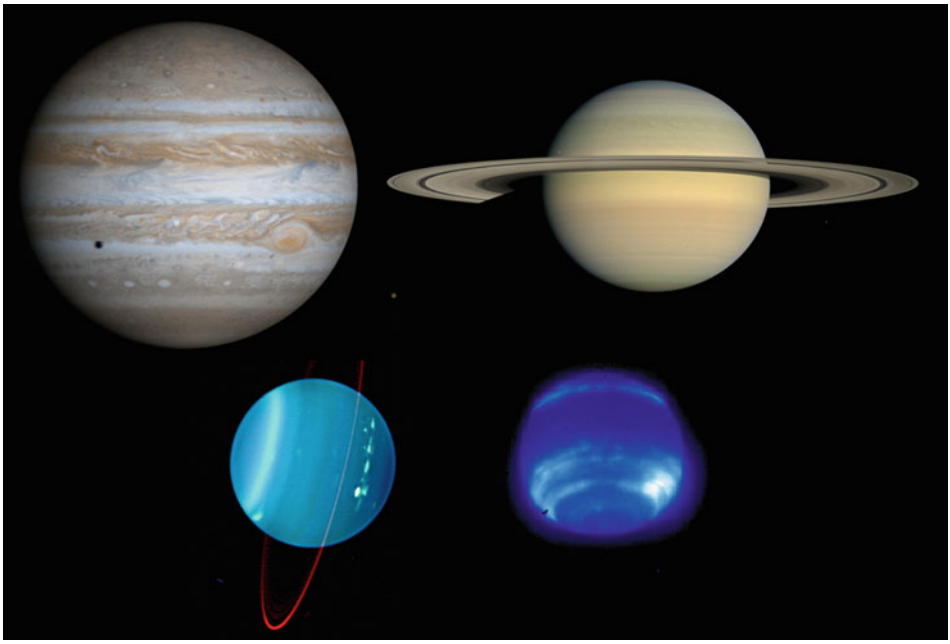


Fig. 5-1
Montage of the four giant planets of our solar system (not shown to scale). *Top left*: Jupiter image taken by the Cassini Imaging Science Subsystem on December 7, 2000 (image courtesy of NASA). *Top right*: image of Saturn taken by the Cassini Imaging Science Subsystem on July 23, 2008 (image courtesy of NASA). *Lower left*: Uranus image acquired with the Keck NIRC2 near-infrared camera on July 11–12, 2004 (courtesy of W. M. Keck Observatory). *Lower right*: image of Neptune acquired with the Keck II adaptive optics system and the KCAM near-infrared imager (Max et al. 2003)

well as the general distinctions between the gas giants and the ice giants. The gravitational instability model posits that the protostellar disk became dense enough to be gravitationally unstable and the giant planets formed directly through gravitational collapse of the circumstellar disk. The core accretion model proposes the growth of ice-rock giant planet cores through the collision of planetesimals, followed by gas accretion from the nearby regions of the protosolar nebula. The core accretion model provides a better explanation for the observed nonsolar (enriched) abundances of heavy elements in the outer solar system, but there are still problems with both models that have been highlighted by the discovery of numerous extrasolar planetary systems that do not resemble our own. Despite these new discoveries, the giant planets of our own solar system remain the best laboratories for studying solar system formation and evolution, and these areas have been driving scientific themes in the history of planetary exploration.

The giant planets of the outer solar system have been explored by spacecraft beginning in 1973 with the Pioneer 10 mission to Jupiter. Since that time, all four giant planets have been visited by the Voyager 2 spacecraft, and both Jupiter and Saturn have been explored by orbiters. ❯ *Table 5-2* lists the past, present, and planned missions to the giant planets in our solar system.

Table 5-1

Orbital, physical, and atmospheric data for the giant planets of the solar system. All data are from de Pater and Lissauer (2001), unless otherwise noted

Parameter	Jupiter	Saturn	Uranus	Neptune
Orbital semimajor axis (AU)	5.20	9.54	19.19	30.07
Mass (10^{24} kg)	1898.6	568.46	86.832	102.43
Equatorial radius[a] (km)	71,492	60,268	25,559	24,766
Oblateness[b]	0.065	0.098	0.023	0.017
Rotation period[c] (h)	9.924	10.543[d]	17.24	16.11
Obliquity (deg)	3.12	26.73	97.86	29.56
Effective temperature (K)	124.4 ± 0.3	95.0 ± 0.4	59.1 ± 0.3	59.3 ± 0.8
Geometric albedo	0.52	0.47	0.51	0.41
Energy balance[e]	1.67 ± 0.09	1.78 ± 0.09	1.06 ± 0.08	2.61 ± 0.28
H_2 volume mixing ratio	0.864 ± 0.006	0.963 ± 0.03	0.85 ± 0.05	0.85 ± 0.05
He volume mixing ratio[f]	0.157 ± 0.030	0.034 ± 0.03	0.18 ± 0.05	0.18 ± 0.05
CH_4 volume mixing ratio[f]	$(2.10 \pm 0.4) \times 10^{-3}$	$(4.5 \pm 2.2) \times 10^{-3}$	0.024 ± 0.01	0.035 ± 0.010

[a] Radius is defined to be from the planet center to the 1-bar pressure level

[b] Oblateness is a function of the equatorial and polar radii and is defined as $(R_e\text{-}R_p)/R_e$

[c] The rotation periods of Jupiter and Saturn are given in the System III system, which is referenced to the radio emissions of those planets

[d] Updated value from Cassini observations (Anderson and Schubert 2007); previous value was 10.656 h

[e] Energy balance is defined as the ratio of the energy radiated to space to the amount of solar energy absorbed

[f] From Niemann et al. (1998)

Table 5-2

Missions to the outer solar system

Spacecraft	Target body	Year
Pioneer 10	Jupiter	1973
Pioneer 11	Jupiter	1974
Pioneer 11	Saturn	1979
Voyager 1	Jupiter	1979
Voyager 2	Jupiter	1979
Voyager 1	Saturn	1980
Voyager 2	Saturn	1981
Voyager 2	Uranus	1986
Voyager 2	Neptune	1989
Ulysses	Jupiter	1992[a]
Galileo	Jupiter	1995–2003
Cassini	Jupiter	2000
Cassini	Saturn	2004–present
New Horizons	Jupiter	2007
Juno	Jupiter	2016[b]

[a] Heliophysics mission used to study Jupiter's magnetosphere

[b] Expected arrival at Jupiter based on its August 2011 launch

The spacecraft exploration of these large giant worlds of the outer solar system, along with the concomitant ground-based telescopic observations, has yielded numerous answers to outstanding questions about the giant planets and generated at least as many new unanswered questions. However, there are several fundamental areas of study related to giant planet atmospheres in which significantly increased understanding has emerged over the past three decades. The giant planet atmospheric structure, dynamics, and chemistry are key to understanding the physical processes that govern planetary atmospheres in general and will provide important insight into the atmospheres of the newly discovered planets around other stars, which are discussed further in Chap. 10. Section 2 of this chapter describes the structure and composition of giant planet atmospheres. In Sect. 3, the dynamical processes that influence these atmospheres are reviewed, and in Sect. 4, the role that atmospheric chemistry plays in giant planet atmospheres is discussed. A summary of unanswered questions and future directions in the study of giant planet atmospheres is presented in Sect. 5. The printed form of this chapter is an abridged version; a more extensive version is available in the online volume of this text.

2 Atmospheric Composition and Structure

Clues concerning the composition and vertical structure of the giant planet atmospheres can be found in the protosolar nebula. Based on the chemical abundances of the nebula, out of which the giant planets formed, thermochemical equilibrium models are used to predict the pressure-temperature regimes at which various molecules condense. Using such arguments, Weidenschilling and Lewis (1973) identified the following condensibles on the giant planets: for Jupiter and Saturn, cloud decks of ammonia, ammonium hydrosulfide (NH_4SH), and water are predicted (*Figs. 5-2* and *5-3*, respectively); for Uranus and Neptune, clouds of methane and

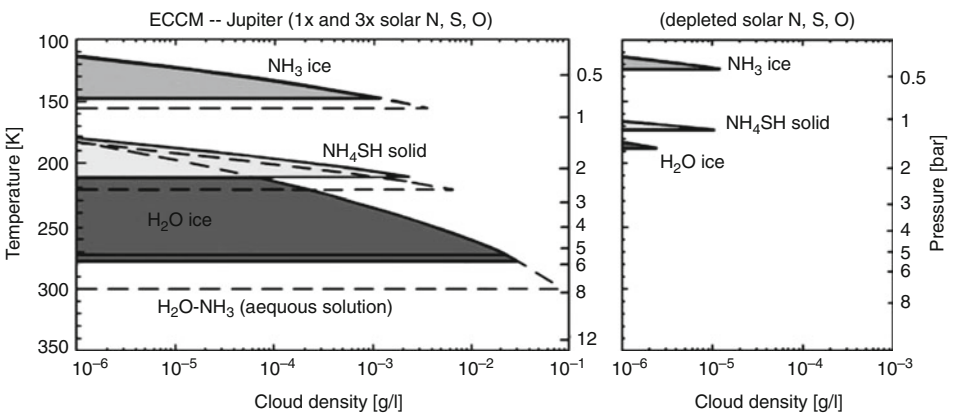


Fig. 5-2

Equilibrium cloud condensation model for Jupiter. *Left*: the pressure levels at which Jupiter's three cloud decks form were computed based on the assumption of 1× solar abundances (*solid area*) and 3× solar (*dashed lines*) values. *Right*: Jupiter's cloud deck locations were computed with the following condensible volatiles depleted relative to solar abundances: H_2O, NH_3, and H_2S (From Atreya and Wong (2005))

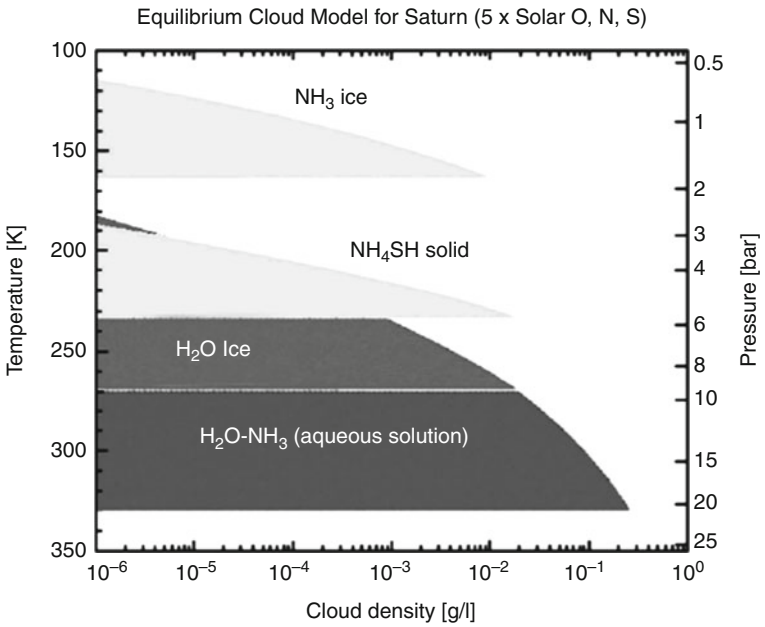


Fig. 5-3

Equilibrium cloud condensation model for Saturn. The pressure levels at which Saturn's cloud decks form were computed based on the assumption of 5× solar abundances of N, S, and O. Note that in Saturn's atmosphere the condensation of the same species occurs at greater pressure levels than on Jupiter due to the colder temperatures of Saturn's atmosphere. An increasingly larger enrichment in heavy elements as one goes from Jupiter to Neptune is consistent with the core accretion model of planet formation (From Atreya and Wong (2005))

water ice are predicted (*Fig. 5-4*). Methane is also present in the atmospheres of Jupiter and Saturn, in smaller abundances than in Uranus and Neptune, but the atmospheric temperatures of the gas giants are warm enough that methane never condenses there.

All of the giant planet atmospheres contain a temperature minimum located at the tropopause; above that, temperatures increase with height due to the absorption of sunlight by stratospheric aerosols and the methane gas absorption bands (*Fig. 5-5*). Below the tropopause, temperatures increase with increasing depth as the most efficient means of energy transport is convection and the atmospheric temperature profile is roughly adiabatic. This portion of the atmosphere is heated from below.

In this discussion, we adopt the atmospheric nomenclature defined by West et al. (2004) for the Jovian atmosphere, where a *haze* is defined as a ubiquitous layer of particles smaller than a micron in size and located in the upper troposphere (200–500 mbar) and in the stratosphere ($P < 100$ mbar). We refer to *clouds* when discussing more spatially and temporally variable assemblages of larger particles at deeper levels.

2.1 Cloud Locations

For several decades, the true chemical identities of the clouds in the giant planet atmospheres remained unconfirmed. Although on Jupiter and Saturn the uppermost cloud is assumed to be

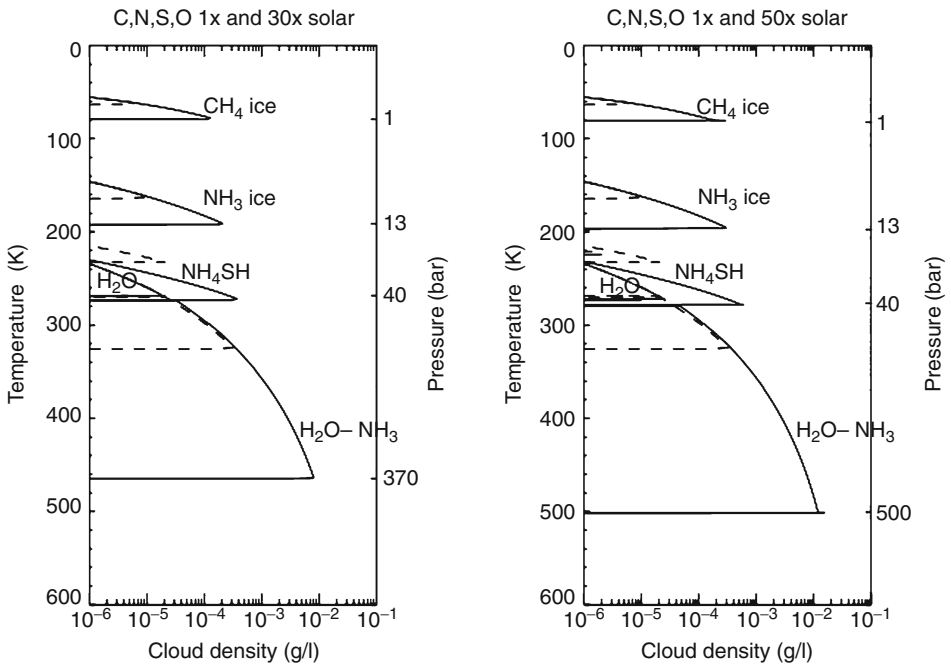


Fig. 5-4

Equilibrium cloud condensation model for Neptune, comparing the pressure levels at which Neptune's cloud decks are expected to form when computed assuming 1× (*dashed lines*) and 30× solar abundances (*left panel*) and 50× solar values (*right panel*) for the following condensible volatiles: H_2O, NH_3, H_2S, and CH_4. The atmospheric structure for Uranus is likely very similar to that of Neptune since both planets have very similar thermal structures and atmospheric compositions (From Atreya and Wong (2005))

made of ammonia ice based on the thermochemical equilibrium calculations described above, spectroscopic evidence of this remained elusive. It was not until the Galileo mission provided detailed views of the Jovian cloud deck with the Near-Infrared Mapping Spectrometer (NIMS) that spectrally identified ammonia clouds (SIACs) were detected in localized regions on Jupiter (Baines et al. 2002). A 3-μm absorption feature seen in both space-based and ground-based infrared spectra of Jupiter was initially attributed to ammonia (Brooke et al. 1998), but more recent improved model fits suggest a layer of small ammonia-coated particles overlying an optically thicker layer of larger NH_4SH particles (Sromovsky and Fry 2010). The fact that Jupiter's SIACs are short-lived, i.e., last on the order of a few days, seems at odds with the prediction that the upper cloud decks of Jupiter and Saturn are composed of ammonia ice, and suggests that some other process such as photochemical "tanning" or a coating of the pure ammonia ice is commonplace in the giant planet atmospheres.

Confirmation of Jupiter's water cloud also remained challenging. Water vapor was detected on Jupiter using spectroscopic observations from the Voyager and Galileo spacecraft (Carlson et al. 1992; Roos-Serote et al. 1998) as well as airborne telescopes flying high in Earth's atmosphere (Larson et al. 1975; Bjoraker et al. 1986). Since these observations probed to pressure levels ~5 bars in the jovian atmosphere, the presence of water vapor was not surprising given

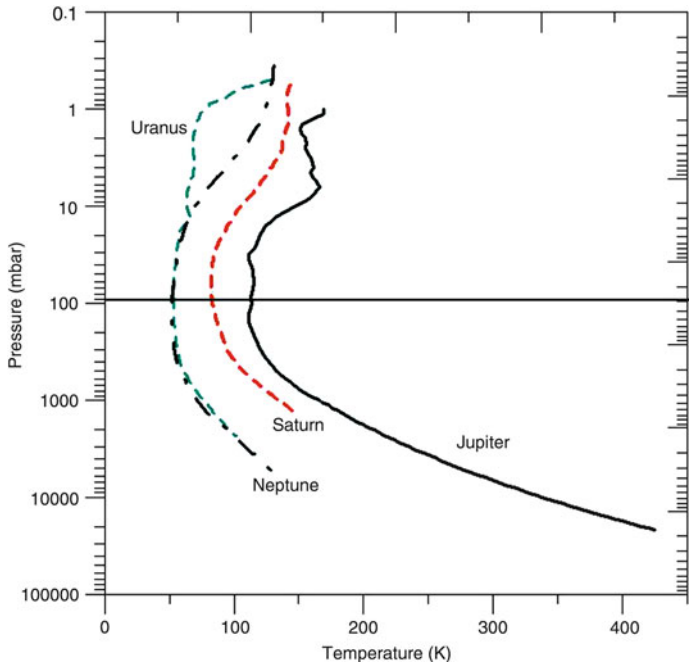


Fig. 5-5

Pressure-temperature profiles for all four giant planets. Figure from Sánchez-Lavega (2010); data for each profile were taken from Voyager occultation measurements by Lindal et al. (1981) (Jupiter), Lindal et al. (1985) (Saturn), Lindal et al. (1987) (Uranus), and Lindal et al. (1990) (Neptune). The Jupiter profile also includes data from the Galileo probe measurements by Seiff et al. (1998)

the thermochemistry believed to have taken place in Jupiter's atmosphere. However, it was also expected that Jupiter's vigorous convection should be strong enough to loft water vapor to higher and colder altitudes, where it would condense into ice and be seen above the ammonia cloud deck. Yet no ubiquitous water ice signatures were detected by the Voyager Infrared Interferometer Spectrometer and Radiometer (IRIS) instrument (Hanel et al. 1979).

The presence of water ice on Jupiter was not confirmed until vertical structure studies using Galileo Orbiter imaging data (Banfield et al. 1998; Gierasch et al. 2000) motivated a reexamination of Voyager data. The identification of convective thunderstorm clouds in Galileo and Voyager imagery prompted Simon-Miller et al. (2000) to reexamine the Voyager IRIS data. They found that approximately 1% of the spectra analyzed contained a far-infrared absorption feature attributable to water ice but that the optical depth of the water ice was not large enough to distinguish this cloud against the denser overlying ammonia cloud deck.

The vertical locations, thicknesses, and compositions of the cloud layers in the giant planet atmospheres are determined through vertical structure modeling, which has generally been conducted using two different approaches. Forward modeling techniques assume a basic cloud structure with some number of free parameters that describe the aerosol physical properties, layer thicknesses, and locations. These models use a radiative transfer code to generate synthetic spectra or center-to-limb (CTL) brightness variation curves. These spectra or CTL curves are

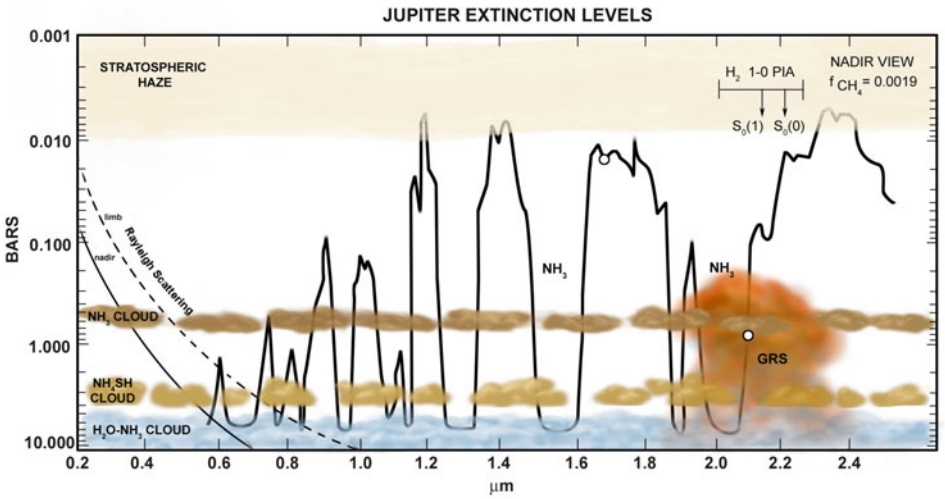


Fig. 5-6

Extinction levels in Jupiter's atmosphere as a function of wavelength from UV to near IR. Ammonia and H_2 pressure-induced absorption (PIA) features are indicated (From Baines (personal communication))

then compared with observations, and the free parameters in the model are subsequently varied iteratively until the synthetic curves or spectra that best match observational data are achieved. An alternative approach is an inversion technique, where inferences about an atmosphere's structure can be made from observational data.

Aerosol optical and radiative properties are determined from remote sensing measurements. Multiwavelength atmospheric radiance measurements are used as inputs to radiative transfer modeling codes, and the physical properties of the aerosol layers (e.g., particle size, optical depth) can be computed. *Figures 5-6* and *5-7* show the depths to which one can sound in the atmospheres of Jupiter and Saturn, respectively, as a function of wavelength.

2.1.1 Limitations of Remote Sensing

The atmosphere of Jupiter is the best-studied among the four giant planets. Jupiter's large size and relative proximity to Earth have enabled amateur and professional astronomers alike to conduct long-term monitoring of the Jovian atmosphere with first photographic and now CCD imaging techniques. A historical record of Jupiter observations can yield important insight into long-term changes in the giant planet atmospheres (Beebe et al. 1989), but such observations are limited in scope. As technologies such as adaptive optics and infrared imaging have improved, so has the demand for observing facilities with such capabilities. Thus, regular access to these new advances over decadelong time scales has remained a challenge for ground-based giant planet observers.

Historically amateur astronomers have made significant contributions to the study of giant planet atmospheres, earlier through their careful drawings and now more commonly with their photography and digital imagery of Jupiter and Saturn. With modest-sized telescopes

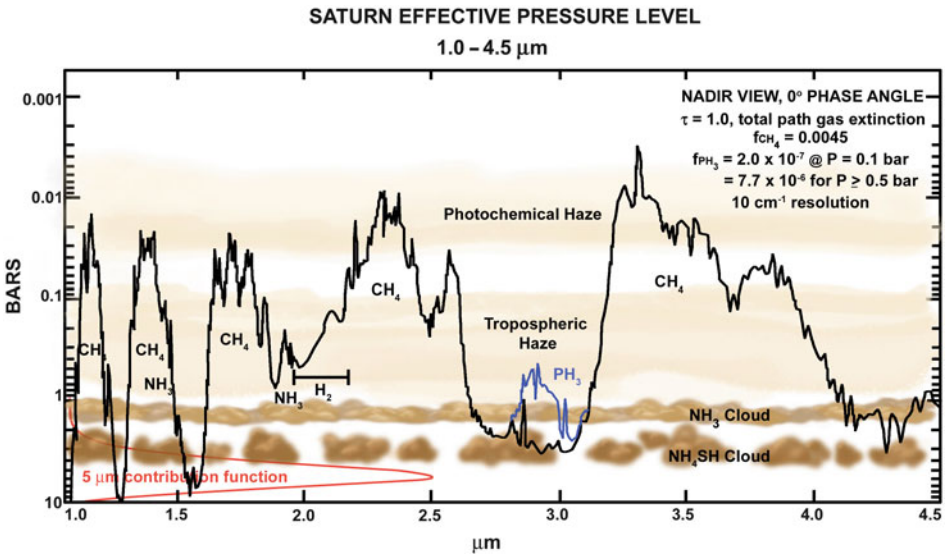


Fig. 5-7

Extinction levels in Saturn's atmosphere as a function of near-infrared wavelength. Ammonia, methane, phosphine, and H_2 pressure-induced absorption features are indicated (From Baines (personal communication))

(~25–35 cm mirror diameter) and rapid imaging cameras, amateur observers can produce very high-quality images by taking observations at a very high-time cadence. A small subset of these images is taken in very brief moments of excellent seeing. (This is sometimes referred to as the "lucky imaging" technique, cf. Law et al. 2006.) Postprocessing software is then used to identify the best images out of the gigabytes of data acquired in one night. The amateur observers regularly contribute their images to the Atmospheres Node of the International Outer Planets Watch through its online database, the Planetary Virtual Observatory and Laboratory (PVOL), where the data can be viewed by professional and amateur astronomers worldwide (Hueso et al. 2010). In addition to the long-term monitoring of atmospheric phenomena enabled by this rich historical record, amateur astronomers have been the first to report many significant phenomena in giant planet atmospheres, such as color changes (e.g., the reddening of White Oval BA) and bolide impacts. The PVOL images generally lack absolute photometric calibration and spectral information other than that afforded by broadband and sometimes a methane absorption filter, so their strength primarily lies in their extensive temporal coverage, which can elucidate dynamical changes in the atmospheres of Jupiter and Saturn.

The geometry of the solar system also limits the information one can glean from Earth concerning the physical characteristics (size, shape, scattering properties) of the giant planet aerosols. Large variations in phase angle, or the observer-target-Sun angle, are necessary to accurately quantify the scattering properties of the giant planet aerosols, yet from Earth we are limited to maximum phase angles of ~11°, 6°, 3°, and 2° for Jupiter, Saturn, Uranus, and Neptune, respectively. Only spacecraft observations can achieve much larger phase angles, and as can be seen from *Table 5-2*, those opportunities have been limited in number.

Finally, remote sensing observations of the giant planet atmospheres can probe a variety of pressure levels and aerosol layers by capitalizing on the vertical discrimination afforded

by various molecular absorption bands. It remains difficult to accurately sound below the visible cloud deck, though, due to the increasing opacity with depth. Models used to determine atmospheric structure based on cloud reflectivity measurements are often plagued with nonuniqueness problems because variations in several different parameters can produce the same observed effect. For example, to explain the contrasts seen in Jupiter's clouds at continuum wavelengths, two different modeling approaches have been used. Forward modeling studies such as those by West et al. (1986) and Chanover et al. (1997) find that the continuum contrasts are mostly due to spatial variations in the single-scattering albedo of the upper cloud, whereas models used by Banfield et al. (1998) and Sromovsky and Fry (2002) suggest that optical depth variations are largely the cause of cloud contrasts (West et al. 2004). More spacecraft observations over a wide range of wavelengths and phase angles are needed to break this degeneracy.

2.2 In Situ Measurements

The atmosphere of Jupiter is currently the only one of the four giant planets that has been explored via in situ measurements made from an entry probe. The Galileo probe entered Jupiter's atmosphere on December 7, 1995 and descended through the Jovian atmosphere, transmitting data for 61 min before contact was lost when the probe was near the 24-bar pressure level (Young et al. 1996) (*Fig. 5-8*). The probe was equipped with an instrument suite designed to characterize Jupiter's inner magnetosphere as well as the atmospheric cloud structure, chemical environment, and wind field. The lightning and radio emission detector (LRD) and the energetic particle instrument (EPI) were used to characterize the energetic particles in the innermost regions of Jupiter's radiation environment. The EPI data were acquired

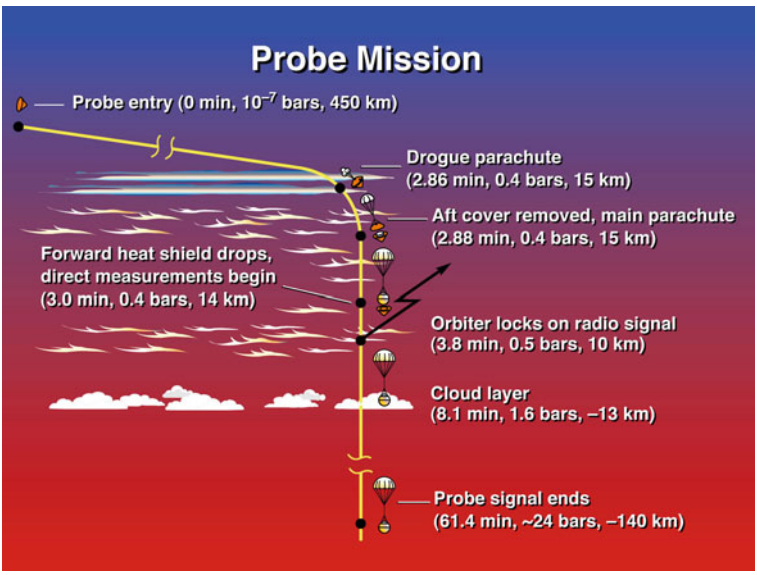


Fig. 5-8
Timeline of events associated with the Galileo probe entry and descent through Jupiter's atmosphere (Image courtesy of NASA)

during the preentry phase of the mission, whereas the LRD operated throughout the descent phase as well. The chemical composition of Jupiter's atmosphere was directly measured with two instruments. The helium abundance detector determined Jupiter's relative helium abundance with greater accuracy than previous estimates, which has important implications for giant planet and solar system formation theories. The neutral mass spectrometer measured mixing ratios of both major and minor species in the Jovian atmosphere as well as isotopic ratios. These in situ measurements were critical for determining the relative contributions of the protosolar nebula and icy planetesimals of the outer solar system to the giant planet atmospheres (Young 2003).

The vertical structure and optical properties of Jupiter's atmosphere were quantified using three instruments: the nephelometer, which was designed to measure the aerosol scattering properties and locations; the atmospheric structure instrument (ASI), which measured the atmospheric thermal structure and vertical winds; and the net flux radiometer (NFR), which was used to determine the vertical distribution of atmospheric heating and cooling and atmospheric opacity throughout the probe's descent phase. Finally, the Galileo probe mission included a Doppler wind experiment (DWE), which was not a separate instrument on board the probe. Rather, the Doppler delay of the probe relay carrier frequency was used to determine the deep zonal winds, i.e., below the cloud decks, in Jupiter's atmosphere.

The Galileo probe entered Jupiter's atmosphere at a latitude of 6.5° N in the North Equatorial Belt, a region characterized by bright white convective plumes and darker, relatively cloud-free, areas of subsidence between the plumes. It descended through Jupiter's atmosphere in one of the interplume downwelling regions (Orton et al. 1998), thus providing an unexpected and likely somewhat atypical picture of the Jovian atmosphere. The Galileo probe measured a helium abundance lower than that expected from the protosolar nebular predictions, which indicates that helium may have gravitationally settled deeper in Jupiter's interior (von Zahn et al. 1998). Neon was found to be severely depleted, while C, N, S, Ar, Kr, and Xe were all measured with ~3× solar abundance. Condensible species such as H_2O, NH_3, and H_2S were all depleted, which is likely due to the unique nature of the probe entry site (Niemann et al. 1998). The jovian cloud structure sensed by the Galileo probe contained a tenuous cloud with a base at 0.5 bar, a thin but well-defined cloud with a base at 1.4 bar, and another tenuous aerosol layer between ~2.4 and 3.6 bars (Ragent et al. 1998; Sromovsky et al. 1998). The lack of detection of a deep water cloud suggests that a dynamical process, i.e., a strong downdraft, was responsible for clearing this region of Jupiter's atmosphere. Jupiter's zonal (east-west) winds appeared to extend down at least to the depth where the probe data collection ceased (Atkinson et al. 1998), but this again may be due to the localized conditions in this downwelling region rather than a result that is broadly applicable to the entire planet. The Galileo probe results were key for improving the understanding of giant planet formation and evolution, and they highlighted the need for in situ measurements of giant planet atmospheres that are complementary to the richer orbiter and ground-based data sets.

3 Atmospheric Dynamics

The dynamics of the giant planet atmospheres are driven by a combination of solar input (which for Saturn, Uranus, and Neptune varies seasonally) and internal heat. Below the cloud decks, the atmospheres of the giant planets are believed to be adiabatic. One of the fundamental questions of giant planet atmospheric dynamics is the manner in which their atmospheric circulations are linked to the abyssal circulations.

3.1 Winds

The winds on the giant planets are the fastest in the solar system. *Figure 5-9* shows the mean zonal (east-west) wind profiles for each of the four Jovian planets. The profiles are all characterized by east-west jets, with prograde equatorial jets for Jupiter and Saturn, which also have a greater total number of jets, and retrograde equatorial jets for Uranus and Neptune. All four of the giant planets exhibit remarkable north-south symmetry in their zonal wind profiles. *Figure 5-10* shows Jupiter's zonal wind profile overlaid on a cylindrical map made from HST images. We see a strong correlation between the zonal wind jet maxima and minima and the banding of Jupiter's clouds, indicating that these jets are responsible for Jupiter's visible appearance.

The meridional (north-south) winds on the giant planets are several orders of magnitude weaker than the zonal winds, indicating that the giant planets, circulation is dominated by east-west motions. As shown in *Table 5-1*, there is considerable variation among the gas giants in terms of both solar insolation and internal heat. This suggests that the forcing of the atmospheric circulation in these planetary atmospheres is largely a function of their formation and rapid rotation.

Uranus is unique among the giant planets in that its equilibrium temperature is quite close to its true disk-averaged temperature (*Table 5-1*). This implies that any internal heat, if it exists, is inefficiently transferred to the observable weather layer. In contrast with Voyager 2

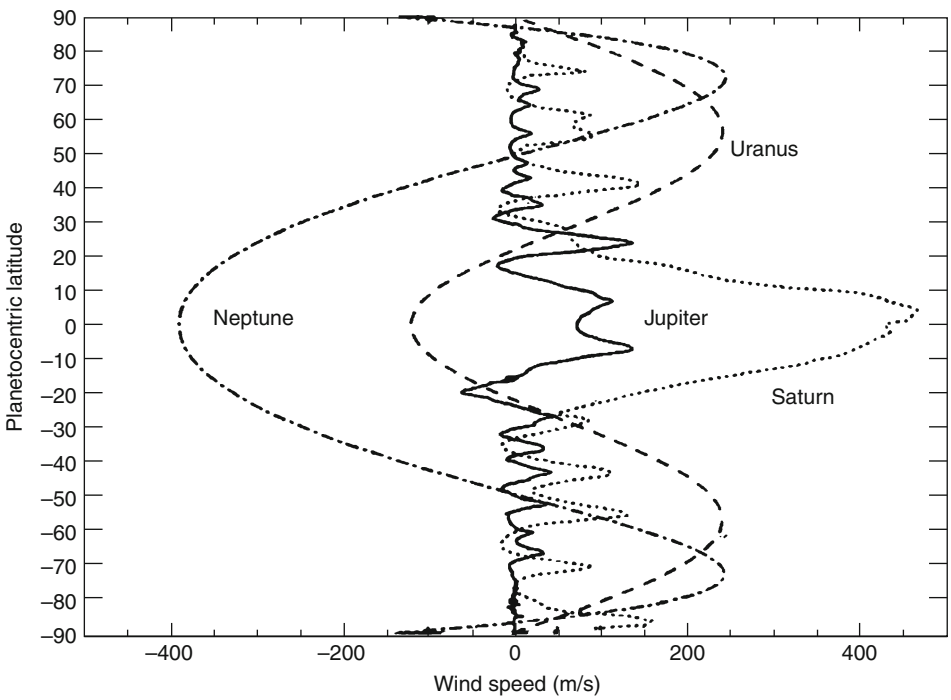


Fig. 5-9
Zonal wind profiles for all four giant planets (Image from Irwin (2009), Fig. 5.2)

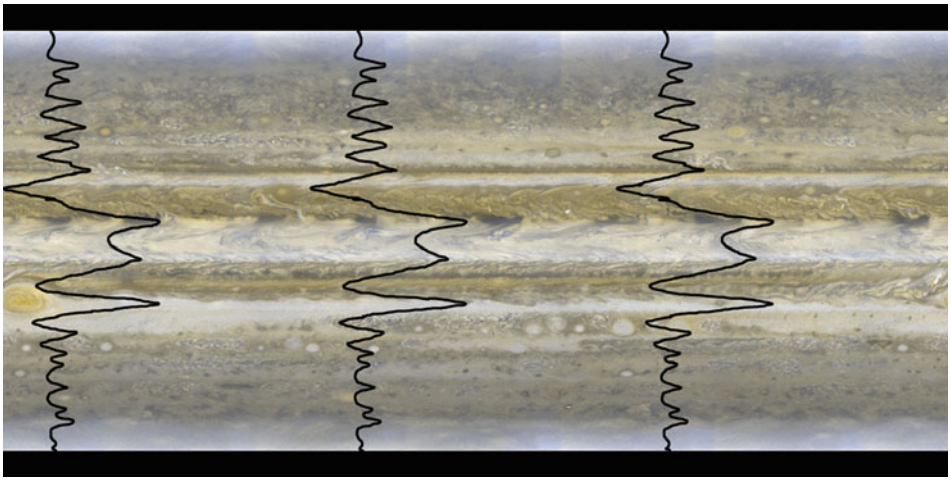


Fig. 5-10
Jupiter's zonal wind profile overlaid on a cylindrical map made from HST images of Jupiter acquired in October 1996 (From Simon-Miller (personal communication))

measurements, which show an adiabatic lapse rate below approximately the 0.1 bar level, deep interior models of Uranus suggest strong molecular gradients that should inhibit convection (Guillot 1999). Thus, even if internal heat within Uranus is substantial, the planet is prevented from following the usual path of heat transfer from its interior used by other gas giants.

It is expected that solar insolation plays a considerable role in the energy budget of Uranus' weather layer. Uranus' ratio of total emitted infrared irradiance to that of absorbed solar heating, as shown in *Table 5-1*, is significantly less than that found on the other giant planets. Combined with Uranus' extreme axial tilt, seasonal variations of atmospheric dynamics are expected to be significant in the radiative boundary layer and potentially deeper. The current epoch represents a unique opportunity for us to analyze seasonal forcing of atmospheric dynamics, as Uranus reached its equinox in December 2007. Uranus' last equinox took place in 1965, before the advent of modern detectors, thus we are currently in a "new" age for Uranus atmospheric studies, since both its northern and southern hemispheres are receiving roughly equal amounts of sunlight for the first time in 42 years. Neptune, on the other hand, has the largest internal heat contribution to its outgoing flux of all the giant planets. Thus, a comparative study of the role that insolation plays in driving the atmospheric dynamics on the planets with the two extrema in energy balance can reveal important clues to their divergent evolution.

3.1.1 Observational Evidence for Seasonal Changes on Uranus and Neptune

At first glance, a comparison between recent images of Uranus taken near equinox and Voyager 2 images of the planet taken near solstice immediately reveals obvious changes in Uranus' atmosphere (*Fig. 5-11*). Recent near-infrared imaging with the Keck AO system and the Hubble Space Telescope (Sromovsky and Fry 2005; Hammel et al. 2005) shows an increase in the number of discrete cloud features when compared with 1986 Voyager 2 visible imaging, particularly in the late winter northern hemisphere. However, this comparison is likely

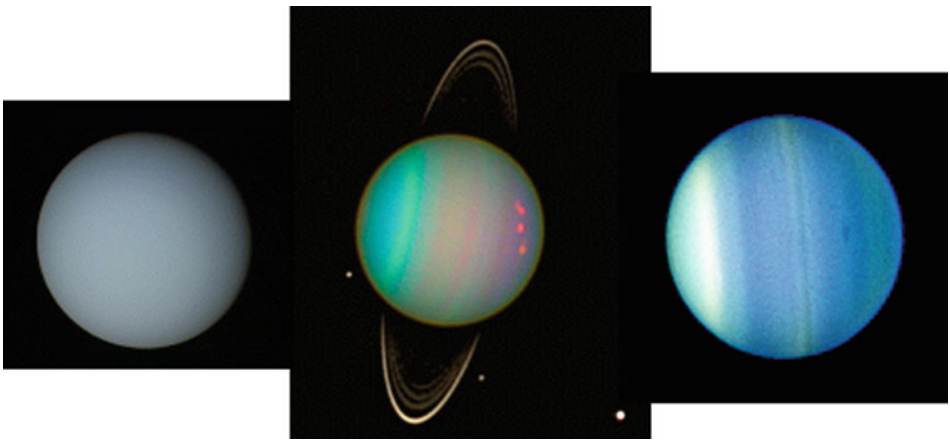


Fig. 5-11
***Left*: image of Uranus taken by the Voyager 2 spacecraft in 1986. *Middle*: image of Uranus taken with ACS on Hubble Space Telescope in 2003. *Right*: image of Uranus taken with ACS on Hubble Space Telescope in 2006 (Images courtesy of NASA)**

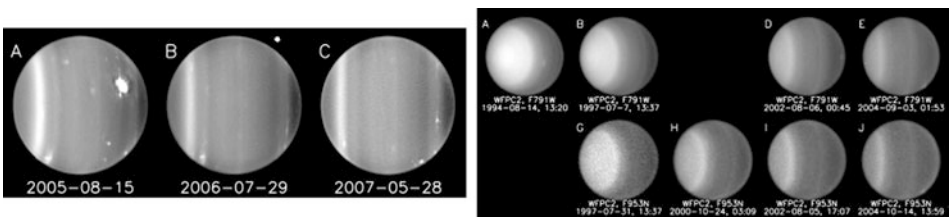


Fig. 5-12
***Left*: near-IR Keck images of Uranus showing a new bright band forming in Uranus' northern hemisphere, which is seen on the right side of each image. *Right*: HST images illustrating the evolution of Uranus' south polar cap from 1994 to 2004 (All images courtesy of Sromovsky (personal communication))**

oversimplified: some of this increase is certainly due to the higher contrast of clouds in the NIR as well as improved visibility of the equatorial latitudes and northern hemisphere as Uranus approaches equinox (Karkoschka 2001). Nonetheless, seasonal change may also play a role in the increase of observable cloud features.

A more compelling argument that seasonal change is occurring on Uranus is the observed asymmetry in the cloud bands, as seen in the right panel of *Fig. 5-11* as well as in *Fig. 5-12*. The bright band in the southern hemisphere is probably not a permanent feature of Uranus' atmosphere; any fixed pattern would most likely be symmetric about the equator to match the symmetry of the annual average distribution of solar heating within the atmosphere. In fact, a north-south asymmetry is just what would be expected for a seasonal response to solar forcing when that response has a time constant that is long enough to cause a large delay in response but not so long that the response is completely washed out. If the cloud pattern

on Uranus does indeed manifest seasonal change, it should have begun to reverse just before equinox. In fact, that reversal is being seen: the southern band is declining in brightness and a new band is forming in the northern hemisphere (Rages et al. 2007; *Fig. 5-12*, left panel). The bright polar cap that used to be an obvious feature of the southern hemisphere began to decline even earlier (Rages et al. 2004; *Fig. 5-12*, right panel), which may indicate a seasonal response with an even shorter time constant.

In addition to the more frequent appearance of discrete clouds and changes in large-scale banded cloud patterns, Uranus' zonal winds may also demonstrate seasonal change. While the Sromovsky and Fry (2005) observations, acquired in 2003 and 2004, show no clear evidence of seasonal change in the zonal winds, Hammel et al. (2005) do find significant changes in wind speeds from 2000 to 2003, with an increase of approximately 30 m/s in the 20–50°N latitude range. This issue requires further study with new observations spanning a wider seasonal coverage, but the possible linkage between zonal winds and seasonally varying insolation is tantalizing.

Radio observations of Uranus at 2 and 6 cm were also suggested to show evidence of seasonal change (Hofstadter and Butler 2003). These wavelengths probe much deeper than visible imaging, down to approximately 50 bars. A pole-to-equator radio brightness gradient has been known to exist even prior to the Voyager 2 flyby. However, while this gradient remained relatively constant from 1981 to 1989, it may have undergone an increase in magnitude between 1989 and 1994. Such a gradient is generally interpreted as an equatorial enhancement in condensable gases such as NH_4SH rather than a true temperature gradient (de Pater et al. 1991). This may imply a Hadley-like circulation in which upwelling gases near the equator condense out while depleted gases are subsiding near the pole (*Fig. 5-13*). Thus, an increase in the magnitude of this brightness gradient may indicate a seasonal invigoration of this circulation.

Observations of Uranus between 1 mm and 20 cm do not reveal any *hemispheric* asymmetries, only pole-to-equator variations (Hofstadter, personal communication), suggesting that the processes governing ice giant radio brightnesses differ from those causing the observed variations in reflected sunlight. Neptune, on the other hand, does exhibit hemispheric asymmetries

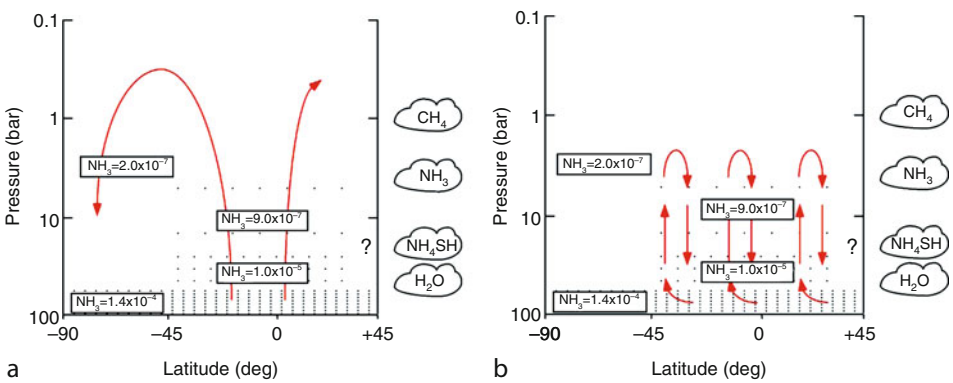


Fig. 5-13

(a) The distribution of absorbers in Uranus' atmosphere required to fit 1994 radio observations. The suggested Hadley-type circulation cell can explain the observed latitudinal gradient. (b) An alternative proposed circulation pattern (Figure from Hofstadter and Butler (2003))

at radio wavelengths. In the early 1990s both of Neptune's polar regions were brighter than its equator. Today, close to southern summer solstice, Neptune's south pole is still radio bright, but the north pole is not visible (Hofstadter, personal communication). The link between seasonal insolation and ice giant atmospheric dynamics clearly warrants further study.

Uranus' extreme axial tilt results in an unusual hemispheric asymmetry in solar energy deposition. Neptune, on the other hand, is an ice giant planet with a more moderate axial tilt of nearly 30°. Although its energy balance is more than twice that of Uranus (*Table 5-1*), Neptune also exhibits long-term atmospheric variations, as evidenced by an increase in disk-averaged brightness (Lockwood and Thompson 2002; Sromovsky et al. 2003). A lagged seasonal model was invoked by Sromovsky et al. (2003) to explain Neptune's long-term brightness changes, which they suggested were caused by Neptune's changing subsolar latitude. However, the long-term photometric record suggests that something other than seasonal change is at work; previous investigators also have considered Neptune's response to the 11-year solar cycle or its changing heliocentric distance (Lockwood and Jerzykiewicz 2006; Hammel and Lockwood 2007) as causes for Neptune's observed temporal variations. It is clear that with a much larger internal heat source than Uranus, along with a weaker solar flux at its greater orbital distance, the role of solar insolation in driving atmospheric dynamics on Neptune may be different than for Uranus.

3.2 Storm Features

Convective activity in the giant planet atmospheres is manifested through the apparition of storm features over a wide range of size scales. These features vary in diameter, latitudinal and longitudinal extent, and longevity on each of the giant planets due to the unique aspects of each atmosphere.

3.2.1 Jupiter

The Jovian atmosphere is teeming with cyclonic and anticyclonic storm features. Several of the largest of these features are visible even with low-power ground-based telescopes. The most famous of these storms is the Great Red Spot (GRS), which is more than twice the size of Earth and has been in existence at least since it was first observed more than 150 years ago (*Fig. 5-14*). It is a high-pressure, anticyclonic, counterclockwise-rotating storm system that is maintained at a planetocentric latitude of roughly −25° by a strong prograde jet to its south and a strong retrograde jet to its north.

South of the GRS at approximately −30° latitude, three smaller anticyclonic white ovals, named FA, BC, and DE, formed out of the South Temperate Zone (STZ) in 1939. They were first seen following a period when the STZ rapidly clouded over and became a white band; once the clouds subsided, they coalesced into three discrete, rotating storm systems (Peek 1953). In 1998, BC and DE merged to form Oval BE, and in 2000 BE and FA merged to form a single White Oval, BA. The color of BA changed from white to red in 2005 (*Fig. 5-15*). The cause of this color change is not clear, but it may be due to changing dynamics within the oval that resulted in the exposure of reddish condensation nuclei that were lofted from Jupiter's deeper atmosphere Cheng et al. (2008), Perez-Hoyos et al. 2009.

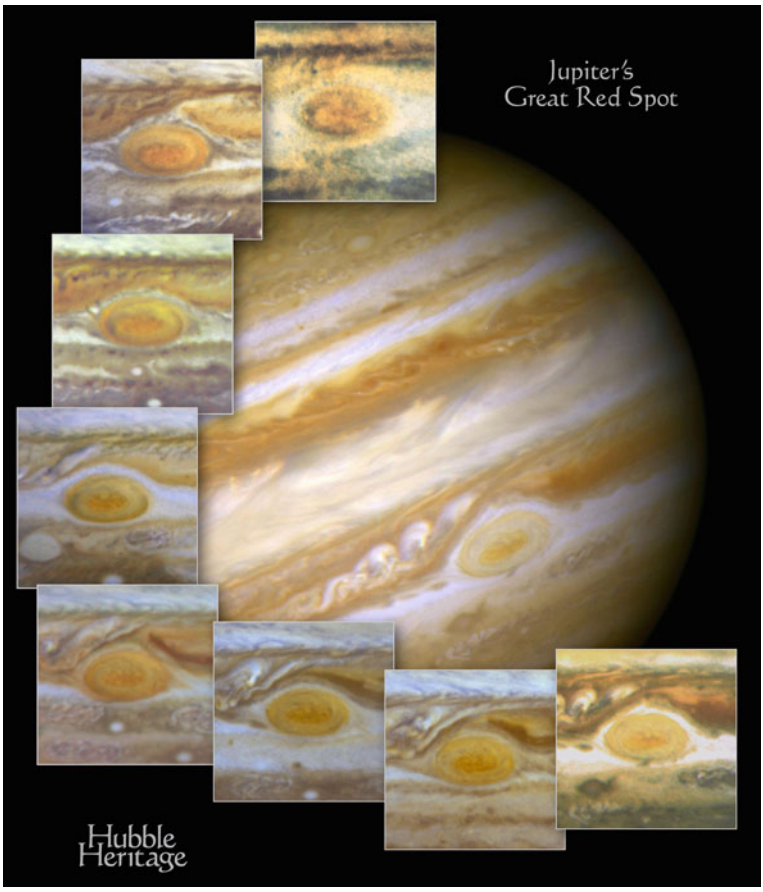


Fig. 5-14
Montage of image of Jupiter's Great *Red Spot* acquired with the Wide Field Planetary Camera 2 on Hubble Space Telescope (Image courtesy of NASA)

The coloration of the GRS and, more recently, Oval BA remains an outstanding mystery in studies of the Jovian atmosphere. West et al. (1986) summarized the state of knowledge at that time by providing two lists of candidate compounds – both organic and inorganic – that could be responsible for the coloration of Jupiter's clouds. These included hydrogen sulfides, allotropes of phosphorus and sulfur, and irradiated mixtures of hydrogen, methane, and ammonia ices. There has been relatively little laboratory work done at conditions appropriate for Jovian pressures and temperatures to confirm or rule out these candidate materials as coloring agents in Jupiter's atmosphere. Two competing hypotheses have been invoked to explain the sources of the coloring agents in the giant planet atmospheres: compounds containing sulfur, phosphorus, hydrogen, and nitrogen that have been convectively transported upward (West et al. 1986) and the coating of ammonia ice particles by photochemically produced hydrocarbons (Atreya et al. 2005; Kalogerakis et al. 2008). High quality multispectral observations of Jupiter's atmosphere, analyzed with numerical techniques such as Principal Component Analysis, indicate

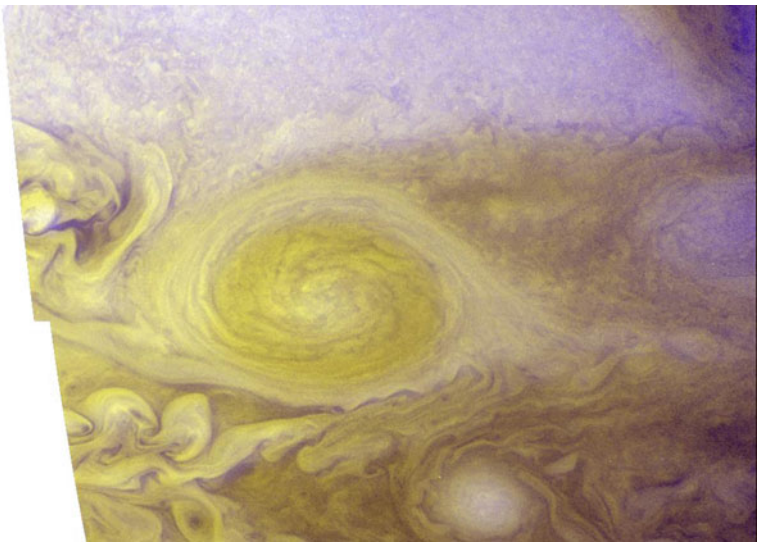


Fig. 5-15
Image of Jupiter's "*Little Red Spot*," or Oval BA, acquired by the LORRI imager on New Horizons on February 27, 2007 (Image courtesy of NASA)

that several chromophores are needed to explain the spatial variations in the Jovian cloud colors (Simon-Miller et al. 2001a, 2001b; Strycker et al. 2011). However, the exact nature or identity of these chromophores remains unclear. Further analysis that relates particle microphysics, atmospheric dynamics, and local radiation field will shed light on this issue. It is important that we understand the nature of the giant planet chromophores because they are almost surely linked to the dynamics and photochemistry of those atmospheres and may reveal a production mechanism for organic molecules.

3.2.2 Saturn

Imaging Saturn's atmospheric features at high spatial resolution presents a challenge for ground-based telescopes. Until the advent of the CCD cameras in the 1980s, only about a dozen features were clearly detected visually or photographically over a century of observations (Sánchez-Lavega 1982). Typically ground-based CCD imaging in the visual range can detect cloud features on Saturn larger than about 3,000 km if they are located at the sub-Earth point. In addition, they must have sufficient contrast relative to surrounding clouds to be detected. This is difficult on Saturn due to the extinction of reflected sunlight by a dense high-altitude haze layer. The optimal contrast in the optical/CCD wavelength range is found in blue-green and 890 nm methane-band images.

The Voyager 1 and 2 encounters in 1980 and 1981 provided high-resolution images of Saturn's cloud morphology but in a limited spectral range from violet to red wavelengths. Voyager images showed that isolated features are rare in Saturn's atmosphere, but other interesting atmospheric features were seen in the north such as the "ribbon wave" (Sromovsky et al. 1983) and

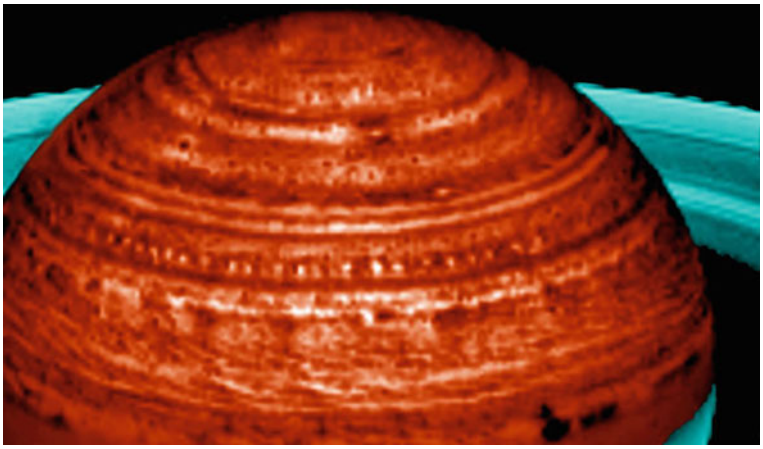


Fig. 5-16
Image of Saturn's "string of pearls" cloud formation, taken with Cassini's visual and infrared mapping spectrometer on April 27, 2006 (Image courtesy of NASA)

the "polar hexagon" (Godfrey 1988). Newer Cassini data revealed high quality observations of Saturn's north polar hexagon and vortices at both poles (Baines et al. 2009, Dyudina et al. 2009), as well as additional phenomena such as the "string of pearls" (*Fig. 5-16*), which shows regularly spaced clearings in the clouds (the bright regions in the infrared image in *Fig. 5-16*). This appears to be a manifestation of a planetary-scale wave and is likely linked to the deeper circulation in Saturn's atmosphere.

Saturn's atmosphere also undergoes localized convective activity roughly every 30 years, which corresponds to approximately once per Saturnian year. These convective outbursts are manifested by the development of gigantic storm systems known as Great White Spots (Sánchez-Lavega 1982). A major event occurred in the northern equatorial region in 1990 and was studied both from the ground and with Hubble Space Telescope (Sánchez-Lavega et al. 1991; Beebe et al. 1992; Barnet et al. 1992; Westphal et al. 1992). The cloud tops of this storm were convectively lofted to more than a scale height above the surrounding cloud deck, and the resultant increase in cloud optical depth of Saturn's equatorial region persisted for several years after the outbreak.

More recently, a major storm outbreak occurred in the northern midlatitudes. It was first detected through ground-based observations on December 5, 2010, at a planetographic latitude of 38° and, within the span of 1 week, expanded to a linear size of ~8,000 km (Sánchez-Lavega et al. 2011). This time, the Cassini spacecraft was wellpositioned in its orbit around Saturn to study the storm and its evolution with its imaging and spectroscopic instrument suite. *Figure 5-17* shows the storm as imaged by Cassini's Imaging Science Subsystem roughly 2.5 months after the storm was first detected. Thermal infrared imaging and spectroscopy revealed that this storm penetrated vertically well into Saturn's stratosphere, which resulted in heating as the material fell back onto Saturn's upper atmosphere (Fletcher et al. 2011). Outbreaks such as these provide a unique opportunity to study the zonal wind structure and vertical energy transport in Saturn's atmosphere.

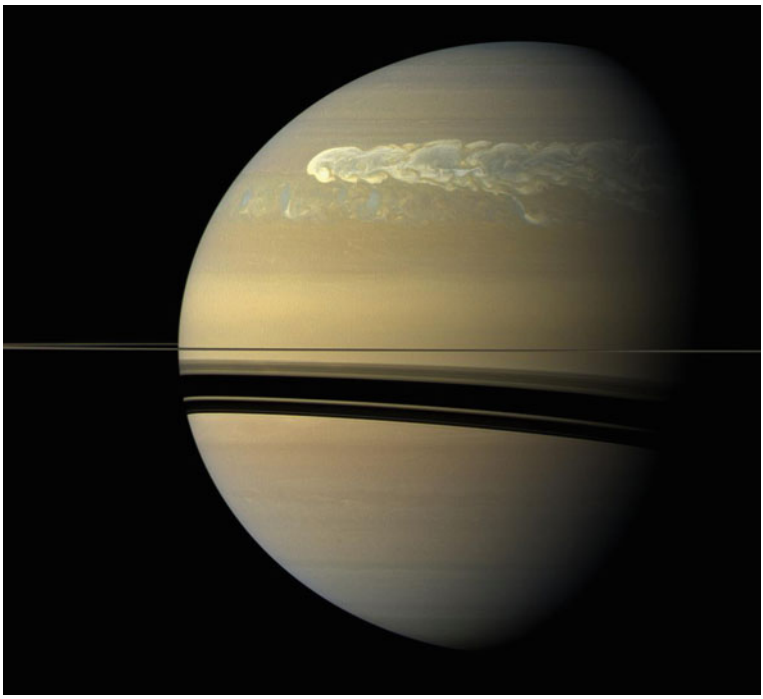


Fig. 5-17
Image of the recent storm in Saturn's northern hemisphere, taken with Cassini's Imaging Science Subsystem wide-angle camera on February 25, 2011 (Image courtesy of NASA)

3.2.3 Uranus and Neptune

When Voyager 2 flew by Uranus in 1986 during southern summer solstice, the planet lacked significant storm clouds and atmospheric features. Only through a radical adjustment of image contrast and dynamic range could several faint storm systems be discerned in the Voyager 2 data. However, over the past decade the atmosphere of Uranus has become much more active, now revealing numerous short-lived convective storm features. The cause of these changes is not entirely clear, although seasonal variation in insolation is likely a significant factor. Since Uranus reached equinox in December 2007, over the past decade its northern hemisphere began to receive direct sunlight after a period of several decades in darkness.

The application of new technologies to the study of Uranus' atmosphere also aided in the detection of new small storm features. The Keck telescopes, equipped with adaptive optics system, have enabled diffraction-limited imaging of Uranus in the nearIR, where the contrast between convective storm features and the methane clouds is greatest. The Hubble Space Telescope has also provided high spatial resolution imagery of the Uranian atmosphere from optical through near-IR wavelengths, including the first detection of a dark spot on Uranus in 2006 (Hammel et al. 2009). A large bright complex of clouds was seen to migrate in latitude with an inertial oscillation superimposed on the zonal flow (Sromovsky et al. 2007), and another long-lasting southern feature migrated toward the equator, where it met its eventual demise

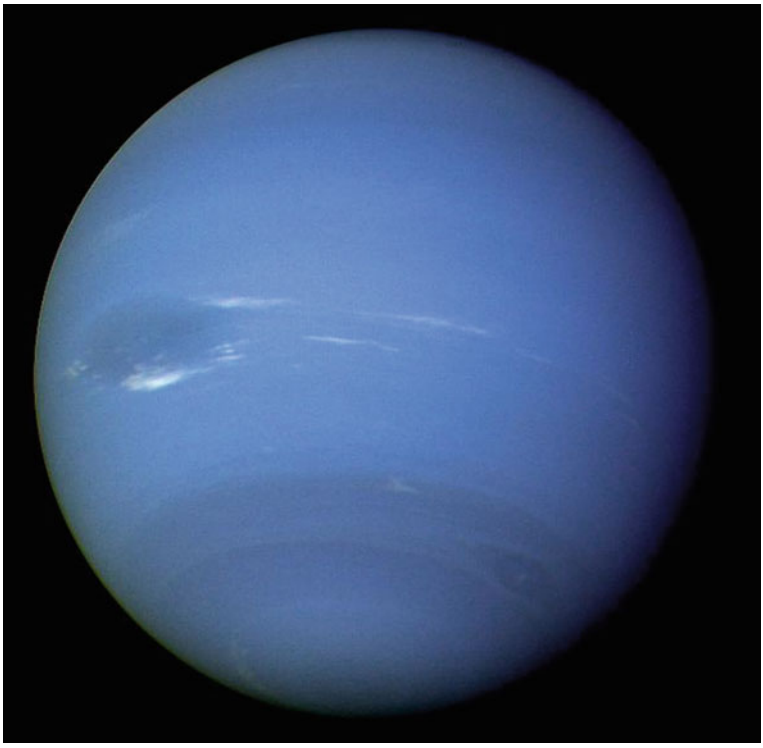


Fig. 5-18
Image of Neptune taken by the Voyager 2 narrow-angle camera on August 16 and 17, 1989 (Image courtesy of NASA)

(de Pater et al. 2011). These interactions hint at a coupling between giant planet zonal circulation and convective motion, which should be explored further as the planet moves toward its next solstice.

Neptune's atmosphere has demonstrated convective activity regularly since the time of the Voyager 2 encounter in 1989. A large disturbance dubbed the Great Dark Spot was seen in Voyager 2 imagery (*Fig. 5-18*), but this disturbance along with its accompanying white clouds did not persist for decades; they were gone by the time HST observed Neptune in 1994 (Hammel et al. 1995). Neptune is also expected to undergo seasonal variations due to its nonzero axial tilt, and continued observations over the next several years will likely reveal new features and changes in zonal wind structure due to its slowly varying insolation.

4 Atmospheric Chemistry

The distribution of chemical species in the atmospheres of the giant planets is governed by several different processes depending on altitude. In the troposphere, where the temperature gradient is negative, convection is the dominant means of energy transport. Species transport

is relatively rapid; molecules that are formed at depth through thermochemical equilibrium reactions and condensation can be lofted to altitudes where their interactions with ultraviolet sunlight can change their properties. An example of this kind of reaction is the proposed "tanning" of ammonia ice particles by solar UV radiation after they have been transported vertically from deeper in the atmosphere (discussed above as a proposed mechanism for the coloration of Jupiter's reddish clouds).

In the stratosphere, above the temperature minimum, the temperature gradients in the giant planets are positive and/or isothermal, which inhibits vertical motions and results in slow vertical transport of species. In this case, photochemical disequilibrium processes play a significant role in determining the composition of the giant planet stratospheres. For example, gaseous methane in Jupiter's atmosphere undergoes photochemical reactions to produce hydrocarbons such as ethane (C_2H_6), acetylene (C_2H_2), and ethylene (C_2H_4).

4.1 Energy Balance

The temperatures in the stratospheres of the giant planets can be derived using both indirect and direct techniques. A commonly used indirect method is the observation of occultations. An atmospheric occultation is where the attenuation of a source, such as a background star or a spacecraft radio signal, is observed as the planetary atmosphere passes in front of (or occults) the source. Atmospheric occultations provide measurements of the density structure of an atmosphere through the inversion of the observed light curves. From the density information, a temperature profile can be inferred. This indirect method of obtaining stratospheric temperatures was highly effective for obtaining temperature profiles of all four giant planets during the Voyager missions (shown in *Fig. 5-5*). Stellar occultations continue to be employed by orbiters (e.g., Cassini) and ground-based observers as a means of exploring temporal and spatial variations in the atmospheric temperature structure of the giant planets.

Temperatures in the giant planet stratospheres can also be obtained through observations of emission spectra in both the UV and IR spectral regimes. Such observations reveal that the stratospheric temperatures of the giant planets are remarkably uniform with latitude, which suggests that the meridional transport of radiation at these altitudes is very efficient. The energy balance of the giant planet stratospheres is controlled by the heating and cooling by photochemically produced species. For example, on Jupiter, ethane emission is the dominant cooling mechanism between pressures of ~0.2 and 20 mbar.

4.2 Case Study: Shoemaker-Levy 9 Impacts on Jupiter

Photochemical models indicate that methane photochemistry dominates the production of hydrocarbons in the giant planet stratospheres. Yet the detection of some species, such as H_2O and CO, cannot be explained through traditional photochemical pathways, suggesting that the source of these materials must be exogenic. Delivery mechanisms such as comets and interplanetary dust particles have been suggested as a means of supplying the giant planet stratospheres with these molecules (Moses et al. 2000 and references therein).

In 1994, planetary scientists worldwide had a unique opportunity to view such a delivery event in real time. Comet Shoemaker-Levy 9, which had passed close enough to Jupiter in 1992 to be broken up into 22 fragments, impacted Jupiter between July 16 and 22, 1994.

Observations of these impacts were made with all available ground-based and space-based telescopic assets, including HST and Galileo, which resulted in new insight concerning cometary structure and composition, ballistic impact physics, and Jupiter's atmospheric response to these impacts. The temporal evolution of the particulates associated with each impact confirmed the decay of Jupiter's zonal winds with height, while the spectroscopic measurements confirmed that cometary impacts are likely significant contributors to Jupiter's stratospheric CO budget (Lellouch et al. 1997). Although impact events such as this – where astronomers worldwide had more than a year to prepare for their observations – are likely exceedingly rare, the unique physics and chemistry probed by the SL9 impacts highlighted the importance of regular monitoring observations, rapid response capabilities, follow-up observations, and international coordinated efforts where possible.

5 Future Directions

5.1 Unanswered Questions

Fueled in part by the fantastic new discoveries made in the field of giant planet atmospheres over the past several decades, numerous outstanding questions remain. In particular, we hope that over the next few decades, progress will be made in addressing the following issues:

- The water abundance in the atmospheres of the gas giants
- The source of the coloration in the clouds of Jupiter
- The linkage between the giant planets' zonal circulations at the cloud-top level and the deep abyssal circulation
- The role that seasonally varying insolation plays in the vertical structure, cloud chemistry and microphysics, and dynamics of giant planet atmospheres

Advances in understanding these issues can be made through additional observations as we continue to push the instrumental limits of spectral grasp, spectral resolution, spatial resolution, and temporal coverage. Computational modeling of atmospheric dynamics on all scales, ranging from the smallest vortices to the global circulations, will also be critical for advancing the understanding of giant planet atmospheres.

5.2 Future Missions to the Outer Solar System

The Cassini spacecraft reached Saturn in 2004 and continues to provide magnificent views of Saturn's atmosphere. The nominal Cassini mission was scheduled to operate from 2004 to 2008. Due to its extraordinary success, NASA extended it for two more years for the Cassini Equinox mission (2008–2010). It was recently extended again for the Cassini Solstice mission and is currently scheduled to operate in this third phase from 2010 to 2017. The second extension of the Cassini mission will enable planetary astronomers to study the atmosphere of Saturn over half of a year, from northern winter solstice to northern summer solstice, with unprecedented detail. Seasonal changes in Saturn's atmosphere are thought to be caused by the rings, which cast shadows on the atmosphere and shield some of the planet from direct sunlight. The instruments on board Cassini can asses seasonal changes in atmospheric temperatures, composition, and dynamics, and further study new questions about Saturn's atmosphere that arose during the

Cassini Prime mission (e.g., looking for seasonal change in lightning activity rates or in the structure of Saturn's south polar vortex). The final 42 orbits of the 155-orbit Solstice mission will be spent in the proximal orbit phase, where the orbit of the spacecraft will transition to a more elliptical, polar orbit where the spacecraft regularly passes through Saturn's ring plane. This will enable the study of Saturn's internal structure in a fashion analogous to the future Juno mission to Jupiter and will enable a comparative study of the interiors of the two large gas giant planets in our solar system.

The Juno mission to Jupiter was launched in August 2011 and will reach Jupiter in 2016, at which point the spacecraft will enter a highly elliptical polar orbit. Juno will orbit Jupiter 32 times over the 15-month span of the mission; the instrument suite on board the spacecraft is designed to provide new insight into Jupiter's internal structure and global water abundance. The internal mass distribution of the solar system's largest planet will be determined by precisely mapping Jupiter's gravitational field. Obtaining a better estimate for the mass of Jupiter's core will enable planetary scientists to distinguish between the two prevailing theories of giant planet formation: whether Jupiter started with a massive core whose gravity attracted all of the nearby gas or whether the planet formation was triggered by the collapse of an unstable region in the protosolar nebula.

Knowledge of Jupiter's deep water abundance is another key ingredient for solar system and giant planet formation theories, as it has implications for the methods by which volatile compounds were distributed throughout the solar system by icy planetesimals or protoplanets. Juno will measure Jupiter's deep water abundance with a microwave radiometer, which is sensitive to emission coming from Jupiter's atmosphere at depths much greater than those sounded by the Galileo probe.

Jupiter's polar magnetosphere and auroral emissions will be characterized for the first time by Juno, using infrared and ultraviolet remote sensing instruments while simultaneously directly sampling the charged particles and magnetic field near the poles. The entire magnetic field of Jupiter will be mapped over the course of the mission, which will provide an indication of how deep in the planet the magnetic field is generated and will yield new insight into dynamo physics.

The European Space Agency (ESA) recently announced the selection of the Jupiter Icy Moons Explorer, or JUICE mission, for launch early in the next decade. This will be an orbiter whose primary objective will be to study Ganymede, Europa, and Callisto, the large icy satellites of Jupiter. Although the science objectives are focused on satellites, observations of Jupiter's atmosphere would also be possible and would provide new insight into the structure and dynamics of the Jovian atmosphere. Assuming a somewhat standard instrument suite of a wide- and narrow-angle camera, a visible-infrared spectrometer, a thermal infrared and/or ultraviolet spectrometer, and a submillimeter or microwave sounder, this mission may be able to characterize the abundances of minor species (particularly ammonia and water) in Jupiter's troposphere and stratosphere, which has implications for the origin and evolution of giant planet atmospheres. The Jovian atmospheric dynamics and structure also can be characterized through studies of winds, cloud features, atmospheric waves, and tropospheric and stratospheric temperatures.

The James Webb Space Telescope (JWST), which is scheduled to be launched in the next decade, will provide unprecedented infrared observations of Uranus and Neptune. Both Jupiter and Saturn may be too bright to be observed with JWST in its standard imaging and spectroscopic modes, although perhaps through clever techniques such as binning and/or subframing new observations of the gas giant atmospheres will be possible. The spectral coverage afforded by JWST extends that of HST into the infrared, and its 6.5-m diameter

primary mirror will offer a significant increase in spatial resolution over the infrared Spitzer Space Telescope. Through observations of the ethane, methane, and acetylene emissions in the stratospheres of Uranus and Neptune, we will gain an improved understanding of the heating mechanisms in ice giant atmospheres.

Finally, there are several giant planet missions that were recommended as part of the National Academy of Sciences planetary science decadal survey entitled *Vision and Voyages for Planetary Science in the Decade 2013–2022*. This new decadal survey identifies the most important scientific questions in planetary science and provides a prioritized list of flight investigations that can address these fundamental science questions. After the top priority Mars sample return flagship mission, two outer solar system missions were recommended as second and third priorities for flagships in the next decade: a EuropaJupiter mission and a Uranus system. The decadal survey also recommended five possible missions for the New Frontiers class, including a shallow Saturn probe mission that will yield fundamental new insight into Saturn's atmospheric structure and water abundance.

5.3 Links to Exoplanets

Over the last 15 years, the study of giant planets has been completely revolutionized by the discovery of hundreds of large (Jupiter-sized or larger) extrasolar planets. These planets, many of which orbit their host stars at distances closer than 1 AU, have forced a reexamination of old ideas of giant planet formation and evolution. A deeper understanding of our own solar system and its evolution since the time of formation is critical for understanding these exciting, newly discovered planetary systems elsewhere in the Milky Way galaxy.

Acknowledgments

It is a pleasure to thank R. Beebe, M. Sussman, R. Carlson, P. Strycker, and C. Miller for valuable discussions.

Cross-References

Exoplanet Detection Methods

References

Atkinson, D. H., Pollack, J. B., & Seiff, A. 1998, J. Geophys. Res., 103, 22911

Atreya, S. K., & Wong, A. S. 2005, Space Sci. Rev., 116, 121

Atreya, S. K., Wong, A. S., Baines, K. H., Wong, M. H., & Owen, T. C. 2005, Planet. Space Sci., 53, 498

Anderson, J. D. & Schubert, G. 2007, Science, 317, 1384

Baines, K. H., Carlson, R. W., & Kamp, L. W. 2002, Icarus, 159, 74

Baines, K. H., Momary, T. W., Fletcher, L. N., Showman, A. P., Roos-Serote, M., Brown, R. H., Buratti, B. J., Clark, R. N., & Nicholson, P. D. 2009, Planet. Space Sci. 57, 1671

Banfield, D., Gierasch, P. J., Bell, M., Ustinov, E., Ingersoll, A. P., Vasavada, A. R., West, R. A., & Belton, M. J. S. 1998, Icarus, 135, 230

Barnet, C. D., Westphal, J. A., Beebe, R. F., & Huber, L. F. 1992, Icarus, 100, 499
Beebe, R. F., Orton, G. S., & West, R. A. 1989, in Time-Variable Phenomena in the Jovian System, ed. M. J. S. Belton et al. (Washington, DC: NASA SP-494), 245
Beebe, R. F., Barnet, C., Sada, P. V., & Murrell, A. S. 1992, Icarus, 95, 163
Bjoraker, G. L., Larson, H. P., & Kunde, V. G. 1986, Astrophys. J., 311, 1058
Brooke, T. Y., Knacke, R. F., Encrenaz, T., Drossart, P., Crisp, D., & Feuchtgruber, H. 1998, Icarus, 136, 1
Carlson, B. E., Lacis, A. A., & Rossow, W. B. 1992, Astrophys. J., 388, 648
Chanover, N. J., Kuehn, D. M., & Beebe, R. F. 1997, Icarus, 128, 294
Cheng, A. F., Simon-Miller, A. A., Weaver, H. A., Baines, K. H., Orton, G. S., Yanamandra-Fisher, P. A., Mousis, O., Pantin, E., Vanzi, L., Fletcher, L. N., Spencer, J. R., Stern, S. A., Clarke, J. T., Mutchler, M. J., & Noll, K. S. 2008, AJ, 135, 2446
de Pater, I., & Lissauer, J. J. 2001, Planetary Sciences (Cambridge, UK: Cambridge University Press)
de Pater, I., Romani, P. N., & Atreya, S. K. 1991, Icarus, 91, 220
de Pater, I., Sromovsky, L. A., Hammel, H. B., Fry, P. M., LeBeau, R. B., Rages, K., Showalter, M., & Matthews, K. 2011, Icarus, 215, 332
Dyudina, U. A., Ingersoll, A. P., Ewald, S. P., Vasavada, A. R., West, R. A., Baines, K. H., Momary, T. W., Del Genio, A. D., Barbara, J. M., Porco, C. C., Achterberg, R. K., Flasar, F. M., Simon-Miller, A. A., & Fletcher, L. N. 2009, Icarus, 202, 240
Fletcher, L. N., Hesman, B. E., Irwin, P. G. J., Baines, K. H., Momary, T. W., Sanchez-Lavega, A., Flasar, F. M., Read, P. L., Orton, G. S., Simon-Miller, A., Hueso, R., Bjoraker, G. L., Mamoutkine, A., del Rio-Gaztelurrutia, T., Gomez, J. M., Buratti, B., Clark, R. N., Nicholson, P. D., & Sotin, C. 2011, Science, 332, 1413
Gierasch, P. J., Ingersoll, A. P., Banfield, D., Ewald, S. P., Helfenstein, P., Simon-Miller, A., Vasavada, A., Breneman, H. H., Senske, D. A., & Galileo Imaging Team 2000, Nature, 403, 628
Godfrey, D. A. 1988, Icarus, 76, 335
Guillot, T. 1999, Science, 286, 72
Hammel, H. B., & Lockwood, G. W. 2007, Icarus, 186, 291
Hammel, H. B., Lockwood, G. W., Mills, J. R., & Barnet, C. D. 1995, Science, 268, 5218
Hammel, H. B., de Pater, I., Gibbard, S. G., Lockwood, G. W., & Rages, K. 2005, Icarus, 175, 534
Hammel, H. B., Sromovsky, L. A., Fry, P. M., Rages, K., Showalter, M., de Pater, I., van Dam, M. A., LeBeau, R. P., & Deng, X. 2009, Icarus, 201, 257
Hanel, R. A., Conrath, B., Flasar, M., Kunde, V., Lowman, P., Maguire, W., Pearl, J., Pirraglia, J., Samuelson, R., Gautier, D., Gierasch, P., Kumar, S., & Ponnamperuma, C. 1979, Science, 204, 972
Hofstadter, M. D., & Butler, B. J. 2003, Icarus, 165, 168
Hueso, R., Legarreta, J., Pérez-Hoyos, S., Rojas, J. F., Sánchez-Lavega, A., & Morgado, A. 2010, Planet. Space Sci., 58, 1152
Irwin, P. G. J. 2009, Giant Planets of Our Solar System: Atmospheres, Composition, and Structure (2nd ed.; Chichester, UK: Praxis)
Kalogerakis, K. S., Marschall, J., Oza, A. U., Engel, P. A., Meharchand, R. T., & Wong, M. H. 2008, Icarus, 196, 202
Karkoschka, E. 2001, Icarus, 151, 84
Larson, H. P., Fink, U., Treffers, R., & Gautier, T. N., III 1975, Astrophys. J., 197, L137
Law, N. M., Mackay, C. D., & Baldwin, J. E. 2006, Astron. Astrophys., 446, 739
Lellouch, E., Bézard, B., Moreno, R., Bocklelée-Morvan, D., Colom, P., Crovisier, J., Festou, M., Gautier, D., Marten, A., & Paubert, G. 1997, Planet. Space Sci., 45, 1203
Lindal, G. F., Wood, G. E., Levy, G. S., Anderson, J. D., Sweetnam, D. N., Hotz, H. B., Buckles, B. J., Holmes, D. P., Doms, P. E., Eshleman, V. R., Tyler, G. L., & Croft, T. A. 1981, J. Geophys. Res., 86, 8721
Lindal, G. F., Sweetnam, D. N., & Eshleman, V. R. 1985, Astron. J., 90, 1136
Lindal, G. F., Lyons, J. R., Sweetnam, D. N., Eshleman, V. R., & Hinson, D. P. 1987, J. Geophys. Res., 92, 14987
Lindal, G. F., Lyons, J. R., Sweetnam, D. N., Eshleman, V. R., & Hinson, D. P. 1990, Geophys. Res. Let., 17, 1733
Lockwood, G. W., & Jerzykiewicz, M. 2006, Icarus, 180, 442
Lockwood, G. W., & Thompson, D. T. 2002, Icarus, 156, 37
Max, C. E., Macintosh, B. A., Gibbard, S. G., Gavel, D. T., Roe, H. G., de Pater, I., Ghez, A. M., Acton, D. S., Lai, O., Stomski, P., & Wizinowich, P. L. 2003, Astron. J., 125, 364
Moses, J. I., Lellouch, E., Bézard, B., Gladstone, G. R., Feuchtgruber, H., & Allen, M. 2000, Icarus, 145, 166
Niemann, H. B., Atreya, S. K., Carignan, G. R., Donahue, T. M., Haberman, A., Harpold, D. N., Hartle, R. E., Hunten, D. M., Kasprzak, W. T., Mahaffy, P. R., Owen, T. C., & Way, S. H. 1998, J. Geophys. Res., 103(E10), 22831
Orton, G. S., Fisher, B. M., Baines, K. H., Stewart, S. T., Friedson, A. J., Ortiz, J. L., Marinova, M., Ressler, M., Dayal, A., Hoffmann, W., Hora, J.,

Hinkley, S., Krishnan, V., Masanovic, M., Tesic, J., Tziolas, A., & Parija, K. C. 1998, J. Geophys. Res., 103, 22791
Peek, B. M. 1953, The Planet Jupiter: The Observer's Handbook (Boston, MA: Fabe and Faber)
Perez-Hoyos, S., Sanchez-Lavega, A., Hueso, R., Garcia-Melendo, E., & Legarreta, J. 2009, Icarus, 203, 516
Ragent, B., Colburn, D. S., Rages, K. A., Knight, T. C. D., Avrin, P., Orton, G. S., Yanamandra-Fisher, P. A., & Grams, G. W. 1998, J. Geophys. Res., 103, 22891
Rages, K. A., Hammel, H. B., & Friedson, A. J. 2004, Icarus, 172, 548
Rages, K. A., Hammel, H. B., & Sromovsky, L. 2007, Bull. Am. Astron. Soc., 39, 425
Roos-Serote, M., Drossart, P., Encrenaz, T., Lellouch, E., Carlson, R. W., Baines, K. H., Kamp, L., Mehlman, R., Orton, G. S., Calcutt, S., Irwin, P., Taylor, F., & Weir, A. 1998, J. Geophys. Res., 103, 23032
Sánchez-Lavega, A. 1982, Icarus, 49, 1
Sánchez-Lavega, A. 2010, An Introduction to Planetary Atmospheres (Boca Raton, FL: CRC)
Sánchez-Lavega, A., Colas, F., Lecacheux, J., Laques, P., Parker, D., & Miyazaki, I. 1991, Nature, 353, 397
Sánchez-Lavega, A., del Río-Gaztelurrutia, T., Hueso, R., Gómez-Forrellad, J. M., Sanz-Requena, J. F., Legarreta, J., García-Melendo, E., Colas, F., Lecacheux, J., Fletcher, L. N., Barrado-Navascués, D., Parker, D., & The International Outer Planet Watch Team 2011, Nature, 475, 71
Seiff, A., Kirk, D. B., Knight, T. C. D., Young, R. E., Mihalov, J. D., Young, L. A., Milos, F. S., Schubert, G., Blanchard, R. C., & Atkinson, D., 1998, J. Geophys. Res., 103(E10), 22857
Simon-Miller, A. A., Conrath, B., Gierasch, P. J., & Beebe, R. F. 2000, Icarus, 145, 454
Simon-Miller, A. A., Banfield, D., & Gierasch, P. J. 2001a, Icarus, 149, 94
Simon-Miller, A. A., Banfield, D., & Gierasch, P. J. 2001b, Icarus, 154, 459
Sromovsky, L. A., & Fry, P. M. 2002, Icarus, 157, 373
Sromovsky, L. A., & Fry, P. M. 2005, Icarus, 179, 459
Sromovsky, L. A., & Fry, P. M. 2010, Icarus, 210, 230
Sromovsky, L. A., Revercomb, H. E., Krauss, R. J., & Suomi, V. E. 1983, J. Geophys. Res., 88, 8650
Sromovsky, L. A., Collard, A. D., Fry, P. M., Orton, G. S., Lemmon, M. T., Tomasko, M. G., & Freedman, R. S. 1998, J. Geophys. Res., 103, 22929
Sromovsky, L. A., Fry, P. M., Limaye, S. S., & Baines, K. H. 2003, Icarus, 163, 256
Sromovsky, L. A., Fry, P. M., Hammel, H. B., de Pater, I., Rages, K. A., & Showalter, M. R. 2007, Icarus, 192, 558
Strycker, P. D., Chanover, N. J., Simon-Miller, A. A., Banfield, D., & Gierasch, P. J. 2011, Icarus, 215, 552
von Zahn, U., Hunten, D. M., & Lehmacher, G. 1998, J. Geophys. Res., 103, 22815
Weidenschilling, S. J., & Lewis, J. S. 1973, Icarus, 20, 465
West, R. A., Strobel, D. F., & Tomasko, M. G. 1986, Icarus, 65, 161
West, R. A., Baines, K. H., Friedson, A. J., Banfied, D., Ragent, B., & Taylor, F. W. 2004, in Jupiter: The Planet, Satellites, and Magnetosphere, ed. F. Bagenal et al. (Cambridge, UK: Cambridge University Press), 79–104
Westphal, J. A., Baum, W. A., Ingersoll, A. P., Barnet, C. D., de Jong, E. M., Danielson, G. E., & Caldwell, J. 1992, Icarus, 98, 94
Young, R. E. 2003, New Astron. Rev., 47, 1
Young, R. E., Smith, M. A., & Sobeck, C. K. 1996, Science, 272, 837

6 Planetary Magnetospheres

Fran Bagenal
Astrophysical and Planetary Sciences, Dept. and Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA

1 Introduction ***252***

2 Magnetospheric Principles ***254***
2.1 Planetary Magnetic Fields 254
2.2 Scales of Planetary Magnetospheres 260
2.3 Plasma Sources 263
2.4 Plasma Dynamics 266
2.4.1 Energetic Particles and Radiation Belts 267
2.4.2 Rotational Flows 271
2.4.3 Global Solar-Wind-Driven Convection 273
2.4.4 Plasmoid Ejection 277

3 Magnetospheres of the Outer Planets ***278***
3.1 Jupiter 279
3.2 Saturn 285
3.3 Uranus and Neptune 290

4 Small Magnetospheres ***292***
4.1 Mercury 293
4.2 Ganymede 293

5 Induced Magnetospheres ***295***
5.1 Venus 295
5.2 Mars 297
5.3 Titan 298
5.4 Io 298
5.5 Pluto and Comets 302

6 Outstanding Questions ***302***

References ***303***

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_6, © Springer Science+Business Media Dordrecht 2013

Abstract: The nature of interaction between a planetary object and the surrounding plasma depends on the properties of both the object and the plasma flow in which it is embedded. A planet with a significant internal magnetic field forms a magnetosphere that extends the planet's influence beyond its surface or cloud tops. There are seven objects in the solar system that presently have internally generated magnetic fields: Mercury, Earth, Jupiter, Saturn, Uranus, Neptune, and the satellite Ganymede. A planetary object without a significant internal dynamo can interact with any plasma flowing past via remanent magnetization of the crust and/or currents associated with local ionization or induced in an electrically conducting ionosphere or layer of water. Venus, Mars, Titan, Io, Enceladus, and Europa have strong interactions with their surroundings. Planetary magnetospheres span a wide range of sizes but involve similar basic principles and processes.

Keywords: Aurora; Bow shock; Compression; Conductivity; Convection; Corotation; Diffusion; Dipole; Dynamo; Flux rope; Instability; Interchange; Io Plasma torus; Ion escape; Ionosphere; Kelvin–Helmholtz; Magnetic moment; Magnetization; Magnetopause; Magnetosheath; Magnetosphere; Magnetotail; Multipole; Non-dipole magnetic fields; Plasma sources; Plasmapause; Plasmasphere; Plasmoids; Radiation belts; Radio emission; Reconnection; Remanent magnetization; Rotation; Solar wind; Sputtering; Stagnation

1 Introduction

As the name suggests, a planet's *magnetosphere* is the region of space influenced by the planet's magnetic field. The nature of the interaction between a planetary object and the surrounding plasma depends on the properties of both the object and the plasma flow in which it is embedded. A planet with a significant internal magnetic field forms a magnetosphere that extends the planet's influence beyond its surface or cloud tops. A planetary object without a significant internal dynamo can interact with any surrounding plasma via remanent magnetization of the crust and/or currents induced in an electrically conducting ionosphere or layer of water. *Figure 6-1* is a schematic of the archetypical magnetosphere of Earth, illustrating the general anatomy.

All solar system objects are embedded in the *solar wind* that streams radially away from the Sun. The flow speed of the solar wind exceeds the speed of the fastest wave mode that can propagate in the interplanetary plasma. The interaction of the supersonic solar wind with a planetary magnetic field (either generated by an internal dynamo or induced externally) produces a *bow shock* upstream of the planet. Behind the bow shock, the subsonic wind – the *magnetosheath* – is deflected around the magnetospheric obstacle. The magnetospheric boundary – the *magnetopause* – was usually regarded to first order as an impenetrable boundary. However, the amount of mass, momentum, and magnetic flux exchanged across the magnetopause has become an active area of research at Earth and other magnetospheres. The distance between the center of the planet and the magnetopause in the direction of the Sun (approximately the closest distance) is labeled R_{MP}, generally described in units of the planetary radius (R_P). Whatever the details of the interaction, in nearly all cases, the interaction region has a "wake" or "tail" – the *magnetotail* – that can extend for several hundred times R_{MP} downstream in the solar wind.

Venus, Mars, and (likely) Pluto do not have dynamos generating an internal field at present, though strong remanant magnetization of crustal rocks is evidence that Mars certainly had a dynamo in the past and this is also quite possible for Venus too. At present, the solar wind

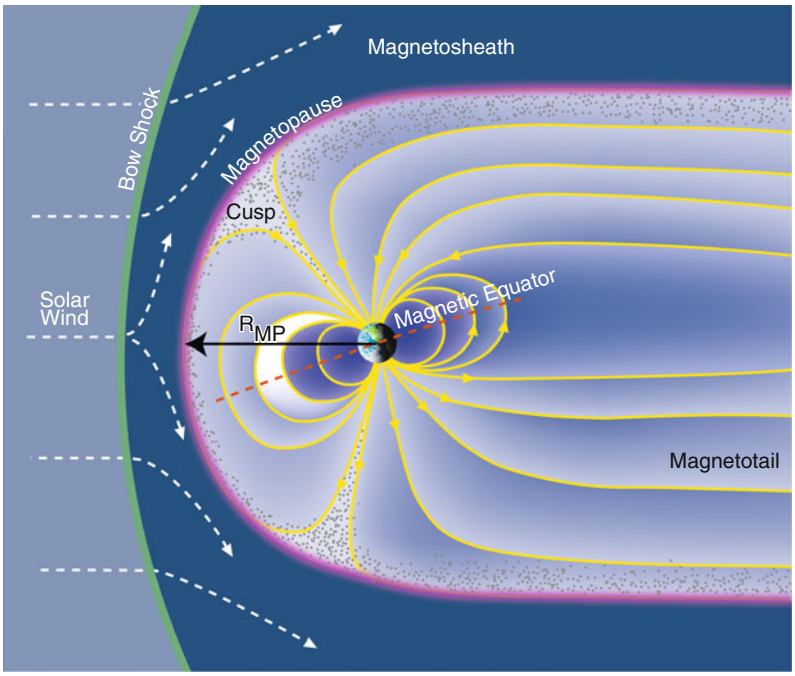


Fig. 6-1
Anatomy of a magnetosphere, applied here to Earth

interacts with the substantial ionospheres of these planets, induction currents deflecting the bulk of the flow around the planet, acting as the obstacle to the supersonic solar wind, and producing an upstream bow shock. Any neutral atoms or molecules escaping from planetary atmospheres often become ionized either by solar photons or charge exchange with solar wind protons. These atmospheric ions are then entrained in and extract momentum from the solar wind. The slowing of the solar wind around these obstacles carries the Sun's magnetic field which is then temporarily draped around the planet and stretched back into a comet-like tail.

Objects such as the Earth's Moon that have no appreciable atmosphere and a low-conductivity surface have minimal electrodynamic interaction with the surrounding plasma and just absorb the impinging solar wind with no upstream shock. Interactions between planetary satellites and magnetospheric plasmas are as varied as the moons themselves: Ganymede's significant dynamo produces a mini-magnetosphere within the giant magnetosphere of Jupiter; the electrodynamic interactions of magnetospheric plasma flowing past volcanically active Io (Jupiter) and Enceladus (Saturn) generate substantial currents and supply extended clouds of neutrals that become ionized to supply more plasma to the system; plasma interactions with Titan's thick atmosphere and substantial ionosphere are likened to Venus; in the absence of an atmosphere, charged particles bombard the moon surfaces, sometimes sputtering significant exospheres (e.g., Europa, Dione, Callisto). The flow within magnetospheres tends to be subsonic, so that none of these varied interactions forms a shock upstream of the moon.

Reviews of planetary magnetospheres range in their approach to the subject from considering it a topic in space plasma physics (exploiting the range of planetary environments as a laboratory to explore space plasmas) to a branch of planetary science (presenting the space

environment as a component in understanding the planetary objects). A basic, qualitative introduction is given in van Allen and Bagenal (1999). Deeper studies of comparative magnetospheres range from the abstract to the specific (Siscoe 1979; Vasyliūnas 2004; Vasyliunas 2009, 2010; Kivelson 2007; Walker and Russell 1995; Bagenal 1992; Russell 2004, 2006; Kivelson and Bagenal 2007; Bagenal 2009). This chapter takes an intermediate path, with the goal of applying the general principles to specific planets but also providing a qualitative appreciation of the different characters of our local family of magnetospheres.

The general principles of the structure and dynamics of planetary magnetospheres are presented in Sect. 2. The Earth is the nominal case with which to compare the basic properties between the planets. Section 3 introduces the magnetospheres of the outer planets, magnetospheres that are large, dominated by rotation, and strongly influenced by the moons that are embedded within. By contrast, Sect. 4 discusses the mini-magnetospheres of Mercury and Ganymede. Section 5 returns to plasma interactions with nonmagnetized objects where the varied plasma interactions with planets such as Venus, Mars, and Pluto are discussed, as well as moons Titan, Io, Enceladus, and Europa.

2 Magnetospheric Principles

The interaction of a planetary object with its surroundings depends on the properties of both the planetary body and the impinging plasma. For the nine major planetary bodies, *Table 6-1* lists the properties of the interplanetary medium (the strength and direction of the interplanetary magnetic field (IMF) and the speed, density, and temperature of the solar wind), as well as the strength of any planetary magnetic field, the planetary rotation rate, and the scale of the observed magnetospheres. In *Table 6-2*, the properties of the planetary dynamos are listed: the strength and direction of the planet's magnetic field and the direction of the planet's spin. Below, how these properties affect the characteristics and behavior of planetary magnetospheres is discussed.

2.1 Planetary Magnetic Fields

Spacecraft carrying magnetometers have flown to and characterized the magnetic fields of all the planets except (dwarf) Pluto. All four of the giant planets have strong magnetic fields. The smaller terrestrial planets have weaker fields, Mercury's being much weaker than Earth's. The upper limit on an internally-generated field of Venus is less than 10^{-5} times Earth's magnetic moment. While strong magnetization of surface rocks show that Mars' internal dynamo was active in the past, geological evidence shows that the dynamo shut down around 4 billion years ago.

The history of space exploration of planetary magnetism is given by Ness (2010), while Balogh (2010) reviews techniques that have been employed to measure planetary magnetic fields. Thorough reviews of planetary magnetic field observations and their analysis are presented for all planets by Connerney (2007) and for the giant planets by Russell and Dougherty (2010). Anderson et al. (2010, 2011) present recent determinations of Mercury's magnetic field from the MESSENGER spacecraft. Magnetic field measurements from orbit allow the separation of the internally generated field from the effects of external currents in the magnetosphere (see review by Olsen et al. 2010 of the techniques for doing this).

Table 6-1

Properties of the solar wind and scales of planetary magnetospheres

	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
Distance from Sun, a_p(AU)[a]	0.39	0.72	1[b]	1.52	5.2	9.5	19	30	40
Solar wind density[b] (cm^{-3})	53	14	7	3	0.2	0.07	0.02	0.006	0.003
IMF strength[c] (nT)	41	14	8	5	1	0.6	0.3	0.2	0.1
IMF azimuth angle[c]	23°	38°	45°	57°	80°	84°	87°	88°	88°
Radius, R_P(km)	2,440	6,051	6,373	3,394	71,400	60,268	25,600	24,765	1,170 (±33)
Sidereal spin period (day)	58.6	−243	0.9973	1.026	0.41	0.44	−0.72	0.67	−6.39
Magnetic moment[d] (M_E)	$3-6\times10^{-4}$	$<10^{-5}$	1	$<10^{-5}$	20,000	600	50	25	?
Surface magnetic field[e] B_o(nT)	195	-	30,600	-	430,000	21 ,400	22,800	14,200	?
R_{CF}[f](R_P)	1.6 R_M	-	10 R_E	-	46 R_J	20 R_S	25 R_U	24 R_N	?
Observed R_{MP} (R_P)	1.5 R_M	-	8–12R_E	-	63–92 R_J	22–27 R_S	18 R_U	23–26 R_N	?

[a]Semimajor axis of orbit. 1 AU = 1.5×10^8 km

[b]The number density of the solar wind fluctuates by about a factor of 5 about typical values of n_{sw} ~7 (cm^{-3})/a_p^2. The mass density of the solar wind is $\rho_{sw} = 1.04\ n_{sw}$ (amu cm^{-3})

[c]Mean values for the interplanetary magnetic field (IMF) in units of nano-Tesla with spherical components *Br*, Bθ, Bϕ. The azimuth angle is $\tan^{-1}(B\phi/Br)$. The radial component of the IMF, *Br*, decreases as $1/a_p^2$ while the transverse component, $B\phi$, increases with distance

[d]$M_{Earth} = 7.9\times10^{25}$ G $cm^3 = 7.9\times10^{15}$ T m^3

[e]Magnitude of dipole (see text for references)

[f]R_{CF} is calculated using $R_{CF} = \xi\ (B_o^2/2\mu_o\rho\ V_{sw}^2)^{1/6}$ for typical solar wind conditions of ρ_{sw} given above and $V_{sw} \sim 400$ km s^{-1} and ξ an empirical factor of ~1.4 to match Earth observations (Walker and Russell 1995)

Table 6-2
Planetary magnetic fields

	Ganymede	Mercury	Earth	Jupiter	Saturn	Uranus	Neptune
B_{Dipole}[a](nT)	719	195	30,600	430,000	21,400	22,800	14,200
Maximum/minimum[b]	2	~2	2.8	4.5	4.6	12	9
Dipole tilt and sense[c]	−4°	<+3°	+9.92°	−9.4°	−0.0°	−59°	−47°
Dipole offset[d] (R_P)		~0.2	–	0.119	0.038	0.352	0.485
Obliquity[e]	0°	0°	23.5°	3.1°	26.7°	97.9°	29.6°
Range in solar wind angle[f]	90°	90°	67 – 114°	87 – 93°	64–117°	8–172°	60–120°

[a]Surface field at dipole equator. Values derived from modeling the magnetic field as an offset, tilted dipole (OTD)
[b]Ratio of maximum surface field to minimum (equal to 2 for a centered dipole field). This ratio increases with larger non-dipolar components and tends to increase with the planet's oblateness
[c]Angle between the magnetic and rotation axes. Positive values correspond to magnetic field directed north at the equator
[d]Values for the giant planets come from dipole (OTD) models of Connerney (1993, 2007). The Earth's dipole is from the International Geomagnetic Reference Field, while the magnetic dip poles of the Earth's field are located (in 2010) at 85°N and 64°S latitudes and moving over 10° per century (Finlay et al. 2010). Mercury's magnetic field is from Anderson et al. (2011)
[e]The inclination of a planet's spin equator to the ecliptic plane
[f]Range of angle between the radial direction from the Sun and the planet's rotation axis over an orbital period. In Ganymede's case, the angle is between the corotational flow and the moon's spin axis

As planetary magnetic field measurements are improved (in coverage, accuracy, and proximity to the planet), they tend to show increasing complexity. The standard technique is to describe the internal magnetic field as a sum of multipoles or spherical harmonics (e.g., Walker and Russell 1995; Connerney 1993, 2007; Merrill et al. 1996), the higher orders being functions that drop off increasingly rapidly with distance so that one needs to get very close to the planet to see any effects of these high-order multipoles. The amplitude of each multipole is derived by fitting magnetic field observations obtained by magnetometers on spacecraft flying past the planet (e.g., Connerney 1981; Russell and Dougherty 2010). The extensive coverage provided by surface explorers over the centuries and afforded by low-orbiting spacecraft at Earth in the past 50 years not only allows the present Earth's field to be described with hundreds of terms but also allows description of the variation of the Earth's field over time (e.g., reviewed by Hulot et al. 2010). Moreover, paleomagnetic data extend the Earth's record back in geological time, revealing many polarity reversals of the magnetic field. By contrast, the sampling of planetary magnetic fields is too poor to determine any temporal variation over the few decades of space exploration, and currently there is no evidence of whether other planetary dynamos reverse polarity.

For magnetospheric purposes, where one is relatively far from the planet, a simple dipole description (equivalent to a bar magnetic placed inside the planet) has proved to be very valuable. The formula for a dipole magnetic field vector (measured in units of Tesla) **B** as a function of position vector **r** is

$$B = [3r(M \cdot r) - Mr^2]/r^5, \tag{6.1}$$

where M is the planetary magnetic moment (in units of T m^3). It is easier to understand this vector field if one looks at the components (B_r, B_θ, B_ϕ) in a spherical coordinate system (r, θ, ϕ), centered and aligned with the dipole. For a dipole centered on the planet, the field is azimuthally symmetric and $B_\phi = 0$, while the other components are

$$B_r = 2B_o \cos\theta / R^3 \tag{6.2}$$

$$B_\theta = B_o \sin\theta / R^3 \tag{6.3}$$

$$|B| = B_o (1 + 3\cos^2\theta)^{1/2} / R^3. \tag{6.4}$$

where B_o is the value of the magnetic field at the equator, $R = |r|/R_P$, and θ is colatitude. Note that the field strength at the poles ($\theta = 0$) is twice that at the equator, *Bo.*

The planetary magnetic fields are generally not aligned with the planet's spin axis. Thus, the simplest description is often a dipole that is offset from the center of the planet and tilted. *Figure 6-2* illustrates the offset, tilted dipole (OTD) that approximates planetary fields and the values are given in *Table 6-2*. Except for Uranus and Neptune, the tilts are modest (<10°) and the offset a small fraction of a planetary radius. The large tilts and offsets of Uranus' and Neptune's magnetic fields indicate the highly non-dipolar nature of these fields. Further indication of deviation of the true magnetic field from the simplicity of centered dipole is illustrated by the ratio of maximum to minimum surface field strength being much larger than value of 2 expected for a centered dipole. The range of complexities from Saturn's highly symmetric, dipolar field to the irregular fields of Uranus and Neptune are illustrated in the surface maps of magnetic field strength shown in *Fig. 6-3* and the high values of the max/min ratio in *Table 6-2*.

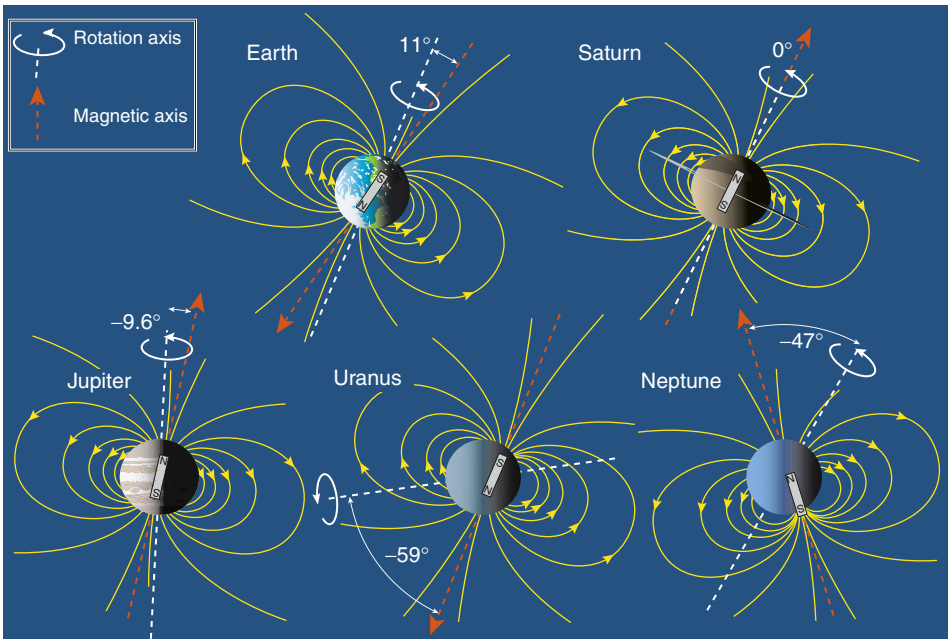


Fig. 6-2

The tilt angles between the spin and magnetic axes are shown for the five main magnetized planets. Considering the horizontal direction of the diagram to be parallel to the ecliptic plane and the vertical direction the ecliptic normal, then the spin axis is shown for conditions of maximum angle from the ecliptic normal (i.e., at solstice). Each planet's magnetic field can be approximated as a dipole where the orientation and any offset from the center of the planet is illustrated by a bar magnet located at the center of the planet

While the theory of planetary dynamos has yet to reach the level of sophistication where it could predict with accuracy the presence (let alone the specific characteristics) of an internally generated magnetic field, it is generally understood that, for such a field to be present, planets need to have an interior that is sufficiently electrically conducting and that is convecting with sufficient vigor. Various simple scaling laws have been derived over the years (e.g., reviewed by Christensen 2010) that relate the strength of the planet's magnetic field to other properties, but these laws seem to be more useful for testing theoretical ideas about dynamos than for predictions. Extensive geophysical measurements have revealed substantial information about the distribution of density, temperature, and flows inside the Earth. Moreover, the remanent magnetization of surface rocks tells us how the Earth's field has changed over geological time. These geophysical data are powerful constraints on the geodynamo and numerical dynamo models are beginning to show consistent behavior (Christensen 2010; Wicht and Tilgner 2010). In addition, laboratory experiments test ideas about parameters that control geodynamo behavior (Verhille et al. 2010). For other planetary objects, where information about the interior properties is much more limited, the presence or absence of a magnetic field is an important constraint on the interior conditions. Dynamo models are now being developed that vary the size of the planetary core, rotation rate, conductivity, heat flux, etc., to match the wide range of conditions at different planets (reviewed by Stanley and Glatzmaier 2010).

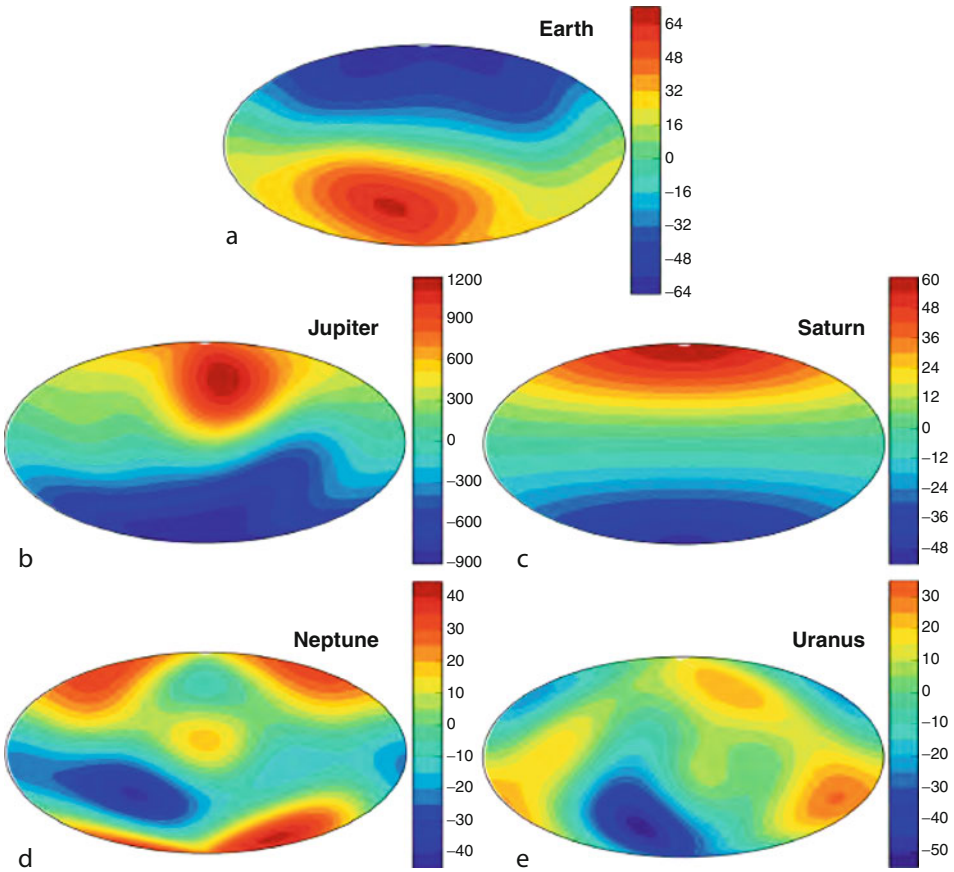


Fig. 6-3

Surface maps of the strength of the radial component of magnetic field for (a) Earth, (b) Jupiter, (c) Saturn, (d) Neptune, and (e) Uranus. For the gaseous planets, the surface is taken to mean the 1 bar pressure level (Data from Merrill et al. (1996) for Earth, Connerney (1993) for Jupiter and Saturn and Holme and Bloxham (1996) for Uranus and Neptune. Based on Stanley and Bloxham (2006))

Given the disparity in scale between the giant and terrestrial planets (e.g., the volume of Jupiter is 1,400 times that of the Earth), it is perhaps not surprising that the four terrestrial planets have far weaker magnetic fields generated in their interiors than the giant planets (Russell 1993; Connerney 1993; Stevenson 2003; Breuer et al. 2010). The iron cores are potential dynamo regions of terrestrial planets. The high pressures inside the giant planets Jupiter and Saturn put the hydrogen into a phase where it has the electrical conductivity of liquid metal (see Chap. 4). Jupiter's three times higher mass than Saturn produces a much larger volume of metallic hydrogen, responsible for ~×20 times stronger dynamo. Inside Uranus and Neptune, the pressures are too weak to make hydrogen metallic, and it is postulated that their dynamos must be generated in regions of liquid water where, as in Earth's ocean, small concentrations of ions provide sufficient conductivity. Stanley and Bloxham (2006) show that confining the dynamos of these water giants to a relatively thin conducting shell can produce highly irregular, non-dipolar magnetic fields.

Careful analysis by Phillips and Russell (1987) produced an upper limit to Venus' magnetic moment of ~1×10^{-5} M_E and revealed no evidence of crustal remanent magnetization. The apparent lack of an active dynamo inside Venus puts interesting constraints on the thermal evolution of that planet (Stevenson et al. 1983; Schubert et al. 1988). A common misconception is that it is the slowness of the rotation of Venus that prevents a dynamo. In fact, very little rotation is needed for a dynamo, and all objects in the solar system have sufficient rotation (Stevenson 2003). So, the question becomes why is Venus' core not convecting? Stevenson et al. (1983) proposed that Venus' core temperature is too high for a solid iron inner core to condense (the differentiation of solid iron from an outer liquid sulfur–iron alloy drives Earth's dynamo). The lack of plate tectonics at Venus may be limiting the cooling of the planet's upper layers, further suppressing internal convection. Another possibility is that Venus may be in a state of transition following the period of global volcanism that resurfaced the planet about 700 million years ago (see ❯ Chap. 3). Thus, Venus might have had an active dynamo in the past and may well develop one in the future. Why neighbor twin planets should have suffered such different internal histories is a major mystery of planetary geophysics (Smrekar et al. 2007).

Measurements of strong remanent crustal magnetism (surface fields of up to 1,500 nT) suggest that Mars has had an active dynamo and experienced changes in polarity over geological time scales (Acuña et al. 2001; Connerney et al. 2004) but stopped generating an internal field some 4 billion years ago. Stevenson (2010) summarizes the three main contending theories of how Mars' dynamo operated for a few hundred million years and then ceased: (1) cooling of the core slowed down to the point where conductive heat loss dominated, without an inner core forming (Stevenson et al. 1983); (2) Nimmo and Stevenson (2000) suggest that after a period of efficient convection (perhaps driven by plate tectonics), Mars underwent a change in convective style to the currently observed stagnant lid mode, causing the mantle and core to stop cooling and turning off core convection and the dynamo (note that this model would work irrespective of whether Mars has an inner core); or (3) the inner core of Mars froze sufficiently so that the remaining fluid region of the outer core was too thin to sustain a dynamo (Stewart et al. 2007). These theories span a wide range of states of the core of Mars. Future geophysical sounding (in particular, seismology) will hopefully reveal the state of Mars' interior.

Having radii of ~40% of the Earth's radius, Mercury and Ganymede were originally expected to have cooled off, shutting down any internal dynamo. But spacecraft flybys showed each object to have a significant magnetic field. Thermal models of the particularly large iron core (>70% of the radius) of Mercury suggest that at least an outer region is likely to be liquid and possibly convecting. However, the observed field is much weaker than standard dynamo theory would predict (Stevenson 2003). Efforts to reconcile models of thermal evolution (Breuer et al. 2010) and dynamo models (Stanley and Glatzmaier 2010) of these small bodies is an active area of research.

2.2 Scales of Planetary Magnetospheres

The term "magnetosphere" was coined by Gold (1959) to describe "the region above the ionosphere in which the magnetic field has dominant control over the motions of gas and fast charged particles." ❯ *Figure 6-1* presents a schematic of the Earth's magnetosphere showing the major regions.

Planetary magnetospheres are embedded in the solar wind, which is the outward expansion of the solar corona. At the Earth's orbit, the solar wind has an average speed of about 400 km/s. The density of particles (mainly electrons and protons with a few percent alpha particles) is observed to decrease (from values of about 3–10 cm^{-3} at Earth) as the inverse square of the distance from the Sun, consistent with a steady radial expansion of solar gas into a spherical volume. The solar wind speed, while varying between 300 and 700 km/s, always greatly exceeds the speed of waves characteristic of a low density, ionized, and magnetized gas (Alfven waves). The planetary bow shock formed upstream of an obstacle in the super-Alfvenic solar wind flow can be described in fluid terms as a discontinuity in bulk parameters of the solar wind plasma in which mass, momentum, and energy are conserved. Entropy, however, increases as the flow traverses the shock with the solar wind being decelerated and heated so that the flow can be deflected around the magnetosphere. Thus, a shock requires dissipative processes, and the presence of a magnetic field allows dissipation to occur on a scale much smaller than the scale length for collisions between solar wind particles. Although planetary bow shocks do not play a significant role in processes occurring inside the magnetosphere, the crossings of spacecraft through planetary bow shocks have provided an opportunity to study the exotic plasma physics of high Mach number collisionless shocks that cannot be produced in the laboratory (see reviews by Lembege et al. 2004; Balogh et al. 2005; Treumann 2009).

Well before (Biermann 1957) provided cometary evidence of a persistent solar wind, Chapman and Ferraro (1930) considered how a strongly magnetized body would deflect a flow of particles from the Sun. They estimated the location of the stagnation point where the dynamic pressure of the solar wind ($\rho_{sw} {V_{sw}}^2$) is balanced by the internal pressure of the planet's magnetic field (treating the boundary as impenetrable and ignoring any contributions to internal pressure from particles). ❯ Equation 6.4 shows that the dipole field strength as a function of radial distance (in the equatorial plane) is

$$B(r) = B_o(R_P/r)^3 \tag{6.5}$$

so that the planetary magnetic pressure varies as

$$B^2/2\mu_o = (B_o^2/2\mu_o)(R_P/r)^6. \tag{6.6}$$

For the case of a dipolar magnetic field (with poles perpendicular to the solar wind direction), the Chapman-Ferraro stagnation distance R_{CF} is

$$R_{CF} = \xi(B_o^2/2\mu_o\rho_{sw}{V_{sw}}^2)^{1/6}, \tag{6.7}$$

where ξ is a numerical factor that corrects for the effects of electrical currents that flow along the magnetopause (discussed in textbooks such as Cravens 1997; Kivelson and Russell 1995). Some prefer to define R_{CF} with $\xi = (2)^{1/3}$ (e.g., Vasyliunas 2009). Empirically, ξ is found to be about a factor of 1.4 to be consistent with the actual distance of the subsolar magnetopause distance (R_{MP}) at Earth (Walker and Russell 1995).

❯ *Table 6-1* shows that R_{CF} is a reasonable approximation to the observed magnetospheric scale R_{MP} except in the case of Jupiter (and a lesser extent Saturn), where substantial plasma pressure inside the magnetosphere expands the magnetopause to roughly twice the standoff distance of a dipole (discussed in ❯ Sect. 2.4.1). ❯ *Figure 6-4* illustrates the huge range in scale of the planetary magnetospheres. The magnetospheres of the giant planets encompass most of their extensive moon systems, including the four Galilean moons of Jupiter as well as Titan (Saturn) and Triton (Neptune). Earth's Moon, however, resides almost entirely outside

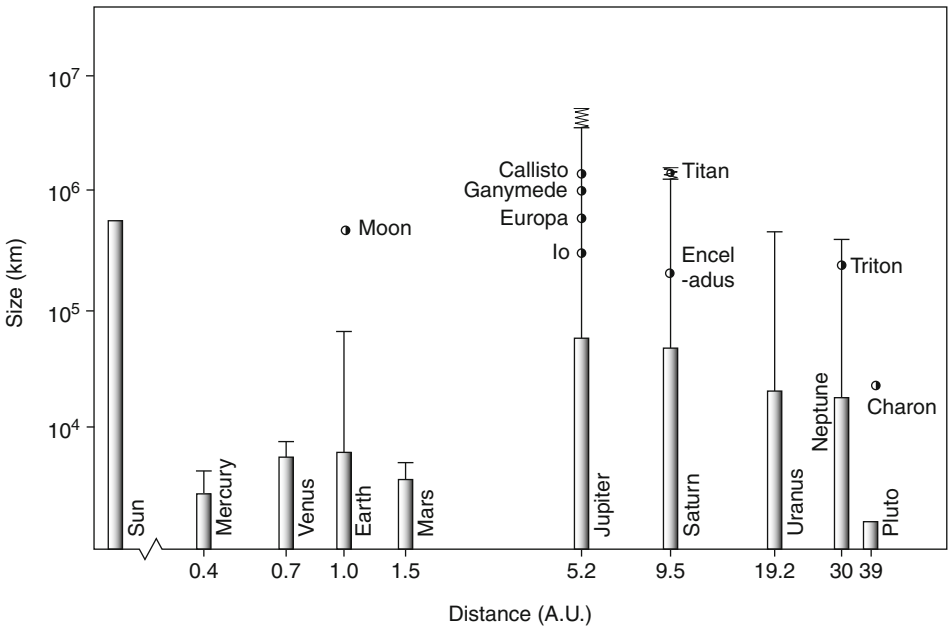


Fig. 6-4

A logarithmic plot of size of object vs. distance from the Sun for the planets (*solid bars*), their magnetospheres (RMP, *thin bars*) and the orbital radii of their primary moons. The range in sizes of the magnetospheres of Jupiter and Saturn are shown by *zig-zag lines*

the magnetosphere, spending less than 5% of its orbit crossing the magnetotail. *Figure 6-5* illustrates the vast range in scales: Each magnetosphere fits into the volume of the next-larger planet. Earth tends to be considered as the standard of comparison for other magnetospheres. It is natural that our home planet's magnetosphere is better explored and its vicissitudes studied in detail, but it is also important to test our understanding of the magnetospheric principles derived at Earth by applying these concepts to other planets.

Finally, when we discuss the dynamics of magnetospheres, it will be clear that an important factor is the orientation of the planet's magnetic field relative to the interplanetary magnetic field (*Table 6-2*). The obliquity is the angle of the planet's spin axis relative to the ecliptic plane normal. As a planet orbits the Sun, if it has a large obliquity, it will experience not only large seasonal changes but also a wide range in angles between the upstream solar wind (and embedded IMF) and the planet's magnetic field. Moreover, the large tilt of Uranus' and Neptune's magnetic fields with respect to their spin axes means that these magnetospheres also see a modulation of this solar wind angle over their spin period (i.e., a planetary day). While the solar wind remains flowing within a few degrees of radial from the Sun, the IMF forms a spiral of increasingly tangential field. At Earth the average spiral angle is 45°, at Jupiter it averages 80°, and at farther planets the field is basically tangential to the planet's orbit. The polarity changes several times during the ~25-day solar rotation (more frequently during solar maximum). Important for the influence of the solar wind on magnetospheric dynamics are the variations in the north–south component of the IMF (which fluctuates about the ecliptic plane) and changes in solar wind ram pressure

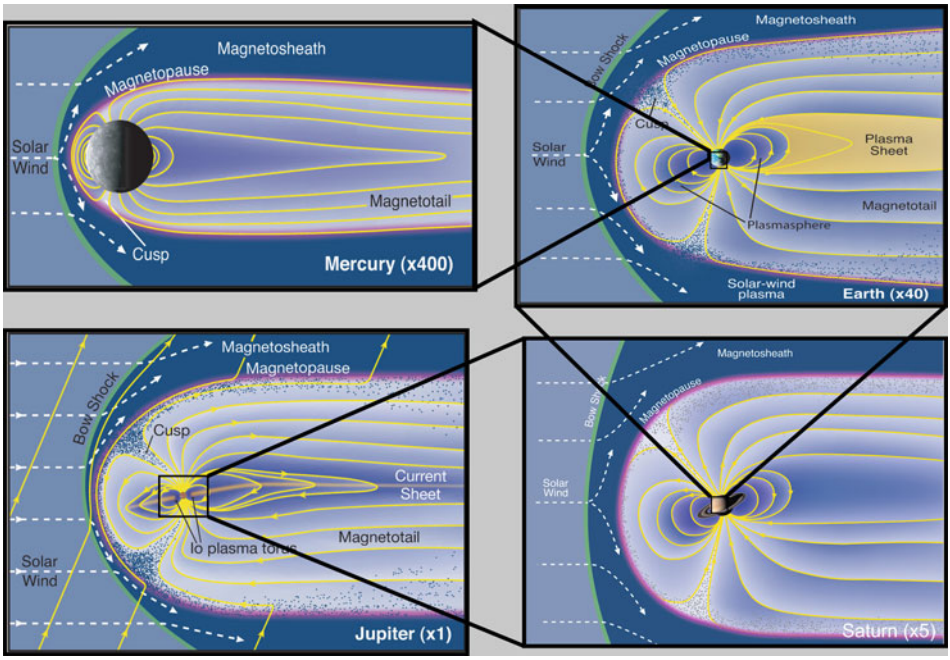


Fig. 6-5
Scaling of the magnetospheres from Mercury, Earth, Saturn, to Jupiter

impacting the magnetosphere. Other factors such as the Alfven Mach number of the incoming solar wind and plasma pressure on either side of the magnetopause also play roles that are being explored in current research (La Belle-Hamer et al. 1995; Swisdak et al. 2003; Cassak and Shay 2011).

2.3 Plasma Sources

The plasma found in a planetary magnetosphere could have a variety of sources: it could have leaked across the magnetopause from the solar wind, it may have escaped the planet's gravity and flowed out of the ionosphere, or it may be the result of the ionization of neutral material coming from satellites or rings embedded in the magnetosphere. The study of the origin of plasma populations and their evolution as they move through the magnetosphere is a detective story that becomes more complex the deeper one delves (e.g., review of Earth's plasma sources by Moore and Horwitz 2007).

The clearest indicator of which source is responsible for a particular planet's magnetospheric plasma is chemical composition (*Table 6-3*). For example, the O^+ ions in the Earth's magnetosphere must surely have come from the ionosphere, and the sulfur and oxygen ions at Jupiter have an obvious origin in Io's volcanic gases. But the source of protons is not so clear – protons could be either ionospheric (particularly for the hydrogen-dominated gas giants), dissociation of water ejected from icy satellites, or from the solar wind. One might consider that a useful source diagnostic would be the abundance of helium ions. Emanating from the hot (millions

Table 6-3

Plasma characteristics of planetary magnetospheres

	Ganymede	Mercury	Earth	Jupiter	Saturn	Uranus	Neptune
Max. plasma density (cm^{-3})	~400	~1	4,000	~3,000	~100	3	2
Neutral density (cm^{-3})	–	–	–	~50	~1,000	–	–
Major ion species	O^+, H^+	H^+	O^+, H^+	O^{n+}, S^{n+}	O^+, W^+,[a] H^+	H^+	N^+, H
Minor ion species		O^+,[b] Na^+		H^+,[c] H_3^+	H^+,[c] H_3^+		
Dominant source	Ganymede	Solar wind	Ionosphere[d]	Io	Enceladus	Atmosphere	Triton
Neutral source[e] (kg/s)			600–2,600	70–750			
Plasma source[f] (kg/s)	5	~5	5	260–1,400	12–250	0.02	0.2
Plasma source (ions/s)	10^{26}	10^{26}	2×10^{26}	$>10^{28}$	$3–5 \times 10^{26}$	10^{25}	10^{25}
Lifetime	Minutes	Minutes	hours-days[g]	20–80 days	30–50 days	1–30 days	~1 day

[a] Water-group ions from ionization, dissociation, and recombination of water (OH^+, H_2O^+, H_3O^+)

[b] Mercury's tenuous atmosphere is a likely source of heavy ions

[c] There probably are ionospheric and solar wind sources but how they compare to satellites sources is not known

[d] Ionospheric plasma dominates the inner magnetosphere with solar wind sources being significant in the outer regions

[e] Net loss of neutrals from satellite/ring sources (Bagenal and Delamere 2011)

[f] Net production of plasma production (Bagenal 1992; Bagenal and Delamere 2011)

[g] Typical residence time in the magnetosphere. Plasma stays inside the plasmasphere for days but is convected through the outer magnetosphere in hours

of kelvins, a few hundred eV) solar corona, helium in the solar wind is fully ionized as He^{++} ions and comprises ~3–5% of the number density. Ionospheric plasma is much cooler (thousands of kelvins, <0.1 eV), so that ionospheric helium ions are mostly singly ionized. Thus, a measurement of the abundance ratios He^{++}/H^{+} and He^{+}/H^{+} would clearly distinguish the relative importance of these sources. Unfortunately, measuring the composition to such a level of detail is difficult for the bulk of the plasma, with energies in the range 1 eV–1 keV (e.g., Young 1997a, b, 1998). Measurement of composition is more feasible at higher energies, but then one needs to consider whether the process that has accelerated the ions within the magnetosphere, since they left the source region, is mass or charge dependent.

The temperature of a plasma can also be an indicator of its origin. Plasma in the ionosphere has characteristic temperatures of <0.1 eV; the ionization of neutral gases produces ions with energies associated with the local plasma flow speed, while material that has leaked in from the solar wind tends to have energies of a few keV. But, again, we need to consider carefully how a parcel of plasma may have heated or cooled as it moved through the magnetosphere to the location at which it is measured. *Figure 6-6* illustrates various ways in which ionospheric plasma enters the Earth's magnetosphere and evolves by different processes. As we explore other magnetospheres, we should expect similar levels of complexity.

Table 6-3 summarizes the main plasma characteristics of the six planetary magnetospheres. To a first approximation, one can say that escape of material from the satellites dominates the magnetospheres of Jupiter, Saturn, and Neptune, with ionospheric sources being secondary. Uranus having fewer, smaller, satellites; its weak ionospheric source probably supplies the main contribution. With only the most tenuous of exospheres, Mercury's magnetosphere contains mostly solar wind material, but energetic particle and photon bombardment of the surface may be a significant source of O^{+}, Na^{+}, K^{+}, Mg^{+}, etc. (Slavin 2004). At Earth, the net sources from the solar wind and ionosphere are probably comparable, though the most recent studies suggest that the ionospheric contribution seems to be dominant (e.g., Moore and Horwitz 2007).

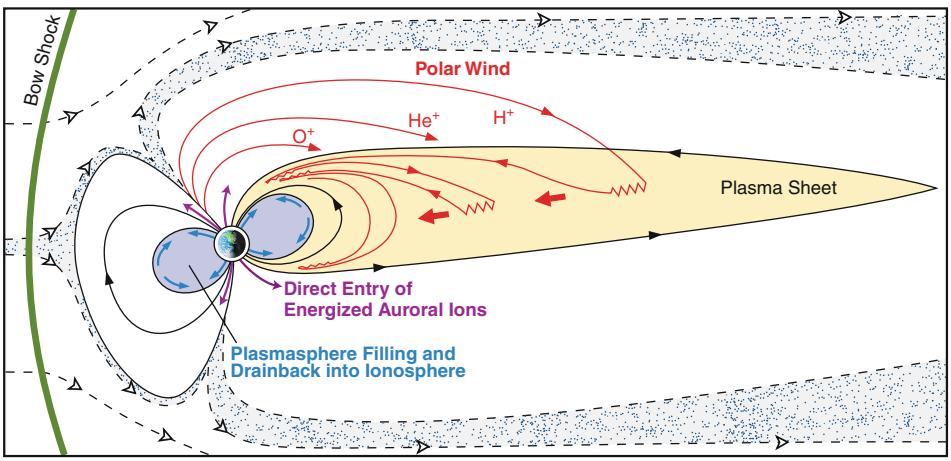


Fig. 6-6

Sources of plasma the Earth's magnetosphere (after Chappell 1988). The *shaded, dotted area* illustrates the boundary layer through which the solar wind plasma enters the magnetosphere

2.4 Plasma Dynamics

First we describe how charged particles move in response to specified electric **E** and magnetic **B** fields. Depending on the situation, a range of approaches can be taken, from treating each particle separately to regarding the plasma as a magnetized fluid, plus hybrid approaches in between. The particle approach is usually appropriate for very energetic particles (e.g., trapped in radiation belts). To model plasma behavior over larger spatial and temporal scales (e.g., global magnetospheric models), it is generally appropriate to use magnetohydrodynamics (MHD). Then there are some situations where electrons can be treated as a fluid but ions need to be treated as particles (e.g., in modeling the interaction with some of the satellites). The basic physics of space plasmas is described in textbooks such as Kivelson and Russell (1995), Gurnett and Bhattacharjee (2005), Gombosi (1998), or Cravens (1997).

After describing the motions of energetic particles in dipole magnetic fields close to the planet, the radiation belts of the major magnetospheres are compared and their properties listed in *Table 6-4*. Moving farther away from the planet, the magnetic field becomes weaker and can be modified from a simple dipole by electrical currents flowing through the plasma. Theoretical ideas are applied to the different planetary magnetospheres to determine where and when plasma flows are predominantly rotation with the planet vs. controlled by the interaction of the solar wind with the magnetosphere. *Table 6-5* lists various dynamical parameters of the different planetary magnetospheres that quantify the relative importance of rotational vs. solar wind influences in each case.

When comparing the dynamics of different magnetospheres, the traditional approach has been to compare electric fields and electric currents (i.e., **E, J**). Over the past decade, the case has been made (Parker 2007; Vasyliūnas 2001, 2011) that such an "electrical circuit" approach is only valid for quasi-static situations (specifically, where temporal changes occur over time scales that are long compared with the transit time of Alfven waves across the system) and that one should derive the flows and magnetic fields (**B**, **v**) resulting from the various stresses on the system. This review presents the traditional (**E, J**) approach partly because it is the one that dominates the current literature but also because it is perhaps easier to explain the interactions between different components of a complex system.

Table 6-4
Energetic particle characteristics in planetary magnetospheres

	Earth	Jupiter	Saturn	Uranus	Neptune
Phase space density[a]	20,000	200,000	60,000	800	800
Ring current[b] ΔB (nT)	10–23	200	10	<1	<0.1
Plasma β[c]	<1	10–100	1–5	~0.1	~0.2
Auroral power (W)	10^{10}	10^{12}	10^{11}	5×10^9	2–8 $\times 10^7$

[a]The phase space density of energetic particles (in this case 100 MeV/G ions) is measured in units of $c^2(cm^2 s\ sr\ MeV^3)^{-1}$ and is listed near its maximum value (Cheng et al. 1987; Mauk et al. 1995)

[b]The magnetic field produced at the surface of the planet due to the ring current of energetic particles in the planet's magnetosphere

[c]The ratio of the thermal pressure to magnetic pressure of a plasma, $\beta = nkT/(B^2/2\mu_o)$. These values are typical for the body of the magnetosphere. Higher values are often found in the tail plasma sheet and, in the case of the Earth, at times of enhanced ring current

Table 6-5
Dynamical characteristics of planetary magnetospheres

	Mercury	Earth	Jupiter	Saturn	Uranus	Neptune
R_{MP}[a] (km)	4,000	6.5×10^4	6×10^6	1×10^6	6×10^5	6×10^5
V_{sw}[b] (km/s)	370	390	420	430	450	460
t_{N-T}[c]	10 s	3 min	4 h	45 min	20 min	20 min
R_T[d] (R_p)	3	20	170	40	50	50
R_T[d] (km)	8,000	1.3×10^5	1.2×10^7	2.3×10^6	1.3×10^6	1.2×10^6
$V_{rec,1}$[e] (km/s)	40	22	16	16	16	16
$V_{rec,2}$[f] (km/s)	37	39	42	43	45	46
t_{rec}[g]	3 min	1 h	80 h	15 h	8 h	7 h
d_X[h] (R_p)	30	200	1,700	400	500	500
$V_{co}/V_{rec,2}$[i]	4×10^{-5}	0.04	8	1.3	0.4	0.4
R_{pp}[j] (R_p)	0.03	6.7	350	95	70	70

[a] Subsolar magnetopause distance
[b] $V_{sw} = 387(a_p/a_E)^{0.05}$ (km/s) from Belcher et al. (1993)
[c] Solar wind nose-terminator time: $t_{N-T} = R_{MP} / V_{sw}$
[d] Radius of cross section of magnetotail, approximated as $R_T = 2R_{MP}$
[e] Reconnection speed assuming 20% reconnection efficiency and $v_{rec} \sim 0.2\, v_{sw}B_{sw}/ B_{MP}$ km/s (e.g., Kivelson 2007)
[f] Reconnection speed assuming 10% reconnection efficiency and $v_{rec} \sim 0.1\, v_{sw}$ km/s
[g] Reconnection time $t_{rec} = R_T/ v_{rec,2}$(s)
[h] Distance to X-line $d_x \sim v_{sw}\, t_{rec}$
[i] Assumes rotation speed at the magnetopause is ~30% of rigid corotation
[j] Distance to plasmapause, where corotation is comparable to reconnection flow (e.g., Kivelson 2007)

2.4.1 Energetic Particles and Radiation Belts

A particle with charge q, mass m, and velocity v in an electric field E and magnetic field B experiences a Lorentz force which causes the particle to accelerate

$$F = q(E + v \times B) = mdv/dt, \tag{6.8}$$

Solving (6.8) is relatively straightforward if E and B are specified. For the case of a dipole magnetic field, charged particles exhibit motions on three temporal and spatial scales, as illustrated in *Fig. 6-7a*. On the shortest time scale, particles gyrate about the magnetic field with a gyroradius of

$$R_g = mv_\perp/qB, \tag{6.9}$$

where $v_\perp$ is the speed of the particle perpendicular to the magnetic field. Positively and negatively charged particles gyrate in opposite directions. As a particle moves along the magnetic field, it experiences stronger magnetic field as it approaches the poles. In stronger fields, the gyromotion is increased and, through conservation of total energy, the particle motion along the field decreases. Thus, the particle is trapped and "bounces" between polar regions of stronger fields. On longer time scales, the particles experience drifts around the planet (*Fig. 6-7a*) producing a "belt" of trapped particles.

The particle source and loss processes act on much longer time scales (hours–years at Earth, months–years at Jupiter). Radial (cross-field) motions are diffusive – mostly scattering by small-scale perturbations in the magnetic field. Ultimately, the particles eventually escape

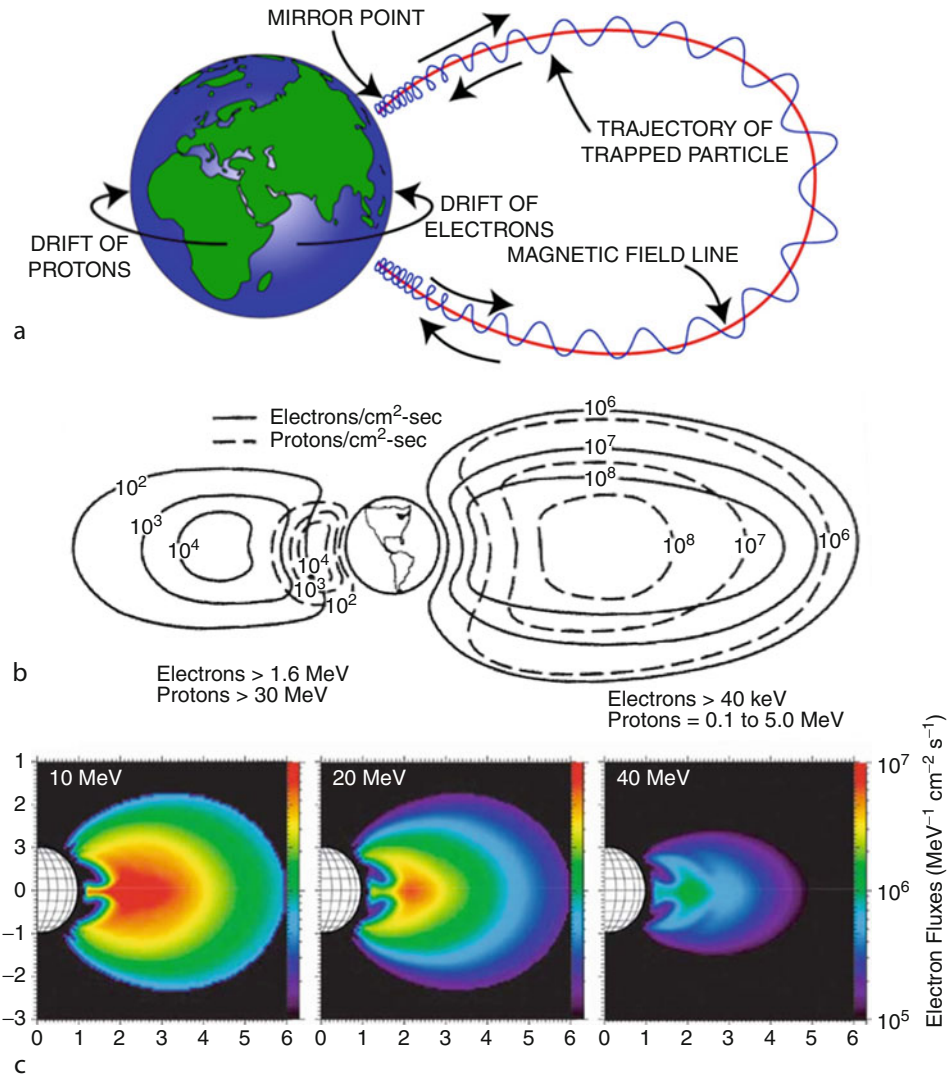


Fig. 6-7

(a) Radiation belt dynamics. Flux of radiation belt particles at (b) Earth (from Walt 2005) and (c) Jupiter (From Santos-Costa and Bourdarie 2001)

the magnetosphere, charge exchange with neutral particles (producing energetic fast neutrals that escape the system), or are lost by hitting moons or the planet's atmosphere.

Figure 6-7b shows typical fluxes of energetic electrons and protons measured in the Earth's radiation belts (for introductory text see Walt 2005). The higher energy particles (few-100s MeV, left of *Fig. 6-7b*) are produced via a process called cosmic ray albedo neutron decay (CRAND, whereby neutrons generated by cosmic rays bombarding the atmosphere decay to produce protons and electrons) and are confined closer to the planet. The lower energy particles (10s–100s KeV, right of *Fig. 6-7b*) are "injected" into the inner magnetosphere from the

magnetotail during magnetic storms, dominate the particle energy density, and carry the ring current. Luckily, the most useful orbits for satellites (low Earth orbit and geosynchronous orbit at 6.6 R_{Earth}) generally avoid the radiation belts, but at times of high geomagnetic activity, sensitive electronics as well as astronauts can be exposed to significant fluxes of damaging energetic particles.

As the discovery of Earth's radiation belts marked the dawn of the space age, the nearly simultaneous detection of radio emission from Jupiter in the late 1950s started the exploration of the jovian magnetosphere. The radiation belts of Jupiter have been observed via radio emission from MeV electrons which generate intense synchrotron emission at decimetric wavelengths observed from Earth-based radio telescopes as well as by the Cassini spacecraft on its way to Saturn (see the review by Bolton et al. 2004). *Figure 6-7c* shows the distribution of energetic electron fluxes at 10, 20, and 40 MeV derived from modeling emissions at different wavelengths (Santos-Costa 2001). These high fluxes of energetic electrons (and the accompanying energetic ions) provide a quick, lethal dose to sensitive electronics so that spacecraft aimed close to Jupiter need to avoid and/or be protected against high radiation doses. Such a mission, Juno, will go into polar orbit in the summer of 2016, skimming over Jupiter's clouds and ducking under the radiation belts.

The smaller physical scale and shorter time scales of the Saturn system result in less net acceleration and weaker fluxes of energetic particles. Absorption by the majestic ring system further prevents the buildup of comparable fluxes close to the planet, so that there are no belts emitting synchrotron radiation at Saturn. Significant populations of energetic particles were detected at Uranus and Neptune, but the fluxes were much lower than at Jupiter and Saturn. It could be that the shorter residence times in these smaller magnetospheres limit the amount of acceleration, or it may be much harder for particles to be stably trapped in such non-dipolar fields. The trapped populations of energetic particles in the magnetospheres of the major planets are compared by Cheng et al. (1987), Mauk et al. (1995), and Mauk and Fox (2010). In the mini-magnetospheres of Ganymede and Mercury, the time scales for energetic particles to drift around these objects are only minutes, suggesting that particles are not stably trapped (see review by Kivelson 2007).

In the cases of all the giant planets, the observed high fluxes of energetic particles in the middle magnetosphere and compositional evidence imply that some fraction of the thermal plasma is accelerated to high energies, either by tapping the rotational energy of the planet (in the cases of Jupiter and Saturn) or by processes in the non-dipolar fields of the tail (at Earth and probably Uranus and Neptune). If the energy density of the energetic particle populations is comparable to the magnetic field (i.e., $\beta > 1$ where $\beta = nkT/(B^2/\mu_0)$, as shown in *Table 6-4* for Jupiter and Saturn), then particle pressures inflate and stretch out the magnetic field and generate strong currents in the equatorial plasma disk. While Uranus and Neptune have significant radiation belts, the energy density remains small compared with the magnetic field (i.e., $\beta \ll 1$, see *Table 6-4*).

Figure 6-8 compares the magnetospheres of Mercury, Earth, Saturn, and Jupiter which are scaled to the Chapman–Ferraro distance (Sect. 2.2) that assumes the internal plasma pressure is negligible and that the planetary field is a dipole. *Figure 6-8* illustrates how the substantial plasma pressure inside at Jupiter (and to a lesser extent at Saturn) expands the magnetosphere. At Jupiter, the high plasma pressures in the plasma sheet dominate the local magnetic field pressure, producing values of β greater than unity beyond ~ 15 R_J, increasing to greater than 100 at 45 R_J (Mauk et al. 2004). Not only does the plasma pressure dominate the magnetic pressure, but the radial profile of plasma pressure is also considerably flatter than the $R^{-1/6}$ variation in

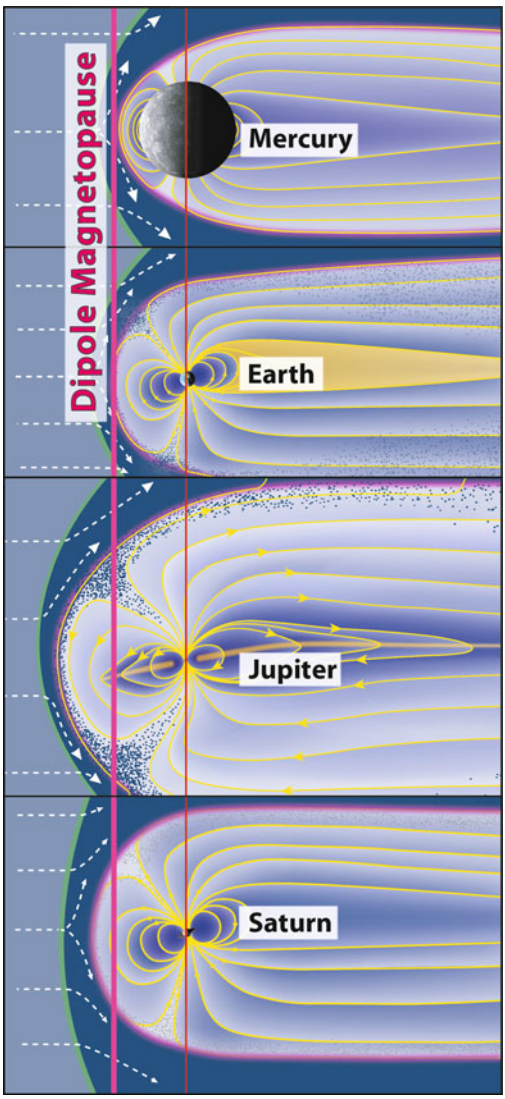


Fig. 6-8

Magnetospheres of Mercury, Earth, Jupiter, and Saturn scaled to the distance of the magnetopause for a dipole field (based on Chapman and Ferraro 1930). Jupiter and Saturn have extended magnetospheres due to the significant plasma pressure inside

magnetic pressure for a dipole field. It is the high plasma pressure in the plasma disk that doubles the scale of Jupiter's magnetosphere from the dipolar stand-off distance of ~42 R_J to 65–90R_J. A simple pressure balance between the ram pressure of the solar wind and the magnetic pressure of a dipole produces a weak variation in the terrestrial dayside magnetopause distance R_{MP} for a solar wind density ρ and speed v_{sw} such that $R_{MP} \propto (\rho v_{sw}^2)^{-1/6}$. Measurements of the magnetopause locations at Jupiter indicate a much stronger variation, $R_{MP} \propto (\rho v_{sw}^2)^{-1/4.5}$

(Slavin et al. 1985; Huddleston et al. 1998; Joy et al. 2002; Alexeev and Belenkaya 2005). Consequently, a factor 10 variation in ram pressure at Earth changes the magnetopause distance by only 70%, while at Jupiter, the tenfold variations in solar wind pressure often observed at 5 AU cause the dayside magnetopause to move between ~100R_J and ~50R_J. At Saturn, the plasma pressures are less than Jupiter but the plasma beta is still greater than unity beyond 8 Rs (e.g., Sergis et al. 2010) and has values of 2–5 in the plasma sheet. The more modest values of beta at Saturn are consistent with the magnetopause standoff distance varying as −1/5 power of solar wind pressure, as found by Kanani et al. (2010).

2.4.2 Rotational Flows

Magnetospheric configuration is generally well described by MHD in which the magnetic field can be considered frozen into the plasma flow. Thus, we need to consider the processes controlling magnetospheric flows. The two largest sources of momentum in planetary magnetospheres are the planet's rotation and the solar wind. The nature of large-scale circulation of material in the magnetosphere depends on which momentum source is tapped. For planetary magnetospheres, corotation of plasma with the planet is a useful first approximation with any departures from strict corotation occurring when certain conditions (described below) break down. It may be helpful to think of plasma in the magnetosphere as mass that is coupled by means of magnetic field lines to a giant flywheel (the planet) with the ionosphere (or magnetosphere just above) acting as the clutch. *Figure 6-9* illustrates the dynamical process whereby the magnetospheric plasma is coupled to the angular momentum of the spinning planet (for a detailed mathematical description and further references see Vasyliunas 1983). The region within a magnetosphere where the flow is predominately rotational is called the *plasmasphere*.

For a magnetospheric plasma to rotate with the planet, the upper region of the neutral atmosphere must corotate with the planet and be closely couple to the ionosphere by collisions. The electrical conductivity of the ionosphere σ^i is large so that in a corotating ionosphere (with velocity V^i), any horizontal currents (perpendicular to the local magnetic field) are given by Ohm's law

$$J_\perp^i = \sigma^i (E^i + V^i \times B). \tag{6.10}$$

Just above the ionosphere where the conductivity perpendicular to the magnetic field in the (collision-free) magnetosphere, $\sigma^m{}_\perp$ is essentially zero and $E^m = -V^m \times B$. Because the plasma particles are far more mobile in the direction of the local magnetic field, the parallel conductivity $\sigma_{//}^m{}_\perp$ is large and the field lines can be considered to be equipotentials ($\mathbf{E} \bullet \mathbf{B} = 0$). Thus, the electric field in the magnetosphere can be mapped into the ionosphere (*Fig. 6-8a*). Because the ionosphere is relatively thin, the electric field E^m just above the ionosphere is the same as E^i so that we can write

$$J_\perp^i = \sigma^i (V^i - V^m) \times B, \tag{6.11}$$

The condition for the corotation of the magnetospheric plasma is that the ratio $J_\perp^i / \sigma^i$ be sufficiently small so that

$$V^m = V^i = \Omega \times R. \tag{6.12}$$

The corotational electric field is therefore

$$E_{\text{cor}} = -(\Omega \times R) \times B_{\text{planet}}. \tag{6.13}$$

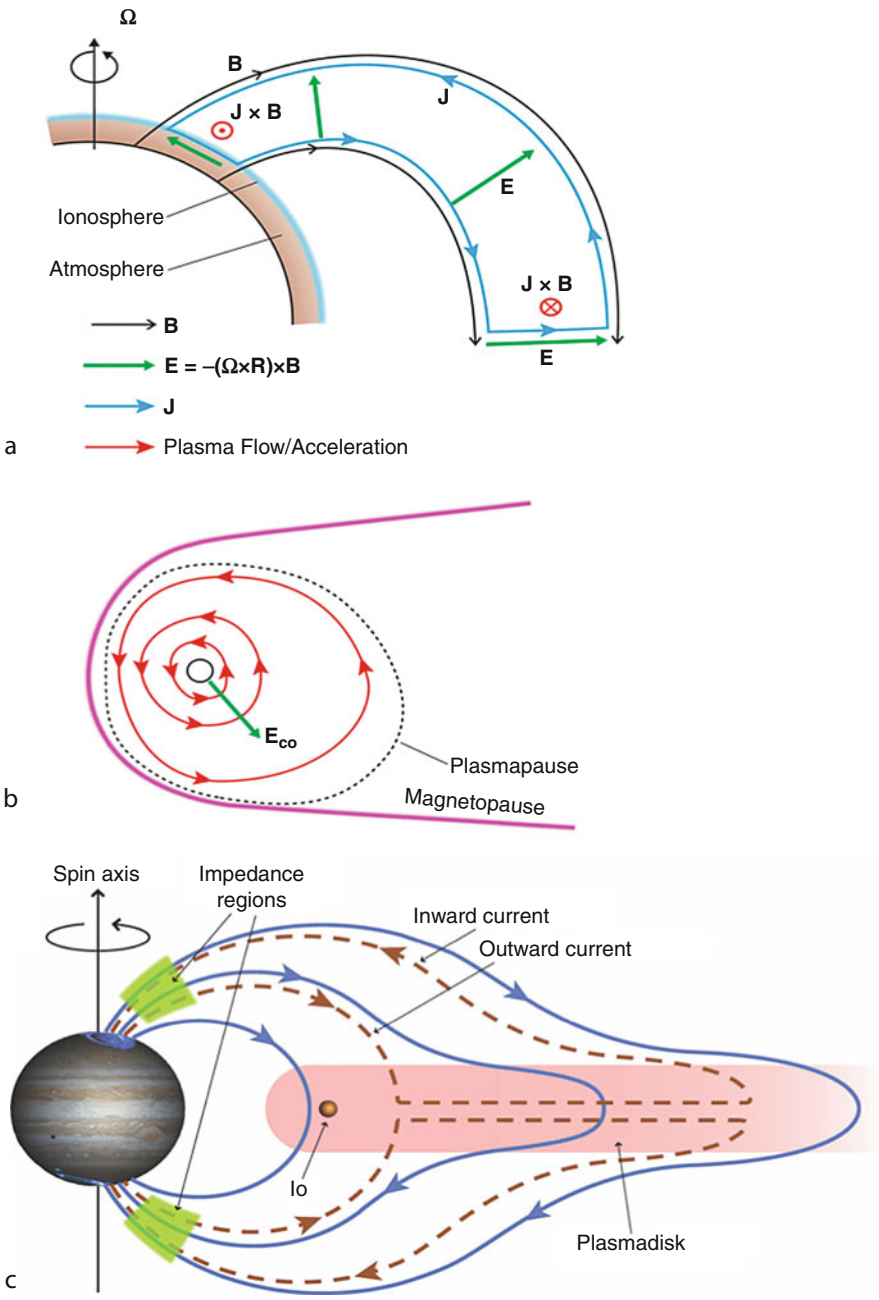


Fig. 6-9

Dynamics of rotation-dominated magnetospheres. (a) The electrodynamics in the meridional plane. (b) Dynamics in the equatorial plane. The plasmapause separates rotation-driven flows (in the plasmasphere) from solar-wind-driven flows outside. (c) In a magnetosphere where rotation confines the plasma to the equator (as illustrated at Jupiter) there is a lack of charged particles able to carry electrical currents along the magnetic field. Based on experience at Earth, it is expected that this leads to the development of regions of high impedance at high latitudes

For a dipolar magnetic field that is aligned with the rotation axis, the corotational electric field (in the equatorial plane, see Fig. 6-9b) is radial with magnitude

$$E_{\mathrm{cor}} = \Omega Bo/r^2. \tag{6.14}$$

It is clear that large ionospheric conductivities facilitate corotation. A large $\sigma^m_{//}$ also means that any currents in the magnetosphere that result from mechanical stresses on the plasma are directly coupled by field-aligned currents to the ionosphere. Thus, corotation breaks down when mechanical stresses on the magnetospheric plasma drive ionospheric currents that are sufficiently large for the ratio $J^i_\perp/\sigma^i$ to become significant. Such conditions might occur in regions where there are large increases in mass density due to local ionization of neutral material, where there are strong radial motions of the plasma, or where external stresses begin to influence the magnetospheric plasma. When the magnetosphere imposes too large a load, the ionospheric clutch begins to slip.

The above argument, originally made by Brice and Ioannidis (1970), quantified by Hill (1979), and reviewed by Vasyliunas (1983) and Mauk et al. (2009), assumes that there are plenty of particles around to carry any currents between the magnetosphere and the ionosphere. In magnetospheres that are dominated by rotation, centrifugal forces confine the ions (which pull in the electrons electrostatically) to the region along a magnetic flux tube that is farthest from the rotation axis. The net results are to stretch the magnetic field in the equator (see Fig. 6-9c) and to limit the electrical conductivity along the field line ($\sigma^m_{//}$). Consequently, significant parallel electric fields develop in the magnetosphere, probably confined to small regions quite close to the planet, labeled impedance regions in Fig. 6-8c (Mauk et al. 2002; Ergun et al. 2009; Ray et al. 2009, 2010). Such parallel electric fields accelerate electrons into the atmosphere (in regions of upward current) where they trigger strong auroral emissions. Thus, the region where corotation with the planet begins to break down can be associated with bright aurora (see review by Clarke et al. 2005).

Because the plasma is magnetically trapped in a rotation-dominated magnetosphere, transport away from the source implies either inward or outward radial transport across the magnetic field. In a magnetosphere that is dominated by rotation, outward transport is energetically favored over inward transport. As plasma builds up (e.g., from ionization of material coming from moons), it becomes energetically favorable for magnetic flux tubes laden with plasma to interchange with outward neighbors that contain less plasma. This process of flux tube interchange is thought to be responsible for transport of plasma on times scales of weeks through the giant magnetospheres, but the exact process and the mechanisms that control the radial transport rate are far from understood in detail (see review in Thomas et al. 2004). Furthermore, one expects plasma that expands into a larger volume to become colder as it moves outward. Yet the plasmas at both Jupiter and Saturn are hotter at larger radial distance. The issue of what is heating the plasmas of these magnetospheres remains a major conundrum of planetary magnetospheres (see review by Bagenal and Delamere 2011).

2.4.3 Global Solar-Wind-Driven Convection

Next, let us consider how the momentum of the solar wind is harnessed by processes occurring near the magnetopause. In the early 1960s, there was a debate about how these processes operate. Axford and Hines (1961) proposed a viscous interaction at the magnetopause boundary.

This idea was dismissed for the Earth because (a) people could not see how collisionless plasmas could have "viscosity" and (b) the observations supported the alternative idea. Recently, there has been a revival of interest in small-scale processes that might act like an effective viscosity, and we shall return to such ideas at the end of this section. In the meantime, Dungey (1961) showed how, under certain conditions, the solar magnetic field interconnects with the planetary magnetic field. *Figure 6-10* shows how reconnection of the planet's magnetic field with the interplanetary field harnesses the momentum of the solar wind and drives the circulation of plasma within the magnetosphere; this circulation is sometimes called the Dungey cycle.

The first task is to quantify the spatial and temporal scales over which the Dungey cycle operates at each planet. The actual process of reconnection (where adjacent magnetic field lines of different orientations are "cut and reconnected" as in steps 1 and 6 of *Fig. 6-9*) is a plasma process that occurs on very small scales. Reconnection proceeds when the IMF brought to the magnetopause by the solar wind has a component of the embedded magnetic field that is antiparallel to the planetary magnetic field just inside the magnetopause (step 1). The reconnected flux tubes is limited to relatively small structures – flux transfer events (FTEs) – whose recurrence and fractional scale on the magnetopause decreases as the Alfven Mach number of the flow seems to increase, consistent with a lower rate of magnetic flux being convected into the magnetopause (see Jia et al. (2010b) for a comparison of FTEs at different planets). Magnetic reconnection efficiency can be strongly reduced in high (>10) magnetosonic Mach solar wind flows due to the dominance of plasma pressure forces over magnetic forces (Scurry and Russell 1991; Scurry et al. 1994), though (Grocott et al. 2009) found, no evidence for a significant reduction. The factors controlling the reconnection efficiency of the IMF and planetary fields are active areas of research (La Belle-Hamer et al. 1995; Swisdak et al. 2003; Cassak and Shay 2011).

Figure 6-10 shows that the reconnected field lines (e.g., 2 and 7) are "bent" indicating strong currents and tension forces. While the microscale process that initiates the reconnection is dissipative, the net result is the release of considerable magnetic tension that accelerates plasma from the reconnection point, generating beams of energetic particles. Reconnection is a major source of energy in the solar corona as well as a source of energetic particles in the Earth's magnetosphere.

Consider the situation where some fraction of the time there is a component of the IMF that is opposite to the direction of the planetary magnetic field at the magnetopause (e.g., a negative *Bz* for Earth and a positive *Bz* for Jupiter and Saturn, ignoring the complexities of Uranus and Neptune for the moment). Such a configuration allows the reconnection of planetary and interplanetary fields at the dayside magnetopause (see step 1 of *Fig. 6-10*). There is now one end of the flux tube attached to the planet and the other is out in the solar wind. To estimate how long it takes the section of flux tube in the solar wind to move to the plane of the planet's terminator (step 3), the subsolar magnetopause distance R_{MP} is divided by the local solar wind speed V_{sw}. For *Table 6-5*, an empirical fit to Voyager data is used that includes a modest increase in the solar wind speed with distance from the Sun, but the basic results would not be very different if a constant value for the solar wind (say ~400 km/s) were used. One immediately sees the effect of the vast scale of the giant magnetospheres of the outer planets: the nose-terminator time scale, τ_{N-T}, is a mere 10 s at Mercury, 3 min at Earth, and as much as 4 h at Jupiter.

The next step is to calculate how long the open flux tube would take to convect to the equator or central plane of the magnetotail (from steps 3 to 6 in *Fig. 6-10*). For simplicity, the radius of

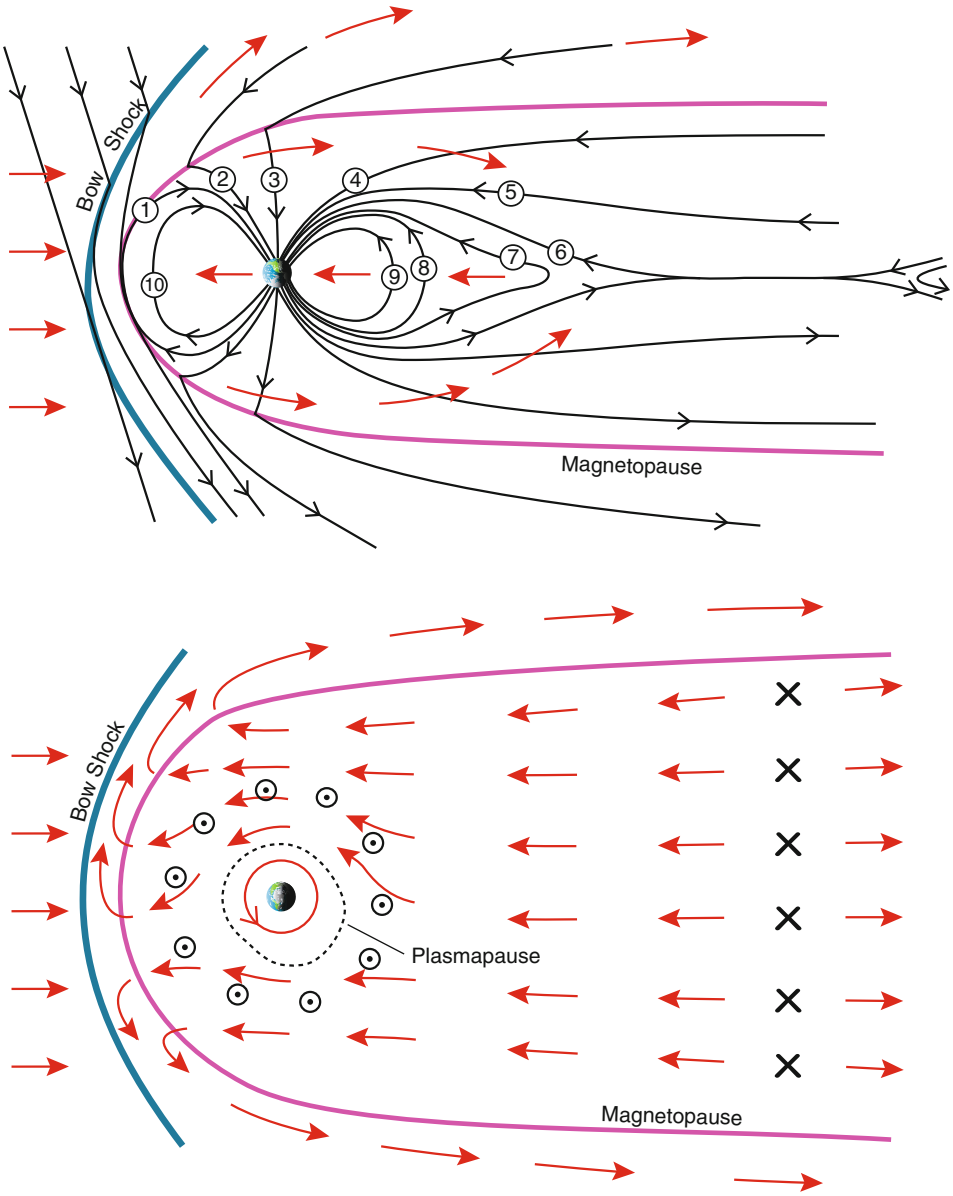


Fig. 6-10

Magnetospheric dynamics associated with the Dungey cycle driven by the solar wind. *Top*: view in the noon-midnight meridian plane. The numbers show the time sequence for a flux tube being reconnected at the dayside magnetopause and convected through the magnetosphere. *Bottom*: view in the equatorial plane (After Dungey (1961))

each magnetotail has been approximated as twice the subsolar standoff distance (i.e., $2R_{MP}$). We need to divide this distance by a convective speed to estimate a minimum convective time scale. The traditional approach to calculating the speed of circulation in the magnetosphere driven by

solar wind was to calculate the electric field associated with an object moving with the planet relative to the solar wind, $E_{sw} = -V_{sw} \times B_{IMF}$, assume that some fraction (say, 10–20%) of this electric field permeates the whole magnetosphere (i.e., the convective electric field $E_{con} \bullet 0.1–0.2\, E_{sw}$), and then estimate how magnetospheric plasma would drift in this convection electric field and the local planetary magnetic field ($V_{con} = E_{con} \times B_{planet}$) (e.g., Cravens 1997).

In the meantime, to obtain a rough upper estimate for a reconnection-driven convection speed, we have taken 10% of the solar wind speed (roughly 40 km/s at all planets), corresponding to a ~10% reconnection efficiency. Again, the large scales of the giant planet magnetospheres mean that even with generous values for the convection speed, one obtains long time scales for flux tubes to convect to the equator from the upper and lower magnetopause boundaries. At Jupiter, this time scale is 80 h, equivalent to eight full rotation periods. The time scales for steps 3–6 of the Dungey cycle for the other giant planets are much less, but they are still several hours and comparable with the planetary rotation rate. By contrast, this convection time scale is just an hour at Earth and a few minutes at Mercury.

The Dungey cycle time scale mentioned above can also be used to estimate the length of the magnetotail, by multiplying the reconnection time scale and the solar wind speed. More accurately, it gives us the distance down the tail to the X-line, where further reconnection closes the open magnetic flux (hence conserving, on average, the total magnetic flux emanating from the planet). The re-closed magnetic flux tube then convects sunward (steps 7–10 in *Fig. 6-10*) to begin the Dungey cycle again at the dayside magnetopause. *Table 6-5* shows that values for this X-line (often called, for obscure reasons, the distant Earth neutral line). This X-line distance is about 20 R_{MP} if one takes the simplest formular for reconnection-driven convective speed V_{con} to be 10% of V_{sw} and the tail radius to be $2R_{MP}$. Lower estimates of V_{con} (e.g., derived including field compression by Kivelson 2007) give larger distances to the tail X-line. In practice, we know that the Earth's tail extends for several thousand R_E, while Jupiter's magnetotail was encountered by Voyager 2 as it approached Saturn at a distance greater than $9{,}000R_J$ or 4 AU downstream of Jupiter. The estimates of distances to magnetotail X-lines derived from simple Dungey cycle principles shown in *Table 6-4* illustrate the vast scales of the magnetospheres of the outer planets, and the huge distances that flux tubes reconnecting (re-closing) in the tail would need to travel back to the planet if these magnetospheres were driven by Earth-like processes.

We compared the corotation speed $V_{cor} = \Omega \times R$ with our upper estimate of the convection flows driven by reconnection, V_{con}. The very low values in *Table 6-5* of V_{cor}/V_{con} for Mercury and Earth confirm that the dynamics of these magnetospheres are dominated by coupling to the solar wind, while it is clearly the case that rotation dominates Jupiter and Saturn. Uranus and Neptune, once again, are not simple cases with speed ratios of order unity that would suggest the comparable importance of rotation and solar-wind-driven circulation.

In a general sense, close to the planet where the magnetic field is strong and rotation speeds are low, one expects strong coupling to the planet's rotation. At larger distances from the planet, one expects decreasing corotation and an increasing influence of the solar wind. Finally, we can estimate the size R_{pp} of a region (called the *plasmapause* at Earth) within which rotation flows dominate and outside of which the solar wind interaction drives flows. The values for R_{pp} in the bottom row of *Table 6-5* further illustrate how the planets' magnetospheres span the range between the extremes of Jupiter (where $R_{pp} \gg 1$ and rotation dominates throughout) and Mercury (where $R_{PP} \ll 1$ means that there is no region of corotating plasma in the tiny magnetosphere).

Fundamentally, dawn-dusk asymmetry of the magnetic field and plasma flows inside a planetary magnetosphere is the result of the solar wind interaction with the magnetosphere. To account for these solar-wind-driven dawn-dusk asymmetries, either a mechanism such as reconnection-driven global convection (as in the Dungey cycle above) is evoked or one might consider a mechanism more akin to the original (Axford and Hines 1961) model of a viscous interaction of the solar wind with the magnetopause boundary. Rather than a Dungey-style global cycle of reconnection (that opens planetary magnetic flux on the dayside, carries the flux tube over the poles, and closes the flux in the tail), the magnetic flux could be opened and closed intermittently in small-scale structures in turbulent interaction regions on the flanks of the magnetosphere.

So what might be responsible for viscous processes at the magnetosphere? Except in the densest locations (e.g., ionospheres), space plasmas are generally collisionless (i.e., the mean free path is larger than the typical scale of the system). Thus, it is necessary to find other dissipative processes occurring on a scale comparable to the magnetopause thickness. But there are a variety of waves, perhaps driven by shear (Kelvin–Helmoltz) instabilities or by nonequilibrium particle velocity distributions, that might act as a means for the solar wind plasma to interaction with the plasma at the boundary of the magnetosphere, particularly when the magnetic field on either side of the boundary is weak. With strong contrast in flows across the magnetopause, the Kelvin–Helmholtz instability (KHI) – analogous to "wind over water" instability of hydrodynamics – is a good possibility. Observations at Earth and numerical models show that KHI vortices generate twisted magnetic fields, strong currents, and small-scale reconnections (both opening and closing flux intermittently) that allow plasma transport across the boundary (see review by Delamere and Bagenal 2010). Such a viscous process could be considered a comet-like interaction, with the IMF being temporarily "hung up" on the magnetopause, is to stretch the IMF out behind the object in an extended tail. This would mean that the magnetic field in the magnetotail would not be attached to the planet (as implied by ❯ *Fig. 6-10* for the Dungey cycle) but has each end in the solar wind and is "kinked" where flux tubes are dragged over the magnetopause.

2.4.4 Plasmoid Ejection

The Dungey cycle of opening magnetic flux on the dayside of a magnetosphere and subsequent closing in the magnetic tail produces a pinching off of the nightside plasma sheet and ejection of magnetospheric plasma down the magnetotail (see far right of ❯ *Fig. 6-10*). The ejection of such a blob of plasma – a *plasmoid* – involves rapid conversion of energy stored in the stretched magnetic field into kinetic energy of the ejected plasmoid as well as beams of energetic particles. Such an explosive ejection of material and the associated phenomena is called a "substorm" at Earth. Substorms occur frequently (on average several times per day) at Earth from an X-line that ranges from 8 to 20 R_E.

For a magnetosphere that is driven primarily by rotation rather than the solar wind (i.e., R_{pp} in listed in ❯ *Table 6-5* is large), as the plasma rotates around onto the nightside, it is no longer confined by magnetopause currents, moves farther from the planet, and stretches the magnetic field with it (field line (1) in ❯ *Fig. 6-11*). At some point, either the coupling to the planet breaks down completely (e.g., because the Alfven travel time between the equator and the poles becomes a substantial fraction of a rotational period) or the field becomes so radially

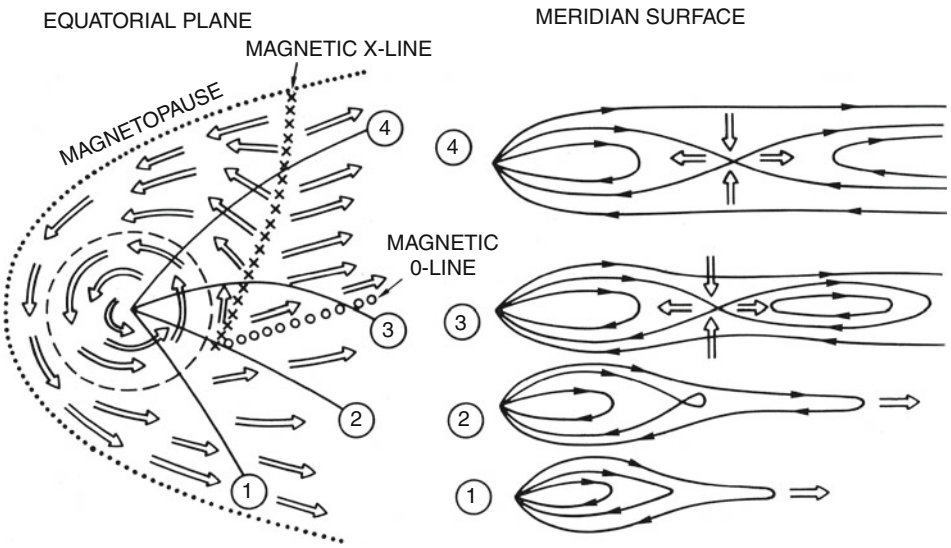


Fig. 6-11
Qualitative sketch of plasma flow (*left*) in the equatorial plane and (*right*) in a sequence of meridian surfaces (locations 1, 2, 3, and 4) expected from the planetary wind model (From Vasyliunas 1983)

extended that an x-point develops and a blob of plasma detaches and escapes down the magnetotail (field lines (2), (3), and (4) in *Fig. 6-11* from Vasyliunas 1983). Kivelson and Southwood (2005) point out that the stretched, equatorial magnetic field becomes so weak that the gyroradii of the heavy ions becomes comparable to scales of local gradients. It is possible that the plasma diffuses across the magnetic field and "drizzles" down the magnetotail. If the process were entirely diffusive, then the magnetic flux would remain connected to the planet. The flux tubes would become unloaded and presumably shrink ("dipolarize") as they swung around to the dayside.

It is quite possible that Dungey cycle transport of flux toward the center of the magnetotail acts in combination with rotationally driven expulsion of plasmoids (sometimes called the Vasyliunas cycle), depending on how much flux is opened by large-scale dayside reconnection and whether the opened flux tubes penetrate deep into the tail vs. being closed by reconnection on the flanks of the magnetosphere.

3 Magnetospheres of the Outer Planets

The Voyager flybys of all four giant planets allowed comparison of their magnetospheres (e.g., Bagenal 1992). While all four are dominated by rotation, they can be separated into large, regular, and fast rotators (Jupiter and Saturn) vs. irregular oblique rotators (Uranus and Neptune). We discuss each planetary magnetosphere in turn below.

3.1 Jupiter

Jupiter is a planet of superlatives: the most massive planet in the solar system, which rotates the fastest, has the strongest magnetic field and has the most massive satellite system of any planet. These unique properties lead to volcanoes on Io and a population of energetic plasma trapped in the magnetic field that provides a physical link between the satellites, particularly Io, and the planet Jupiter. For those seeking further details, the jovian magnetosphere is reviewed in seven chapters of the book *Jupiter: The Planet, Satellites and Magnetosphere* (Bagenal et al. 2004), and only subsequent research is cited in this section.

Clear indications that Jupiter traps electrons in its magnetic field were apparent as soon as astronomers turned radio receivers to the sky. Early radio measurements showed that Jupiter has a strong magnetic field tilted about 10° from the spin axis, that energetic (MeV) electrons were trapped at the equator close to the planet, and that Io must be interacting with the surrounding plasma and triggering bursts of emission. The magnetometers and particle detectors on Pioneer 10 (1973) and Pioneer 11 (1974) revealed the vastness of Jupiter's magnetosphere and made in situ measurements of energetic ions and electrons. The Voyager 1 flyby in 1979 revealed Io's prodigious volcanic activity, thus explaining why this innermost Galilean moon plays such a strong role. Additional data came from subsequent traversals by the Ulysses (1992), Cassini (2000), and New Horizons (2007) spacecraft, but it was the 34 orbits of Galileo (1995–2003) around Jupiter that mapped out magnetospheric structures and monitored their temporal variability. As at Earth, magnetospheric activity is projected onto the planet's atmosphere via auroral emissions; this has been observed from X-rays to radio wavelengths with ground- and space-based telescopes. Jupiter has the advantage for us over the rest of the outer planets of not just being very large but also being much closer, allowing high-quality measurements to be made from Earth.

The magnetosphere of Jupiter extends well beyond the orbits of the Galilean satellite system (*Fig. 6-4*), and it is these moons that provide much of the plasma (*Table 6-3*) and some interesting magnetospheric phenomena. In particular, Io loses about 1 ton/s of atmospheric material (mostly SO_2 and dissociation products), which, when ionized to sulfur and oxygen ions, becomes trapped in Jupiter's magnetic field (*Fig. 6-12*). Coupling to Jupiter causes the magnetospheric plasma to corotate with the planet. Strong centrifugal forces confine the plasma toward the equator. Thus, the densest plasma forms a torus around Jupiter at the orbit of Io.

Compared with the local plasma, which is corotating with Jupiter at 74 km/s, the neutral atoms are moving slowly, close to Io's orbital speed of 17 km/s. When a neutral atom becomes ionized (via electron impact), it experiences an electric field, resulting in a gyromotion of 57 km/s. Thus, new S^+ and O^+ ions gain 540 and 270 eV in gyro-energy. The new "pickup" ion is also accelerated up to the speed of the surrounding plasma. The necessary momentum comes from the torus plasma, which is in turn coupled, via field-aligned currents, to Jupiter – the jovian flywheel being the ultimate source of momentum and energy for most processes in the magnetosphere. About one-third to one-half of the neutral atoms are ionized to produce additional fresh plasma, while the rest are lost via reactions in which a neutral atom exchanges an electron with a torus ion. On becoming neutralized, the particle is no longer confined by the magnetic field and flies off as an energetic neutral atom. This charge-exchange process adds gyro-energy to the ions and extracts momentum from the surrounding plasma, but it does not add more plasma to the system.

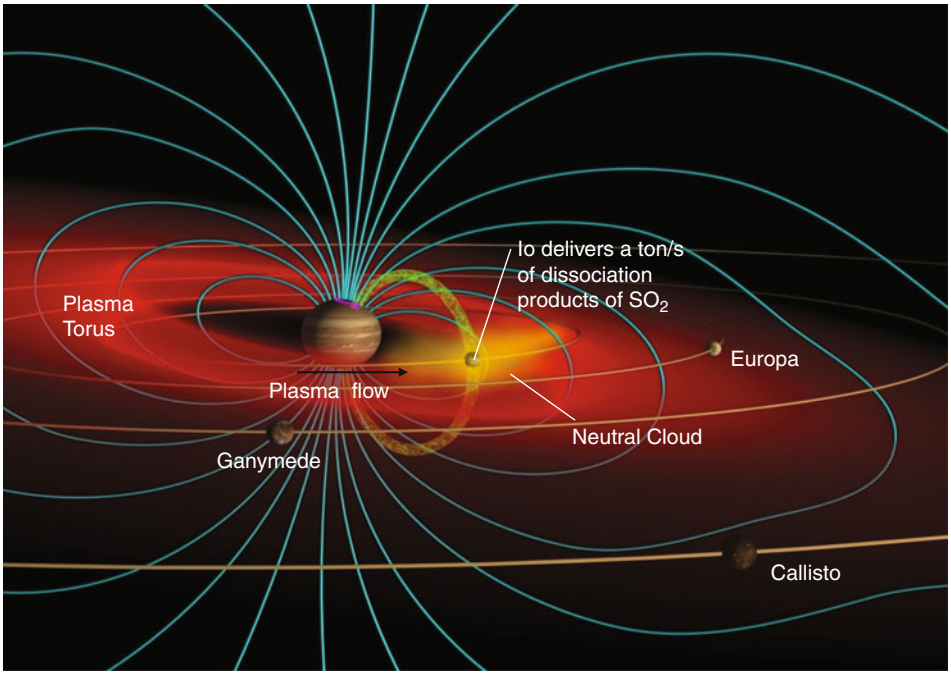


Fig. 6-12
The main components of the Jupiter-Io system

The Io plasma torus has a total mass of ~2 megaton, which would be replenished by a source of ~1 ton/s in ~23 days. Multiplying by a typical energy (Ti ~ 60 eV, Te ~ 5 eV), we obtain ~6×10^{17} J for the total thermal energy of the torus. The observed UV power is about 1.5 TW, emitted via more than 50 ion spectral lines, most of which are in the EUV. This emission would drain all the energy of the torus electrons in ~7 h. Ion pickup replenishes energy, and Coulomb collisions feed the energy from ions to electrons but not at a sufficient rate to maintain the observed emissions. A source of additional energy, perhaps mediated via plasma waves, seems to be supplying hot electrons and a comparable amount of energy as ion pickup. The 20–80 day time scale (equivalent to 50–200 rotations) for the replacement of the torus indicates surprisingly slow radial transport that maintains a relatively strong radial density gradient. Flux tubes laden with denser, cooler, plasma move outward and relatively empty flux tubes containing hotter plasma from the outer magnetosphere move inward.

Voyager, Galileo, and, particularly, Cassini observations of UV emissions from the torus show temporal variability (by about a factor 2) in torus properties (Steffl et al. 2004, 2006). Models of the physical chemistry of the torus match the observed properties in regard to the production of neutral O and S atoms, a radial transport time, and a source of hot electrons (Delamere and Bagenal 2003). Steffl et al. (2008) showed that a small (<1%) hot electron population that varies with longitude and drifts by a few percent with respect to corotation could explain modulations in ionization state and emissions. The source of these hot electrons is not understood, but the discussions of what processes might be causing periodicities observed at Saturn (see next section) suggest that perhaps ionospheric winds might be driving currents through the jovian magnetosphere, carried by these hot electrons. On longer time scales, the

variation in torus emissions observed over several months by Cassini reflect the observed changes in the output of Io's volcanic plumes (Delamere et al. 2004; Bagenal and Delamere 2011).

The earliest theoretical studies concluded that the magnetosphere of Jupiter is "all plasmasphere" with little influence of solar-wind-driven convection. Indeed, rotation dominates the plasma flows observed in the jovian magnetosphere out to distances ~70 R_J. Yet, the presence of sulfur and oxygen ions in the middle magnetosphere, far from Io, indicates that plasma is transported outward, in directions transverse to the magnetic field.

The net radial transport is thought to be slowest near Io's orbit (~15 m/s) and to speed up farther out (~50 m/s beyond $10R_J$). Plasma from the Io torus spreads out from Jupiter as a ~$5R_J$-thick plasma sheet throughout the magnetosphere. While the flow direction remains primarily rotational, both a lag behind corotation and local time asymmetries increase steadily with distance from the planet. Bursts of flow down the magnetotail are observed and also, on the dawn flanks, occasional strong bursts of super-rotation. Below we return to these deviations from corotation and discuss how they relate to auroral structures.

As the equatorial plasma rotates rapidly, it exerts a radial (centrifugal) stress on the flux tubes. Additional stress is provided by the radial pressure gradient of the plasma, inflating the magnetic field (see *Fig. 6-13*). The net result is a stretching of the initially dipolar field lines away from the planet, in a configuration that implies an azimuthal current in the near-equatorial disk (*Fig. 6-13a*). The lower two panels of *Fig. 6-13* show magnetic field lines derived from models that include the internally generated field plus the effects of currents on the magnetopause and in the plasma sheet. *Figure 6-8d* shows magnetic field lines projected onto the equatorial plane and illustrates how the field lines also bend or "curl" in the azimuthal direction, which means that there are also radial currents in the equatorial plasma sheet (*Fig. 6-13b*). Alternatively one can think of sub-corotating plasma pulling the magnetic field away from radial. At Jupiter, the field is more or less azimuthally symmetric out to about $50R_J$, but *Fig. 6-13d* shows that strong local time asymmetries develop in the outer magnetosphere (Khurana 2001; Khurana and Schwarzl 2005).

Just as at Earth, the auroral emissions at Jupiter are important indicators of magnetospheric processes. With limited spacecraft coverage of these magnetospheres, auroral activity is a projection of magnetospheric processes, communicated via precipitating energetic particles, onto the atmosphere; thus, it allows us to study global processes not yet accessed by spacecraft. *Figure 6-14* illustrates the three main types of aurora at Jupiter (see the reviews by Bhardwaj and Gladstone 2000; Clarke et al. 2005). There is a fairly steady main auroral oval that produces approximately 10^{14} W globally and that can exceed 1 Wm^{-2} locally. This oval is quite narrow, corresponding to about 1° in latitude or a few hundred kilometers horizontally in the atmosphere of Jupiter and mapping along magnetic field lines to (20–30)R_J at the equator in the magnetosphere, well inside the magnetopause. Auroral emissions are also observed at the feet of flux tubes at Io, Europa, and Ganymede. While the magnetosphere interaction with Callisto is thought to be much weaker than for the other satellites, any Callisto aurora would be difficult to separate from the main aurora. The Io-related aurora includes a "wake" signature that extends halfway around Jupiter. The third type of jovian aurora is the highly variable polar aurora, which occurs at higher latitudes than the main aurora, corresponding to greater magnetospheric distances.

The fact that the shape of the jovian main auroral oval is constant and fixed, in magnetic coordinates (including an indication of a persistent magnetic anomaly in the northern hemisphere), tells us that the auroral emissions correspond to a persistent magnetospheric process

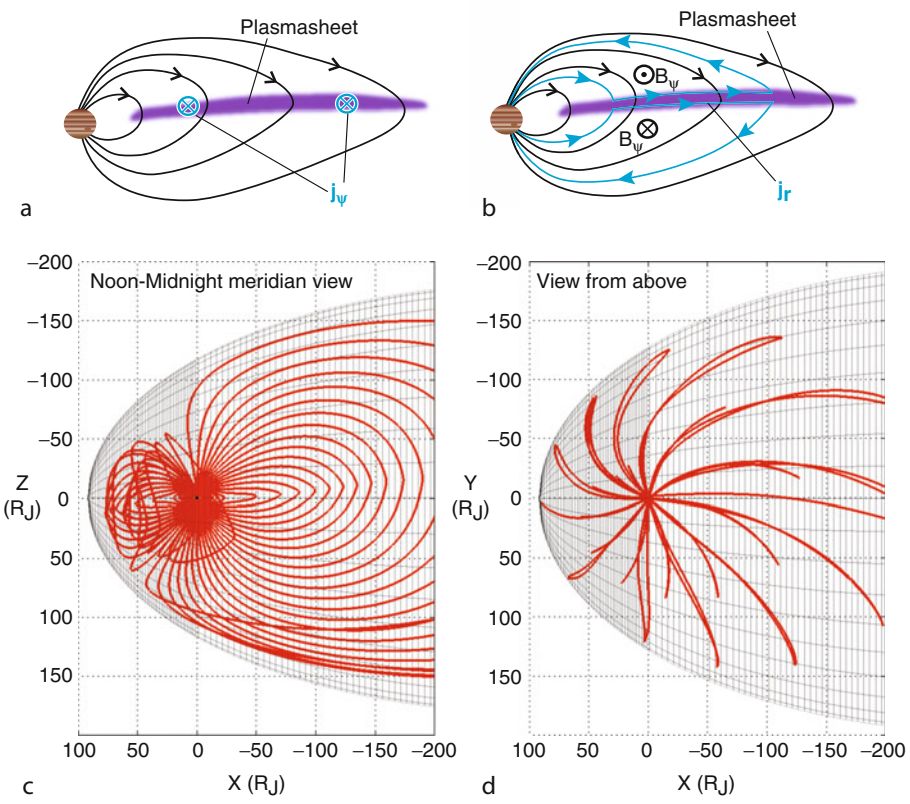


Fig. 6-13

Magnetic field configuration and current systems in Jupiter's magnetosphere. The *top* diagrams show the (a) azimuthal and (b) radial current systems. The *lower* diagrams show the magnetic field configuration (c) in the noon-midnight meridian plane and (d) in the equatorial plane derived from in situ magnetic field measurements (Khurana and Schwarzl 2005)

that causes a more or less constant bombardment of electrons onto Jupiter's atmosphere. Unlike the terrestrial auroral oval, the jovian oval has no relation to the boundary between open and closed field lines of the polar cap; it maps to regions well within the magnetosphere. It is difficult to map the magnetic field lines accurately because of the strong equatorial currents, which are variable and imprecisely determined. But it has become clear that the main aurora is the signature of Jupiter's attempt to spin up its magnetosphere or, more accurately, Jupiter's failure to spin up its magnetosphere fully.

Figure 6-8a shows the simple current system proposed by Hill (1979). As the Iogenic plasma moves outward, the conservation of angular momentum would suggest that the plasma should lose angular speed. In a magnetized plasma, however, electrical currents easily flow along magnetic fields and couple the magnetospheric plasma to Jupiter's flywheel. Hill (1979) argued that at some point the load on the ionosphere increases to the point where the coupling between the ionosphere and corotating atmosphere – manifested as the ionospheric conductivity – is not sufficient to carry the necessary current, causing the plasma to lag behind corotation. Using a

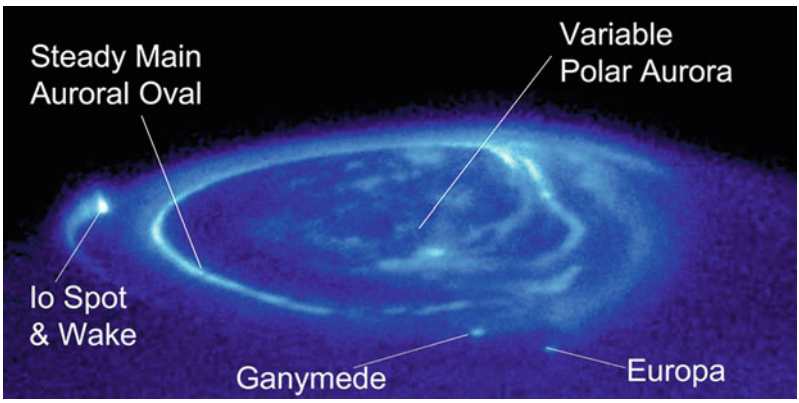


Fig. 6-14
Three main regions of Jupiter's aurora: Main oval, satellite footprints, polar emission (From Clarke et al. 2005)

simple dipole magnetic field, Hill (1979) obtained an expression for the critical distance for corotation lag that depended on the mass production and transport from Io and the (poorly determined) ionospheric conductivity. Matching his simple model to the Voyager observations of McNutt et al. (1979) and Hill et al. (1981) found he could model the observed profiles of azimuthal flow with a source giving 2–5 ton/s and an onospheric conductivity equal to 0.1 mho. Over the past decade, Jupiter's main aurora has become an active area of study. Researchers have considered the effects of the non-dipolar nature of the magnetic field, the narrowness of the auroral emissions, realistic mass-loading rates, the nonlinear feedback of ionospheric conductivity responding to electron precipitation, and the development of electrostatic potential drops in the region of low density between the ionosphere and torus (Cowley and Bunce 2001, 2003; Nichols and Cowley 2005; Ray et al. 2010). The understanding of plasma processes developed in the terrestrial magnetosphere is being applied to the different regimes at Jupiter and will ultimately be tested when the Juno spacecraft goes into a close polar orbit in 2016.

The auroral emissions poleward of the main auroral oval (see *Fig. 6-14*) are highly variable; they are modulated by the solar wind and controlled in local time, being usually dark on the dawn side and brighter on the dusk side (see the reviews by Grodent et al. 2003; Clarke et al. 2005). The region of magnetic field lines that is open to the solar wind in the polar cap is thought to be relatively small (Vogt et al. 2011). Thus, much polar auroral activity reflects activity in the outer magnetosphere, occurring on closed magnetic field lines. Polar auroral activity has been associated with polar cusps (Waite et al. 2001; Pallier and Prangé 2004; Bunce et al. 2004) as well as tail plasma sheet reconnection and the ejection of plasmoids down the magnetotail (Grodent et al. 2004; Radioti et al. 2008, 2010, 2011; Ge et al. 2010). Spectral observations of auroral X-ray flares suggest that energetic ions are bombarding the polar atmosphere and may be the signature of the plasma sheet return (downward) current (Waite et al. 1994; Cravens et al. 1995; Hui et al. 2010; Ozak et al. 2010).

A major interest in studying the aurora is to explore how the various emissions are related to the dynamics of the outer magnetosphere. The innermost region, which we will call the Hill region, comprises the equatorial plasma disk where rotation dominates the flow. At a distance

of about 20 R_J, the lag of plasma in the equatorial plasma sheet behind strict corotation drives upward currents, and the associated electron bombardment of the atmosphere causes the main aurora.

The middle magnetosphere is a compressible region (sometimes called the "cushion" or Vasyliunas region, after his seminal article (Vasyliunas 1983) in which the dynamics of the outer magnetosphere was first addressed in a substantial fashion). On the dayside of the magnetosphere, the ram pressure of the solar wind compresses the magnetosphere. Inward motion on the dawn side reduces the load on the ionosphere, producing a correspondingly dark region in the dawn polar aurora (*Fig. 6-14*). On the dusk side, the plasma expands outward and strong currents try to keep the magnetospheric plasma corotating. These strong currents produce the active dusk polar aurora. Kivelson and Southwood (2005) argued that the rapid expansion of flux tubes in the afternoon to dusk sector means that the second adiabatic invariant is not conserved, which results in the heating and thickening of the plasma sheet.

Pursuing evidence for Vasyliunas' argument that plasmoids are ejected down the jovian magnetotail, Grodent et al. (2004) found evidence of spots of auroral emission poleward of the main aurora connected to the nightside magnetosphere that flashed with an approximately 10-min duration. Such events were rare, recurring only about once per 1–2 days. These flashes seemed to occur in the pre-midnight sector, and Grodent et al. (2004) estimated that they are coupled to a region of the magnetotail that was about $5R_J$–$50R_J$ across and located further than $100R_J$ down the tail. Studies of in situ measurements (Russell et al. 2000; Woch et al. 2002; Vogt et al. 2010; Ge et al. 2010) led to the conclusion that plasmoids on the order of ~$25R_J$ in scale were being ejected every 4 h–3 days, with a predominance for the post-midnight sector and distances of 70–$120R_J$. Could such plasmoids account for most of the plasma loss down the magnetotail? Bagenal (2007) approximated a plasmoid as a disk of plasma sheet $2R_J$ thick having diameter $25R_J$ and density of 0.01 cm^{-3}, so that each plasmoid has a mass of about 500 ton. Ejecting one such plasmoid per day is equivalent to losing 0.006 ton/s. Increasing the frequency to once per hour raises the loss rate to 0.15 ton/s. Thus, on the one hand, even with optimistic numbers, the loss of plasma from the magnetosphere due to such plasmoid ejections cannot match the canonical plasma production rate, 0.5 ton/s. On the other hand, a steady flow of plasma of density 0.01 cm^{-3}, in a conduit that is $5R_J$ thick and $100R_J$ wide, moving at a speed of 200 km/s would provide a loss of 0.5 ton/s. Such numbers suggest that a quasi-steady loss rate is feasible. The question of the mechanism remains unanswered. Bagenal (2007) proposed three options: a diffusive "drizzle" across weak, highly stretched, magnetotail fields; a quasi-steady reconnection of small plasmoids, below the scale detectable via auroral emissions; or a continuous but perhaps gusty magnetospheric wind.

In the spring of 2007, the New Horizons spacecraft flew past Jupiter, getting a gravitational boost on its way to Pluto, and made an unprecedented passage down the core of the jovian magnetotail, exiting on the northern dusk flank. For over 3 months, while covering a distance of $2{,}000R_J$, the spacecraft measured a combination of iogenic ions and ionospheric plasma (indicated by H^+ and H^{3+} ions) flowing down the tail (McComas et al. 2007; McNutt et al. 2007). The fluxes of both thermal and energetic particles were highly variable on time scales of minutes to days. The tailward fluxes of internally generated plasma led (McComas and Bagenal 2007) to argue that perhaps Jupiter does not have a complete Dungey cycle but that the large time scale for any reconnection flow (see *Table 6-4*) suggests that magnetic flux that is opened near the subsolar magnetopause re-closes on the magnetopause before it has traveled down the tail. They suggested that the magnetotail comprises a pipe of internally generated plasma that disconnects from the planetary field and flows away from Jupiter in intermittent surges or bubbles,

with no planetward Dungey return flow. Delamere and Bagenal (2010) argue that, due to the viscous processes on the magnetopause boundary, along the flanks of the magnetotail, solar wind plasma becomes entrained and mixed with the ejected iogenic material.

An overview sketch of the dynamics of the magnetosphere as proposed by Delamere and Bagenal (2010) is presented in *Fig. 6-15*. Alternative views combine the Vasyliunas rotationally driven ejection of plasmoids of *Fig. 6-11* with the Dungey cycle of *Fig. 6-9* (e.g., Cowley et al. 2007, 2008b; Kivelson and Southwood 2005). Hopefully, observations by the Juno mission will distinguish between these different ideas.

3.2 Saturn

Before the Cassini mission, it was tempting to dismiss the magnetosphere of Saturn as merely a smaller, less exciting, version of the jovian magnetosphere. However, Cassini measurements of the particles and fields in Saturn's neighborhood have shown processes similar to those at Jupiter (e.g., satellite sources, ion pickup, flux tube interchange, corotation, etc.), but they have also revealed substantial intriguing differences (for reviews of initial results of the Cassini mission see Dougherty et al. 2009). The magnetosphere of Saturn is strongly dominated by neutral atoms and molecules. The number-density ratio of neutrals to ions is 12:1 in the Enceladus torus compared with 1:20 in the Io torus. In contrast with Jupiter's steady main aurora, Saturn's auroral emissions are strongly modulated by the solar wind, particularly the solar wind ram pressure. While one might expect the alignment of Saturn's magnetic axis with the planet's spin axis to produce an azimuthally symmetric magnetosphere, observations show an intriguing rotational modulation. Even more mysteriously, the rotational modulation varies with time (on time scale of ~ years) and is different for the northern and southern hemispheres. The magnetosphere of Saturn is shown in *Fig. 6-16*. Below, we provide a brief summary of current ideas about these topics, which are under active research as the Cassini spacecraft continues to orbit Saturn.

One of the great discoveries of the Cassini mission to Saturn has been the active volcanism of the small icy moon Enceladus. While Enceladus is a mere one-seventh the size of Io, this small moon suffers tidal heating that drives the eruption of geysers from the south polar region. The geyser plumes, extending over 500 km from the surface, seem to be mostly ice particles with water vapor and minor quantities of molecular nitrogen, methane, and carbon dioxide (Porco et al. 2006; Hansen et al. 2006; Waite et al. 2006).

Estimates of the total neutral production rate of water molecules (presumably ultimately coming from Enceladus' plumes) vary around the initial value of 300 kg/s, determined from the initial UV occultation of the plume by Hansen et al. (2006) which is the same as from the earlier (Jurac and Richardson 2005) model constructed to match HST observations of the OH neutral cloud. Sittler et al. (2008) preferred 600 kg/s but only claimed a factor of 2 accuracy, so this value is still consistent with Hansen et al. (2006)'s 300 kg/s. Saur et al. (2008) modeled the electrodynamics of the plume deriving values as high as 1,600 kg/s for the E0 flyby and as low as 200 kg/s for E1 and E2. Meanwhile, a value of ~200 kg/s was derived from a second UV occultation reported by Hansen et al. (2008). Similarly, Fleshman et al. (2010a, b) found 100–180 kg/s was consistent with their physical chemical modeling of the Enceladus torus. Finally, Smith et al. (2010) have analyzed INMS data from three Cassini flybys of Enceladus from which they conclude that the net production has increased from <72 kg/s (at the time of E2) to 190 kg/s (at E3) to 750 kg/s (at E5). Thus, the Enceladus neutral source rate could have varied by a factor of 10 between July 2005 and October 2008.

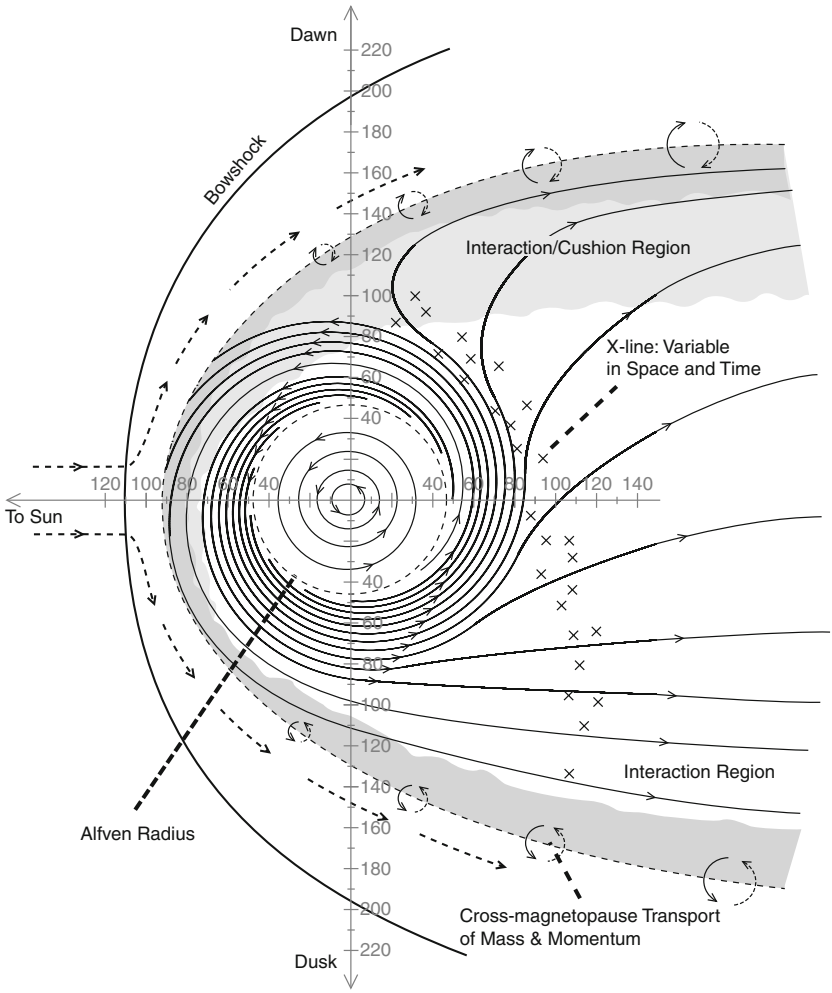


Fig. 6-15

Composite sketch of the structure and dynamics of the jovian magnetosphere. Inside ~60 R_J the plasma flow is corotational and the plasma sheet has little local time asymmetry. Beyond ~60 R_J, radial outflow combines with rotation to produce spiral flow that removes the plasma from the magnetosphere within about a day. Beyond ~80–100 R_J, blobs of plasma detach (at an x point) and are shed down the magnetotail. Between midnight and dawn the region of x-points is well defined by in situ observations and is consistent with estimates of the location where tailward Maxwell stresses dominate over confinement by the planetary magnetic field. In the dusk to premidnight region the location of x points is not well determined, but observational evidence suggests that it could be as far as 150–200 R_J downtail. Strong velocity shear across the magnetopause drives instabilities that act as a viscous-like interaction between the draped solar wind and largely closed magnetosphere, intermittently transferring mass and momentum across the magnetopause boundary. This interaction region is particularly wide on the dawnside of the magnetosphere, corresponding to what is sometimes called the "cushion region". After Delamere & Bagenal (2011)

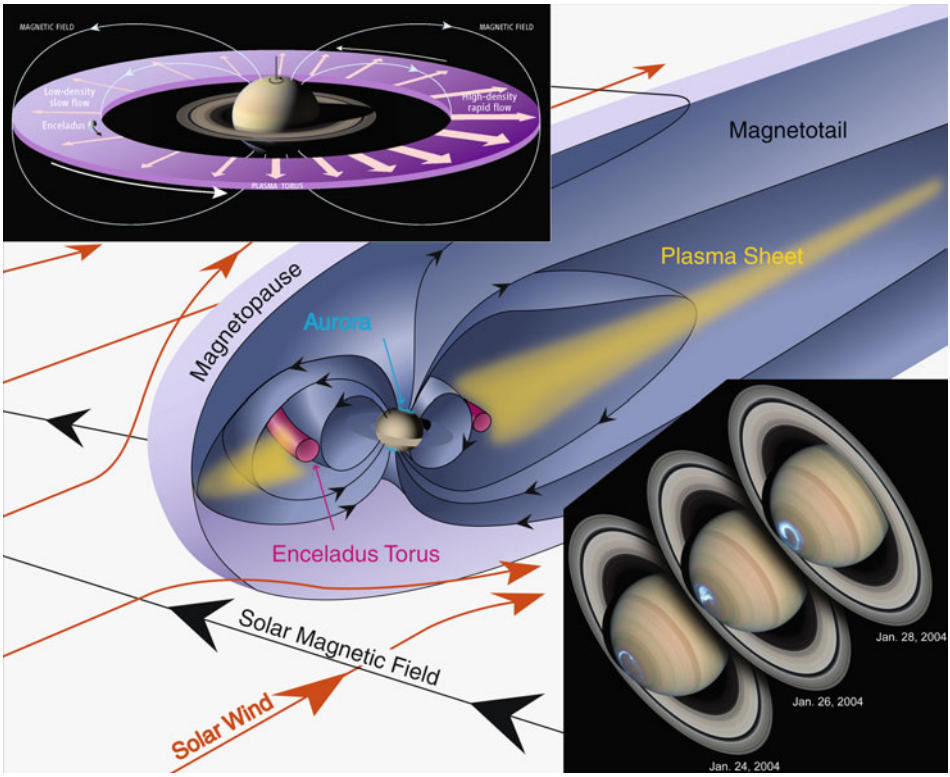


Fig. 6-16

***Center*: Three-dimensional schematic representation of the magnetosphere of Saturn. *Top left*: Sketch of asymmetric plasma disk where Gurnett et al. (2007)) propose that the observed density variations are caused by a pattern of asymmetric radial outflows. *Bottom right*: Hubble Space Telescope observations of Saturn's auroral emissions (Clarke et al. 2005)**

The fate of the neutrals is more complicated at Saturn than Jupiter. The high neutral-to-ion density ratio at Saturn is a result of lower ionization rates (caused as much by photoionization at Saturn as electron-impact ionization that dominates at Jupiter). Only a fraction of the neutral material is transported out into the plasma sheet. Some of the corotating ions charge exchange with neutrals to become escaping fast neutrals but other collisional processes such as photo- and electron-dissociation, neutral–neutral collisions, and low-velocity charge exchange "puff" up the neutral cloud, spreading it beyond Enceladus' orbit ($4R_S$) as well as sending a substantial flux of neutrals into the planet Saturn.

It is not clear that the rate of ionization vs. other neutral loss processes would be maintained at the modeled fractions if the neutral source increases to Smith et al. (2010)'s E5 values of ~750 kg/s or Saur et al. (2008)'s E0 value of 1,600 kg/s. Neutral production increases the amount of neutral–neutral collisions that would cause more of the material to spread out from Enceladus' orbit. One might expect that more material would escape as neutrals rather than be ionized. Electron-impact ionization would be reduced due to collisional cooling of the electrons. In fact, Tokar et al. (2009) do not report higher-than-average plasma densities around

the time of E5. Estimates of the plasma source range between 12 and 250 kg/s (see review by Bagenal and Delamere 2011).

The nearly three orders of magnitude difference in the ion–neutral density ratios of the two magnetospheres can be been explained in terms of a much lower energy input into the Saturn system (Delamere et al. 2007). At Saturn, the plasma flowing past Enceladus (at an orbital distance of ~4 saturnian radii) has a slower speed than the plasma flow past Io (at ~6 jovian radii). A factor 2 difference in relative motion (i.e., 26 km/s at Enceladus as against 57 km/s at Io) means that new ions pick up a factor 4 less energy. With less pickup energy, the ions deliver less energy to the electrons. At low electron temperatures, the ionization rates plummet and, correspondingly, plasma production drops. In fact, Delamere et al. (2007) showed (backed up by an extensive study by Fleshman et al. 2010a, b) that without an additional source of hot electrons (similar to that in the Io plasma torus), the Enceladus plasma torus would not be sustained.

The weaker plasma source at Saturn results in weaker centrifugal stresses and weaker magnetospheric currents. Thus, the field structure at Saturn is similar to that shown in ❯ *Fig. 6-13* for Jupiter but with less pronounced distortion from dipolar. The plasma pressure is also much reduced, so that Saturn's magnetosphere is less compressible than Jupiter's and shows a less dramatic response to changes in solar wind dynamic pressure (Kanani et al. 2010).

❯ *Figure 6-16* shows Hubble Space Telescope (HST) images of Saturn's aurora (Clarke et al. 2005). In contrast with Jupiter's large main auroral oval, which maps to regions deep inside the magnetosphere, Saturn's small auroral oval and strong variations in auroral intensity with solar wind conditions indicates that Saturn's aurora, like Earth's, marks the boundary of open and closed regions of magnetic flux. The picture was clarified during a campaign of combined Hubble and Cassini observations as the spacecraft approached Saturn in late 2000. For 22 days, Cassini's instruments measured the magnetic field, plasma density, and plasma velocity in the solar wind, while Hubble cameras and the Cassini radio antennas monitored Saturn's auroral activity. Nature cooperated and provided a couple of interplanetary shock waves that passed the Cassini spacecraft on January 15 and 25, 2001, and then hit the magnetosphere of Saturn some 17 h later. Clarke et al. (2005) reported HST observations of the subsequent brightening of auroral emission, and Kurth et al. (2005) reported accompanying increases in radio emission. Crary et al. (2005) show a correlation of auroral intensity with solar wind dynamical pressure, supporting the view that the solar wind has an Earth like role at Saturn.

But further study showed that it was compression of the magnetopause by the solar wind that correlates with auroral intensity rather than reconnection of the solar and planetary magnetic fields. Crary et al. (2005) pointed out that, at Saturn's orbit, the solar magnetic field is essentially tangential so that the solar and planetary fields are largely orthogonal to each other: far from optimal conditions for magnetic reconnection. Clarke et al. (2005, 2009) showed that the brightest auroral emissions occurred after the passage of a solar wind pressure pulse. Cowley et al. (2005) suggested that the rapid compression of Saturn's magnetosphere induces enhanced tail reconnection, which would explain the subsequent shrinking of the auroral oval (see ❯ *Fig. 6-9*). The explanation for why Saturn's aurora responds to compression rather than the direction of the IMF (as at Earth) principally involves the longer time scale for the solar wind to flow past the larger magnetosphere. Fluctuations of Bz component of the IMF are similar at Earth and Saturn (~10s min to an hour or two). Similarly, the rate of dayside reconnection is thought to be about the same at each planet. But the amount of open flux in Saturn's tail is thought to be much larger than Earth's (about a factor of ~100) so the buildup of open flux in the magnetotail (stages 1–4 in ❯ *Fig. 6-9*) could be much longer, typically ~1 week instead of ~1 h. So, Saturn's tail will almost never respond to individual intervals of positive *Bz*, but

instead inflates on time scales comparable to the time between recurrent solar wind pressure pulses. Thus, compression-induced tail reconnection, while rather rare at Earth, may be the usual mode at Saturn (Jackman et al. 2005; Badman et al. 2005).

The magnetospheric processes driving Saturn's aurora began to be better understood after Cassini moved to higher magnetic latitudes in 2007. Observations by Cassini particle and field instruments show a large-scale field-aligned current present at the open–closed field line boundary (Cowley et al. 2008a; Bunce et al. 2008, 2010; Talboys et al. 2009a, b), in the same region as auroral radio emissions were generated (Lamy et al. 2009, 2010).

In the mean time, the difficulties in measuring Saturn's rotation rate have wreaked havoc with our simple ideas of magnetospheric dynamics. So how could one establish how fast the interior of a gas planet is spinning? The usual trick is to measure the periodicity of radio emissions modulated by the planet's internal magnetic field. In this method, it is assumed that the magnetic field is tilted and that the dynamo region where the field is generated spins at a rate representative of the bulk of the planet. Recent Cassini data indicate that apparent changes in Saturn's spin could in fact be caused by processes external to the planet. This raises new questions about how we measure and understand the rotation of the large gas planets. Saturn at first dumbfounded planetary theorists who study dynamo models by being observed to have a highly symmetric internal magnetic field. A field that is symmetric about the rotation axis violates a basic theorem of magnetic dynamos (Cowling 1933). The second puzzle came with the detection of a systematic rotational modulation of the radio emission similar to a flashing strobe, which should not occur for a symmetric magnetic field. Meanwhile, radio measurements have revealed that Saturn's day appears to have become about 6–8 min longer – it is now roughly 10 h and 47 min – since the 1980s when measured by the Voyager missions (Kurth et al. 2008). Furthermore, the spin rate seems to keep changing and may be modulated by the solar wind speed (Zarka et al. 2007) and is different in the northern and southern hemispheres (Gurnett et al. 2009), the rotation rates switching hemispheres over equinox (Gurnett et al. 2010; Southwood 2011). Auroral UV emissions are also modulated at the same rate at the radio emissions (Nichols et al. 2010), as are oscillations in the magnetic fields (Andrews et al. 2010a, b). The variation in modulation with season evokes an atmospheric driver. To drive periodic modulations with thermospheric winds via currents in the ionosphere will require realistic models of Saturn's ionosphere (Smith 2011; Galand et al. 2011).

A fundamental issue is whether the magnetospheric observations, including the radio emissions, do actually require the magnetic field emanating from the interior of Saturn to be asymmetric. Nearly 30 years ago, Stevenson suggested that strong shear motions in an electrically conducting shell surrounding the dynamo might impose symmetry around the rotational axis (Stevenson 1982). That the rotational modulation of magnetospheric phenomena seems to be fairly constant with radial distance, that dynamic changes occur in the external plasma structures around Saturn, and that there is an apparent modulation by the solar wind speed indicate that an external explanation for Saturn's apparently erratic spin rate seems far more plausible than perturbations in the massive interior of the planet. Yet, localized magnetic anomalies (i.e., high-order multipoles) at high latitudes remain possible and may be affecting the currents that couple the magnetosphere to the planet (Southwood and Kivelson 2007). Gurnett et al. (2007) showed how Saturn's radio emission, the magnetic field measured in the magnetosphere, and the density of the plasma trapped in the magnetic field are all modulated with the same drifting period. They argued that the process that transports plasma radially outward could be stronger on one side of Saturn than the other, as illustrated in the top left of *Fig. 6-16*. Gurnett et al. (2007) suggested that this circulation pattern also produces higher plasma densities in the region of stronger outflow and proposed that plasma production stresses the

electrodynamic coupling between the magnetosphere and the planet, causing the pattern of weaker or stronger outward flow to slowly slip in phase relative to Saturn's internal rotation. What causes the proposed asymmetric convection pattern? In the 1980s, researchers tried to explain variations in the Io plasma torus (Hill et al. 1981) by invoking a convection pattern that rotated with the planet; however, evidence of such a flow pattern in the jovian magnetosphere remains elusive. Alternatively, a system of neutral winds in Saturn's atmosphere could drag the ionosphere around, which would stir up the magnetosphere electrodynamically and provide a source of hot electrons. Could small variations in the high-energy electron population in the Enceladus torus, similar to those in the Io torus, be causing the dramatic changes in plasma density observed by Cassini? If so, large-scale convection patterns in the magnetosphere may not be necessary, just minor modulations in the electrical currents that flow along the magnetic field between the equatorial plasma disk and the planet's ionosphere, bringing small fluxes of ionizing high-energy electrons to the torus. Delamere and Bagenal (2008) showed that a modulation in the small hot-electron population could produce the factor of 2 variation in plasma density observed by Cassini.

Undoubtedly, the issue of Saturn's rotation rate and its coupling to the magnetosphere will be a vital area of exploration over the next few years. Similarly, it will be important to investigate whether material is ejected down the tail in the manner and to the extent of the jovian system. Only a few plasmoids have been detected to date at Saturn, but this may be a result of limited coverage by the Cassini spacecraft (e.g., Jackman et al. 2008; Hill et al. 2008). The substantial polar cap, marked by the aurora, and the influence of the solar wind on the auroral intensity indicate that the Dungey reconnection cycle plays a substantial role at Saturn. The extent and mechanism whereby any return, planetward, flow operates in the magnetotail awaits further exploration.

3.3 Uranus and Neptune

The Voyager flybys of Uranus (1986) and Neptune (1989) revealed what have to be described as highly irregular magnetospheres. The non-dipolar magnetic fields and the large angle between the magnetic and rotation axes not only pose interesting problems for dynamo theorists but also challenge the ideas of magnetospheric dynamics. Unfortunately, little study has been made of these odd magnetospheres for the past 15 years, and there is only slim hope of further exploration for quite some time. Thus, there is not much to add to the comparative reviews of their fields by Connerney (1993) and of their magnetospheres by Bagenal (1992). Here, we provide a brief précis of these reviews to which the reader should turn for original references.

Tables 6-1 and *Table 6-2* as well as *Fig. 6-4* show Uranus and Neptune to have substantial magnetospheres that envelope most of their satellites. *Figures 6-2* and *Fig. 6-3* give a sense of the irregularity of their magnetic fields, approximated as large tilts and offsets. *Table 6-2* tells us that from just the solar wind and planetary parameters, we should expect both rotation and solar wind coupling to affect the dynamics of these magnetospheres (though the weak IMF of the outer heliosphere suggests that reconnection will be much weaker than at planets closer to the Sun). *Figures 6-17* and *6-18* illustrate how the orientations of these planets' magnetic fields (*Fig. 6-2*), which rotate about the planet's spin axis every 16–17 h, might affect the solar wind coupling process. For Uranus around solstice (the Voyager era of the mid-1980s), when the spin axis is pointed roughly toward the Sun, the large tilt of the magnetic axis will result in a magnetosphere that to first approximation resembles that of the

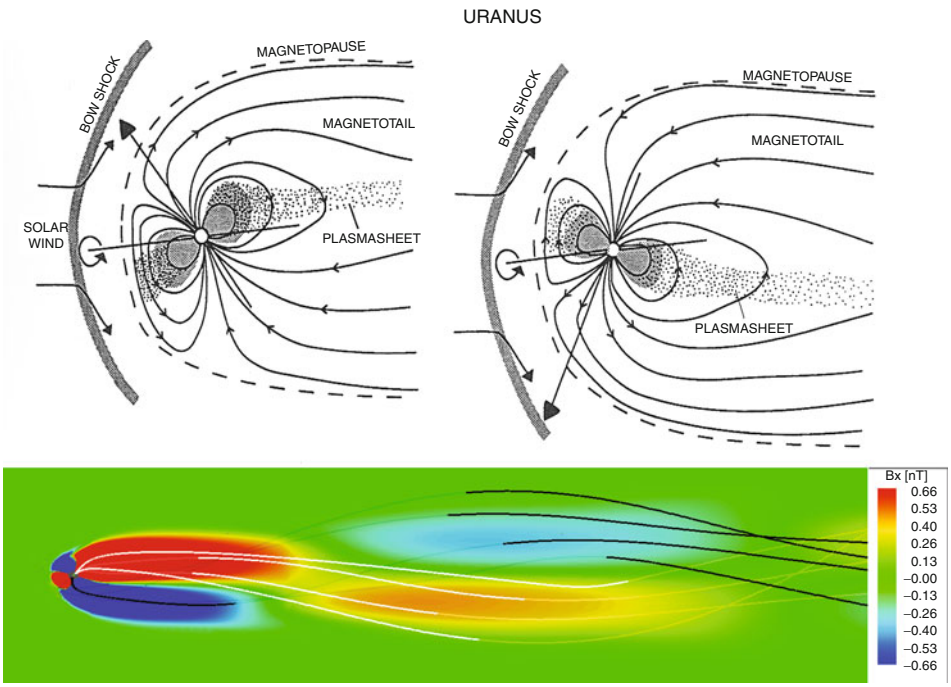


Fig. 6-17

The magnetosphere of Uranus at solstice (time of Voyager 2 flyby). The *top left* and *right* sketches show the configuration at different phases of the planet's 18-h spin period (Bagenal 1992). The *bottom panel* shows a numerical simulation of the helical magnetotail (Toth et al. (2004))

Earth but revolves every 17 h. The finite propagation (at the Alfven speed) of this rotational modulation down the magnetotail produces a helical plasma sheet and braided lobes of oppositely directed magnetic field (*Fig. 6-17*). At Neptune, the planet's obliquity being similar to Earth and Saturn, one might have expected the fairly simple configurations of either of those planet's magnetospheres. But the large tilt angle discovered by Voyager results in a configuration that changes dramatically (the tail current sheet changes from a plane to a cylinder) over the 16-h rotation period (*Fig. 6-18*).

The large range of the "solar wind angle" (see the last row of *Table 6-2*) indicates that substantial changes in orientation of the planet's spin with respect to the radial direction of the solar wind occur over the (long) orbital periods of these planets. Thus, one has the interesting challenge of imagining how the magnetosphere of Uranus was behaving during equinox in 2007, when the spin axis was perpendicular to the solar wind direction (and parallel or antiparallel to the IMF direction). Unfortunately, we are unlikely to have measurements in the near future to test the output of our imaginations. Such speculations are not wasted, however, since it is quite possible that such configurations – and many others – could have occurred in earlier epochs of Earth's history (as modeled by Zieger et al. 2004) or may now be occurring in any of the giant planets detected in other solar systems. Furthermore, keen young scientists are proposing missions to these water giant planets that might test these ideas in future decades (Arridge et al. 2011).

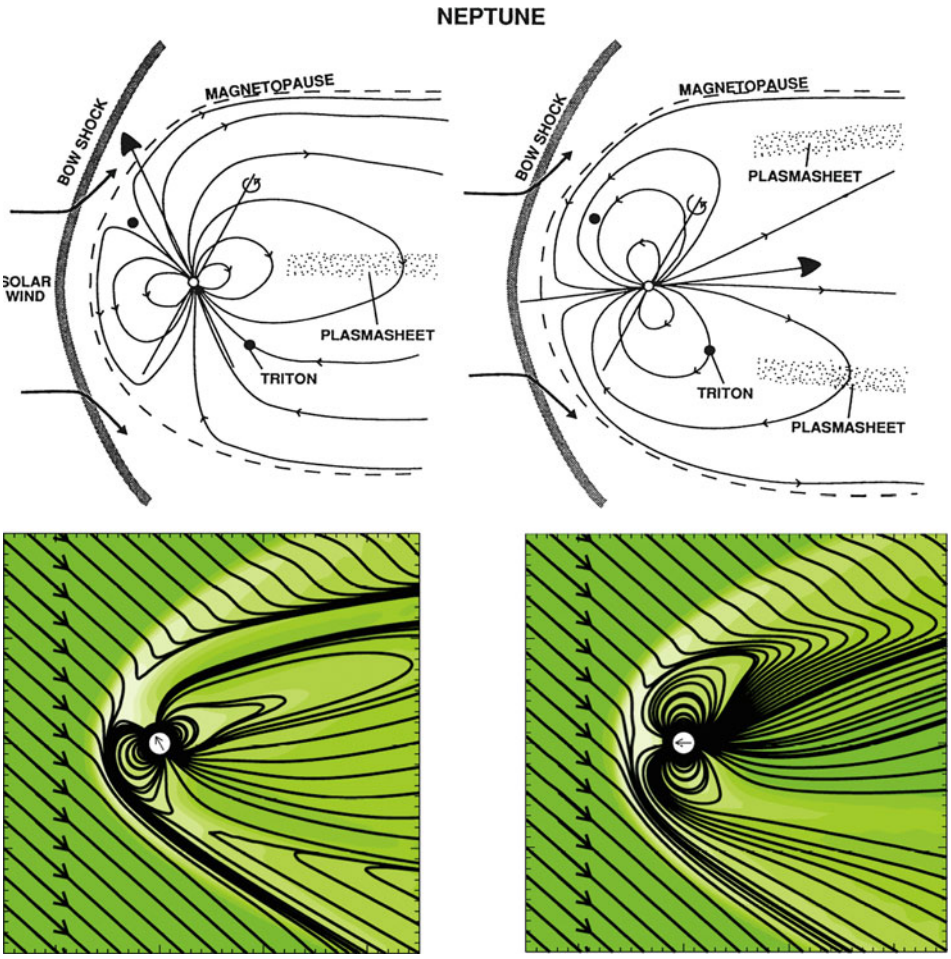


Fig. 6-18

The magnetosphere of Neptune in the configuration corresponding to the time of the Voyager 2 flyby (Bagenal 1992). Over the 19-h spin period the magnetospheric plasma sheet in the tail changes from roughly planar to a cylindrical. From a simulation by Zieger et al. (2004). (*Bottom*) Diurnal variation of the magnetic field configuration and pressure in an equatorial dipolar magnetosphere for dipole axis at 30° to the normal to the ecliptic plane (*left*) and at 90° (*right*). The configurations are close to those relevant to Neptune's magnetosphere at different times during a planetary rotation period

4 Small Magnetospheres

The smallest objects with internal dynamos are Mercury and Ganymede. These mini-magnetospheres were reviewed by Kivelson (2007). The small innermost planet and the solar system's largest moon are about the same size and both are believed to have iron cores. Approximately dipolar magnetic fields have been detected; these hold off the surrounding plasma flow to make small but distinct magnetospheres.

Both Mercury and Ganymede rotate slowly so that neither has a plasmasphere. But in both cases, a significant population of energetic (10s–100s kev) particles have been detected in trapped radiation belts. These energetic particles are likely accelerated in mini-substorms in their magnetotails, but the particles are easily scattered onto the object and probably do not spend more than a few minutes in the magnetospheres. The interaction with these objects with their surroundings is largely between the ambient and dipolar magnetic field. The corresponding time scales for such a Dungey cycle (see *Fig. 6-9*) is seconds to minutes in these mini-magnetospheres, rather than minutes to hours at Earth.

4.1 Mercury

Just two brief flybys by Mariner 10 in the early 1970s gave a glimpse of Mercury's magnetosphere (see review by Slavin et al. 2007). These early observations revealed a magnetosphere that, while small, seemed to have most of the main properties observed at Earth (*Fig. 6-5*), including a small population of trapped energetic particles, mini-substorms, and particle injections from the magnetotail, basically consistent with simple magnetospheric scaling laws (Slavin et al. 2010). The arrival of the MESSENGER spacecraft in 2011 and the future launch of the Bebi Colombo mission have provoked further thought about this largely forgotten little magnetosphere. Preliminary results from MESSENGER have revealed and the dipole is tilted less than 3° but offset northward by 484 ± 11 km, nearly 20% of the radius (Anderson et al. 2011). The small magnetosphere is very dynamic with dramatic changes occurring on time scales of seconds to minutes (Slavin et al. 2010). Slavin et al. (2009) determined that the rate of reconnection at Mercury's dayside magnetopause to be ~10 times that typical at Earth, most probably a result of the low solar wind Alfven Mach number and values of plasma β (see Sect. 2.4.3) typical of the inner heliosphere (Slavin and Holzer 1979). The structure and dynamics of the magnetotail resemble the Earth, responding to changes in IMF direction but on shorter time scales and greater intensity, including quasiperiodic ejection of plasmoids down the tail (Slavin et al. 2010). Issues that MESSENGER will address will be how plasma is trapped and accelerated in this tiny magnetosphere, how it responds to the increasing dynamics of the new solar cycle, and how the magnetosphere couples to a planet that has such a tenuous atmosphere/ionosphere.

4.2 Ganymede

Ganymede's magnetosphere sits deep within the magnetosphere of Jupiter (for the background and discussion of Galileo observations see Kivelson et al. 2004). Unlike the supersonic flows of the solar wind, the magnetospheric plasma impinging on Ganymede is subsonic and sub-Alfvenic. There is no upstream bow shock, therefore, and the flowing magnetospheric plasma convects Jupiter's magnetic field, which is roughly antiparallel to that of Ganymede, toward the upstream magnetopause. The net result is a unique magnetospheric configuration with a region near the equator of magnetic flux that closes on the moon and with polar magnetic flux that connects the moon to Jupiter's north and south ionospheres (*Fig. 6-19a, b*). A Dungey-style reconnection cycle seems to operate: upstream reconnection opens previously closed flux, convects flux tubes over Ganymede's pole, and re-closes the flux downstream (see *Fig. 6-10*).

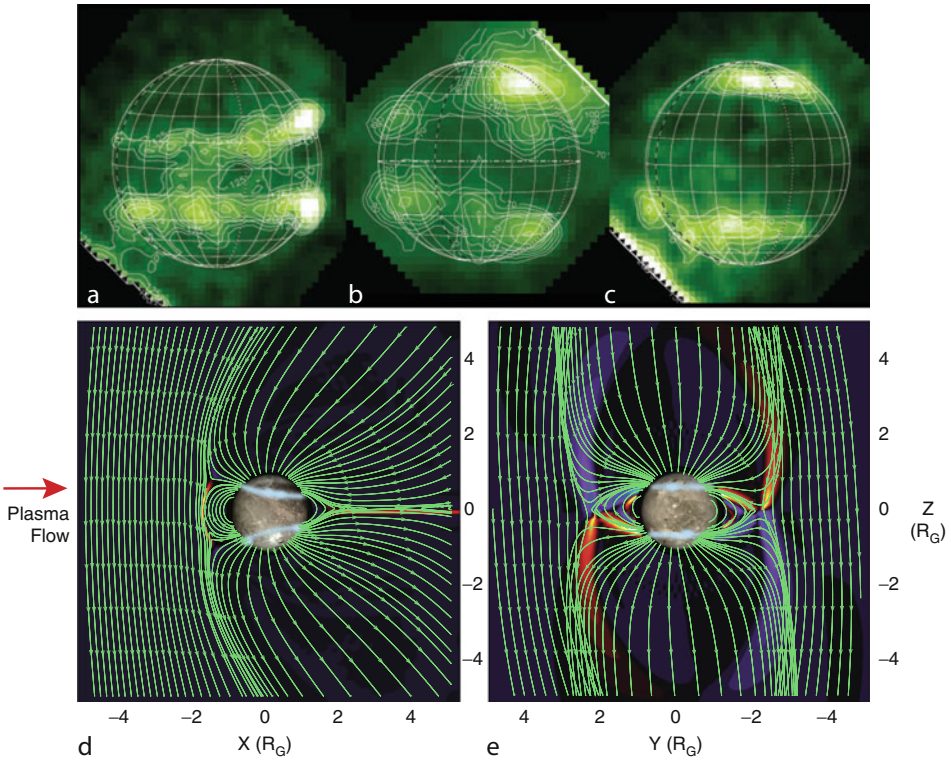


Fig. 6-19

(*Top*) HST/STIS images of Ganymede's aurora due to electron impact excitation of oxygen at OI 1356 A (M. McGrath, private communication). Contours illustrate variations in brightness. (a) The leading (downstream) hemisphere taken on 23 Dec. 2000. (b) Jupiter-facing hemisphere taken on 30 Nov. 2003. (c) Trailing (upstream) hemisphere taken on 30 Oct. 1998. (*Bottom*) Numerical model of the magnetosphere of Ganymede, with the satellite and the location of the auroral emissions superimposed (based on Jia et al. 2008). (d) The view looking at the anti-Jupiter side of Ganymede. (e) The view looking in the direction of the plasma flow at the upstream side (orbital trailing side) of Ganymede, with Jupiter to the left. The *shaded areas* show the regions of currents parallel to the magnetic field

Computer simulations are helpful in visualizing the interaction process (Paty and Winglee 2006; Paty et al. 2008; Jia et al. 2008, 2009, 2010a), but lack of information about the conductivities of Ganymede's tenuous patchy atmosphere and icy surface limit our understanding of the circuit of electrical currents that couple the magnetosphere to the moon. It is clear that electrical currents reach Jupiter, however, because of the strong auroral emissions (*Fig. 6-14*) at the Ganymede footprint (Clarke et al. 2002; Grodent et al. 2009). Short-term (few seconds) variability of aurora at Jupiter associated with the magnetic footprint of Ganymede (Grodent et al. 2009) is perhaps associated with bursty reconnection on the upstream side of Ganymede's magnetosphere (Jia et al. 2010a). The local interaction also bombards electrons into Ganymede's atmosphere, exciting auroral emissions (reviewed by McGrath et al. 2004) as shown in *Fig. 6-19c*. The locations of the aurora on Ganymede are consistent with the boundaries

between regions where the magnetic flux tubes connect at both ends to Ganymede and regions where the flux tubes connect to Ganymede on one end and Jupiter at the other.

5 Induced Magnetospheres

Having discussed the seven objects that have internally generated magnetic fields, we return to the objects without dynamos. The nature of the interaction between such bodies and the plasma in which they are embedded depends on the Mach number of the surrounding flow but is determined principally by the electrical conductivity of the body. If conducting paths exist across the planet's interior or ionosphere, then electric currents flow through the body and into the surrounding plasma, where they create forces that slow and divert the incident flow. In the case of an object sitting in the supersonic solar wind, the flow diverts around a region that is similar to a planetary magnetosphere. Mars and Venus have ionospheres that provide the required conducting paths.

Earth's Moon, with no ionosphere and a very low-conductivity surface, does not deflect the bulk of the solar wind incident on it. Instead, the solar wind runs directly into the surface, where it is absorbed. The absorption leaves the region immediately downstream of the Moon in the flowing plasma (the wake) devoid of plasma, but the void fills in as solar wind plasma flows toward the center of the wake. When the flow impinging on an object is subsonic, no upstream shock forms. But the flow will be absorbed or diverted depending on whether electrical currents flow within the object or within its ionosphere and into the surrounding plasma. Objects interacting with subsonic flow are exemplified by Io; similar processes occur, albeit to a lesser extent, at Enceladus, Titan, Triton, Europa, and several satellites embedded in the giant planet magnetospheres.

5.1 Venus

The magnetic structure surrounding Venus is similar to that around magnetized objects because the interaction causes the magnetic field of the solar wind to drape around the planet (see review by Russell et al. 2006). The draped field stretches out downstream (away from the sun), forming a magnetotail (*Fig. 6-20a*). The symmetry of the magnetic configuration within such a tail is governed by the orientation of the magnetic field in the incident solar wind, and that orientation changes with time. For example, if the interplanetary magnetic field (IMF) is oriented from east to west, then the symmetry plane (and central current sheet) of the tail is in the north–south direction, and the eastern lobe field points toward the sun while the western lobe field points away from the sun. A west-to-east-oriented IMF would reverse these polarities, and other orientations would produce rotations of the tail's plane of symmetry.

The solar wind brings in magnetic flux tubes that pile up at high altitudes at the dayside ionopause where, depending on the solar wind's dynamic pressure, they may either remain for extended times, thus producing a magnetic barrier that diverts the incident solar wind, or penetrate to low altitudes in localized bundles. Such localized bundles of magnetic flux are often highly twisted structures stretched out along the direction of the magnetic field. Such structures are referred to as *flux ropes*. These flux ropes may be dragged deep into the atmosphere, possibly carrying away significant amounts of atmosphere (*Fig. 6-20b*).

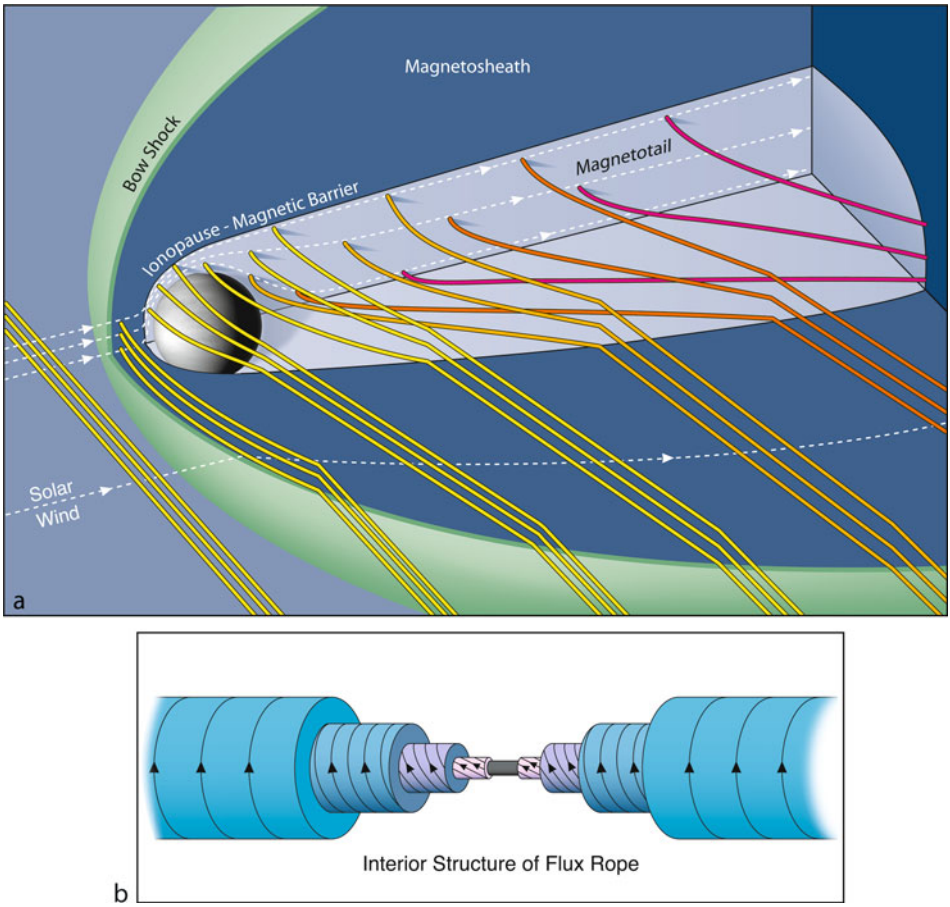


Fig. 6-20

(a) Sketch of the draping of tubes of solar magnetic flux around a conducting ionosphere such as that of Venus. The flux tubes are slowed down and sink into the wake to form a tail (after Saunders and Russell 1986). (b) Schematic illustration of a flux rope, a magnetic structure that has been identified in the ionosphere of Venus. The rope has an axis aligned with the direction of the central field. Radially away from the center, the field wraps around the axis, its helicity increasing with radial distance from the axis of the rope. Structures of this sort are also found in the solar corona and in the magnetotails of magnetized planets

The Venus Express mission has measured the solar wind interaction with Venus revealed some important subtleties. Barabash et al. (2007) report escape of 10^{25} protons/s down Venus' tail. If these ions come from the dissociation and escape of water, then Venus is losing 15 kg/s of water. Delva et al. (2011) use the presence of ion cyclotron waves as evidence of pickup of ionized escaping gases. From analysis of magnetometer data, Zhang et al. (2010) have revealed asymmetries in the magnetotail. By modeling the interaction with a hybrid code (where the electrons are treated as a fluid and the ions as particles), Jarvinen et al. (2010) showed that the large gyroradius of O^+ ions produces an asymmetric tail. Finally, Volwerk et al. (2009, 2010) reports evidence of what could be called substorm activity in Venus' tail.

5.2 Mars

While Mars' remarkably strong remanent magnetism in its crust extends its influence >1,000 km from the surface, the overall interaction of the solar wind with Mars is more atmospheric than magnetospheric (see reviews by Nagy et al. 2004 and Brain 2006). Mars interacts with the solar wind principally through currents that link to the ionosphere, but there are portions of the surface over which local magnetic fields block the access of the solar wind to low altitudes (❯ *Fig. 6-21*). It has been suggested that "mini-magnetospheres" extending up to 1,000 km form above the regions of intense crustal magnetization in the southern hemisphere; these mini-magnetospheres protect portions of the atmosphere from direct interaction with the solar wind. As a result, the crustal magnetization may have modified the evolution of the atmosphere and may still modify energy deposition into the upper atmosphere.

Several processes involved in the solar wind interaction could have contributed to atmospheric losses at Mars. The outer neutral atmospheres of Venus and Mars extend out into the solar wind where neutral atoms are photoionized and carried away by the solar wind. Newly ionized ions pick up substantial energy and correspondingly large gyroradii. These energetic ions bombard the upper atmosphere, causing heating and ionization. At times of particularly high solar wind pressure, the ionosphere can be stripped away in the solar wind. Fresh ionization in the upstream solar wind also generates plasma waves. The solar wind convects

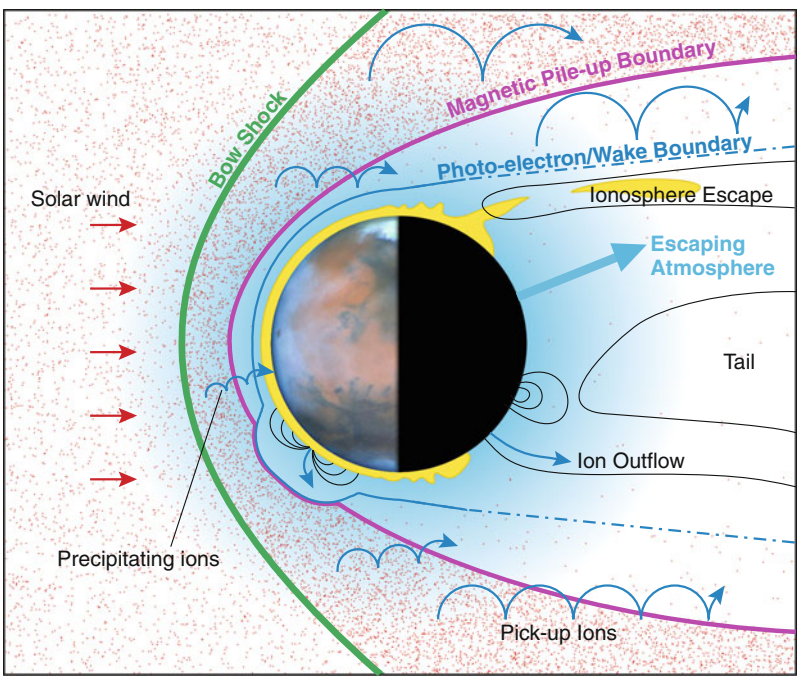


Fig. 6-21

Interaction of the solar wind with the atmosphere, ionosphere, and magnetized crust of Mars illustrating the several processes whereby the planet may have lost much of its atmosphere

the plasma waves toward the planet and into the upper layers of the ionosphere where, funneled and amplified by localized magnetic fields, they heat the ions and drive ion outflows, in a similar way to processes in the polar regions at Earth. Quantitative analyses of these different processes, both currently occurring and in the past, are active areas of research (see Brain et al. 2010 for comparison of different models) and the scientific target of the MAVEN mission to Mars (launch 2013).

5.3 Titan

With a thick atmosphere and a significant ionosphere, one expects Titan to have an induced magnetosphere similar to that of Venus. Orbiting Saturn at 20 R_S, Titan spends most of the time within the magnetosphere, but when the solar wind compresses the magnetosphere, Titan can spend some of its orbit in the magnetospheath or solar wind. The Voyager 2 flyby in 1980 showed that the magnetospheric plasma was deflected and absorbed, the magnetic field draped around Titan pretty much as expected (see review by Neubauer et al. 1984). The multiple flybys of Titan by the Cassini spacecraft are showing that the situation is more complex, as reviewed by Sittler et al. (2009) and illustrated by the sketch in *Fig. 6-22*. There seems to be general agreement that a total of about 300 kg/s of hydrogen is lost from Titan's atmosphere, but estimates of the escape rate of heavier species range from 5 to 85 kg/s (Johnson et al. 2009). The interaction seems to vary significantly with local time and the upstream conditions, and it will require combining the multiple Cassini flybys with models of the interaction before we have a clear consistent picture (Sittler et al. 2009). Nevertheless, it is already clear that the plasma interaction is a significant source of energy as well as a cause of escape for Titan's thick atmosphere (Westlake et al. 2011; Bell et al. 2011).

5.4 Io

The discovery of Io's broad influences on the jovian system predated spacecraft explorations. Bigg (1964) discovered Io's controlling influence over Jupiter's decametric radio emissions. Brown and Chaffee (1974) observed sodium emission from Io, which (Trafton et al. 1974) soon demonstrated to come from extended neutral clouds and not Io itself. Soon thereafter, Kupo et al. (1976) detected emissions from sulfur ions, which Brown (1976) recognized as coming from a dense plasma. With the prediction of volcanism by Peale et al. (1979) just before its discovery by Voyager 1 (Morabito et al. 1979), a consistent picture of Io's role began to emerge. Voyager 1's discovery of Jupiter's aurora and extreme UV emission from the torus (Broadfoot et al. 1979), along with its in situ measurements of the magnetosphere (Bridge et al. 1979), extended our awareness of Io's effect on the larger system.

The ensuing 25 years of observation by interplanetary missions, Earth-orbiting observatories, and ground-based telescopes has deepened our understanding of Io's influences (see the reviews by Thomas et al. 2004 and Schneider and Bagenal 2007). Highlights include Galileo's many close flybys of Io, with detailed fields-and-particle measurements of Io's interaction with the magnetosphere, and Cassini's month-long UV observation of the torus. Progress from Earth-based studies include the Hubble Space Telescope's sensitive UV observations of the footprint aurora and of Io's atmospheric emissions and ground-based observations of new atomic and molecular species in Io's atmosphere and the plasma torus.

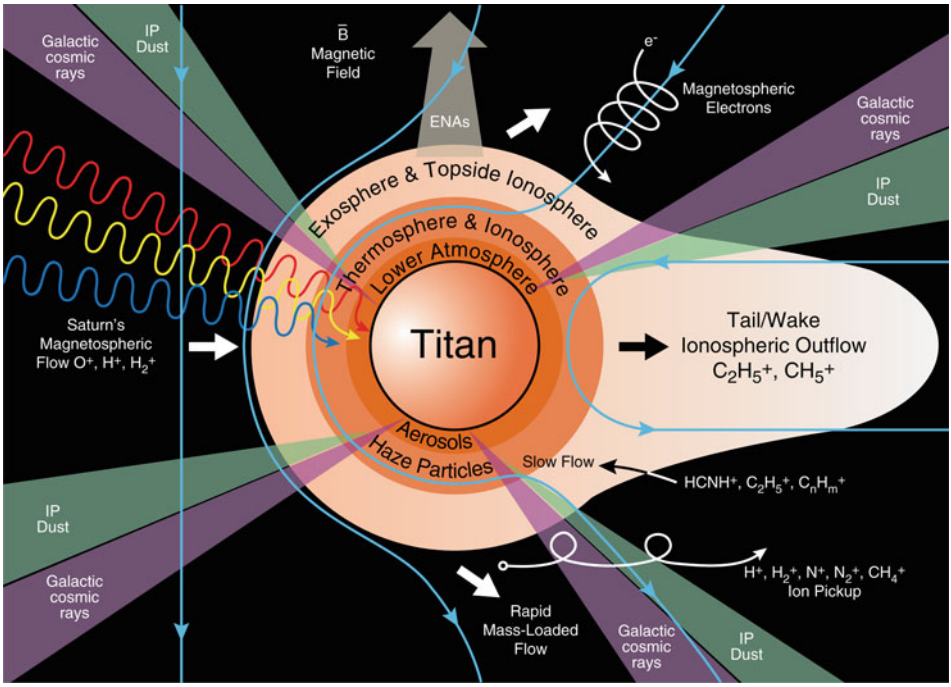


Fig. 6-22

Schematic of the processes contributing to source and loss of mass and energy in the plasma interaction with Titan's atmosphere (From Sittler et al. 2009)

Over the age of the solar system, the ton/s loss of Iogenic material to the magnetosphere accumulates to a net decrease in radius of about 2 km. While this loss is significant, Io is not in danger of running out of SO_2 in the lifetime of the solar system. It is plausible, however, that other volatile species such as H_2O were originally present on Io but were completely lost early in its history through processes now depleting Io of SO_2.

Figure 6-23 presents a sketch of the interaction of Io with the surrounding plasma that illustrates some of the processes. Inelastic collisions of torus ions with Io's atmosphere heat the atmospheric gases, causing a significant population of neutral molecules and atoms to gain speeds above Io's 2.6 km/s gravitational escape speed. These neutrals form an extensive corona circling most of the way around Jupiter. Io loses about 1–3 tons of neutral atoms per second. How much of the neutral escape is in molecular form (SO_2, SO, or S_2) as against atomic O or S is not known.

The various ion–electron–atom interactions each have a key effect on the magnetosphere. Most importantly torus ions collide with neutral atoms in the atmosphere, which in turn collide with other atoms in the process known as sputtering. Typically, one torus ion can transfer enough momentum for several atmospheric atoms or molecules to be ejected into Io's corona or possibly to escape from Io altogether. This is the primary pathway for material to be supplied to the neutral clouds and ultimately to the plasma torus. A second key reaction is electron-impact ionization: a torus electron ionizes an atmospheric atom, which is then accelerated

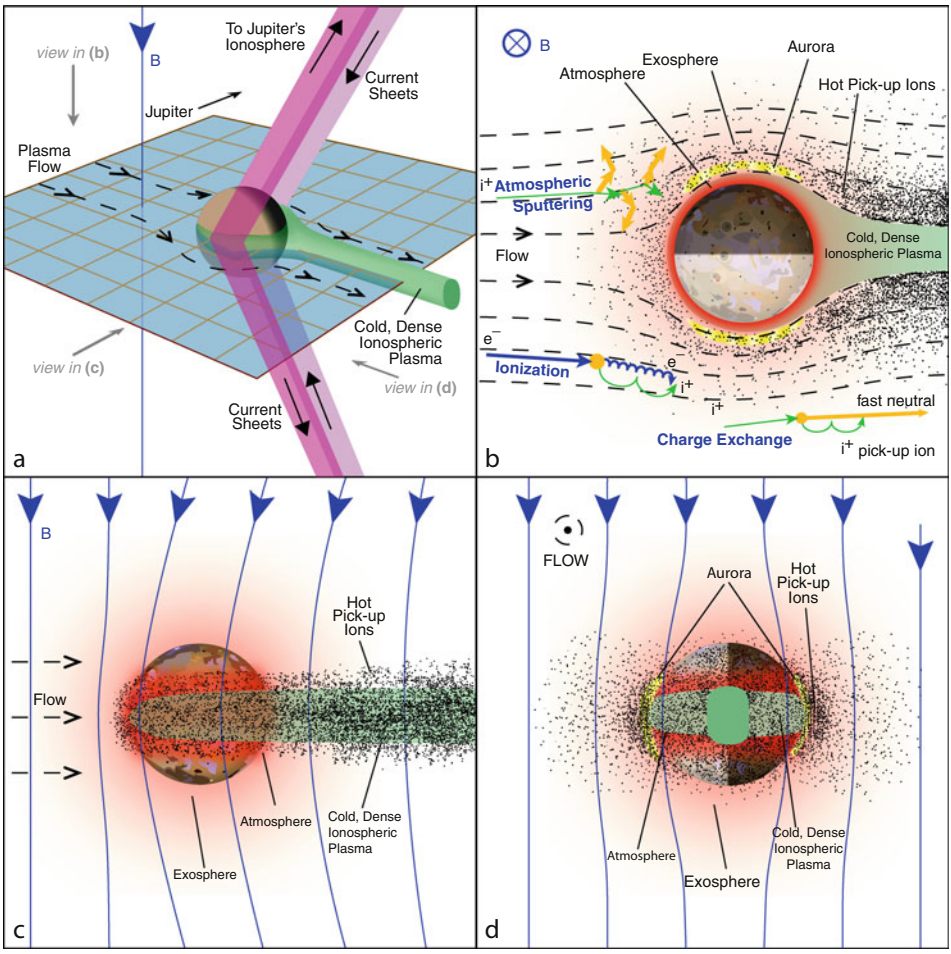


Fig. 6-23

Four views of the interaction between Io and the plasma torus. (a) Is a 3-D view showing the current sheets that couple Io and the surrounding plasma to Jupiter's ionosphere. (b) Is a cross-section of the interaction looking down on the north pole of Io, in the plane of Io's equator, when Io is located between the Sun and Jupiter (orbital phase 180, local noon in magnetospheric coordinates). (c) A projected view of the Io interaction looking from the Sun toward Jupiter. (d) A projected view of the interaction from downstream into the flowing plasma (ahead of Io in its orbit)

up to the speed of the plasma and leaves Io. Torus ions can also charge exchange with atmospheric neutrals, which results in a fresh ion and a high-speed neutral. Elastic collisions between ions and atoms can also eject material at speeds between those resulting from sputtering and charge exchange. Finally, electron-impact dissociation breaks down molecules into their component atoms.

Figure 6-23 shows that the strong magnetic field of Jupiter affects the interaction in such a way that the flow around Io resembles fluid flow around a cylinder. (Note that a strong intrinsic

magnetic field at Io has been ruled out by Galileo flybys over the poles.) Io's motion through the plasma creates an electrical current. While its surface or interior may be modestly conducting, the current is more likely to be carried in other conducting materials surrounding Io, such as its ionosphere and the plasma produced by ionization of its neutral corona. Currents induced across Io are closed by currents that flow along field lines between Io and Jupiter's polar ionosphere in both hemispheres. Observations by the Voyager 1 and Galileo spacecraft indicate that the net current in each circuit is about 3 million amps. The relative contributions from the conduction current through Io's ionosphere and the current generated by ion pickup in the surrounding plasma remains an issue of debate that awaits more sophisticated models (e.g., see the review by Saur et al. 2004).

A major question regarding Jupiter's magnetosphere is whether most mass loading happens in the near-Io interaction or in the broad neutral clouds far from Io. There is no doubt that substantial pickup occurs near Io, simply owing to the exposure of the upper atmosphere to pickup by the magnetosphere. Pickup near Io is also supported by evidence of fresh pickup ions of molecules (SO_2^+, SO^+, S_2^+, H_2S^+) near Io with dissociation lifetimes of just a few hours. But a closer look shows that the bulk of the Iogenic source comes from the ionization of atomic sulfur and oxygen farther from Io. Galileo measurements of the plasma fluxes downstream of Io suggest that the plasma source from the ionization of material in the immediate vicinity (within ~$5R_{Io}$) of Io is less than 300 kg/s, which is ~15% of the canonical net tons-per-second Iogenic source. The remainder must come from ionization of the extended clouds. It is not clear whether the observations were made during a typical situation, nor it is well established how much the net source and relative contributions of local and distant processes vary with Io's volcanic activity.

While most impacting plasma is diverted to Io's flanks, some is locked to field lines that are carried through Io itself. This ~10% of upstream plasma is rapidly decelerated and moves slowly (~3–7 km/s) over the poles. Most particles are absorbed by the moon or its tenuous polar atmosphere, so that the almost-stagnant polar flux tubes are evacuated of plasma. Downstream of Io, the Galileo instruments detected a small trickle of the cold dense ionospheric plasma that had been stripped away. This cold dense "tail" had a dramatic signature (>10 times the background density), but the nearly stagnant flow (~1 km/s) means that the net flux of this cold ionospheric material is at most a few percent of the Iogenic source and quickly couples to the surrounding torus plasma.

The strong electrodynamic interaction generates Alfven waves that propagate away from Io along the magnetic field (reviewed by Saur et al. 2004). Other MHD modes that propagate perpendicularly to the field dissipate within a short distance. The intense auroral emission in Jupiter's atmosphere at each "foot" of the flux tube (*Fig. 6-14*) connected to Io tells us that electrons are accelerated somewhere between Io and the atmosphere. The strong correlation of decametric radio emissions with Io's location also tells us that electrons stream away from Jupiter along the Io flux tube and field lines downstream of Io. But how much of the Alfven wave energy propagates through the torus and reaches Jupiter is not known. MHD models suggest that much of the wave energy is reflected at the sharp latitudinal gradients of density in the torus. Furthermore, how the Alfven wave evolves as it moves through the very low-density region between the torus and Jupiter's ionosphere is far from understood. Early ideas suggested that multiple bounces of the Alfven wave between ionospheres of opposite hemispheres could explain the repetitive bursts of radio emission. More recent studies suggest that the process is more complex, however, with the filamentation of Alfven waves and the development of quasi-static potential structures (see review by Hess et al. 2010 and references therein).

5.5 Pluto and Comets

Last but not least, there are planetary objects with escaping atmospheres that extend well beyond their solid surface. The neutral atoms and molecules are ionized by solar photons or charge exchange with solar wind protons. The fresh ions are picked up by the solar wind and carried downstream. Momentum is extracted from the solar wind and the IMF that is embedded in the slowed-down flow is stretched out behind the object in a magnetotail, similar to that of Venus shown in ❯ *Fig. 6-20*. The weak solar magnetic field at ~30 AU means that the gyromotions of the newly picked up ions are very large compared to the Pluto system so that a kinetic (or hybrid) approach must be applied in modeling the interaction (see Delamere 2009 and references therein). ❯ *Figure 6-24* is a sketch of the extended interaction region. The New Horizons spacecraft will fly past Pluto in 2015 and the particle instruments on board (McComas et al. 2008; McNutt et al. 2008) will determine how close this sketch bears to reality.

6 Outstanding Questions

The tables presented in this chapter quantify the characteristics of the seven magnetospheres of our solar system. The schematics give a glimpse of the diversity of their natures. While magnetospheres must share the same underlying basic physical processes, it is the application to very different conditions at the different planets that makes the study of planetary magnetospheres so interesting and tests our understanding. Below are the major outstanding questions:

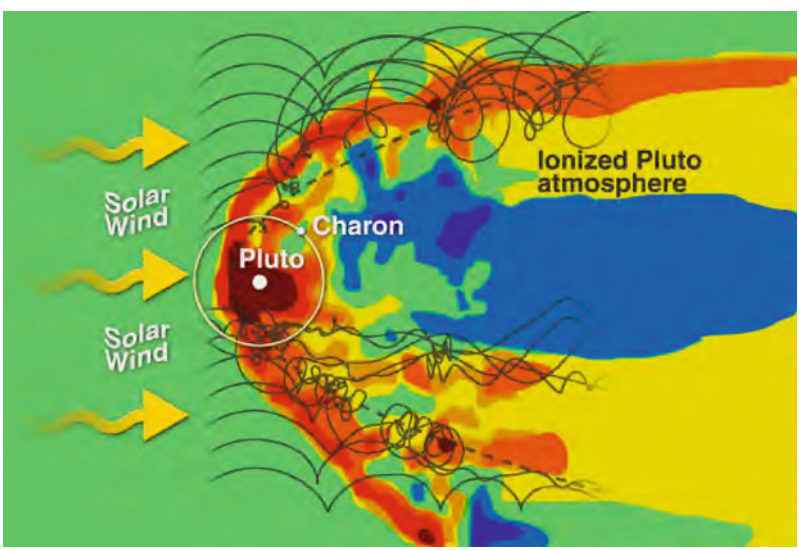


Fig. 6-24
Comet-like interaction of Pluto's escaping atmosphere with the tenuous solar wind at 30 AU. The contours are of ion density. The weak solar magnetic field results in a kinetic process whereby the ions produced by ionization of Pluto's escaping atmosphere exhibit large cycloidal motions, illustrated by sample trajectories in *grey lines*

- How do magnetic dynamos work in the wide range of planetary objects? Why do tiny Mercury and Ganymede have magnetic fields while Earth's sister planet Venus does not? What do the irregular magnetic fields of Uranus and Neptune tell us about their interiors?
- At Saturn, what causes the spin-periodic variability in radio emissions, magnetic field, and plasma properties? What causes the apparent fluctuation in the periodicity?
- How is plasma heated as it moves radially outward in rotation-dominated magnetospheres?
- How is material lost down the magnetotails of Jupiter and Saturn?
- Do Jupiter and/or Saturn have return, planetward, Dungey flows in the magnetotails? If not, how do flux tubes opened by dayside reconnection close and conserve magnetic flux?
- What processes lead to the decoupling of the middle magnetosphere of Jupiter from the planet's rotating ionosphere and cause the narrow auroral oval? What role do parallel potential drops play?
- What processes relate the solar wind variability to the apparent changes in Saturn's main aurora and the polar aurora at Jupiter?
- How do electrical currents couple the magnetospheres of Ganymede and Mercury to these objects with very tenuous atmospheres?
- How are particles accelerated and trapped in the mini-magnetospheres of Ganymede and Mercury?
- What processes have been responsible for removing atmospheric gases (particularly water) over the geological history of Mars and Venus?
- How does the plasma interaction with Titan's atmosphere vary with local time and surrounding plasma conditions? How do these variable conditions affect the fluxes of energy into and material out of Titan's atmosphere?
- What processes are involved in the interactions of Io and Enceladus with their surrounding plasmas? What causes the similarities and differences between the two systems?

References

Acuña, M. H., et al. 2001, J. Geophys. Res., 106, 23403

Alexeev, I. I., & Belenkaya, E. S. 2005, Ann. Geophys., 23, 809

Anderson, B. J., et al. 2010, Space Sci. Rev., 152, 307

Anderson, B. J., et al. 2011, Science, 333, 1859

Andrews, D. J., Coates, A. J., Cowley, S. W. H., Dougherty, M. K., Lamy, L., Provan, G., & Zarka, P. 2010a, J. Geophys. Res., 115, 12252

Andrews, D. J., Cowley, S. W. H., Dougherty, M. K., & Provan, G. 2010b, J. Geophys. Res., 115, 4212

Arridge, C. S., et al. 2011, Exp. Astron., 33, 113

Axford, W. I., & Hines, C. O. 1961, Can. J. Phys., 39, 1433

Badman, S. V., Bunce, E. J., Clarke, J. T., Cowley, S. W. H., Gérard, J.-C., Grodent, D., & Milan, S. E. 2005, J. Geophys. Res., 110, 11216

Bagenal, F. 1992, Annu. Rev. Earth Planet. Sci., 20, 289

Bagenal, F. 2007, J. Atmos. Solar-Terr. Phys., 69, 387

Bagenal, F. 2009, in Comparative Planetary Environmnents, ed. C. J. Schrijver & G. L. Siscoe (Cambridge/New York: Cambridge University Press)

Bagenal, F., & Delamere, P. A. 2011, J. Geophys. Res., 116, 5209

Bagenal, F., Dowling, T. E., & McKinnon, W. B. 2004, in Jupiter, ed. F. Bagenal, T. E. Dowling, & W. B. McKinnon

Balogh, A. 2010, Space Sci. Rev., 152, 23

Balogh, A., et al. 2005, Space Sci. Rev., 118, 155

Barabash, S., et al. 2007, Nature, 450, 650

Belcher, J. W., Lazarus, A. J., McNutt, R. L., Jr., & Gordon, G. S., Jr. 1993, J. Geophys. Res., 98, 15177

Bell, J. M., Westlake, J., & Waite, J. H., Jr. 2011, Geophys. Res. Lett., 38, 6202

Bhardwaj, A., & Gladstone, G. R. 2000, Rev. Geophys., 38, 295

Biermann, L. 1957, Observatory, 77, 109

Bigg, E. K. 1964, Nature, 203, 1008

Bolton, S. J., Thorne, R. M., Bourdarie, S., de Pater, I., & Mauk, B. 2004, in Jupiter's Inner Radiation Belts, ed. F. Bagenal, T. E. Dowling, & W. B. McKinnon, 671–688
Brain, D. A. 2006, Space Sci. Rev., 126, 77
Brain, D., et al. 2010, Icarus, 206, 139
Breuer, D., Labrosse, S., & Spohn, T. 2010, Space Sci. Rev., 152, 449
Brice, N. M., & Ioannidis, G. A. 1970, Icarus, 13, 173
Bridge, H. S., et al. 1979, Science, 204, 987
Broadfoot, A. L., et al. 1979, Science, 204, 979
Brown, R. A. 1976, ApJL, 206, L179
Brown, R. A., & Chaffee, F. H., Jr. 1974, ApJL, 187, L125
Bunce, E. J., Cowley, S. W. H., & Yeoman, T. K. 2004, J. Geophys. Res., 109, 9
Bunce, E. J., et al. 2008, J. Geophys. Res., 113, 9209
Bunce, E. J., et al. 2010, J. Geophys. Res., 115, 10238
Cassak, P. A., & Shay, M. A. 2011, Space Sci. Rev., 265
Chapman, S., & Ferraro, V. C. A. 1930, Nature, 126, 129
Cheng, A. F., Krimigis, S. M., Mauk, B. H., Keath, E. P., & Maclennan, C. G. 1987, J. Geophys. Res., 92, 15315
Christensen, U. R. 2010, Space Sci. Rev., 152, 565
Clarke, J. T., et al. 2002, Nature, 415, 997
Clarke, J. T., et al. 2005, Nature, 433, 717
Clarke, J. T., et al. 2009, J. Geophys. Res., 114, 5210
Connerney, J. E. P. 1981, J. Geophys. Res., 86, 7679
Connerney, J. E. P. 1993, J. Geophys. Res., 98, 18659
Connerney, J. 2007, in Planetary Magnetism, ed. G. Schubert & T. Spohn (Elsevier)
Connerney, J. E. P., Acuña, M. H., Ness, N. F., Spohn, T., & Schubert, G. 2004, Space Sci. Rev., 111, 1
Cowley, S. W. H., & Bunce, E. J. 2001, Planet. Space Sci., 49, 1067
Cowley, S. W. H., & Bunce, E. J. 2003, Planet. Space Sci., 51, 57
Cowley, S. W. H., Nichols, J. D., & Bunce, E. J. 2002, Planet. Space Sci., 50, 717
Cowley, S. W. H., Bunce, E. J., & Nichols, J. D. 2003, J. Geophys. Res., 108, 8002
Cowley, S. W. H., et al. 2005, J. Geophys. Res., 110, 2201
Cowley, S. W. H., Nichols, J. D., & Andrews, D. J. 2007, Ann. Geophys., 25, 1433
Cowley, S. W. H., Badman, S. V., Imber, S. M., & Milan, S. E. 2008a, Geophys. Res. Lett., 35, 10101
Cowley, S. W. H., et al. 2008b, Ann. Geophys., 26, 2613
Cowling, T. G. 1933, MNRAS, 94, 39
Crary, F. J., et al. 2005, Nature, 433, 720
Cravens, T. E., Howell, E., Waite, J. H., & Gladstone, G. R. 1995, J. Geophys. Res., 100, 17153
T. E. Cravens. 1997. Physics of solar system plasmas. SAO/NASA Astrophysics Data System. (Cambridge: Cambridge University Press), B529 C72 1997. http://adsabs.harvard.edu/abs/1997pssp.conf.....C
Delamere, P. A. 2009, J. Geophys. Res., 114, 3220
Delamere, P. A., & Bagenal, F. 2003, J. Geophys. Res., 108, 1276
Delamere, P. A., & Bagenal, F. 2008, Geophys. Res. Lett., 35, 3107
Delamere, P. A., & Bagenal, F. 2010, J. Geophys. Res., 115, 10201
Delamere, P. A., Steffl, A., & Bagenal, F. 2004, J. Geophys. Res., 109, 10216
Delamere, P. A., Bagenal, F., Dols, V., & Ray, L. C. 2007, Geophys. Res. Lett., 34, 9105
Delva, M., Mazelle, C., Bertucci, C., Volwerk, M., Vörös, Z., & Zhang, T. L. 2011, J. Geophys. Res., 116, 2318
Dougherty, M. K., Esposito, L. W., & Krimigis, S. M. 2009, in Saturn from Cassini-Huygens, ed. M. K. Dougherty, L. W. Esposito, & S. M. Krimigis
Dungey, J. W. 1961, Phys. Rev. Lett., 6, 47
Ergun, R. E., Ray, L., Delamere, P. A., Bagenal, F., Dols, V., & Su, Y.-J. 2009, J. Geophys. Res., 114, 5201
Finlay, C. C., et al. 2010, Geophys. J. Int., 183, 1216
Fleshman, B. L., Delamere, P. A., & Bagenal, F. 2010a, J. Geophys. Res., 115, 4007
Fleshman, B. L., Delamere, P. A., & Bagenal, F. 2010b, Geophys. Res. Lett., 37, 3202
Galand, M., Moore, L., Mueller-Wodarg, I., Mendillo, M., & Miller, S. 2011, J. Geophys. Res., 116, 9306
Ge, Y. S., Russell, C. T., & Khurana, K. K. 2010, Planet. Space Sci., 58, 1455
Gold, T. 1959, J. Geophys. Res., 64, 1219
Gombosi, T. I. 1998. Physics of the space environment. SAO/NASA Astrophysics Data System. (Cambridge/New York: Cambridge University Press) http://adsabs.harvard.edu/abs/1998pse..conf.....G
Grocott, A., Badman, S. V., Cowley, S. W. H., Milan, S. E., Nichols, J. D., & Yeoman, T. K. 2009, J. Geophys. Res., 114, 7219
Grodent, D., Clarke, J. T., Waite, J. H., Cowley, S. W. H., Gérard, J.-C., & Kim, J. 2003, J. Geophys. Res., 108, 1366
Grodent, D., Gérard, J.-C., Clarke, J. T., Gladstone, G. R., & Waite, J. H. 2004, J. Geophys. Res., 109, 5201
Grodent, D., Bonfond, B., Radioti, A., Gérard, J.-C., Jia, X., Nichols, J. D., & Clarke, J. T. 2009, J. Geophys. Res., 114, 7212
Gurnett, D. A., & Bhattacharjee, A. 2005
Gurnett, D. A., Persoon, A. M., Kurth, W. S., Groene, J. B., Averkamp, T. F., Dougherty, M. K., & Southwood, D. J. 2007, Science, 316, 442

Gurnett, D. A., Lecacheux, A., Kurth, W. S., Persoon, A. M., Groene, J. B., Lamy, L., Zarka, P., & Carbary, J. F. 2009, Geophys. Res. Lett., 36, 16102
Gurnett, D. A., Groene, J. B., Persoon, A. M., Menietti, J. D., Ye, S.-Y., Kurth, W. S., MacDowall, R. J., & Lecacheux, A. 2010, Geophys. Res. Lett., 37, 24101
Hansen, C. J., Esposito, L., Stewart, A. I. F., Colwell, J., Hendrix, A., Pryor, W., Shemansky, D., & West, R. 2006, Science, 311, 1422
Hansen, C. J., et al. 2008, Nature, 456, 477
Hill, T. W. 1979, J. Geophys. Res., 84, 6554
Hill, T. W., Dessler, A. J., & Maher, L. J. 1981, J. Geophys. Res., 86, 9020
Hill, T. W., et al. 2008, J. Geophys. Res., 113, 1214
Huddleston, D. E., Russell, C. T., Kivelson, M. G., Khurana, K. K., & Bennett, L. 1998, J. Geophys. Res., 103, 20075
Hui, Y., et al. 2010, J. Geophys. Res., 115, 7102
Hulot, G., Finlay, C. C., Constable, C. G., Olsen, N., & Mandea, M. 2010, Space Sci. Rev., 152, 159
Jackman, C. M., Achilleos, N., Bunce, E. J., Cecconi, B., Clarke, J. T., Cowley, S. W. H., Kurth, W. S., & Zarka, P. 2005, J. Geophys. Res., 110, 10212
Jackman, C. M., et al. 2008, J. Geophys. Res., 113, 11213
Jarvinen, R., Kallio, E., Dyadechkin, S., Janhunen, P., & Sillanpää, I. 2010, Geophys. Res. Lett., 37, 16201
Jia, X., Walker, R. J., Kivelson, M. G., Khurana, K. K., & Linker, J. A. 2008, J. Geophys. Res., 113, 6212
Jia, X., Walker, R. J., Kivelson, M. G., Khurana, K. K., & Linker, J. A. 2009, J. Geophys. Res., 114, 9209
Jia, X., Kivelson, M. G., Khurana, K. K., & Walker, R. J. 2010a, Space Sci. Rev., 152, 271
Jia, X., Walker, R. J., Kivelson, M. G., Khurana, K. K., & Linker, J. A. 2010b, J. Geophys. Res., 115, 12202
Johnson, R., Tucker, O., Michael, M., Sitter, E., Smith, H., Young, D., & Waite, J. 2009, Mass loss processes in Titan's upper atmosphere, in Titan from Cassini-Huygens, ed. R. H. Brown, J.-P. Lebreton, & J. H. Waite (Dordrecht/New York: Springer)
Joy, S. P., Kivelson, M. G., Walker, R. J., Khurana, K. K., Russell, C. T., & Ogino, T. 2002, J. Geophys. Res., 107, 1309
Jurac, S., & Richardson, J. D. 2005, J. Geophys. Res., 110, 9220
Kanani, S. J., et al. 2010, J. Geophys. Res., 115, 6207
Khurana, K. K. 2001, J. Geophys. Res., 106, 25999
Khurana, K. K., & Schwarzl, H. K. 2005, J. Geophys. Res., 110, 7227
Kivelson, M. G. 2007, Planetary magnetospheres, in Handbook of the Solar-Terrestrial Environment, ed. Y. Kamide & A. C.-L. Chian (Berlin/New York: Springer), 470
Kivelson, M. G., & Russell, C. T. (ed.) 1995
Kivelson, M. G., & Southwood, D. J. 2005, J. Geophys. Res., 110, 12209
Kupo, I., Mekler, Y., & Eviatar, A. 1976, ApJL, 205, L51
Kurth, W. S., et al. 2005, Nature, 433, 722
Kurth, W. S., Averkamp, T. F., Gurnett, D. A., Groene, J. B., & Lecacheux, A. 2008, J. Geophys. Res., 113, 5222
La Belle-Hamer, A. L., Otto, A., & Lee, L. C. 1995, J. Geophys. Res., 100, 11875
Lamy, L., Cecconi, B., Prangé, R., Zarka, P., Nichols, J. D., & Clarke, J. T. 2009, J. Geophys. Res., 114, 10212
Lamy, L., et al. 2010, Geophys. Res. Lett., 37, 12104
Lembege, B., et al. 2004, Space Sci. Rev., 110, 161
Mauk, B. H., & Fox, N. J. 2010, J. Geophys. Res., 115, 12220
Mauk, B. H., Krimigis, S. M., Cheng, A. F., & Selesnick, R. S. 1995, in Neptune and Triton, ed. D. P. Cruikshank, M. S. Matthews, & A. M. Schumann (Tucson: University of Arizona Press), 169–232
Mauk, B. H., Anderson, B. J., & Thorne, R. M. 2002, Magnetosphere-ionosphere coupling at Earth, Jupiter, and Beyond, in Atmospheres in the Solar System: Comparative Aeronomy, ed. M. Mendillo, A. Nagy, & J. H. Waite (Washington, D.C.: American Geophysical Union), 97
Mauk, B., et al. 2009, Fundamental plasma processes in Saturn's magnetosphere, in Saturn from Cassini-Huygens, ed. S. M. Krimigis, M. K. Dougherty, & L. W. Esposito (Dordrecht/New York: Springer)
McComas, D. J., & Bagenal, F. 2007, Geophys. Res. Lett., 34, 20106
McComas, D. J., Allegrini, F., Bagenal, F., Crary, F., Ebert, R. W., Elliott, H., Stern, A., & Valek, P. 2007, Science, 318, 217
McComas, D., et al. 2008, Space Sci. Rev., 140, 261
McNutt, R. L., Jr., Belcher, J. W., Sullivan, J. D., Bagenal, F., & Bridge, H. S. 1979, Nature, 280, 803
McNutt, R. L., et al. 2007, Science, 318, 220
McNutt, R. L., et al. 2008, Space Sci. Rev., 140, 315
Merrill, R., McFadden, P., & McElhinny, M. 1996, (Academic)
Moore, T. E., & Horwitz, J. L. 2007, Rev. Geophys., 45, 3002
Morabito, L. A., Synnott, S. P., Kupferman, P. N., & Collins, S. A. 1979, Science, 204, 972
Nagy, A. F., et al. 2004, Space Sci. Rev., 111, 33
Ness, N. F. 2010, Space Sci. Rev., 152, 5

Neubauer, F. M., Gurnett, D. A., Scudder, J. D., & Hartle, R. E. 1984, Titan's magnetospheric interaction, in Saturn ed. T. Gehrels & M. S. Matthews (Tucson: University of Arizona Press), 760–787

Nichols, J. D., & Cowley, S. W. H. 2005, Ann. Geophys., 23, 799

Nichols, J. D., et al. 2010, Geophys. Res. Lett., 37, 15102

Nimmo, F., & Stevenson, D. J. 2000, J. Geophys. Res., 105, 11969

Olsen, N., Glassmeier, K.-H., & Jia, X. 2010, Space Sci. Rev., 152, 135

Ozak, N., Schultz, D. R., Cravens, T. E., Kharchenko, V., & Hui, Y.-W. 2010, J. Geophys. Res., 115, 11306

Pallier, L., & Prangé, R. 2004, Geophys. Res. Lett., 31, 6701

Parker, E. N. 2007, Conversations on electric and magnetic fields in the cosmos, in Conversations on Electric and Magnetic Fields in the Cosmos, ed. E. N. Parker (Princeton: Princeton University Press)

Paty, C., & Winglee, R. 2006, Geophys. Res. Lett., 33, 10106

Paty, C., Paterson, W., & Winglee, R. 2008, J. Geophys. Res., 113, 6211

Peale, S. J., Cassen, P., & Reynolds, R. T. 1979, Science, 203, 892

Phillips, J. L., & Russell, C. T. 1987, Adv. Space Res., 7, 291

Porco, C. C., et al. 2006, Science, 311, 1393

Radioti, A., Grodent, D., Gérard, J.-C., Bonfond, B., & Clarke, J. T. 2008, Geophys. Res. Lett., 35, 3104

Radioti, A., Grodent, D., Gérard, J.-C., & Bonfond, B. 2010, J. Geophys. Res., 115, 7214

Radioti, A., Grodent, D., Gérard, J.-C., Vogt, M. F., Lystrup, M., & Bonfond, B. 2011, J. Geophys. Res., 116, 3221

Ray, L. C., Su, Y.-J., Ergun, R. E., Delamere, P. A., & Bagenal, F. 2009, J. Geophys. Res., 114, 4214

Ray, L. C., Ergun, R. E., Delamere, P. A., & Bagenal, F. 2010, J. Geophys. Res., 115, 9211

Russell, C. T. 1993, J. Geophys. Res., 98, 18681

Russell, C. T. 2004, Adv. Space Res., 33, 2004

Russell, C. T. 2006, Adv. Space Res., 37, 1467

Russell, C. T., & Dougherty, M. K. 2010, Space Sci. Rev., 152, 251

Russell, C. T., Khurana, K. K., Kivelson, M. G., & Huddleston, D. E. 2000, Adv. Space Res., 26, 1499

Russell, C. T., Luhmann, J. G., & Strangeway, R. J. 2006, Planet. Space Sci., 54, 1482

Santos-Costa, D., & Bourdarie, S. A. 2001, Planet. Space Sci., 49, 303

Saur, J., Schilling, N., Neubauer, F. M., Strobel, D. F., Simon, S., Dougherty, M. K., Russell, C. T., & Pappalardo, R. T. 2008, Geophys. Res. Lett., 35, 20105

Schneider, N., & Bagenal, F. 2007, Io's neutral clouds, plasma torus and magnetospheric interactions, in Io After Galileo, ed. R. M. C. Lopes & J. R. Spencer (Berlin/New York: Springer)

Schubert, G., Solomatov, V., Tackley, P., & Turcotte, D. 1988, Mantle convection and the thermal evolution of Venus, in Venus II, ed. D. Hunten, R. Phillips, & S. W. Bougher (Tucson: University of Arizona Press)

Scurry, L., & Russell, C. T. 1991, J. Geophys. Res., 96, 9541

Scurry, L., Russell, C. T., & Gosling, J. T. 1994, J. Geophys. Res., 99, 14811

Sergis, N., et al. 2010, Geophys. Res. Lett., 37, 2102

Siscoe, G. L. 1979, Towards a comparative theory of magnetospheres, in Solar System Plasma Physics, ed. E. N. Parker, C. F. Kennel, & L. J. Lanzerotti (Amsterdam/New York: North-Holland), 319–402

Siscoe, G. L., & Summers, D. 1981, J. Geophys. Res., 86, 8471

Siscoe, G. L., Eviatar, A., Thorne, R. M., Richardson, J. D., Bagenal, F., & Sullivan, J. D. 1981, J. Geophys. Res., 86, 8480

Sittler, E. C., et al. 2008, Planet. Space Sci., 56, 3

Sittler, E., Hartle, R., Bertucci, C., Coates, A., Cravens, T., Dandouras, I., & Shemansky, D. 2009, Energy deposition processes in Titan's upper atmosphere and its induced magnetosphere, in Titan from Cassini-Huygens, ed. R. H. Brown, J.-P. Lebreton, & J. H. Waite (Dordrecht/New York: Springer)

Slavin, J. A. 2004, Adv. Space Res., 33, 1859

Slavin, J. A., & Holzer, R. E. 1979, J. Geophys. Res., 84, 2076

Slavin, J. A., Smith, E. J., Spreiter, J. R., & Stahara, S. S. 1985, J. Geophys. Res., 90, 6275

Slavin, J. A., et al. 2007, Space Sci. Rev., 131, 133

Slavin, J. A., et al. 2009, Science, 324, 606

Slavin, J. A., et al. 2010, Science, 329, 665

Smith, C. G. A. 2011, MNRAS, 410, 2315

Smith, H. T., Johnson, R. E., Perry, M. E., Mitchell, D. G., McNutt, R. L., & Young, D. T. 2010, J. Geophys. Res., 115, 10252

Smrekar, S., Elkins-Tanton, L., Leitner, J., Lenardic, A., Mackwell, S., Moresi, L., Sotin, C., & Stofan, E. 2007, Tectonic and thermal evolution of Venus and the role of Volatiles, in AGU Monograph, Vol. 176, Exploring Venus as a Terrestrial Planet, ed. E. R. Stofan, T. E. Cravens, & L. W. Esposito (Washington, DC: American Geophysical Union)

Southwood, D. 2011, J. Geophys. Res., 116, 1201
Southwood, D. J., & Kivelson, M. G. 1987, J. Geophys. Res., 92, 109
Southwood, D. J., & Kivelson, M. G. 2007, J. Geophys. Res., 112, 12222
Stanley, S., & Bloxham, J. 2006, Icarus, 184, 556
Stanley, S., & Glatzmaier, G. A. 2010, Space Sci. Rev., 152, 617
Steffl, A. J., Stewart, A. I. F., & Bagenal, F. 2004, Icarus, 172, 78
Steffl, A. J., Delamere, P. A., & Bagenal, F. 2006, Icarus, 180, 124
Steffl, A. J., Delamere, P. A., & Bagenal, F. 2008, Icarus, 194, 153
Stevenson, D. J. 1982, Geophys. Astrophys. Fluid Dyn., 21, 113
Stevenson, D. J. 2003, Earth Planet. Sci. Lett., 208, 1
Stevenson, D. J. 2010, Space Sci. Rev., 152, 651
Stevenson, D. J., Spohn, T., & Schubert, G. 1983, Icarus, 54, 466
Stewart, A. J., Schmidt, M. W., van Westrenen, W., & Liebske, C. 2007, Science, 316, 1323
Swisdak, M., Rogers, B. N., Drake, J. F., & Shay, M. A. 2003, J. Geophys. Res., 108, 1218
Talboys, D. L., Arridge, C. S., Bunce, E. J., Coates, A. J., Cowley, S. W. H., & Dougherty, M. K. 2009a, J. Geophys. Res., 114, 6220
Talboys, D. L., Arridge, C. S., Bunce, E. J., Coates, A. J., Cowley, S. W. H., Dougherty, M. K., & Khurana, K. K. 2009b, Geophys. Res. Lett., 36, 19107
Thomas, N., Bagenal, F., Hill, T. W., & Wilson, J. K. 2004, The Io neutral clouds and plasma torus, in Jupiter: The Planet, Satellites and Magnetosphere, ed. F. Bagenal, T. E. Dowling, & W. B. McKinnon (Cambridge, UK/New York: Cambridge University Press), 561–591
Tokar, R. L., et al. 2009, Geophys. Res. Lett., 36, 13203
Trafton, L., Parkinson, T., & Macy, W., Jr. 1974, ApJL, 190, L85
Treumann, R. A. 2009, A&AR, 17, 409
van Allen, J. A., & Bagenal, F. 1999, Planetary magnetospheres and the interplanetary medium, in The New Solar System, ed. J. K. Beatty, C. Collins Petersen, & A. Chaikin (Cambridge/New York: Cambridge University Press), 39
Vasyliunas, V. M. 1983, Plasma distribution and flow, in Physics of the jovian Magnetosphere, ed. A. J. Dessler (Cambridge: Cambridge University Press), 395–453
Vasyliūnas, V. M. 2001, Geophys. Res. Lett., 28, 2177
Vasyliūnas, V. M. 2004, Adv. Space Res., 33, 2113
Vasyliunas, V. 2009, Fundamentals of planetary magnetospheres, in Heliophysics: Plasma Physics of the Local Cosmos, ed. C. J. Schrijver & G. L. Siscoe (Cambridge: Cambridge University Press)
Vasyliunas, V. 2010, Energy conversion in planetary magnetospheres, in Heliophysics: Space Storms and Radiation: Causes and Effects, ed. C. J. Schrijver & G. L. Siscoe (Cambridge: Cambridge University Press), 263
Vasyliūnas, V. M. 2011, Space Sci. Rev., 158, 91
Verhille, G., Plihon, N., Bourgoin, M., Odier, P., & Pinton, J.-F. 2010, Space Sci. Rev., 152, 543
Vogt, M. F., Kivelson, M. G., Khurana, K. K., Joy, S. P., & Walker, R. J. 2010, J. Geophys. Res., 115, 6219
Vogt, M. F., Kivelson, M. G., Khurana, K. K., Walker, R. J., Bonfond, B., Grodent, D., & Radioti, A. 2011, J. Geophys. Res., 116, 3220
Volwerk, M., Delva, M., Futaana, Y., Retinò, A., Vörös, Z., Zhang, T. L., Baumjohann, W., & Barabash, S. 2009, Ann. Geophys., 27, 2321
Volwerk, M., Delva, M., Futaana, Y., Retinò, A., Vörös, Z., Zhang, T. L., Baumjohann, W., & Barabash, S. 2010, Ann. Geophys., 28, 1877
Waite, J. H., Jr. et al. 1994, J. Geophys. Res., 99, 14799
Waite, J. H., et al. 2001, Nature, 410, 787
Waite, J. H., et al. 2006, Science, 311, 1419
Walt, M. 2005, in Introduction to Geomagnetically Trapped Radiation, ed. M. Walt (Cambridge/New York: Cambridge University Press)
Westlake, J. H., Bell, J. M., Waite, J. H., Jr., Johnson, R. E., Luhmann, J. G., Mandt, K. E., Magee, B. A., & Rymer, A. M. 2011, J. Geophys. Res., 116, 3318
Wicht, J., & Tilgner, A. 2010, Space Sci. Rev., 152, 501
Woch, J., Krupp, N., & Lagg, A. 2002, Geophys. Res. Lett., 29, 070000
Young, D. 1997a, Ion and neutral mass spectrometry, in Encyclopedia of Planetary Sciences, ed. J. A. Shirley & R. W. Fairbridge (Van Nostrand Reinhold)
Young, D. 1997b, Space plasma particle instrumentation and the new paradigm: faster, cheaper, better, in AGU Monograph 102, Measurement Techniques in Space Plasmas: Particles, ed. D. T. Young, R. F. Pfaff, & J. E. Borovsky (Washington, DC: American Geophysical Union)
Zarka, P., Lamy, L., Cecconi, B., Prangé, R., & Rucker, H. O. 2007, Nature, 450, 265
Zhang, T. L., et al. 2010, Geophys. Res. Lett., 37, 14202
Zieger, B., Vogt, J., Glassmeier, K.-H., & Gombosi, T. I. 2004, J. Geophys. Res., 109, 7205

7 Planetary Rings

Matthew S. Tiscareno
Center for Radiophysics and Space Research, Cornell University, Ithaca, NY, USA

1 Introduction 311
1.1 Orbital Elements 312
1.2 Roche Limits, Roche Lobes, and Roche Critical Densities 313
1.3 Optical Depth 316

2 Rings by Planetary System 317
2.1 The Rings of Jupiter 317
2.2 The Rings of Saturn 319
2.3 The Rings of Uranus 320
2.4 The Rings of Neptune 323
2.5 Unconfirmed Ring Systems 324
2.5.1 Mars 324
2.5.2 Pluto 325
2.5.3 Rhea and Other Moons 325
2.5.4 Exoplanets 327

3 Rings by Type 328
3.1 Dense Broad Disks 328
3.1.1 Spiral Waves 329
3.1.2 Gap Edges and Moonlet Wakes 333
3.1.3 Radial Structure 336
3.1.4 Self-Gravity Wakes 337
3.1.5 Propellers 339
3.1.6 Spokes and Impacts 346
3.2 Dense Narrow Rings 347
3.3 Narrow Dusty Rings 349
3.4 Diffuse Dusty Rings 351
3.5 Ring Arcs and Azimuthal Clumps 356
3.5.1 Neptune's Adams Ring 357
3.5.2 Jupiter's Main Ring and Other Azimuthal Clumps 359
3.5.3 Saturn's G Ring and Other Moon-Embedded Arcs 359
3.6 Rings as Detectors 361

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_7, © Springer Science+Business Media Dordrecht 2013

4 Experimental Rings Science *362*
4.1 Numerical Simulations 362
4.2 Physical Experiments and the Coefficient of Restitution 364
4.3 Spectroscopic Ground Truth 365

5 Age and Origin of Ring Systems *366*

6 Rings and Other Disks *368*

Acknowledgments *368*

References *368*

Abstract: Planetary rings are the only nearby astrophysical disks and the only disks that have been investigated by spacecraft (especially the *Cassini* spacecraft orbiting Saturn). Although there are significant differences between rings and other disks, chiefly the large planet/ring mass ratio that greatly enhances the flatness of rings (aspect ratios as small as 10^{-7}), understanding of disks in general can be enhanced by understanding the dynamical processes observed at close range and in real time in planetary rings.

We review the known ring systems of the four giant planets, as well as the prospects for ring systems yet to be discovered. We then review planetary rings by type. The A, B, and C rings of Saturn, plus the Cassini Division, comprise our solar system's only dense broad disk and host many phenomena of general application to disks including spiral waves, gap formation, self-gravity wakes, viscous overstability and normal modes, impact clouds, and orbital evolution of embedded moons. Dense narrow rings are found both at Uranus (where they comprise the main rings entirely) and at Saturn (where they are embedded in the broad disk) and are the primary natural laboratory for understanding shepherding and self-stability. Narrow dusty rings, likely generated by embedded source bodies, are surprisingly found to sport azimuthally confined arcs at Neptune, Saturn, and Jupiter. Finally, every known ring system includes a substantial component of diffuse dusty rings.

Planetary rings have shown themselves to be useful as detectors of planetary processes around them, including the planetary magnetic field and interplanetary impactors as well as the gravity of nearby perturbing moons. Experimental rings science has made great progress in recent decades, especially numerical simulations of self-gravity wakes and other processes but also laboratory investigations of coefficient of restitution and spectroscopic ground truth. The age of self-sustained ring systems is a matter of debate; formation scenarios are most plausible in the context of the early solar system, while signs of youthfulness indicate at least that rings have never been static phenomena.

1 Introduction

Planetary rings come in a diverse array of shapes and sizes. They may be broad or narrow, dense or tenuous, dusty or not, and they may contain various kinds of structures including arcs, wavy edges, embedded moonlets, and radial variations. Rings share the defining characteristic of a swarm of objects orbiting a central planet with vertical motions that are small compared to their motions within a common plane. The latter arises because planets in our solar system (with the exceptions of Mercury and Venus, which have no known natural material in orbit) are fast-enough rotaters that their shapes are dominated by an equatorial bulge that adds a strong quadrupole moment (J_2) to their gravity fields (see Sect. 1.1).

This is a major contrast between rings and other astrophysical disks, which are not defined by asymmetry in an external gravity field but by the average angular momentum of the disk itself (in both cases, once a preferred plane is established, collisions among particles damp out the motions perpendicular to it). However, rings do have a number of similarities with other astrophysical disks, which add to the motivation for studying them. Unlike other known disk systems that are either many light-years away or (like the early stages of our solar system) far back in time, planetary rings can be studied up close and in real time. Thus, it is worthwhile to consider the parallels that can be drawn between planetary rings and the study of other disks (see Sect. 6).

In this chapter, after some further introductory notes on important concepts (Sects. 1.1, 1.2, and 1.3), we will give an overview of the known ring systems, as well as systems where rings are unconfirmed but plausible, in Sect. 2. More detailed descriptions can be found in Sect. 3, which contains a discussion of various ring structures organized by type, with a focus on finding commonalities among rings in different locations that share certain qualities. In Sect. 4, we will discuss experimental methods of learning about rings, and in Sect. 5, we will discuss the age and origin of ring systems. Finally, in Sect. 6, we will discuss ways that planetary rings can illuminate the study of other astrophysical disks.

1.1 Orbital Elements

Rings are fundamentally populations of orbiting material. Therefore, to discuss structure within rings, we will occasionally refer to one or more of the six *orbital elements* that describe the orbit of any object around another object. These six parameters are simply a transformation of the six Cartesian parameters for position and velocity (x, y, z, $\dot{x}$, $\dot{y}$, and $\dot{z}$) under the assumption that the object moves in the gravity field of a point mass (hereafter the "planet"). The derivation can be found in any textbook on orbital mechanics (e.g., Sect. 2.8 in Murray and Dermott 1999).

A diagram of the orbit in space is found in *Fig. 7-1*. The size of the orbit, and its gravitational potential energy, is described by the semimajor axis a, which is the mean distance between the orbiting particle and the planet. The shape of the orbit is described by the eccentricity e; for a circular orbit $e = 0$, but real orbits that remain bound to the planet take elliptical shapes with $0 < e < 1$, with most ring particles having $e \ll 1$. Unbound orbits, either parabolic or hyperbolic, have $e \geq 1$. The orbit plane may be inclined with respect to the reference plane (for ring applications, this is often the planet's equatorial plane), by an angle known as the inclination I. A nonzero inclination requires an account of the orbit plane's orientation, and thus its line of intersection with the reference plane (the "line of nodes") is described by the longitude of the ascending node Ω, measured with respect to a reference axis. Similarly, a nonzero eccentricity requires an account of the orientation of the ellipse within the orbit plane, and thus the line connecting the planet to the location of the particle's closest approach (its "periapse") is described by the argument of periapse ω. Finally, once the orbit has been defined by the five

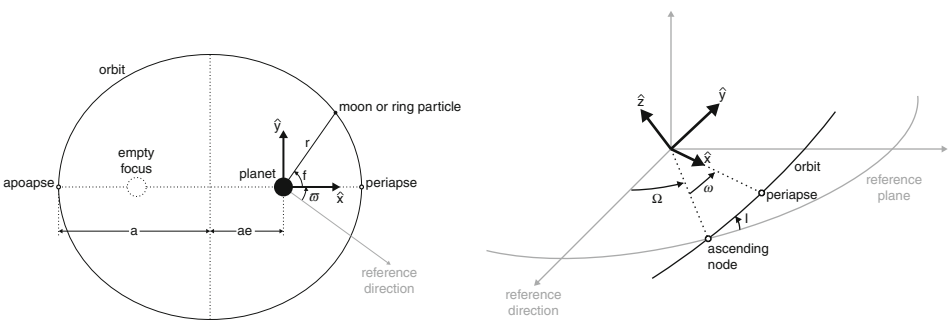


Fig. 7-1
The geometry of (*left*) an elliptical orbit within the orbit plane and (*right*) the orbit plane within 3-D space

parameters already mentioned, the particle's position along the orbit can be given by its actual position (the true anomaly f) or its time-averaged position (the mean anomaly M) relative to periapse. Also commonly used are the longitude of periapse $\varpi = \Omega + \omega$ and the mean longitude $\lambda = \Omega + \omega + M$, which are not physical angles since they are the sums of angles not necessarily in the same plane, but they have the virtue of being reckoned from a stationary reference axis rather than a moving line and are useful as long as I is not too large.

The osculating orbital elements, which are most simply calculated and most often used, assume that the planet's gravity field is that of a point mass. But for ring applications, the known planets are oblate (or bulged at the equators) due to their fast rotation. This is adequately described by adding to the account of the gravity field a positive quadrupole moment J_2 (for details see, e.g., Sect. 4.5 of Murray and Dermott 1999), though it may be necessary to further include higher moments for applications requiring great precision. The presence of a nonzero J_2, in addition to defining the Laplace plane[1] for orbits near the planet, causes orbits to precess, in the prograde direction for apses ($\dot{\varpi} > 0$) and in the retrograde direction for nodes ($\dot{\Omega} < 0$).

A nonzero J_2 also compromises the physical meaningfulness of the osculating elements, especially for low-eccentricity orbits, introducing fast (i.e., orbit-frequency) variations in all six elements. The physical meaningfulness of orbital elements can be restored using a revised system of *epicyclic orbital elements* (Borderies and Longaretti 1987; Longaretti and Borderies 1991; Borderies-Rappaport and Longaretti 1994), which are based on the geometrical shape of streamlines. These put the orbit-frequency variations back into an analogue of λ, leaving the other five elements to again describe a static (or at least slowly varying) orbit. A useful algorithm for converting Cartesian coordinates into epicyclic orbital elements was devised by Renner and Sicardy (2006).

1.2 Roche Limits, Roche Lobes, and Roche Critical Densities

The "Roche limit" is the distance from a planet within which its tides can pull apart a compact object. Simply speaking, a ring would be expected to reside inside its planet's Roche limit, while any disk of material beyond that distance would be expected to accrete into one or more moons. However, the Roche limit does not actually have a single value but depends particularly on the density and internal material strength of the moon that may or may not get pulled apart (Canup and Esposito 1995). A simple value for the Roche limit can be calculated from a balance between the tidal force (i.e., the difference between the planet's gravitational pull on one side of the moon and its pull on the other side) that would tend to pull a moon apart, and the moon's own gravity that would tend to hold it together. This works out to (e.g., Eq. 4.131 in Murray and Dermott 1999)

$$a_{\mathrm{Roche}} = R_{\mathrm{p}} \left(\frac{4\pi\rho_p}{\gamma\rho} \right)^{1/3}, \tag{7.1}$$

where R is radius and ρ is internal density, and the subscript "p" denotes the central planet.

The dimensionless geometrical parameter $\gamma = 4\pi/3 \approx 4.2$ for a sphere but is smaller for an object that takes a nonspherical shape with its long axis pointing toward Saturn, as one would expect for an actively accreting body and as at least several of Saturn's ring-moons appear to

[1] The Laplace plane is the plane about which orbits precess. When the vertical motions of objects are damped by mutual collisions, material will settle into a ring centered on the Laplace plane.

do (Porco et al. 2007; Charnoz et al. 2007). Simply distributing the moon's material into the shape of its Roche lobe, with uniform density, yields $\gamma \approx 1.6$ (Porco et al. 2007). However, fully accounting for the feedback between the moon's distorted shape and its (now non-point-mass) gravity field (Chandrasekhar 1969; see, e.g., Murray and Dermott 1999) leads to a rather smaller value, $\gamma \approx 0.85$; on the other hand, some central mass concentration and the failure of a rubble pile to exactly take its equilibrium shape will likely prevent γ from becoming quite this low.

Note that the moon's internal density ρ appears in (7.1). Thus, the Roche limit is variable; a denser object can venture closer to the planet without danger than can an object that is less dense. The moon's diameter, on the other hand, does not appear in (7.1). Why then do we commonly imagine a large object getting pulled into smaller pieces when it ventures inside the Roche limit? This is because the Roche limit has been defined here as the distance within which an object can no longer be held together *by its own gravity*. Size becomes relevant for objects small enough to be held together by their internal material strength (which is not considered in (7.1)) in spite of the tidal forces that enter into the Roche calculation.

In fact, it is often more useful in the context of rings to consider the limit from planetary tides not as a critical distance but as a critical density. At any given distance a from the planet, there is a Roche critical density ρ_{Roche} at which the moon's size entirely fills its region of gravitational dominance (its "Roche lobe" or "Hill sphere" of characteristic radius r_{Hill}). We can rearrange (7.1) to obtain

$$r_{\mathrm{Hill}} = a\left(\frac{m}{3M_{\mathrm{p}}}\right)^{1/3} \tag{7.2}$$

and

$$\rho_{\mathrm{Roche}} = \frac{4\pi\rho_p}{\gamma(a/R_{\mathrm{p}})^3} = \frac{3M_{\mathrm{p}}}{\gamma a^3}, \tag{7.3}$$

where m and M_{p} are the masses of the moon and planet, respectively. The first expression for ρ_{Roche} is the most useful for interpreting *Fig. 7-2*.

Within a ring, where material for accretion is plentiful, any preexisting solid chunk with internal density greater than ρ_{Roche} should accrete a mantle of porous ring material until its density decreases to match ρ_{Roche}. This process should govern the size and density of the largest disk-embedded objects (Porco et al. 2007; Charnoz et al. 2007). On the other hand, the density naturally achieved by transient clumps should be compared to ρ_{Roche} in order to predict whether disruption (rings) or accretion (discrete moons) will dominate in a particular location, and the persistent existence of a ring implies that the densities of transient clumps do not exceed ρ_{Roche} (i.e., we expect $\rho \lesssim \rho_{\mathrm{Roche}}$). As seen in *Fig. 7-2*, Saturn's rings extend outward to significantly lower values of ρ_{Roche}, approaching $0.4\,\mathrm{g\,cm^{-3}}$, than are seen in any of the other three known ring systems, probably reflecting their much lower rock fraction (and higher fraction of water ice) as already known from spectroscopy and photometry (Cuzzi et al. 2009). That ρ_{Roche} for Saturn's rings reaches values much lower even than the density of solid water ice indicates a high degree of porosity, which is not surprising for a system in balance between disruption and accretion.

If the outer edge of a ring system is taken to be the transition between disruption-dominated and accretion-dominated regions, which is probably true at least for Saturn and Uranus given the large number of moons immediately outward of their main rings, and if the porosity of accreting objects is relatively constant among the different systems, then differences in ρ_{Roche} at the transition location probably reflect differences in bulk composition. Since Uranus has a transitional ρ_{Roche} three times that of Saturn (*Fig. 7-2*), we may well infer that its rings are

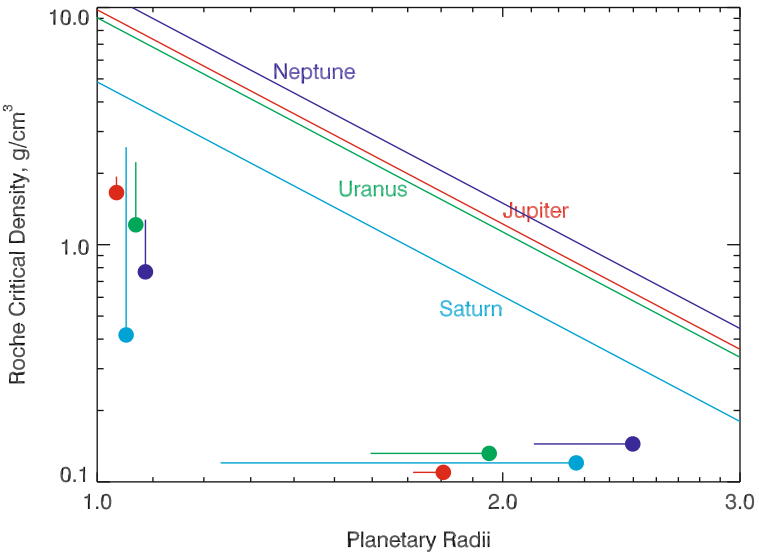


Fig. 7-2

Roche critical density ρ_{Roche} ((7.3), with $\gamma = 1.6$) plotted against planetary radii for Jupiter (*red*), Saturn (*cyan*), Uranus (*green*), and Neptune (*blue*). An object must have density higher than ρ_{Roche} to be held together by its own gravity; conversely, in the presence of abundant disk material, an embedded object will actively accrete as long as its density remains higher than ρ_{Roche}. The colored bars along the bottom and along the left-hand side show the extent of each planet's main ring system. For each, a solid circle indicates the outermost extent, and the corresponding minimum ρ_{Roche}, of the main rings

made of material with a higher grain density, i.e., a significantly higher rock/ice ratio. Neptune's transitional ρ_{Roche} is intermediate between Saturn's and Uranus', possibly indicating an intermediate rock/ice ratio. Our inference, from the Roche critical density at the ring/moon transition, that the Uranus system is rockier overall than the Saturn system is consistent with the fact that the average density of Saturn's mid-size moons (Matson et al. 2009) is 1.2 g cm^{-3}, while that of Uranus' major moons (Jacobson et al. 1992) is 1.6 g cm^{-3}. We cannot test our inference that Neptune's rock/ice ratio is intermediate in this way, as Neptune has no indigenous major moons due to the cataclysm of Triton's capture (Goldreich et al. 1989).

The extent of Jupiter's Main ring, in contrast to the other three ring systems, is clearly limited by the availability of material (which originates at source moons Metis and Adrastea and evolves inward, and which is not abundant) rather than by a disruption/accretion balance. However, its high value of $\rho_{\text{Roche}} \sim 1.7$ g cm^{-3} places the only known limit on the densities (and thus masses) of the source moons. However, it may not be valid to assume that Metis and Adrastea are held together by gravity, as accreting masses must be, given the large gap in particle size between the ~10-km moons and other Main ring particles, which observationally cannot be larger than 1 km (Showalter et al. 2007). This large gap in particle size might be explained if Metis and Adrastea are solid bodies originating further out, now held together by material strength, while no bodies of similar size are now able to form through in situ accretion.

1.3 Optical Depth

The amount of material in a system with general disk morphology can be measured in several ways. The most straightforward is the surface density, the mass per unit surface area of the disk, though it must be borne in mind that a disk with greater vertical thickness will have proportionately lower volume density than a vertically thin disk, even if both have the same surface density. However, surface density can be difficult to measure directly. A much more common observable is the optical depth τ, which can be thought of as the attenuation of a beam of light passing through the disk, measured in e-folding terms. That is,

$$\tau = -\ln\left(\frac{I}{I_*}\right) \equiv -\ln T, \tag{7.4}$$

where I is the observed intensity, I_* is the unocculted intensity (e.g., from a background star), and the ratio between them is defined as the ring's transparency T.

The optical depth is sensitive to both the number density and the size of ring particles, which can be obtained when both the optical depth and the surface density are known (e.g., Colwell et al. 2009a). For a given value of the surface mass density, the optical depth scales inversely with particle size, that is, with the ratio of volume to surface area.

For a vertically thick homogeneous disk, the optical depth is proportional to the path length through the disk, which in turn is proportional to $\mu \equiv \sin B$, where the elevation angle B is the angle between the line-of-sight and the ring plane (❯ *Fig. 7-3*a). In order to compare observations taken over a range of elevation angles, the normal optical depth, $\tau_\perp \equiv \mu\tau$, is often used. However, this parameter must be used with caution for disks that lack homogeneity and/or are close to a single layer thick. Colwell et al. (2007) found that $\tau_\perp$ varies strongly with μ in the B ring and that the uncorrected τ is a more robust parameter in that case, indicating that the B ring is composed of vertically thin nearly opaque clumps with nearly transparent gaps between them (❯ *Fig. 7-3*b), and that the optical depth is controlled by the relative abundance of clumps and gaps (see ❯ Sect. 3.1.4).

In numerical simulations (see ❯ Sect. 4.1), the photometric optical depth τ (❯ 7.4) is cumbersome to calculate, but a useful proxy known as the dynamical optical depth $\tau_{\rm dyn}$ can be found by summing the total cross-section area of all simulated particles and dividing by the area of the simulation patch. This quantity turns out to be equal to the photometric optical depth as long as particles are randomly distributed, as the Gaussian probability of particle overlap plays the

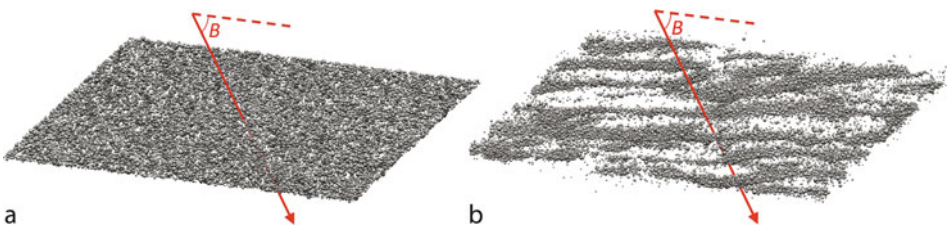


Fig. 7-3
Schematic showing a light ray path, slanted from the horizontal by an angle *B*, passing through (*left*) a homogeneous ring and (*right*) flattened self-gravity wakes (SGWs). The measured optical depth τ is proportional to sin *B* in the first case, but is relatively insensitive to elevation angle in the second. Figure created with the `rebound` software package (Rein and Liu 2012)

same role when using τ_{dyn} to calculate the total transparency that the exponential plays when using (7.4). However, for high values of τ, when the distance between particles becomes comparable to the particle size, particles become constrained as to the locations in space they can occupy and the two quantities diverge. Specifically,

$$\tau/\tau_{dyn} \simeq 1 + kD, \tag{7.5}$$

where the volume filling factor D is calculated from particle radius, disk scale height, and τ_{dyn}, and k is a scalar of order unity (Salo and Karjalainen 2003; Tiscareno et al. 2010a). Furthermore, the existence of microstructure such as self-gravity wakes causes the distribution of particles to be strongly nonrandom, and can cause τ_{dyn} to diverge strongly from the photometrically observed τ.

2 Rings by Planetary System

2.1 The Rings of Jupiter

Jupiter is adorned by the simplest of the known ring systems. All of its rings are tenuous and composed of dust-sized[2] particles. As the only confirmed ring system without any dense component, and by far the least massive (Burns et al. 2004), Jupiter's is the only ring system to have been discovered by spacecraft without having previously been seen from Earth either directly (as Saturn's) or by stellar occultations (as Uranus' and Neptune's). The Main ring was first clearly described from *Voyager 1* images (Owen et al. 1979) after initial hints from charged-particle detectors aboard *Pioneer 11* (Fillius et al. 1975; Acuna and Ness 1976; Burns et al. 2004).

The basic structure of Jupiter's rings is well understood (see Burns et al. 2004, for a recent comprehensive review). The Main ring and the two Gossamer rings are like three nested "tuna cans" (*Fig. 7-4*), with radius set by the semimajor axis (a) of the ring's source moon and vertical height by the moon's vertical excursions relative to Jupiter's equatorial plane ($a \sin I$, for

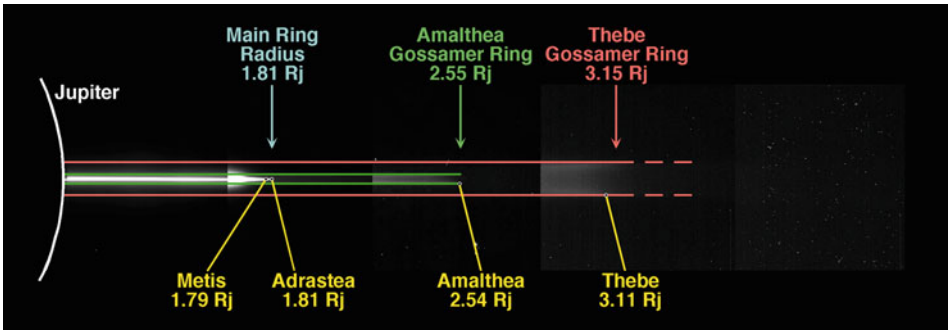


Fig. 7-4

***Galileo* image mosaic of Jupiter's rings, seen nearly edge-on at very high phase angle, annotated to show the primary components of the ring system. Image sensitivity increases from left to right, in order to show the increasingly faint structure (Figure from Ockert-Bell et al. 1999)**

[2]Throughout this work, we will use the word "dust" to refer to μm-sized particles regardless of their composition.

inclination I). Particles enter the ring as ejecta from micrometeoroid impacts onto the moon (Burns et al. 1999) and begin with orbital parameters a, e, and I (see Sect. 1.1) similar to the moon's. The tuna-can structure arises as the orientations of particles' orbit planes (Ω) become quickly randomized due to small variations in a and thus in the precession rate. Particles evolve inward under Poynting-Robertson drag (Burns et al. 1999), thus filling out the cylindrical shape. A double-layered vertical structure arises dynamically because any orbiting particle spends more time at its vertical maxima than it does in the midplane. The increasing vertical thickness of the rings arises because Amalthea's inclination is larger than that of Metis or Adrastea, while Thebe in turn has an even larger inclination.

The two Gossamer rings consist entirely of material evolving inward in this way, though the Thebe Gossamer ring has an additional segment extending slightly *outward* from Thebe's orbit, which has been attributed to charging and discharging of grains as they pass into and out of the planet's shadow (Hamilton and Krüger 2008). Further inward, the ~1,000-km-wide core of the Main ring contains a significant population of cm-size and larger particles lying between the orbits of Adrastea and Metis. This core, which is the only component of the ring system not composed of dust and the only one that appears bright at low phase angles,[3] is composed of several ringlets (including one lying just outside Adrastea's orbit), whose cause is not known. *New Horizons* imaging limits the sizes of objects in this belt to be <1 km (Showalter et al. 2007) – other than Adrastea and Metis themselves, which have mean radii of 8 and 22 km, respectively. However, Showalter et al. (2007) did find several azimuthal clumps in a ringlet just inward of Adrastea (see Sect. 3.5). A dusty component of the Main ring extends inward of its core, also evolving under Poynting-Robertson drag (*Fig. 7-5*).

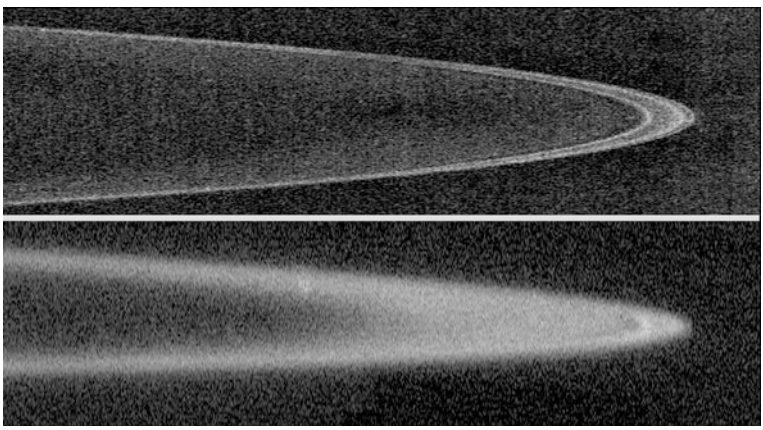


Fig. 7-5
***New Horizons* images of Jupiter's Main ring at low phase (*upper panel*) and at high phase (*lower panel*), respectively showing the structures composed of macroscopic particles and the dusty envelope (Credit: NASA)**

[3] The phase angle is formed by the Sun-object-observer lightpath. Dust-sized particles, having size comparable to the wavelength of visible light, tend to diffract light forward and are brightest at high phase angles. Larger objects tend to reflect light and are brightest at low phase angles.

The dusty component of the Main ring has a vertical scale height that increases monotonically with decreasing distance to Jupiter (Ockert-Bell et al. 1999). When the inward-moving material reaches a radius of 122,800 km from Jupiter's center, it becomes strongly affected by a 3:2 "Lorentz resonance" (Burns et al. 1985) between its orbital period and the rotation period of Jupiter's magnetic field. Inward of this location, the vertical extent of the ring increases dramatically, forming the toroidal Halo ring. Material in the Halo ranges tens of thousands of kilometers above Jupiter's ring plane, though most of its material is concentrated within just a few hundred kilometers (Burns et al. 2004).

2.2 The Rings of Saturn

Saturn possesses by far the most massive and the most diverse of the known ring systems (*Fig. 7-6*). The only ring system known before recent decades, Saturn's rings were among the first objects observed through a telescope, by Galileo Galilei in 1610, explicated as a disk by Christiaan Huygens, proved to consist of individual particles on independent orbits by James Clerk Maxwell, and have in general been the focus of much productive study by astronomers over the past four centuries (Alexander 1962; Van Helden 1984; Miner et al. 2007). The main part of the rings comprises the solar system's only known broad and dense disk (Sect. 3.1), which was found by G. D. Cassini to be divided into two parts – now called the A and B rings, separated by what is now called the Cassini Division. Furthermore, the latter is now known to

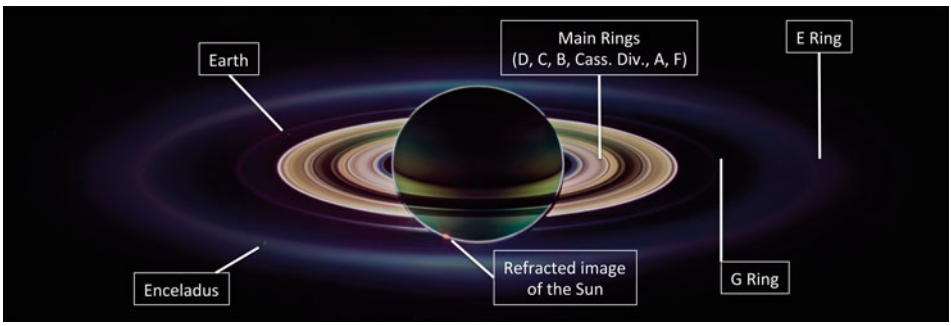


Fig. 7-6

This *Cassini* image mosaic shows Saturn's tenuous D, E, and G rings with comparable brightness to the main disk, which occurs because the viewing geometry is at high phase angle (in fact, in eclipse) and also views the unlit face of the main disk. The *darkest part* of the main disk is actually the densest and most opaque, namely, the mid- to outer-B ring. The Cassini Division is difficult to distinguish from the A ring in this view. The markings on the planet do not line up with those on the rings because the latter are due to sunlight filtering directly through the rings while the former are due to light reflected off the rings, then reflected off the planet, and then filtered through the rings again. The Sun, which is actually behind Saturn, can be seen refracted through the planet's atmosphere at 7 o'clock. Enceladus (actually, only its geyser plume is bright in this geometry) can be seen embedded in the E ring at 8 o'clock. The Earth can be seen as a pinpoint of light between the F and G rings at 10 o'clock (Credit: NASA and M. Hedman, annotated by the author)

be not an empty gap but simply a region of the disk with more moderate surface density, similar in character to the C ring, which lies inward of the B ring and was discovered in 1850.

There are a small number of truly empty radial gaps in the dense disk of the main rings, most of them in the C ring and Cassini Division but two in the outer A ring, all of them sharp-edged. These gaps are named for scientists who have made contributions to the study of Saturn's rings. The two A-ring gaps are held open by moons at their centers, and the C ring's Colombo Gap is known to be held open by a resonance with Titan, but most of the gaps remain unexplained. Recent work by Hedman et al. (2010a) suggested that all the Cassini Division gaps are due to a secondary resonance associated with the Mimas 2:1 resonant mechanism that defines the nearby outer edge of the B ring, though this idea has yet to be successfully worked out in detail (Spitale and Porco 2010). A diverse array of narrow rings and ringlets resides within ring gaps, some of them dense and sharp-edged and others diffuse and/or dusty. They often are given the same name as the gap within which they reside, though several have been given nicknames. These structures, which can be compared with narrow rings around other planets, are discussed in Sect. 3.2.

The Saturn system also contains the most diverse retinue of tenuous dusty ring structures known in the solar system, discussed in Sect. 3.4. The main components, given letters in order of their discovery, are the D ring situated innermost between the main rings and Saturn's atmosphere, the dense F ring (Sect. 3.3) just off the edge of the A ring, and the G and E rings farther out. The region between the A and F rings, now known as the Roche Division,[4] contains a tenuous dusty sheet, and several other rings or ring arcs are named for moons whose orbits they share.

The largest known ring in the solar system, the Phoebe ring, was recently discovered by the Spitzer Space Telescope (Verbiscer et al. 2009). This ring is also the only known ring to be tilted from its planet's equatorial plane (it lies in the plane of Saturn's orbit, as solar perturbations are much more important than Saturn's J_2 at its distance) and is likely the only known ring whose particles orbit in the retrograde direction (if indeed its material is primarily derived from Phoebe, which orbits retrograde). As particles in the Phoebe ring spiral inward under Poynting-Robertson drag, they preferentially impact the leading hemisphere of Iapetus (Soter 1974; Tamayo et al. 2011), which, together with solar-driven thermal processing, appears likely to explain the strong brightness dichotomy on the surface of that moon (Spencer and Denk 2010; Denk et al. 2010).

The end of *Cassini*'s initial 4-year mission at Saturn (though its extended mission continues) has occasioned several recent reviews of Saturn's rings, including articles by Cuzzi et al. (2010) and Esposito (2010). A recent comprehensive review in five parts discussed the rings' structure (Colwell et al. 2009b), dynamics (Schmidt et al. 2009), particle sizes and composition (Cuzzi et al. 2009), diffuse rings (Horányi et al. 2009), and origins (Charnoz et al. 2009a), in addition to a review of pre-*Cassini* understanding (Orton et al. 2009).

2.3 The Rings of Uranus

All of Uranus' main rings are narrow, and many are eccentric and/or inclined, unlike the broad disk of Saturn. On the other hand, many of Uranus' rings are dense and sharp-edged, unlike

[4] As recently formalized by the IAU, a "division" is defined as a region between two lettered rings that contains a sheet of material, while a "gap" is a clear region within a lettered ring that may or may not contain one or more ringlets (http://planetarynames.wr.usgs.gov/append8.html).

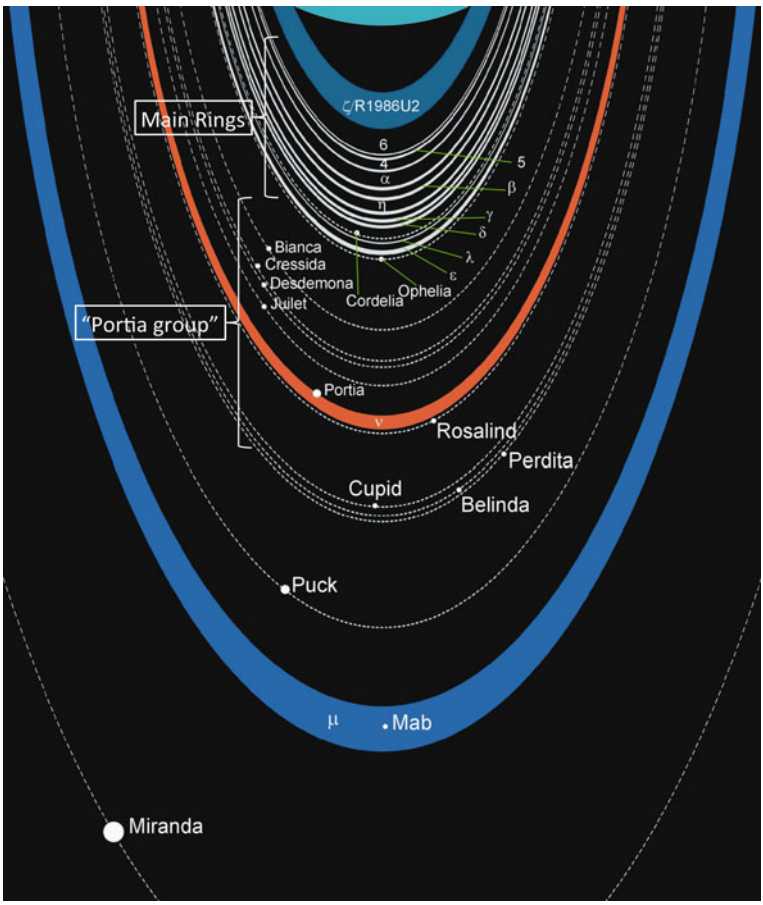


Fig. 7-7

Uranus' main rings are situated immediately inward of a retinue of small moons. If one were to spread the mass of Uranus' "Portia group" of moons evenly over the annulus they occupy, the surface density would be similar to that of Saturn's A ring. This moon system may be very similar in origin to the known ring systems, except that the natural density of accreted objects is larger than the Roche critical density (i.e., it is beyond the "Roche limit," see *Fig. 7-2*) so that any moon that gets disrupted by a collision (which ought to have happened many times over the age of the solar system) will simply reaccrete (Credit: Wikimedia Foundation, annotated by the author)

the diffuse rings of Jupiter. Thus, the Uranian system represents a third paradigm for ring systems, one that truly deserves the label of "rings" in the plural. The main set of ten narrow rings (including all the named rings except dusty ζ, ν, and μ) occupies a fairly small radial range from 1.64 to 2.00 R_U from Uranus' center (*Fig. 7-7*), inward of Uranus' 13 small inner moons except that the innermost moon Cordelia is inward of the ϵ ring. A panoply of unnamed dusty rings was seen interspersed with the main rings in the single high-resolution high-phase image taken by *Voyager 2* (*Fig. 7-8*), and gaps in these have been cited as evidence for additional moons (Murray and Thompson 1990, 1991). The so-called "Portia group" of eight moons packed into an annulus from 59,100 to 76,500 km from Uranus' center (i.e., 2.31–2.99 R_U) appears from

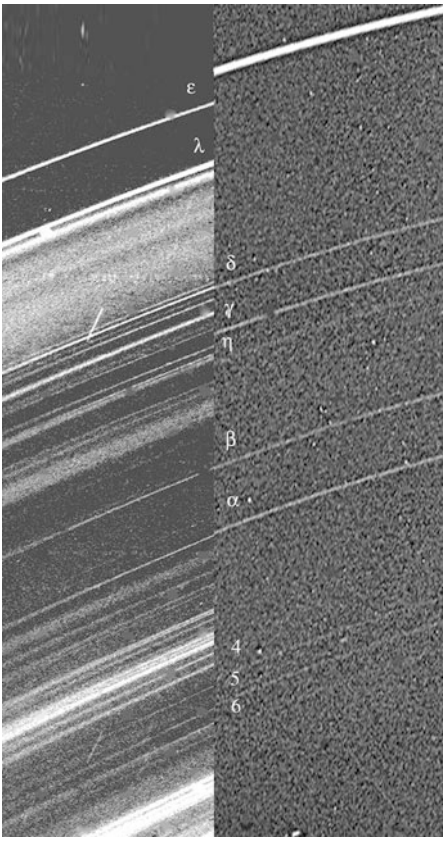


Fig. 7-8

This composite image of Uranus' main rings in forward-scattered (*left*) and back-scattered (*right*) light shows that a network of dust structures is interleaved with the planet's dense main rings. The disjoint in the ϵ ring is due to its eccentricity. As the left-hand image is the only high-phase image ever successfully taken of Uranus' rings (by the post-encounter *Voyager 2*), the detailed workings of the dust structures remain largely unknown (Credit: NASA and Wikimedia Foundation)

orbital simulations to be dynamically unstable on timescales of 10^6–10^8 years (Duncan and Lissauer 1997; Showalter and Lissauer 2006). The dusty ν ring, which lies between two of the moons in this group, may well be the detritus of a recent significant collision, perhaps the disruption of a moon. Outward beyond the Portia group, the μ ring is centered on the orbit of Mab (Showalter and Lissauer 2006; see ❯ Sect. 3.4).

Given their dynamical instability, the Portia group of Uranian moons are probably constantly evolving by means of occasional collisions followed by the reaccretion of material into new moons (Showalter and Lissauer 2006; Dawson et al. 2010; French and Showalter 2011) and may have looked rather different from its present configuration over most of solar system history, though what effect this might have on the main rings is unknown. The contrast between the main Uranian rings and the Portia group may simply be the difference between a particle population dominated by disruption and one dominated by accretion. The mean surface density of

the Portia group region is ~45 g cm^{-2}, calculated by spreading the moons' mass evenly over the annulus containing them. This is comparable to the surface density of dense rings such as Saturn's A ring, though direct comparisons to the Uranian rings are difficult as the masses of the latter are poorly known. The Roche critical density (see Sect. 1.2) for the boundary between the two regions is 1.2 g cm^{-3}, which possibly indicates a high rock fraction, especially considering the high internal porosity inherent in accretion of small bodies with low central pressures.

The composition of the Uranian rings is almost entirely unknown, as *Voyager* did not carry an infrared spectrometer with enough spatial resolution to detect the rings. However, it is clear from their low albedo that at least the surfaces of the ring particles cannot be primarily water ice. Color imaging indicates that the Uranian rings are dark at all visible wavelengths, indicating a spectrum similar to that of carbon.

Most of Uranus' rings have been given Greek letters (α, β, γ, δ, ϵ, η, λ, ζ, μ, and ν) in order of their discovery, except for three which are labeled with numbers (4, 5, and 6). This idiosyncratic system can be traced back to their simultaneous discovery by two different research groups, one of which (Elliot et al. 1977) labeled the rings with Greek letters while the other (Millis et al. 1977) numbered them. The former system was given priority for future use, but three of the rings had not been observed by the former research group and thus retained as their labels the numbers given to them by their discoverers (Miner et al. 2007).

A comprehensive review of the Uranian rings, up to and including the *Voyager 2* encounter, was published in two parts discussing the rings' structure (French et al. 1991) and particle properties (Esposito et al. 1991).

2.4 The Rings of Neptune

Neptune's ring system, like that of Uranus, consists primarily of a few narrow rings, though Neptune's are generally less dense, higher in dust, less sharp in their edges, and farther from their planet than those of Uranus. Neptune's rings are named for individuals associated with the 1846 discovery of Neptune. The Le Verrier, Arago, and Adams rings are narrow, while the Galle and Lassell rings are tenuous sheets of dust (*Fig. 7-9*). The Adams ring, Neptune's most substantial, is best known for its series of arcs, the first ever discovered, which are discussed in Sect. 3.5.

Unlike those of Saturn and Uranus, Neptune's ring system is thoroughly interwoven with known moons. This is at least partly enabled by the fact that Neptune's rings are the farthest from their planet, in terms of planetary radii, and have the lowest value of ρ_{Roche} at the *inner* edge of the ring system (*Fig. 7-2*). Still, the presence of rings, rather than accreted moons, must indicate that the natural density of accreted objects is lower than ρ_{Roche}. Thus, it may be that the ring-moons Naiad, Thalassa, Despina, and perhaps Galatea, either originated farther from the planet than their current position, or accreted in a formerly more dense ring.

The composition of the Neptunian rings, like that of the Uranian rings, is unknown due to *Voyager*'s inability to detect them in the infrared. However, again like the Uranian rings, the low albedo of Neptunian ring particles makes it clear that at least their surfaces cannot be primarily water ice.

A comprehensive review of the Neptunian rings, up to and including the *Voyager 2* encounter, was given by Porco et al. (1995).

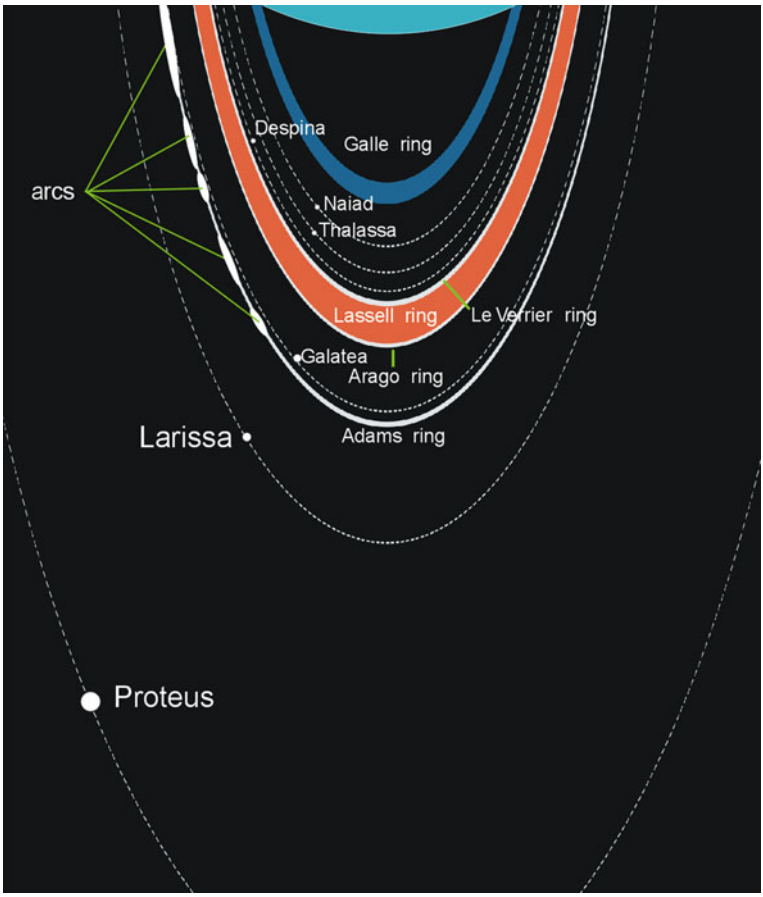


Fig. 7-9
In Neptune's ring system, uniquely, narrow rings and diffuse rings and moons are all interspersed together (Credit: Wikimedia Foundation)

2.5 Unconfirmed Ring Systems

The four giant planets are the only bodies known to have rings, as just described. Here we discuss bodies for which rings have been seriously discussed but not observed.

2.5.1 Mars

Mars has been predicted to have a tenuous ring system comprising dust grains ejected from its moons Phobos and Deimos by meteoroid impactors (Hamilton 1996; Krivov and Hamilton 1997, and references therein). Simulations by Burns et al. (2001) indicate that Deimos' ring should be offset away from the Sun and tilted out of Mars' equatorial plane by the Sun's perturbations. However, attempts to observe rings around Mars have been unsuccessful to

date (Showalter et al. 2006), and the image quality has progressed to the point that some models can now be observationally excluded. The lack of dust could be due to dust production rates being lower than expected, or the lifetimes of dust particles being shorter than expected. Solar radiation pressure limits the lifetimes of dust particles (especially smaller ones) by driving their orbital eccentricity to values so high that they impact the planet. In situ observations by Mars-orbiting spacecraft of anomalies in the solar wind magnetic field were interpreted in the 1980s as being due to Martian rings, but more extensive measurements by the magnetometer aboard *Mars Global Surveyor* showed that observable fluctuations are likely due to well-known solar wind or bow shock phenomena (Øieroset et al. 2010).

Looking for rings around a solid planet like Mars at low phase is more difficult than similar observations at gas giant planets because of the former's lack of atmospheric methane. Because methane has very strong absorption bands (e.g., at 2.2 μm), images of gas giants taken at selected wavelengths will see a greatly darkened planet, facilitating the detection of faint rings. On the other hand, looking for Martian rings at high phase is less effective than for other dusty rings because particles smaller than ~50 μm are expected to be depleted due to radiation pressure (Hamilton 1996; Showalter et al. 2006).

2.5.2 Pluto

Pluto, like Mars, could harbor a tenuous ring system of dust derived from its small moons Nix and Hydra and P4 (Stern et al. 2006). Charon, as discussed in Sect. 3.4, is paradoxically less likely to be a major source of dusty rings because its gravitational field will more efficiently retain any dust ejected from its surface. Observations to date (Steffl and Stern 2007) are not sensitive enough even to rule out the conservatively estimated normal optical depth $\tau_{\perp} < 10^{-6}$ suggested by Stern et al. (2006), though more sensitive observations were very recently obtained (Showalter et al. 2011a).

Further Earth-based observations may improve on this sensitivity, and a clearer picture of Pluto's rings (or lack thereof) should come from imaging during the planned *New Horizons* flyby in 2015, especially during the post-encounter period when the phase angle will be high. However, both of the handicaps discussed above for Mars, the lack of atmospheric methane and the loss of smaller dust grains due to radiation pressure, are also likely to hamper the detection of any Plutonian rings.

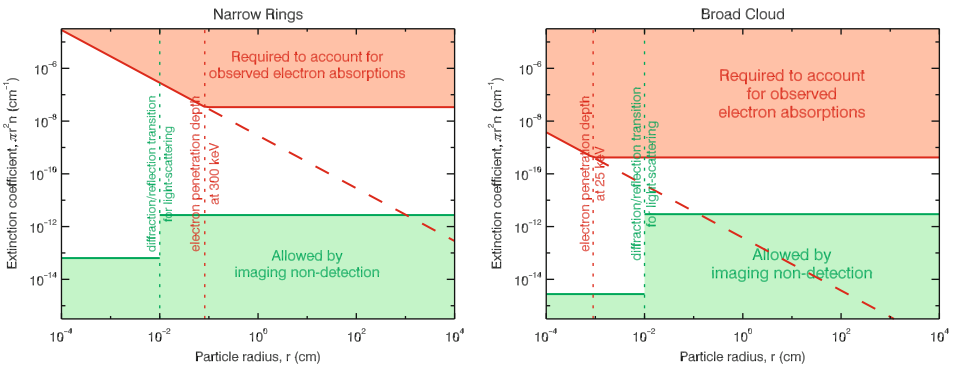
2.5.3 Rhea and Other Moons

Rhea, the largest of Saturn's airless moons, is in the opposite situation from Mars and Pluto, with no clear theoretical prediction but a claim that rings have been observed. Jones et al. (2008) reported unusual electron-absorption signatures detected by the Magnetospheric Imaging Instrument (MIMI) during Rhea flybys of the *Cassini* spacecraft and attributed these signatures to rings. What would be the first known ring system around a moon includes several narrow bands embedded within a broad diffuse cloud. The MIMI instrument detects changes in the miasma of charged particles through which the spacecraft is constantly passing, enabling inferences regarding solid and magnetospheric structures that shape the plasma environment.

A good analogy is the way a person driving in a blinding rain can perceive having driven under a bridge by the sudden cessation of raindrops hitting the windshield.[5]

However, an extensive search for Rhea rings using *Cassini* ISS images (Tiscareno et al. 2010b) places severe limits on any possible Rhea rings. Using the calculations of Jones et al. (2008), which assume that a particle's ability to block electrons increases linearly with its mass (and thus its volume), the ISS observations limit the narrow rings to a characteristic particle size $r > 10$ m, and μm-sized dust in any significant abundance is unequivocally excluded. This minimum particle size is unrealistically large given that such large particles must constantly be eroded to smaller sizes, which then would have been detected in the ISS images. Furthermore, Tiscareno et al. (2010b) pointed out that the electron penetration depth for the low-energy electrons used by Jones et al. (2008) is much smaller than the suggested particle sizes, so that a particle's ability to block electrons should increase linearly with its surface area rather than its volume. Thus, Tiscareno et al. (2010b) recalculate the ring optical depth required to explain the Jones et al. (2008) observations to be several orders of magnitude higher than the values excluded by the ISS observations, both for the narrow rings and the broad cloud (◉ *Fig. 7-10*). It seems most likely, therefore, that the signatures detected by the *Cassini* MIMI instrument are due to some magnetospheric phenomenon, and not to rings of solid material around Rhea.

Following the MIMI announcement that Rhea might have rings, Schenk and McKinnon (2009) reported a clumpy and discontinuous chain of discrete bluish splotches, up to 10 km



◉ Fig. 7-10

Comparison of the radius *r* and number density *n* of particles, the latter expressed in terms of the extinction coefficient $\pi r^2 n$ of putative Rhea rings, inferred from charged-particle absorptions observed by *Cassini* MIMI (Jones et al. 2008) and imaging non-detection by *Cassini* ISS (Tiscareno et al. 2010b), shown in *red* and *green*, respectively. *Dashed lines* indicate requirements previously claimed (Jones et al. 2008) for particle sizes larger than the electron penetration depth. For narrow rings (*left*), even allowing the latter claim, the combined observations require particles larger than 10 m in radius, indicating an unrealistic lack of smaller particles. Furthermore, once allowance is made for the role of the electron absorption length (horizontal lower boundary to red area), the imaging non-detection rules out any form of absorption by solid material as the cause of the observed charged-particle absorptions for both narrow rings and broad cloud (Figure from Tiscareno et al. 2010b)

[5] Credit: G. H. Jones in JPL podcast, 6 March 2008 (http://www.jpl.nasa.gov/podcast/content.cfm?content=671)

wide, aligned along more than half the circumference of a great circle inclined by ~2° to Rhea's equator. Schenk et al. (2011), formally presenting that finding after publication of the ISS non-detection, argued that the band may be a sign of a *former* Rhean ring, even if none exists now. However, it remains unclear whether a ring system is even plausible at Rhea (whose shape is triaxial rather than oblate), what would be the source of its material, whether a tenuous ring would assume the flatness (generally produced only in dense rings by collisional evolution, while tenuous rings generally are vertically thick) implied by the observed narrow band, and whether the observed color variations are the likely result of such impacting material.

Among other Saturnian moons, Mimas and Tethys also have equatorial color features extending over ±40° and ±20° of latitude, respectively, much wider than Rhea's and possibly the result of magnetospheric bombardment (Schenk et al. 2011). Iapetus has a 13-km-high equatorial mountain range extending across >110° of longitude (Porco et al. 2005a; Giese et al. 2008). Several endogenic hypotheses have been suggested for this structure, but Ip (2006) argued that it was created by ring material falling out of Iapetan orbit, and Levison et al. (2011) further developed the idea by proposing a giant impact to create the ring and sketching a scenario that would simultaneously account for Iapetus' surprisingly oblate shape. While Iapetus may be a more hospitable site for rings than Rhea, given its oblate shape and large Hill sphere, it remains unclear just how or whether orbiting material would be incorporated into a mountain range, nor how such a ring would evolve and fall out.

2.5.4 Exoplanets

While hundreds of planets have now been detected in orbit about other stars using methods including radial velocity, transits, astrometry, and direct imaging (Seager 2010), no exoplanet yet has a confirmed ring system. The signature in a transit observation expected from an exoplanet ring system was discussed by Barnes and Fortney (2004), and a few exoplanet detections have been able to set meaningful upper limits on putative ring systems (e.g., Brown et al. 2001), but it is too early to do statistics on the frequency of exoplanet rings, especially since the available observations with sufficient sensitivity are generally for "hot Jupiters" that orbit very close to their stars. There are many factors stacked against the detection of rings around hot Jupiters (Schlichting and Chang 2011), including small Hill spheres, low obliquities[6] that would cause rings to be seen edge-on (see also Ohta et al. 2009), loss of ring particles to Poynting-Robertson drag and viscous drag from the planet's exosphere (see also Gaudi et al. 2003), and equilibrium blackbody temperatures too high for even refractory materials to remain solid. However, each of these factors is mitigated for planets $\gtrsim$0.1 AU from their host stars (Schlichting and Chang 2011). Furthermore, warping of the ring planes (Laplace planes) of "warm Saturns," diagnostic of their planetary oblateness (see Sect. 1.1) may be discernible in transit lightcurves (Schlichting and Chang 2011).

The only known exoplanet for which a ring system has been specifically proposed is Fomalhaut b, which was the first[7] exoplanet to be detected by direct imaging (Kalas et al. 2008). The orbit of Fomalhaut b is ~115 AU from its star and maintains the inner edge of an eccentric

[6] Here, we refer to the inclination of the planet's equatorial plane with respect to the line of sight from Earth.
[7] along with the three planets of the HR 8799 system, announced at the same time

debris belt that was known before the planet was (see ❯ Sect. 6). The brightness of Fomalhaut b in visible light, together with its dimness in the infrared, has led Kalas et al. (2008) to suggest the planet has a large ring system, which would significantly increase its visible brightness via reflection without affecting its emitted infrared radiation. However, a ring with a surface brightness similar to that of Saturn's A ring while extending to the planet's Roche radius would reflect far too little flux to account for the observations, so it is necessary to invoke a disk extending to ~35 planetary radii, approximately the distance from Jupiter to its outermost large moon Callisto. Such a large disk would not be stable against accretion (❯ Sect. 1.2) and thus would perhaps be more of a dynamically evolving protosatellite disk than a stable ring system.

A complex 2-month-long eclipse observed for the star 1SWASP J140747.93-394542.6 may have been due to a disk surrounding an otherwise-unknown planet, though it is also possible that the occulting disk instead adorns a low-mass stellar companion (Mamajek et al. 2012).

3 Rings by Type

3.1 Dense Broad Disks

Saturn:	C ring
	B ring
	Cassini Division
	A ring

Although the idea that Saturn's main ring system is composed of "countless tiny ringlets" continues to appear even in the professional literature, it is inaccurate. It is much better to say that the main ring system is a nearly continuous disk with density that varies radially but only a few true gaps that would separate one "ring" from another. Not only do waves travel radially through the disk (❯ Sect. 3.1.1), but material likely does as well through ballistic transport (Durisen et al. 1989, 1992) and direct migration (Tiscareno et al. 2010c).

Although Saturn's ring system, taken as a whole, is the only dense broad disk known to us, its four main components are quite different from each other and thus still offer room for comparative study. In addition to their wide differences in structure, the components vary in location and density. The B ring is by far the most dense, with measured τ of 5 or more, followed by the A ring at $\tau \sim 0.5$, while the C ring and Cassini Division both have $\tau \sim 0.1$ (❯ *Fig. 7-11*). The B ring's position astride the synchronous distance, at which the orbital period of particles equals the rotation period of the planet (and thus of its magnetic field), is thought to influence at least some of its properties (e.g., spokes), while the A ring's outermost position causes it to be more susceptible to accretion-related processes. The differences between the C ring and the Cassini Division, despite their apparent similarities in optical depth and composition, may also be primarily due to their different locations with respect to Saturn, its moons, and the B ring.

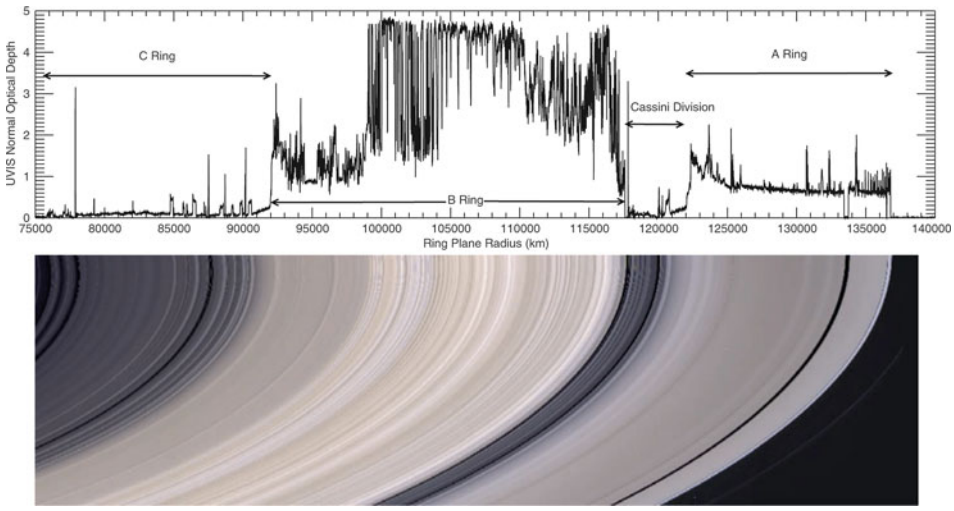


Fig. 7-11
An optical depth profile (*top*) and true-color image (*bottom*) of Saturn's main ring system (Figure from Colwell et al. 2009b)

3.1.1 Spiral Waves

Spiral density waves and spiral bending waves are the most widespread well-understood phenomena in dense rings. First described for the case of galaxies by Lin and Shu (1964), and applied to planetary rings by Goldreich and Tremaine (1978a, b, 1980), they can arise in any disk subject to a periodic perturbation. In Saturn's rings, the predominant mechanism is forcing from a perturbing moon. At discrete locations in the disk, the orbits of individual particles are resonant with the forcing and become excited. For a resonance with a moon, this happens when the resonance argument φ, the quantity that librates about a constant value at the resonant location, is of the form[8]

$$\varphi = (m + k)\lambda' - (m - 1)\lambda - X - kX', \tag{7.6}$$

where m and k are integers, primed quantities refer to the forcing moon, and X is some integer combination of ϖ or Ω (see Sect. 1.1 for orbital elements). For any given value[9] of m and k, and for any particular forcing moon, there is a unique location in the ring plane at which the resonance occurs; this location can be found by differentiating (7.6) and setting the time-derivative $\dot{\varphi} = 0$, since n (= $\dot{\lambda}$), $\dot{\varpi}$, and $\dot{\Omega}$ are all known functions of radial location in the ring plane (a), and the orbital frequencies of the moon are known. In fact, since n and n' are much larger than the others, the approximate location of the resonance can be found from them alone, and the resonance is generally labeled with the ratio $(m+k):(m-1)$. At a Lindblad

[8] For brevity, this discussion is limited to inner Lindblad resonances and inner vertical resonances, where the disk is inward of the forcing moon. Nearly all known spiral waves in rings are of this kind, though Tiscareno et al. (2007) detected inwardly propagating spiral density waves excited by outer Lindblad resonances with Pan.

[9] The azimuthal parameter m gives the number of spiral arms in the resonant wave pattern, while $k + 1$ is the "order" of the resonance, with first-order resonances generally being strongest, followed by second-order, etc.

resonance (LR), where the identity of X (though not necessarily X') is ϖ, the eccentricity of the ring particle is excited, and a spiral density wave (a compression wave) propagates radially outward away from the resonant location. At a vertical resonance (VR), where the identity of X is Ω, the inclination of the ring particle is excited, and a spiral bending wave (a transverse wave) propagates radially inward[10] away from the resonant location. For more details, see, e.g., Sect. 10.3 of Murray and Dermott (1999).

When the perturbation is not too strong, the radial variation in surface density $\Delta\sigma(r)$ for a spiral density wave has the form

$$\Delta\sigma(r) \simeq A\xi\cos(\xi^2/2)e^{-(\xi/\xi_D)^3}, \tag{7.7}$$

where A is an amplitude, ξ_D is a damping constant, and the dimensionless radial distance from resonance ξ is given by

$$\xi \equiv \left(\frac{3(m-1)n_L^2 r_L}{2\pi G\sigma_0}\right)^{1/2} \cdot \frac{r-r_L}{r_L}, \tag{7.8}$$

where r_L and n_L are the radius and mean motion at the location of exact resonance, σ_0 is the unperturbed surface density, and G is Newton's constant. This simplified equation is valid only for the downstream portion of the wave, $\xi \gtrsim 1$. We have also removed the phase term, which causes the wave to have a spiral shape. For the linear theory in its full form, see Goldreich and Tremaine (1982), Shu (1984), or Tiscareno et al. (2007).

Spiral waves, especially weak ones, are useful structures that can be thought of as in situ scientific instruments placed in the rings. Equation 7.7 describes a sinusoid with frequency that increases linearly with distance from the resonance; the rate of increase is inversely proportional to the background surface density σ_0 and thus can be used to measure it. The sinusoid oscillates within an envelope whose amplitude begins by increasing linearly, but then turns over and begins to decay as the exponential damping term takes over; the location of that turnover is governed by the damping constant ξ_D, which can thus be obtained and used to constrain the ring's dynamic viscosity (Goldreich and Tremaine 1978b; Shu 1984). Additionally, the mass of the perturbing moon can be obtained from the amplitude A, its orbital phase from the wave's phase term, and the absolute distance from Saturn's center from the resonance location r_L, though these parameters are often already known too precisely for the density wave to make a significant contribution. An algorithm for extracting these parameters from observed weak density waves (*Fig. 7-12*) was described by Tiscareno et al. (2007), making use of the spatially localized frequency spectra provided by a wavelet transform (Daubechies 1992; Torrence and Compo 1998).

A profile of surface densities and viscosities in the Cassini Division and A ring is shown in *Fig. 7-13*. The *Cassini* measurements (filled circles) have much less scatter than the *Voyager* measurements (open circles) not only because of greater measurement sensitivity but also because they rely on weaker waves that adhere more closely to the linear theory (7.7) but were not detectable in *Voyager* data. Nonlinear spiral density waves occur when the amplitude of the density perturbation $\Delta\sigma$ becomes comparable to the background density σ_0. In this case, instead of the quasi-sinusoidal profile of the linear case ((7.7), *Fig. 7-12*), the oscillation frequency no longer increases linearly with distance from the resonance, and additionally, the wave

[10] The exception is the nodal resonance, labeled −1:0, in which the mean motion of the ring particles is resonant with the forcing moon's nodal precession $\dot{\Omega}$. This peculiar resonance has a negative pattern speed, and its bending wave propagates outward (Rosen and Lissauer 1988).

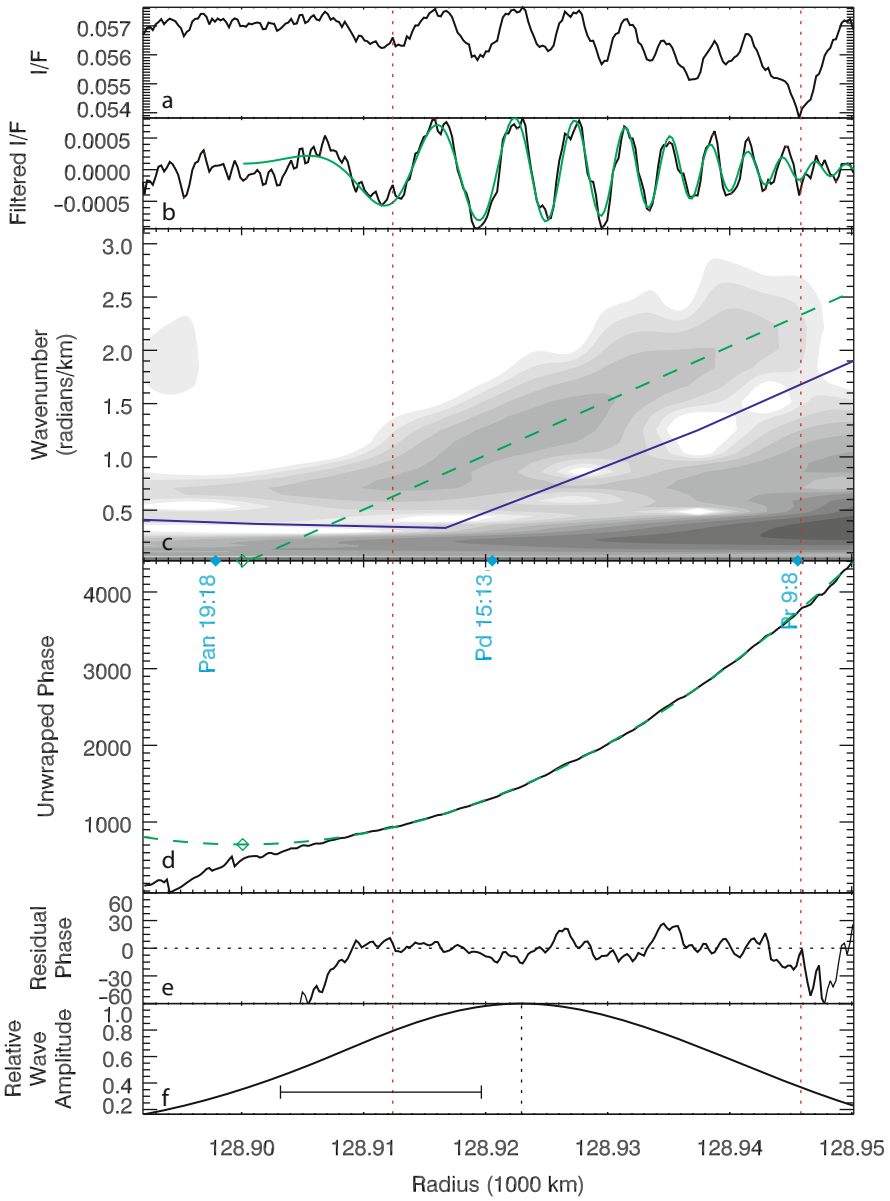


Fig. 7-12

A density wave fitting process carried out on a radial brightness profile from a *Cassini* ISS image. (a) Initial radial scan. (b) High-pass-filtered radial scan, with the final fitted wave shown in *green*. (c) Wavelet transform of radial scan, with *blue line* indicating the filter boundary, and the *green dashed line* indicating the fitted wave's wavenumber. (d) Unwrapped wavelet phase, with *green dashed line* indicating the quadratic fit and *green open diamond* the zero-derivative point. (e) Residual wavelet phase, showing that the interval used for the fit is the interval in which the phase behaves quadratically. (f) Wave amplitude, the local maximum of which (*vertical dotted line*) gives ξ_D; scale bar indicates the smoothing length of the boxcar filter (Figure from Tiscareno et al. 2007, q.v. for details of the wave-fitting process)

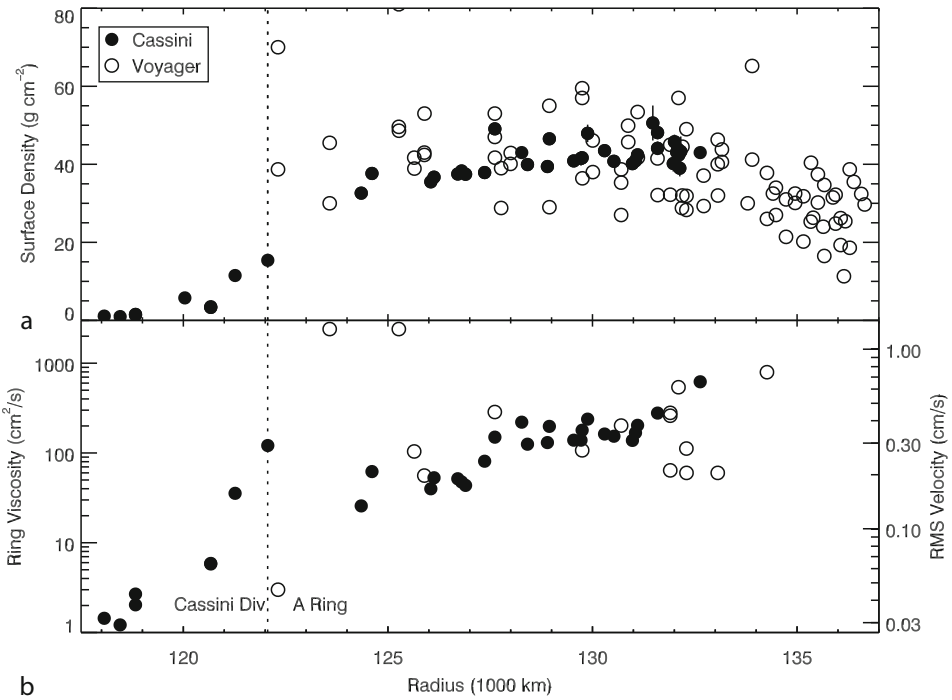


Fig. 7-13

(a) Ring surface mass densities σ_0 from density wave analysis. (b) Local ring viscosities; the associated rms velocity is within the ring plane, and likely enhanced by self-gravity wakes. *Open circles* indicate *Voyager* data, *filled circles Cassini* data. Based on figures from Tiscareno et al. (2007) and Colwell et al. (2009b), q.v. for details of individual measurements

morphology develops sharp narrow peaks and broad flat troughs. Significant progress has been made on developing an analytical model to describe nonlinear density waves (for a description and list of references, see Schmidt et al. 2009). Most recently, Rappaport et al. (2009) extracted model parameters from a series of RSS occultation profiles of the Mimas 5:3 density wave, the strongest (and thus most nonlinear) density wave in Saturn's rings (*Fig. 7-14*).

Spiral bending waves behave similarly to spiral density waves in their wavelength dispersion, but with several differences in other characteristics. Because they are transverse waves, both damping and nonlinearity arise less readily, which in principle makes them even more useful as ring diagnostics. However, they are significantly less abundant than density waves, because they can only be excited by perturbing moons on inclined orbits. Also, bending waves are harder to observe because they do not directly affect the optical depth. Rather, because they appear as corrugations in the ring plane, they are most readily seen when illuminated from a light source nearly in the ring plane and oriented azimuthally so as to shine across waveforms rather than along them, so that the waveforms can maximize the effect of changing the path length and thus the observed τ, even while $\tau_\perp$ remains nearly constant (see Sect. 1.3). In practice, these conditions are met by occultations with low elevation angle and in images taken near in time to Saturn's equinox (which occurs every ~14.5 year). In *Cassini* images taken during the 2009 equinox, some of the strongest bending waves (whose peaks rise the highest out of the mean ring plane) are seen to cast shadows.

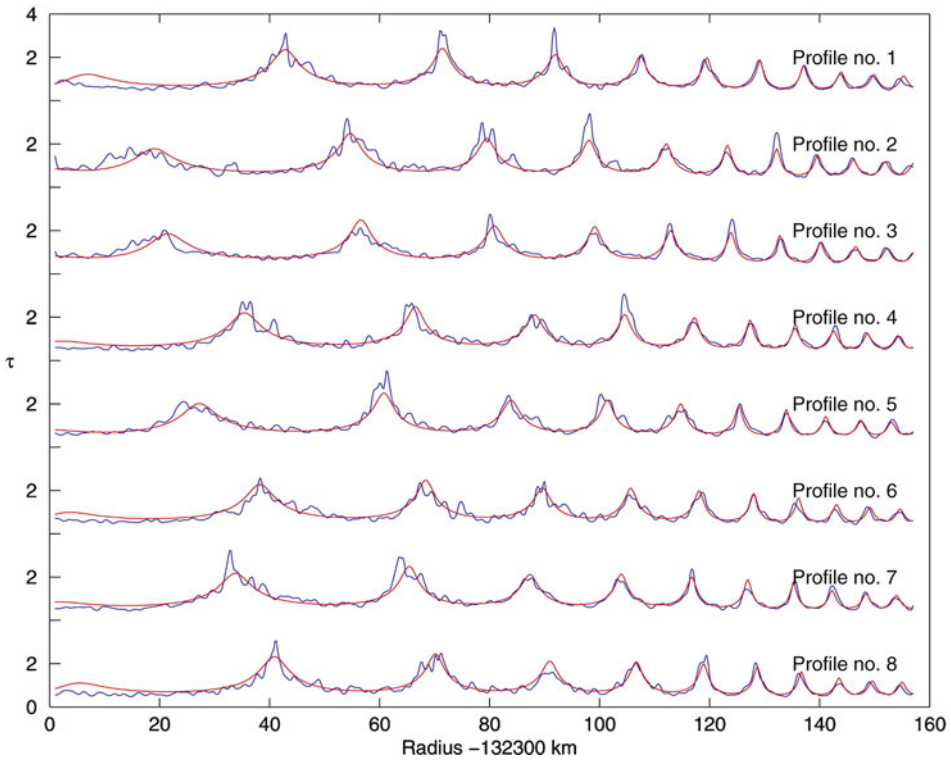


Fig. 7-14
Radial optical depth profiles from RSS occultations of the Mimas 5:3 spiral density wave (*blue*) with model fits (*red*) (Figure from Rappaport et al. 2009)

3.1.2 Gap Edges and Moonlet Wakes

There are 14 named gaps within Saturn's main rings: 4 in the C ring, 8 in the Cassini Division, and 2 in the A ring (*Fig. 7-11*). Both of the A ring's gaps contain known moons that clear them by exerting a torque on nearby ring material, and some gaps in the C ring coincide with known Lindblad resonances, but the other gaps are unexplained. Furthermore, even the two gaps that do contain known moons exhibit a surprising amount of unexpected behavior at their edges.

When a ring particle passes through conjunction with a nearby moon, the moon's gravity imparts an eccentricity as well as very slightly pushing the particle's semimajor axis away from its own. On its now-eccentric orbit, the ring particle passes through periapse and apoapse once per orbit; during the same time period, an inward (outward) particle moves forward (backward) by a distance $3\pi\Delta a$ in the moon's frame of reference (*Fig. 7-15*). To derive this, consider Kepler's Third Law, which states that an object's orbital rate n decreases with increasing semimajor axis a, specifically $n^2 a^3 =$ constant. Differentiating this with respect to a, we find the equation for *keplerian shear*,

$$\frac{\mathrm{d}n}{\mathrm{d}a} = -\frac{3}{2}\frac{n}{a}. \tag{7.9}$$

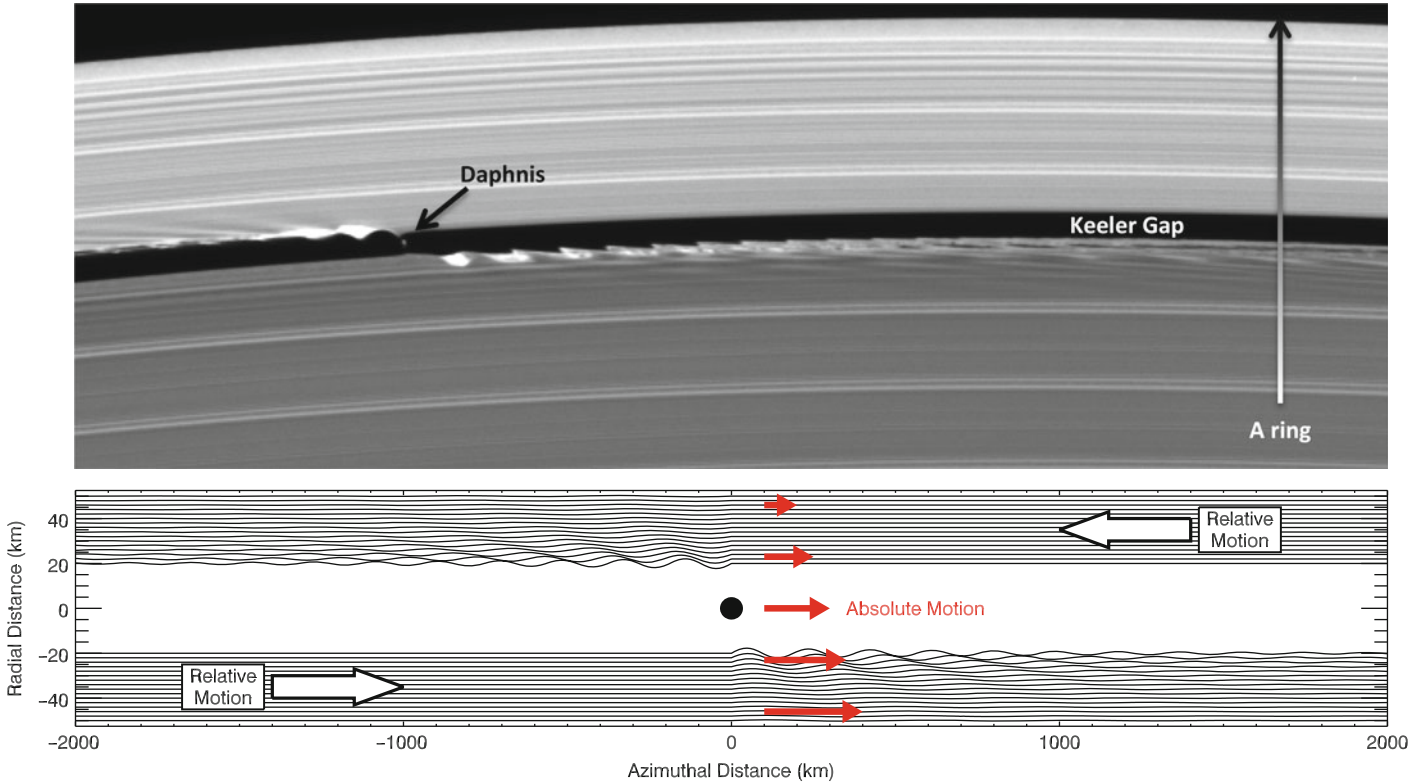


Fig. 7-15

(*top*) Wavy edges and moonlet wakes are seen at the edges of the Keeler Gap, surrounding the position of the gap-moon Daphnis (*bottom*). Streamline diagram showing $3\pi\Delta a$ wavelength set up as ring material passes a gap-moon. Different wavelengths at different radial distances Δa set up the "moonlet wakes" pattern. By Kepler's Third Law, inner material is moving faster than outer material (*red arrows*); by the same token, in the moon's frame of reference, inner material moves forward and outer material moves backward (*white arrows*)

At relative velocity $v = a\Delta n$, over one orbital period $P = 2\pi/n$ the relative motion of a ring particle with respect to a perturbing moon is $vP = 2\pi a\Delta n/n$. Substituting (7.9), this becomes $3\pi\Delta a$.

The characteristic wavelength $3\pi\Delta a$ is ubiquitous in the vicinity of a gap-moon. Not only does the gap edge form waves with that wavelength, but streamlines with that wavelength penetrate into the disk (*Fig. 7-15*). Because of the increasing wavelength, the distance between streamlines (which correlates with surface density) forms a pattern called "moonlet wakes"[11] that rotates with the gap-moon. It should be emphasized that these structures are not properly called "waves," as they do not propagate or otherwise dynamically evolve; rather, they are kinematic phenomena caused by the gap-moon organizing the orbital properties of the material around it into streamlines. However, some of Pan's moonlet wakes do reach high enough densities that the mutual self-gravity of particles enhances the peaks (Porco et al. 2005b) in a manner similar to nonlinear density waves (Sect. 3.1.1). Before Pan or any other gap-moon had been discovered, the wavy edges of the Encke Gap were tracked by Cuzzi and Scargle (1985), and the nearby moonlet wakes by Showalter et al. (1986). Interpretation of these observed features in terms of the simple theory just described allowed these authors to constrain the position of the moon causing them, which led Showalter (1991) to discover Pan in archival *Voyager* images.

Sharp ring edges can also be maintained at the locations of Lindblad resonances (see Sect. 3.1.1). The nature of the perturbation is similar to that in the impulse approximation described above, except that now the resulting streamline wavelength is exactly $1/m$ times the orbital circumference at that location, so that repeated conjunctions combine in a resonant effect. In fact, it can be shown that the resonant streamline wavelength $2\pi a/m$ reduces to $3\pi\Delta a$ in the case of large m (which is to say, small $\Delta a/a$).

For example, the outer edge of the B ring is coincident with the 2:1 LR with Mimas, and the outer edge of the A ring with the 7:6 LR with the co-orbital moons Janus and Epimetheus. Both edges generally exhibit the expected 2-lobed and 7-lobed (respectively) structure, though there are significant deviations that have been attributed to the time-variable orbits of Janus and Epimetheus for the A ring edge (Spitale and Porco 2009) and the excitation of normal modes in the nearby disk for the B ring edge (Spitale and Porco 2010). Of the eight gaps in the Cassini Division, those not containing ringlets all have circular outer edges, and those not associated with known moon resonances all have freely precessing elliptical inner edges (Hedman et al. 2010a). The Keeler Gap in the A ring also follows this pattern, with a nearly circular outer edge and a 32-lobed inner edge due to a resonance with Prometheus, though again the expected pattern is superposed with other structure that may be due to additional free or forced modes (Tiscareno et al. 2005; Torrey et al. 2008).

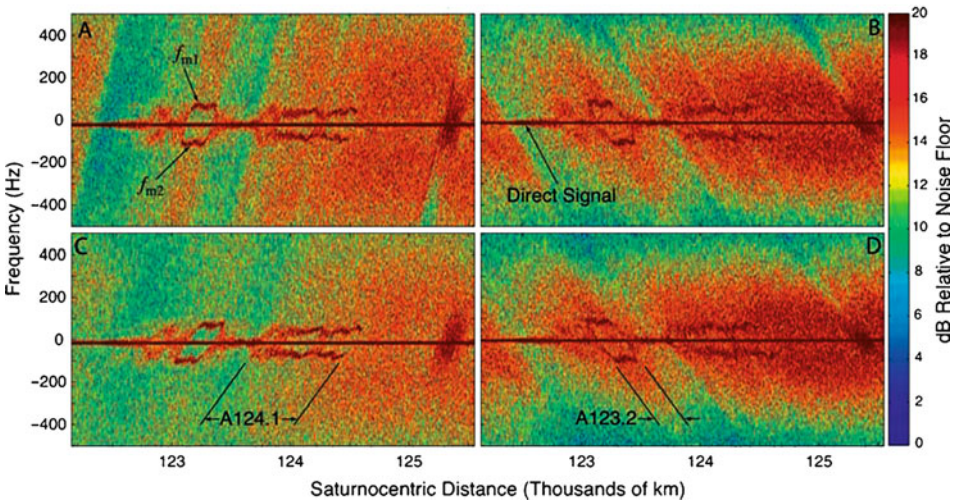
Some edges also exhibit vertical structure. The moon Daphnis, at the center of the Keeler Gap, has an inclined orbit that ventures ~9 km above and below the ring plane (Jacobson et al. 2008). The resulting vertical corrugations of nearby portions of the gap edge were predicted and then seen by their shadows cast during the 2009 equinox (Weiss et al. 2009). A region of vertical structure, probably due to embedded moonlets on inclined orbits, has also been detected and tracked in the outer edge of the B ring (Spitale and Porco 2010).

[11] Moonlet wakes have little in common with self-gravity wakes (Sect. 3.1.4), despite an unfortunate similarity in terminology.

3.1.3 Radial Structure

Several varieties of azimuthally symmetric radial structure are found in Saturn's rings. In addition to the gaps discussed above, the two largest in size are a series of sharp-edged annuli in the outer C ring with optical depths several times higher than the surrounding background, which have been dubbed "plateaux," and an undulating variation in optical depth in the inner B ring. Both of these have radial scales of ~100 km and have remained entirely unchanged during the 25 years between *Voyager* and *Cassini* observations (Nicholson et al. 2008), as well as remaining unexplained.

Regions of short-wavelength axisymmetric waves have been found by occultations in the inner A ring and in the B ring (Thomson et al. 2007; Colwell et al. 2007). The wavelengths range from 0.15 to 0.22 km and are locally monochromatic – the waveforms interacted with the spectrally coherent *Cassini* RSS radio occultation signal as if they were a diffraction grating (Thomson et al. 2007). This structure (◉ *Fig. 7-16*) has been explained in terms of viscous overstability, which occurs when a perturbation triggers an overly strong restoring force that resulting in continuing oscillations (for details and references, see Schmidt et al. 2009). Viscous overstability can occur when the ring's viscosity increases steeply with density, as naturally occurs in dense rings due to increasingly frequent collisions, and is sensitively affected by the strength of mutual self-gravity, the distribution of particle sizes, and the existence of self-gravity wakes. Work is ongoing to characterize the appearance of overstability waves in *Cassini* data, as well as their behavior in response to various environmental factors in simulations.



◘ Fig. 7-16

Spectrograms of *Cassini* RSS data showing periodic microstructure in the inner A ring. The *central horizontal line* in each panel is the direct signal, while the side bands that occur at some locations are coherent diffracted signal from the periodic microstructure acting as a diffraction grating (Figure from Thomson et al. 2007)

Both the A and B rings have sharp drops in optical depth at inner edges, with a "ramp" region of gradually decreasing optical depth (with decreasing distance) inward of them. These ramps have generally been considered to be the outermost portions of the Cassini Division and the C ring, respectively (see *Fig. 7-11*). The similarities in morphology, for the inner boundaries of the only two truly dense ($\tau \gtrsim 0.5$) broad disk structures known, are striking. The morphology of the ramp structure as well as the sharp edge at its outward boundary (which, in both known cases, is not correlated with any known resonance strong enough to explain it), as well as the apparent compositional similarity between the ramp material and the denser ring on the other side of the edge, has been explained in terms of ballistic transport, the radial movement of material due to micrometeoroid bombardment (Durisen et al. 1989, 1992). However, doubt is cast on that hypothesis by recent measurements of the wavelength dispersion of a spiral wave (Sect. 3.1.1) stretching across the sharp edge in optical depth between the Cassini Division ramp and the A ring, which indicates that there may not be a corresponding sharp change in surface density at that location (Tiscareno et al. 2009a).

3.1.4 Self-Gravity Wakes

The boundary between disruption-dominated regions (in which disks are stable) and accretion-dominated regions (see Sect. 1.2) is not a sharp one. In the outer parts of the region of stability for disks, gravitational instabilities drive temporary accretion that is quickly disrupted again by tides, forming a disk microstructure known as self-gravity wakes (SGWs). This balance is characterized by Toomre's Q parameter (for details and references, see Schmidt et al. 2009), defined as

$$Q = \frac{c_r \kappa}{3.36 G \sigma}, \tag{7.10}$$

where c_r is the radial velocity dispersion, σ is the local surface density, and $\kappa \equiv n - \dot{\omega} \approx n$ (see Sect. 1.1). Gravitational instability (i.e., accretion) is generally avoided as long as Q is at least a few times unity, but can become prominent if random velocities are damped (lower c_r) and/or surface density σ is increased and/or at locations further from the planet (lower κ). Balance between accretion and disruption, leading to SGWs, occurs when $Q \sim 2$. Data from the damping of spiral density waves (Sect. 3.1.1) indicate that c_r and σ adjust themselves in order to maintain $Q \sim 2$ over a wide region of the A ring (Tiscareno et al. 2007).

SGWs generally have a webbed structure that is elongated in a characteristic direction (*Fig. 7-17*), usually a few degrees to a few tens of degrees from azimuthal. Perpendicular to the characteristic direction, there is a characteristic spacing given by the Toomre critical wavelength,

$$\lambda_{cr} = \frac{4\pi^2 G \sigma}{\kappa^2}. \tag{7.11}$$

Because of their non-axisymmetric structure and their finite vertical thickness, the brightness of a disk pervaded by SGWs depends on the observer's longitude. An observer looking along the direction of the elongated wake structures will see more of the gaps between the wakes than an observer looking across the wakes (*Fig. 7-18*), especially at low elevation angles. Colombo et al. (1976) were the first to suggest the presence of SGWs in Saturn's rings as an explanation for the observed azimuthal brightness asymmetry. Today, observations of ring photometry combined with simulations of SGWs are further refining understanding of ring properties (see Sects. 4.1 and 4.2).

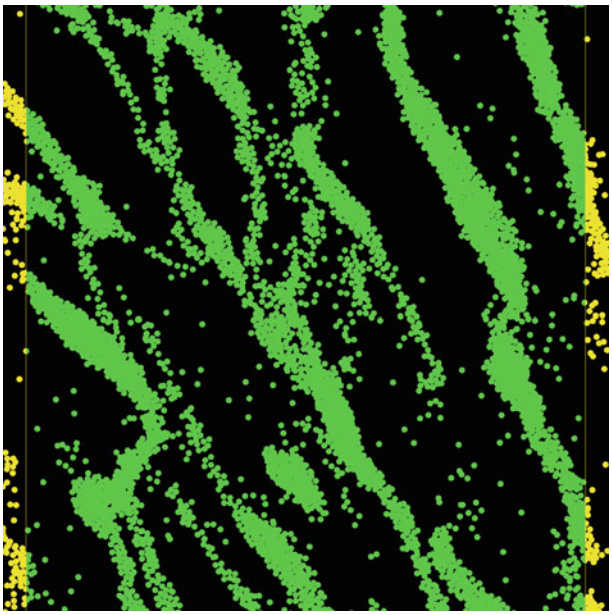


Fig. 7-17
Simulated self-gravity wakes (SGWs). This microstructure develops within dense rings as particles clump together under their mutual self-gravity but are ripped apart again by Saturn's tides (Figure courtesy of R.P. Perrine and D.C. Richardson)

Repeated stellar occultations with varying geometries are another way to gather information about SGW structure. Such observations from *Cassini* UVIS (Colwell et al. 2006, 2007) and *Cassini* VIMS (Hedman et al. 2007a; Nicholson and Hedman 2010) have been interpreted in terms of simple models with a bimodal distribution of optical depths, treating the SGWs as nearly opaque with optical depth τ_{wake}, and the intervening space as characterized by a constant lower optical depth τ_{gap}. Both teams have produced resulting data sets of various wake parameters as a function of radial location in the disk (*Fig. 7-19*). However, Tiscareno et al. (2010a) investigated the density contrast and the distribution of densities in simulated SGWs and found instead a trimodal distribution (*Fig. 7-20*). In their histograms (*Fig. 7-20*), the high-τ peak corresponds to the τ_{wake} assumed in simpler models, while the low-τ peak is practically transparent by contrast. In such a regime, the photometry of SGWs (especially for occultations and for images of the unlit side of the rings) is largely dominated by the mid-τ peak, which simulated movies identify as former wakes in the process of disruption. The areal average of the mid-τ and low-τ regions can be identified with the τ_{gap} values inferred from simpler models, except in the case of very low elevation angle (Tiscareno et al. 2010a), thus preserving the usefulness of the previous UVIS and VIMS studies. Recent preliminary results from a very-high-resolution UVIS occultation appear to give empirical confirmation that the distribution of surface densities in SGW-dominated disk regions is trimodal (Sremčević et al. 2009).

SGWs may have a dramatic effect on the relationship between ring optical depth and surface density. Simulations by Robbins et al. (2010) found that increased surface density merely added

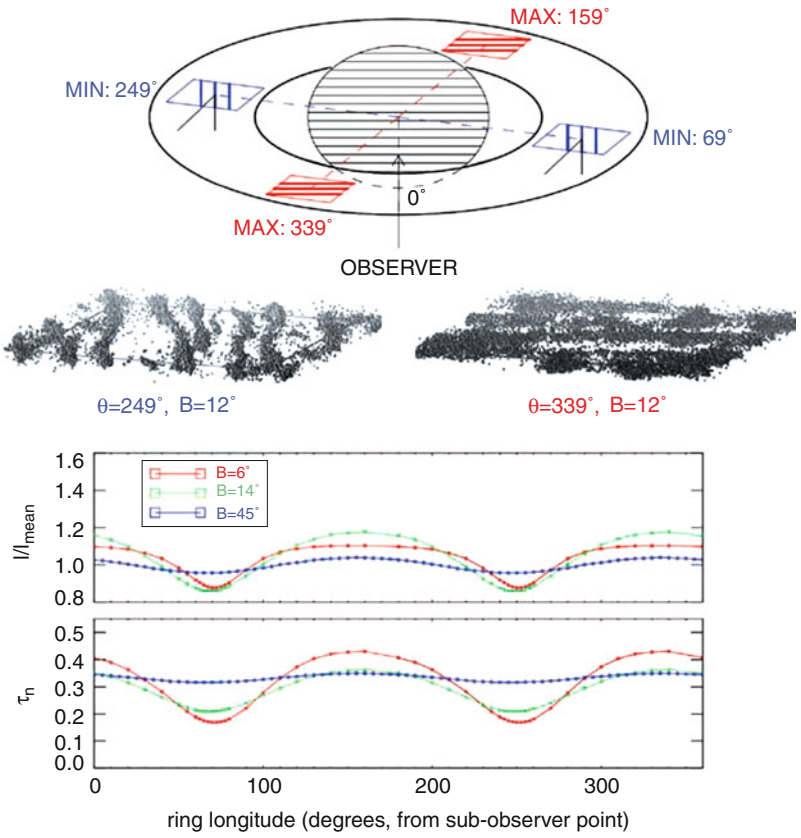


Fig. 7-18

The self-gravity wakes shown here (from the simulations of Salo et al. 2004) are canted at an angle of ~21°. At low elevation angle *B*, the transparency is higher when viewed along the wake structures (*blue*) than when viewed across the wake structures (*red*), leading to an azimuthal brightness asymmetry that depends on longitude relative to the observer (*lower panel*) (Figure from Schmidt et al. 2009)

more mass to the already-opaque wakes and only weakly increased the overall optical depth. They estimated that the mass of the B ring may be higher than *Voyager*-era estimates by a factor of 10 or more, approaching twice the mass of Mimas. See Sect. 5 for a discussion of the impact of this finding on the age and origin of Saturn's rings.

3.1.5 Propellers

A disk-embedded moon that is too small to open a full circumferential gap may still create a local disturbance in the disk. Because of Kepler's Third Law, the radially inward portion of the disturbance is carried forward and leads the moon, while the radially outward portion trails the moonlet (*Figs. 7-21* and *7-22*). Due to this characteristic shape, such moonlet-caused

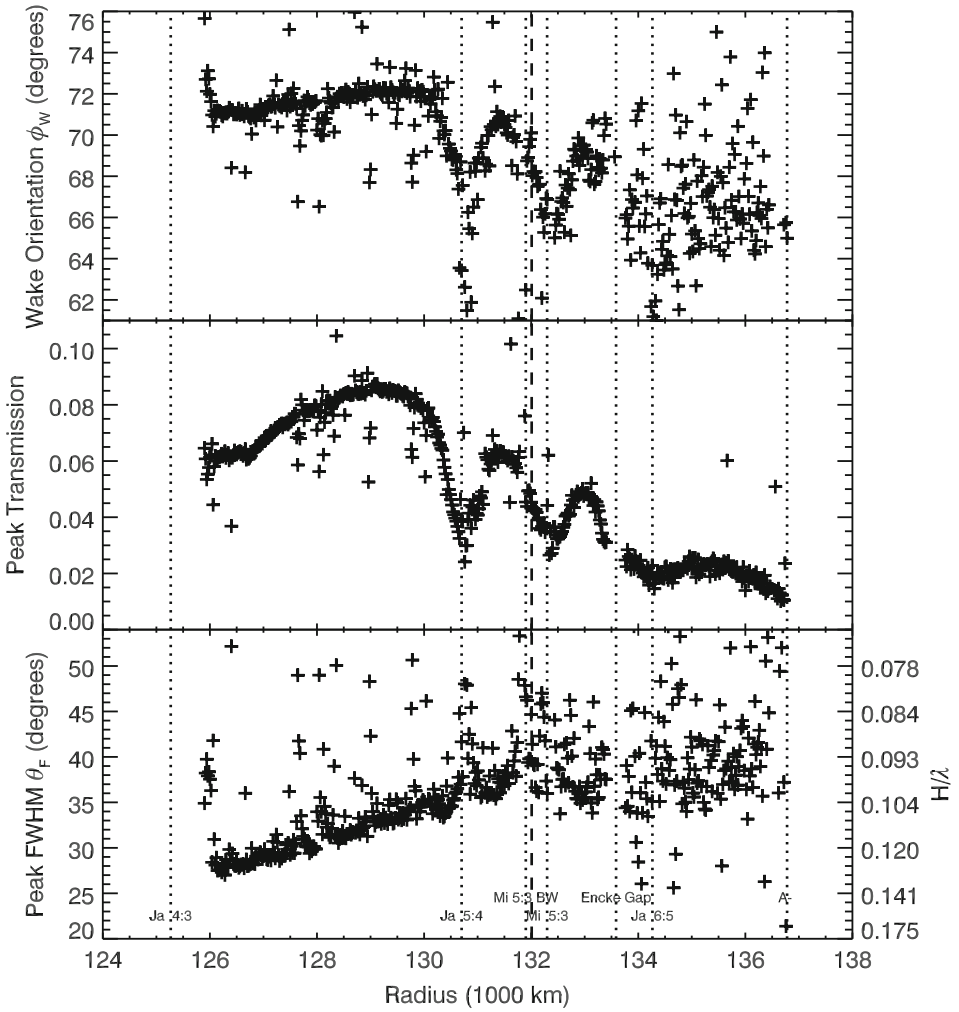


Fig. 7-19

Radial profiles of measured parameters for self-gravity wakes in the A ring. Work is ongoing (Hedman et al. 2011a) to understand the "halos" surrounding density waves (*dotted lines*) but not bending waves (*dashed lines*), within which the intensity and orientation of SGWs are altered, in addition to changes in spectral absorption by water ice and in the abundance of propellers (Sect. 3.1.5) (Figure modified from Hedman et al. 2007a)

disturbances have been named "propellers." A propeller-shaped disturbance can be thought of as a moon's unsuccessful attempt to form a full circumferential gap, which is frustrated by local ring viscous processes that limit the gap's azimuthal extent. Following predictions based on theory and modeling (Spahn and Sremčević 2000; Sremčević et al. 2002; Seiß et al. 2005), propellers in a range of sizes have been observed in *Cassini* images (Tiscareno et al. 2006a);

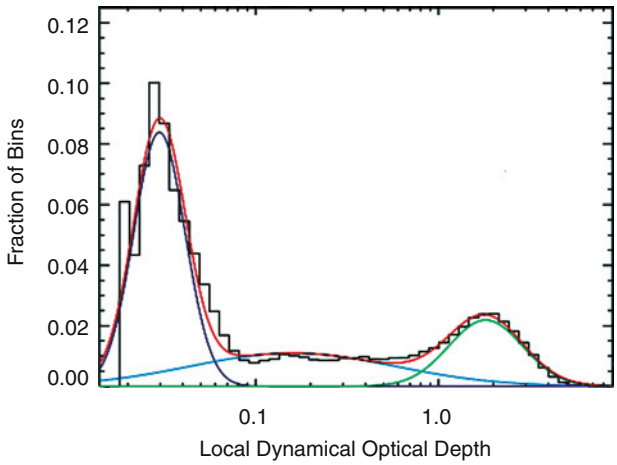


Fig. 7-20

A histogram of local dynamical optical depth τ_{dyn} calculated using a local density estimation method for a ring patch containing self-gravity wakes (Tiscareno et al. 2010a). A least squares fit was made to a sum of three Gaussians. The input surface density was 50 g cm^{-2}, and the coefficient of restitution law (see Sect. 4.2) is that of Borderies et al. (1984) using $v^* = 0.001$ cm s^{-1} (Figure modified from Tiscareno et al. 2010a)

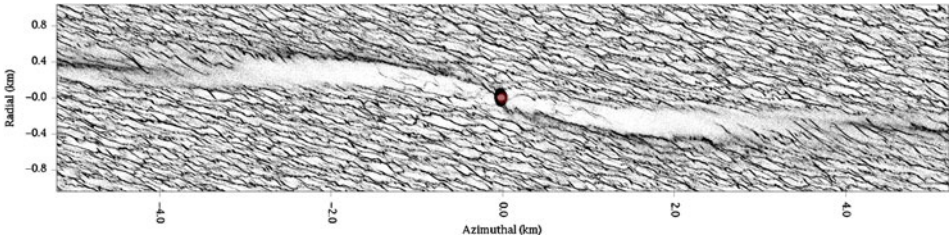


Fig. 7-21

Simulated "propeller" disturbance in the A ring due to an embedded moonlet. Note the texture of self-gravity wakes in the unperturbed regions, and the near-horizontal regions of depletion (*white*) and enhancement (*black*) in density due to the moonlet (Figure courtesy of M.C. Lewis)

(Sremčević et al. 2007; Tiscareno et al. 2008, 2010c). In all cases, only the propeller-shaped disturbance is directly seen, while the responsible moon at the center is inferred.

The occurrence of observed propellers is confined to only a few annular regions within the A ring. The three "Propeller Belts" in the mid-A ring are located between 127,000 and 132,000 km from Saturn's center (*Fig. 7-23*) and are separated by propeller-poor "halo" regions, centered on large density waves, in which other ring properties are also altered (*Fig. 7-19*). It is not known whether the radial variations in observed abundance of propellers are due to variations in their origins, survival, observability, or a combination of the three.

The Propeller Belts contain numerous small propellers with the radial offset parameter Δr (*Fig. 7-22*) ranging from 0.3 to 1.4 km and azimuthal extent of up to several kilometers

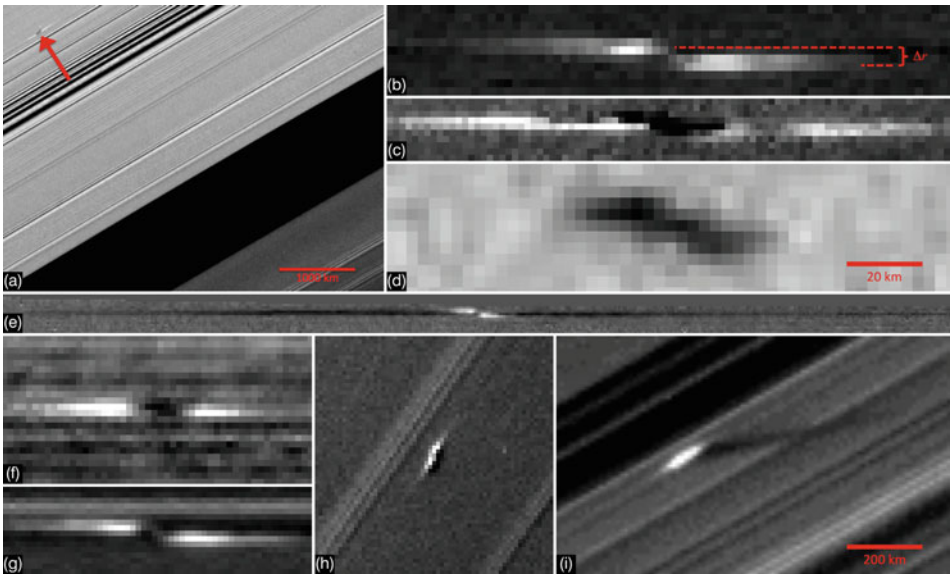


Fig. 7-22

Propellers as seen in selected *Cassini* ISS images. Panel (a) shows a propeller in context of the Encke Gap and several density waves. Panel (b) illustrates the radial offset Δr between the two azimuthally aligned lobes, proportional to the size of the unseen central moonlet. Panels (b, c, and d) show three views of the same propeller at the same scale, demonstrating how its appearance changes with viewing geometry. Non-equinox views are on the lit (b, e, g) or unlit (a, c ,d, f) face of the rings, while panels (h and i) show propellers casting shadows near the Saturnian equinox. The scale bar in panel (d) also applies to panels (b, c, f, and g). The scale bar in panel (i) also applies to panels (e and h) (Figure from Tiscareno et al. 2010c, reproduced by permission of the AAS)

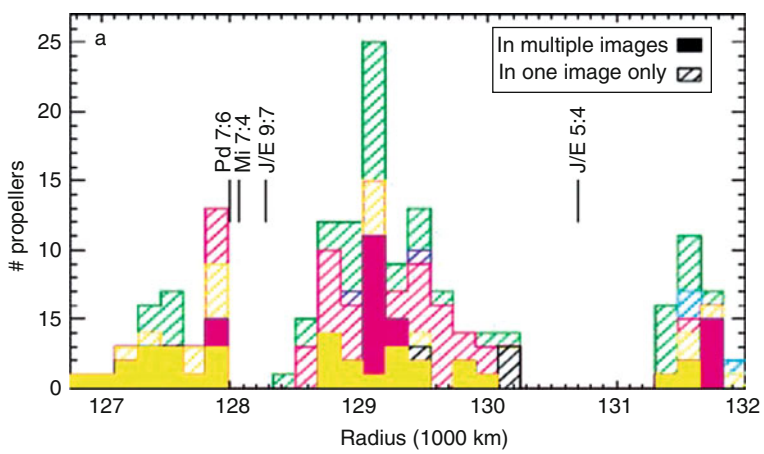


Fig. 7-23

A histogram of the abundance of propellers, as a function of radius, in the Propeller Belts of the mid-A ring (Figure from Tiscareno et al. 2008, reproduced by permission of the AAS)

(Tiscareno et al. 2008). A separate class of "giant propellers" has been found in the outermost A ring, outward of the Encke Gap (133,700 km from Saturn's center) and thus called the "*trans*-Encke" population, with measured Δr as high as 6 km and azimuthal extent of up to several thousand kilometers (Tiscareno et al. 2010c).

Simulated propeller structures have both density-depleted and density-enhanced regions (*Fig. 7-21*), and the interpretation of observed propellers in terms of relative density has been a point of controversy. In simulated propellers, the radial offset Δr is consistently close to four times the central moon's Hill radius (7.2) and thus is diagnostic of its mass. The first observed propellers, which were seen during *Cassini*'s Saturn Orbit Insertion (SOI) maneuver and are thus the smallest known but also the best resolved in the Propeller Belts, are relative-bright with respect to nearby unperturbed regions of the A ring (Tiscareno et al. 2006a). Although this photometry is consistent with a region of enhanced optical depth, the morphology of the SOI propellers is very similar to that of density-depleted regions in propeller simulations. The photometry of larger propellers in the Propeller Belts also turned out to be generally consistent with enhanced optical depth (Sremčević et al. 2007; Tiscareno et al. 2008). Hypotheses to explain the apparent correlation between high optical depth and depleted density include the temporary liberation of ring-particle regolith within the propeller structure (Sremčević et al. 2007; Halme et al. 2010) and the disruption of self-gravity wakes within the propeller structure (Tiscareno et al. 2010a). In giant propellers, for the first time, relative-dark and relative-bright regions are sometimes seen in the same propeller structure (*Fig. 7-22*), enabling density-enhanced and density-depleted regions to be disentangled in some cases (Tiscareno et al. 2010c). However, the structure of giant propellers may be qualitatively different than that of the smaller propellers in the Propeller Belts, so parallels should be drawn cautiously.

The giant *trans*-Encke propellers are larger and more prominent than those in the Propeller Belts, and also much less numerous. Taken together, these factors allow giant propellers to be studied individually and tracked over a period of years. In the Propeller Belts, the swarm of particles is such that, even if the same object were seen on multiple occasions, it would be very difficult to have much certainty that it was the same object due to the many nearby similar objects. This criterion may be useful for drawing a distinction, should one wish to do so, between a "moon" and a "moonlet." Basing such a distinction upon size is arbitrary and lacks any wide agreement, as there is no major physical threshold crossed by objects in the km-size range. We suggest that any object be called a "moon" if it can be singled out for long-term study, while a "moonlet" is a member of a population or swarm that prevents individual tracking. This may still not be a bright line, but at least it is based on a physical property. The question of whether propeller-causing objects deserve to be considered as full-fledged moons is further complicated by the fact that, though their positions have been tracked over long periods, they are not directly seen but hidden within the surrounding propeller-shaped disturbance.

The propeller population follows a very steep particle-size distribution, with a power-law index $q \sim 6$ (Tiscareno et al. 2008, 2010c) for propeller moonlets of size between 30 and 300 m. By contrast, the vast majority of ring mass is concentrated in the continuum particles of size between 1 cm and 10 m, with a much shallower power-law index $q \sim 2.75$ (Zebker et al. 1985; Cuzzi et al. 2009).

Spitale and Porco (2010) reported a single observation of what appears to be an embedded moon of radius 0.3 km in the outermost parts of the B ring, based on the shadow it cast onto the (vertically much thinner) ring while the 2009 equinox brought nearly edge-on ring illumination (*Fig. 7-24*). However, though this object's size is comparable to that inferred for the largest propeller moons, no propeller structure is apparent around this object. Michikoshi and

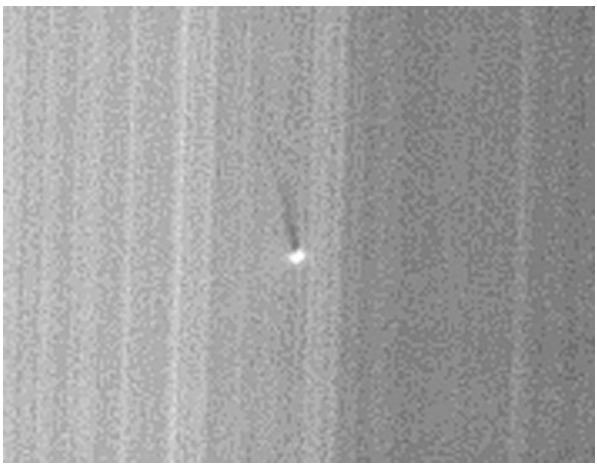


Fig. 7-24
The feature known as S/2009 S 1, identified by the shadow it cast during equinox, was seen only in this image (Figure from Spitale and Porco 2010)

Kokubo (2011) investigated the possibility that the B ring's high surface density, with accompanyingly strong self-gravity wakes, might prevent the formation of a propeller structure. They found that propellers form when the moonlet's Hill radius r_{Hill} (➤ 7.2) is larger than the Toomre critical wavelength λ_{cr} (➤ 7.11). While this condition might hold if a moon of the observed size were in the densest parts of the central B ring, the surface densities inferred by Spitale and Porco (2010) for the outer B ring are not high enough. Possible explanations include that the surface densities in the outer B ring are higher than thought, that the observed object is not a moonlet but perhaps an impact cloud or a large SGW clump, that the observed bright feature does indeed include an unresolved propeller structure, or that propellers require a (not well understood) photometric balance in order to be observed, and that balance is not met in the outer B ring.

Orbital tracking of *trans*-Encke propellers has revealed a surprising non-keplerian component to their motion (Tiscareno et al. 2010c). The best-observed example had its semimajor axis increase (as detected by tracking its longitude as a function of time, not its actual radial position) by a rate as high as $+0.11\,\mathrm{km\,year^{-1}}$ between 2006 and 2007, then decrease by $-0.04\,\mathrm{km\,year^{-1}}$ from 2007 to 2009 (➤ *Fig. 7-25*). Although this profile is similar to a sinusoid, no resonance with a known moon has been found to be capable of producing such behavior. Another mechanism that would produce a sinusoidal residual is the so-called frog resonance (Pan and Chiang 2010), in which the propeller moonlet interacts primarily with the mass at either end of the propeller gap. The propeller moonlet plausibly librates with the observed amplitude and period in a simple version of the model, but it remains unclear whether the moon-formed gap responds sluggishly enough to allow the moon to librate within it.

Other hypotheses for non-keplerian motion of propellers are based on the concept of "Type I migration." As classically formulated for protoplanetary disks (Ward 1986, 1997; Papaloizou et al. 2007), the angular momentum exchange at inner Lindblad resonances between a disk and an embedded mass fails to exactly cancel with that at outer Lindblad resonances,

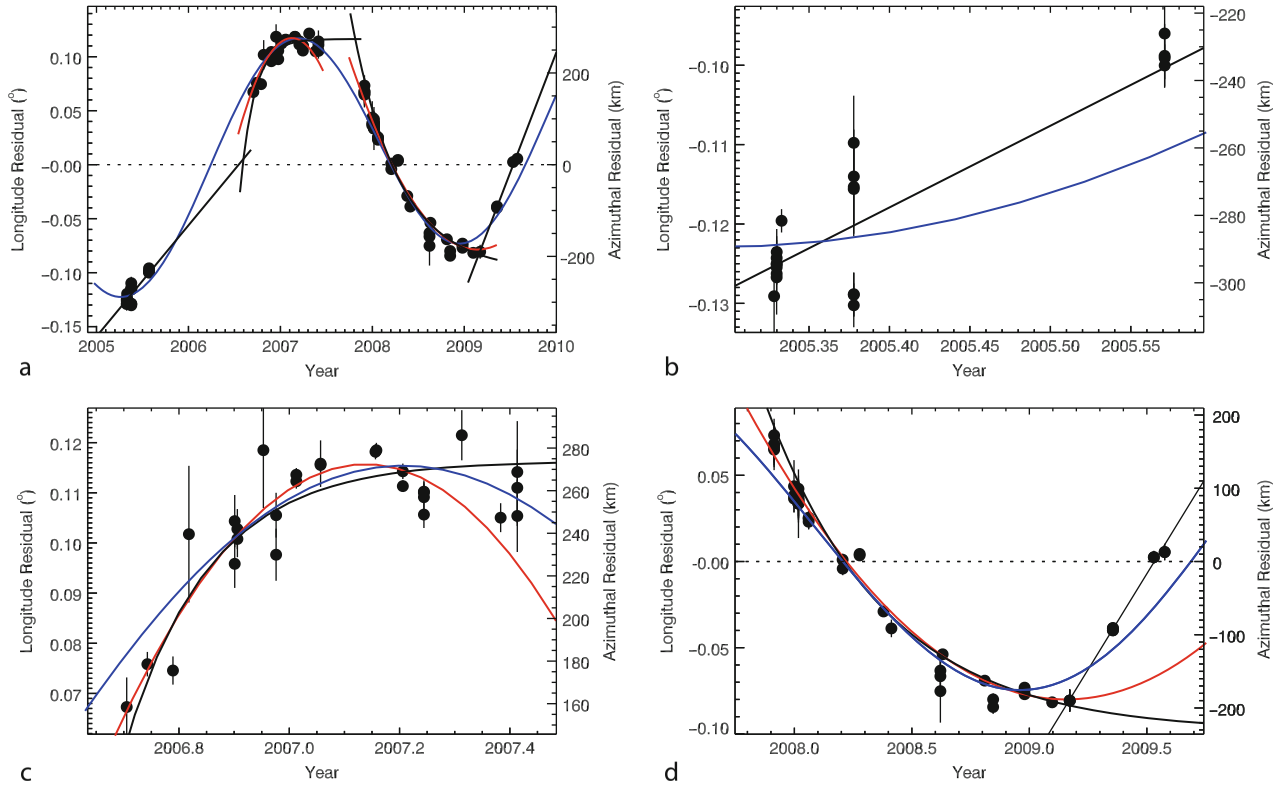


Fig. 7-25

Non-keplerian motion of the propeller "Blériot" over 4 years. Panel (a) contains all the data, while panels (b, c, and d) contain subsets of the data shown in greater detail. The *blue line* indicates a linear-plus-sinusoidal fit to all the data, while the *red lines* indicate piecewise quadratic fits corresponding to a constant drift in semimajor axis and the *black lines* indicate exponential fits (Figure from Tiscareno 2012)

resulting in a differential torque that leads to inward migration of the embedded mass. However, classical Type I migration depends crucially on the gas component of the disk, which causes Lindblad resonance locations to shift asymmetrically. For the case of planetary rings, which are strictly particulate, Crida et al. (2010) rederived the equations for Type I migration from first principles, using analytical arguments and numerical simulations to trace the angular momentum exchange between streamlines of continuum ring particles and the embedded moon. For the case of a homogeneous disk, Crida et al. (2010) found an asymmetric torque that is always inward and is one to two orders of magnitude too weak to explain the observed nonkeplerian motion. Building upon this work, Rein and Papaloizou (2010) considered temporal variations in the disk, proposing that the propeller moon is constantly perturbed by stochastic variations in the disk's surface density due to self-gravity wakes, finding that a random walk in the propeller moon's semimajor axis can result. On the other hand, Tiscareno (2012) considered spatial variations in the disk, proposing that an externally produced radial surface density profile results in an equilibrium semimajor axis to which the propeller moon will return after episodic kicks.

Continued observations should distinguish among these models. The leading models interpret the existing data as a sinusoid (Pan and Chiang 2010), episodically initiated exponentials (Tiscareno 2012), and a pure random walk (Rein and Papaloizou 2010), and thus they differently predict the qualitative nature of future data. However, all of the leading models imply that propeller moons, which are the first objects ever discovered to orbit while embedded in a disk rather than in free space, are directly interacting with the disk, a phenomenon that has long been an integral part of disk models but has never before been directly observed.

3.1.6 Spokes and Impacts

"Spokes" are near-radial markings on Saturn's B ring, likely composed of dust levitating above the ring plane due to electromagnetic forces (for details and references, see Horányi et al. 2009). Spokes usually form on the dawn side of the rings where ring particles are just coming into full sunlight, and have a correlation with periodicities believed to originate with Saturn's magnetic field. Susceptibility to electromagnetic forces surely is connected to the fact that spokes appear at radial locations near to (and often astride) that of synchronous orbit, where the orbital period of a ring particle matches the rotation period of the planet, and thus also of the magnetic field. Spokes appear to be a seasonal phenomenon, correlated with a low elevation angle of the Sun above the ring plane, probably due to variations in the generation of plasma by photocharging near the rings. The Hubble Space Telescope tracked the decline and disappearance of spokes during "Saturnian October/November" in the late 1990s (McGhee et al. 2005), and they remained absent during the first 1.5 year of *Cassini* operations at Saturn before reappearing in "Saturnian January/February" in late 2005 (Mitchell et al. 2006). Since their reappearance, *Cassini* images have tracked the morphology, photometry, evolution, and temporal variability of spokes (Mitchell et al. 2012).

Numerous theories for the formation of spokes have been put forward, but none has gained definitive acceptance. The most popular mechanism is that of Goertz and Morfill (1983), who interpret spokes as dust levitated above the ring plane by an electromagnetic disturbance initiated by a micrometeoroid impact. Following the inference of Smith et al. (1982) from *Voyager 2* images that spokes with a radial dimension of thousands of kilometers form on a timescale of several minutes, the Goertz and Morfill (1983) model includes a propagation speed for

the plasma cloud $\gtrsim$20 km s^{-1}, though this result was criticized by Farmer and Goldreich (2005; see reply by Morfill and Thomas 2005). Other ideas for the rapid appearance of spokes include near-simultaneous impact of a broad assemblage of particles generated elsewhere; the most developed model of this type suggests an electron beam generated by Saturnian lightning (Hill and Mendis 1981; Jones et al. 2006), though a dispersed cloud originating from an impact has also been suggested (Hamilton 2006). On the other hand, despite executing a number of high-frequency imaging sequences designed to do so, *Cassini* has not observed the rapid formation of spokes but did observe spokes growing from negligible to strong over ~hour timescales (Mitchell et al. 2012).

During the 2009 Saturnian equinox, dust clouds evolving under keplerian shear were observed in the A and C rings and attributed to micrometeoroid impacts Tiscareno et al. (2009b; see also ❯ Sect. 3.6). Ongoing analysis of these impact clouds may lead to new constraints on the interplanetary micrometeoroid population, as well as better understanding of the relationship between impacts and spokes.

3.2 Dense Narrow Rings

Saturn:	Titan ringlet
	Maxwell ringlet
	Bond ringlet ("1.470 R_S")
	Huygens ringlet
	"Strange" ringlet
	Herschel ringlet ("1.960 R_S")
	Jeffreys ringlet
	Laplace ringlet ("1.990 R_S")
Uranus:	6 ring
	5 ring
	4 ring
	α ring
	β ring
	η ring
	γ ring
	δ ring
	ϵ ring

Dense narrow rings are a unique assemblage of matter, behaving as a coherent self-contained object on planetary lengthscales yet ephemerally thin (1–100 km in radial width, compared to ~100,000 km in diameter) and not in a gravitational ground state (unlike a planet, one would collapse if it stopped moving). The formation and proximate causes of these highly organized dynamical systems are almost entirely unknown, and even their present dynamics are only partly understood. For Uranian narrow ringlets, which constitute the majority of the known examples, their highly time-variable properties were only dimly revealed by the single snapshot provided by the *Voyager 2* flyby, with temporal resolution provided by Earth-based occultations, while analysis of the more extensive *Cassini* data set of Saturnian narrow ringlets is still in progress.

Nearly all known dense narrow rings are either noncircular or inclined to the main ring plane, or both. Their radial widths range from ~1 km (many examples) up to ~100 km for the Maxwell ringlet and the ϵ ring. Furthermore, nearly all dense narrow rings have edges that are quite sharp; as with dense disks, the existence and stability of such sharp edges requires some confinement mechanism to counteract the natural process of radial viscous spreading. Confinement may be due to an external moon or to the ring's own self-gravity, and in some cases to processes yet to be understood. A detailed table of ringlets and their properties was given by Colwell et al. (2009b) for Saturn's rings and by French et al. (1991) for Uranus'. For general details and references on the dynamics of narrow rings, see French et al. (1991) and Schmidt et al. (2009).

The "shepherding" mechanism by which a moon opens a gap or maintains an edge in a disk (Sect. 3.1.2) can also occur with two shepherd moons on either side of a narrow ringlet (Goldreich and Tremaine 1979a). As previously discussed, this can occur through repeated impulses on nearby material or more distantly through a resonance. The only example of a dense narrow ring with a known shepherd on either side is Uranus' ϵ ring, though Saturn's F ring (see Sect. 3.3) is a variation on that idea. The Titan ringlet in Saturn's C ring is at the location of an apsidal resonance with Titan, where the ring particle's precession frequency $\dot{\omega}$ is commensurate with Titan's orbital motion, so that the eccentric ringlet always keeps its apoapse pointed toward Titan. The Bond ringlet in Saturn's outer C ring is associated with a 3:1 Lindblad resonance with Mimas, and a few edges of Uranian rings coincide with resonances, but the details of the interaction are yet to be understood in all cases.

Any ringlet of finite width ought to have a different precession rate $\dot{\omega}$ at its inner and outer edges, which should smear out the orientation of ring particle orbits and prevent the ringlet from appearing eccentric. However, this effect can be counteracted by the ring's own gravity (Goldreich and Tremaine 1979b, 1981; Borderies et al. 1983), possibly combined with viscous and collisional effects (Dermott and Murray 1980; Chiang and Goldreich 2000; Mosqueira and Estrada 2002). A purely gravitational model requires a positive "eccentricity gradient," which is to say that the eccentricity monotonically increases from the ring's inner edge to its outer edge.

A number of ringlets indeed appear to precess about their planet as a rigid body. Observations of some of these, such as the Maxwell ringlet and the α and β rings, are consistent with a pure freely precessing ellipse (i.e., an $m = 1$ mode) as described by theory, and the Maxwell ringlet even has a clearly positive eccentricity gradient (Spitale and Porco 2006), while other ringlets appear to have additional components to their motion. The Huygens ringlet appears to have an additional $m = 2$ mode, possibly influenced by the Mimas 2:1 resonance that governs the nearby outer edge of the B ring, as well as an $m = 6$ mode of unknown origin (Spitale and Porco 2006). The δ ring also has an $m = 2$ mode, while the γ ring has an $m = 0$ mode, which is to say a radial oscillation (French et al. 1991). Higher-m modes, some corresponding to known moons, have also been found in several Uranian rings by Showalter (2011), who also pointed out that non-detections of shepherd moons at Uranus have become significant enough that shepherding is unlikely to be the dominant mechanism of ring confinement.

Spiral density waves (Sect. 3.1.1) can also occur in dense narrow rings, at least those broad enough (generally a few kilometers) for a wave to develop, but not many examples have been identified. A density wave due to the Pandora 9:7 LR appears to propagate through the Laplace ringlet (Colwell et al. 2009b). A possible density wave was detected in the *Voyager* stellar occultation of Uranus' δ ring, but it cannot be confirmed in the absence of data at multiple longitudes and/or times, and furthermore, there is no known moon at the proper place to raise such a wave.

In Saturn's rings, gaps are named by the IAU but ringlets are not. In most cases, the main ringlet within a particular gap is given the same name as its gap (though many favor the name "Titan ringlet" for the ringlet in the Colombo Gap, for its strong association with a resonance with Titan), but the existence of multiple ringlets in one gap requires some naming creativity. For example, the Huygens Gap contains a second dense narrow ringlet outward of the main Huygens ringlet, which has been informally nicknamed the "Strange" ringlet in part because of its unusually high inclination (Spitale et al. 2008) and in part as a complement to the "Charming" ringlet (Sect. 3.3). Five ringlets with noncircular features were identified in *Voyager*-era publications (e.g., French et al. 1993) only by their distance from Saturn's center in units of Saturn radii; of these, three can now be easily named for the gap containing them as shown in the accompanying table. However, the former "1.495 R_S ringlet" is classified by Colwell et al. (2009b) as a plateau or embedded ringlet because it is adjacent to the continuum C ring rather than fully contained in the Dawes Gap, while the former "1.994 R_S ringlet" is now considered part of the continuum Cassini Division between two narrow gaps, rather than a ringlet that nearly fills its gap, judging from the characteristic pattern of empty gaps having circular outer edges and resonant inner edges (Hedman et al. 2010a, see also Sect. 3.1.2).

3.3 Narrow Dusty Rings

Saturn:	"Charming" ringlet
	Encke ringlets
	F ring
Uranus:	λ ring
Neptune:	Le Verrier ring
	Arago ring
	Adams ring

The smallest particles in dense rings are usually swept up by larger ones and incorporated into their regolith (Cuzzi et al. 2009), and so such rings are largely dust free, which is to say that they have few particles smaller than mm- to cm-size and thus are not strongly forward-scattering in their interactions with light. Therefore, the dustiness of a few structures that do occur within Saturn's and Uranus' main rings is a likely indicator of relative dynamism and/or youthfulness that either prevents dust from being swept up or has not allowed time for it to be swept up yet. In Neptune's rings, on the other hand, the lack of any ring component with optical depth exceeding 0.1 may account for the general dustiness of the system.

In Saturn's main rings, high dust fractions are found exclusively in a small number of narrow dusty ringlets that occur in the larger empty gaps (Horányi et al. 2009). The only gap with multiple dusty ringlets is the Encke Gap, which three ringlets share with the moon Pan. These are riddled with "clumps" (azimuthal brightness variations) and "kinks" (radial offsets) that drift slowly with respect to each other (Hedman et al. 2005, 2011b; Horányi et al. 2009). The Encke ringlets and the "Charming" ringlet within the Laplace Gap are known to be "heliotropic," which is to say that they have an eccentricity forced by solar radiation pressure that causes their apoapses to always point toward the Sun (Hedman et al. 2007b, 2010b). The "Charming" ringlet is smooth, lacking clumps or kinks; the best studied of the heliotropic rings, it also has free eccentricity and free inclination components in addition to its solar-forced eccentricity (Hedman et al. 2010b).

The F ring is the granddaddy of narrow dusty ringlets, being by far the best studied as well as the largest. For a review and references, see Colwell et al. (2009b). Located a few thousand kilometers off the outer edge of Saturn's main rings, and with both significant vertical thickness and inclination ($\gtrsim$10 km in both cases), the F ring effectively frustrates any attempt to see the rest of the main rings in edge-on viewing geometries. Its high fraction of forward-scattering dust makes it by far the brightest component of Saturn's ring system when viewed at high-phase geometries.

The core of the F ring contains a large amount of dust enveloping an unseen belt of km-size moonlets, inferred from their absorption of charged particles (Cuzzi and Burns 1988), from the characteristic "fan" structures they create in surrounding dust (Murray et al. 2008; Beurle et al. 2010), from direct detection by occultations (Esposito et al. 2008; Hedman et al. 2011c), and from shadows cast during the 2009 Saturnian equinox (Beurle et al. 2010). Several dusty lanes or "strands" accompany the core on either side, nearly parallel to it though Charnoz et al. (2005) pointed out that the most prominent strands can be laid end-to-end to form a one-armed kinematic spiral. Nascent strands, or "jets," have been associated with collisions between moonlets and the core (Murray et al. 2008). Despite being entirely composed of "clumps" and "kinks" so numerous that they cannot be individually tracked as they are in the Encke ringlets, the F ring core nevertheless maintains over decadal timescales the shape of a freely precessing eccentric inclined ellipse; the orbital solution formulated to account for *Voyager* and other pre-*Cassini* data (Bosh et al. 2002) has, somewhat surprisingly, remained a good predictor of the core's position through the *Cassini* mission (Murray et al. 2008; Albers et al. 2012). However, decade-scale time variations in the core's internal structure, as well as in the surrounding dust, have clearly taken place between the *Voyager* and *Cassini* visits (Colwell et al. 2009b; Showalter et al. 2009).

The F ring is one of only two narrow rings (along with the ϵ ring) to have known "shepherd" moons orbiting on either side of it. However, it has not been conclusively shown how (or even that) the moons Prometheus and Pandora actually constrain the F ring in its place, and in fact, they appear to stir the ring up at least as much as to maintain it. Simulations by Winter et al. (2007) found the moons to be responsible for both strong chaos and significant radial confinement. Both moons create at each closest approach a "streamer-channel" in nearby dust strands that subsequently moves downstream and shears under keplerian motion (*Fig. 7-26*, Murray et al. 2005). During the 2009 Saturnian equinox, which coincided with the once-in-17-year alignment[12] of Prometheus' apoapse with the F ring's periapse, minimizing their mutual closest approach distance (Chavez 2009), moonlets inferred from their shadows had a clear correlation of abundance with longitude relative to Prometheus (Beurle et al. 2010), indicating that Prometheus is directly triggering accretion within the F ring, the products of which may then be what collides with the core to form new jets and strands (Beurle et al. 2010). All of these interwoven phenomena make the F ring the solar system's foremost natural laboratory for direct observation of accretion and disruption processes.

Like the F ring, the λ ring is by far the brightest component of its planetary ring system when viewed at high phase angles, due to its high fraction of forward-scattering dust. But the λ ring appears to be significantly simpler and more sedate than its Saturnian cousin, and its low detectability at lower phase and in occultations indicates that it is poor in macroscopic particles.

Among dense ring systems, Neptune's has a much higher dust fraction than those of Saturn and Uranus and is unique in having no significant dust-free regions. The three main rings of Neptune lack sharp edges and are generally more tenuous, with only the Adams ring reaching

[12] The periodicity of this alignment was recalculated by Chavez (2009), using updated orbital data.

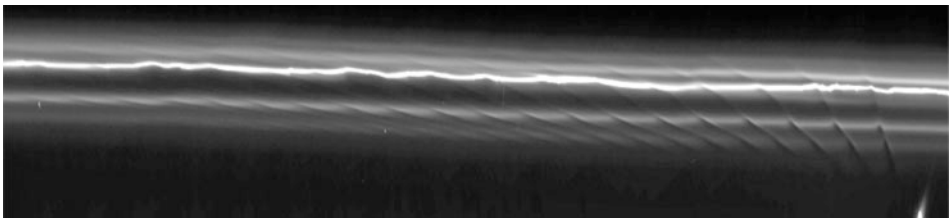


Fig. 7-26
Sheared channels in the F ring created as Prometheus (*lower right*) dips into the ring. The horizontal dimension of this *Cassini* ISS mosaic covers 60° of longitude (~150,000 km) and the vertical (radial) dimension is 1,500 km (Figure from Murray et al. 2005)

optical depths even as high as 0.1, and are consequently less well observed. The only post-*Voyager* observations, other than of the Adams ring and its arcs (see Sect. 3.5), are a marginal detection of the Le Verrier ring reported by Sicardy et al. (1999), which if confirmed would require it to be brighter than expected from *Voyager* data, and a clear detection by de Pater et al. (2005) that found the Le Verrier ring's brightness to be consistent with *Voyager* measurements. Possible explanations include temporal brightening in the Le Verrier ring followed by a return to its previous state, or unexpected spectral properties, or (a possibility they admit) that the Sicardy et al. (1999) detection was affected by image artifacts.

3.4 Diffuse Dusty Rings

Jupiter:	Halo ring
	Main ring
	Amalthea Gossamer ring
	Thebe Gossamer ring
Saturn:	D ring
	Roche Division
	Janus/Epimetheus ring
	G ring
	Methone ring
	Pallene ring
	Anthe ring
	E ring
	Phoebe ring
Uranus:	ζ ring
	ν ring
	μ ring
Neptune:	Galle ring
	Lassell ring

Every known ring system has a diffuse component that is optically thin and composed of μm-size dust particles. The dynamics and evolution of diffuse dusty rings are made more complex by the importance of forces other than gravity (Burns et al. 1979). Dust particles are commonly low enough in mass that static electrical charging makes them susceptible to electromagnetic forces comparable to gravity. They also have high ratios of surface area to volume, which makes them susceptible to pressure from solar radiation, including Poynting-Robertson drag. The evolution induced by these additional forces shortens the lifetimes of dust particles so that the stability of diffuse dusty rings is only of a dynamical variety. The ring consists of particles that originated from a source and are on their way to a sink; it may indeed look the same at a different time, if the sources and sinks have not significantly changed, but the particles comprising it will be different.

A comprehensive review of dusty ring systems was published by Burns et al. (2001).

The primary mechanism for producing orbiting dust particles is micrometeoroid bombardment of larger orbiting bodies. The rate at which a moon of radius R supplies mass to a dusty ring is given by (Burns et al. 1999)

$$\frac{\mathrm{d}M}{\mathrm{d}t} \sim \Phi Y F_\mathrm{e} R^2, \tag{7.12}$$

where Φ is the impactor flux, Y is the yield of ejected mass as a fraction of projectile mass, and F_e is the fraction of ejected mass that achieves planetary orbit. The yield Y depends on the material properties of the moon's regolith, while the fraction F_e of ejecta that can escape the moon's gravity depends on the moon's escape velocity and the distribution of ejecta velocities. Empirically, the fraction of ejecta with velocity higher than v goes as $(v_\mathrm{crit}/v)^{9/4}$, where v_crit is the minimum speed at which material is ejected, between 10 and 100 m s^{-1} where lower values are for softer moon's regoliths (Burns et al. 1984, 1999). The escape velocity of a spherical moon with bulk density ρ goes as $v_\mathrm{esc} \propto \rho^{1/2} R$ and also ranges between 10 and 100 m s^{-1}. So for $v_\mathrm{esc} < v_\mathrm{crit}$, we have $F_\mathrm{e} = 1$ and $\mathrm{d}M/\mathrm{d}t \propto R^2$, but when $v_\mathrm{esc} > v_\mathrm{crit}$, we have $F_\mathrm{e} \propto R^{-9/4}$ and $\mathrm{d}M/\mathrm{d}t \propto R^{-1/4}$. That is, there turns out to be an optimal moon radius $R_\mathrm{crit} \sim 10$ km for supplying dust to a ringlet, with smaller moons intercepting fewer impactors and thus producing less dust, while larger moons allow a smaller fraction of the produced dust to escape (❯ *Fig. 7-27*). However, these trends are only for general estimation. A more detailed treatment must consider the mechanics of ejecta production, including not only surface mechanics but also the changes in the ejecta velocity profile with increasing R as impact formation moves from the strength regime to the gravity regime (e.g., Richardson et al. 2007). Also, larger moons are more efficient at sweeping up escaped particles during subsequent encounters (Agarwal et al. 2008; Kempf et al. 2010).

All four known planetary ring systems appear to have an optically thin dusty ring as their innermost component – Jupiter's Halo ring, Saturn's D ring, Uranus' ζ ring, and Neptune's Galle ring. Dust rings often extend inward of their sources because Poynting-Robertson drag in particular enforces an inward evolution of dust particles, as does resonant charge variation inward of synchronous orbit (Burns et al. 2004). Both Jupiter's Halo/Main ring and Saturn's D ring have likely source material at their outer edges (namely, Metis and Adrastea, and the C ring, respectively), although at least the D ring has internal radial structure (Hedman et al. 2007c) that may also require embedded sources (i.e., undiscovered moons). The sources of the poorly observed ζ ring and Galle ring are not known and may also include embedded moons very close to the planet. Some of these dust sheets may extend all the way down to the planet's cloud tops, although Saturn's D ring appears to have a clear gap of 5,000 km

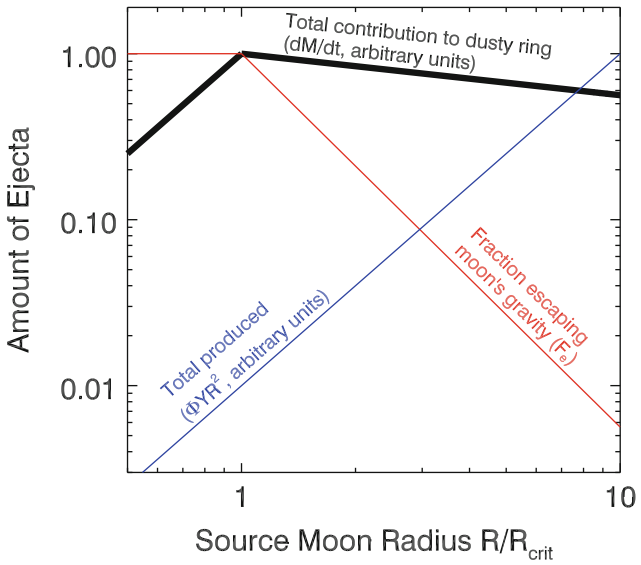


Fig. 7-27

Supply rate to a dusty ring via impact ejecta (d*M*/d*t*) is plotted in black, as a function of source moon radius *R*. This quantity is the product (7.12) of the total ejecta produced (ΦYR^2, plotted in *blue* in arbitrary units) and the fraction of ejecta that escapes the moon's gravity and achieves planetary orbit (F_e, plotted in *red*). Moons smaller than R_{crit} lose all of their ejecta, while larger moons produce more ejecta but allow a smaller fraction of it to escape. The optimal moon radius for supplying dust to a ringlet varies with surface properties but is R_{crit} ~ 10 km for icy moons with soft regolith (Burns et al. 1999)

between its inner boundary and the cloud tops (Hedman et al. 2007c), through which the *Cassini* spacecraft is slated to fly in 2017 during its end-of-mission maneuvers (Seal and Buffington 2009). Models of Jupiter's rings also find an empty region above the cloud tops, through which the *Juno* mission plans to fly in 2016, though this has not been confirmed with definitive observations.

Outward movement of dust is possible and has been invoked to explain an extension of the Gossamer ring beyond the orbit of Thebe (Hamilton and Krüger 2008). The mechanism proposed for outward movement is a "shadow resonance" in which dust particles moving through their planet's shadow temporarily lose their electrical charge and are carried outward by the momentum of their electromagnetically influenced orbits. In other locations, such as Saturn's G ring, dust evolution is primarily outward because of "plasma drag" from abundant charged particles co-rotating with the planet's magnetic field, which orbit at speeds much faster than the dust's keplerian velocity since the G ring is far outward of synchronous orbit (Hedman et al. 2007d).

Not all dusty rings consist of ejecta from a single dominant source moon. Several are tenuous extensions of nearby dense rings, such as the D ring, the dusty sheet in the Roche Division, and the Lassell ring, which all lie inward of their likely sources, respectively, the C ring, the F ring, and the Arago ring. The Phoebe ring may also be derived from multiple sources, as it is difficult

for models to account for the ring's mass from impacts onto Phoebe alone because Phoebe is so large that it should retain much of its ejecta. However, although Phoebe itself has no known collisional family, disruptive impacts are known to have broken other large irregular satellites into pieces that are separately observed today, and it is quite plausible that km-size pieces have been ejected from Phoebe and continue to share similar orbits. Since, as shown above, km-size moons are actually the most efficient sources of orbiting dust, the observed dust densities can be explained by such a distributed population of source bodies (Verbiscer et al. 2009).

The Uranian rings had a very different look (◉ *Fig. 7-28*) when the Uranian equinox (an event occurring every 42 years) was observed in 2007 by the Hubble Space Telescope (de Pater et al. 2007). With the Sun and the Earth on opposite sides of the Uranian ring plane, the brightest features in this rare view of the unlit side of the rings had relatively low optical depth but macroscopic particles. The dense ϵ ring practically disappeared because its high optical depth made it opaque, while the dusty λ ring was also dim due to the low phase angle. The brightest feature was at the location of the η ring and can probably be identified with a low-τ "extension" previously

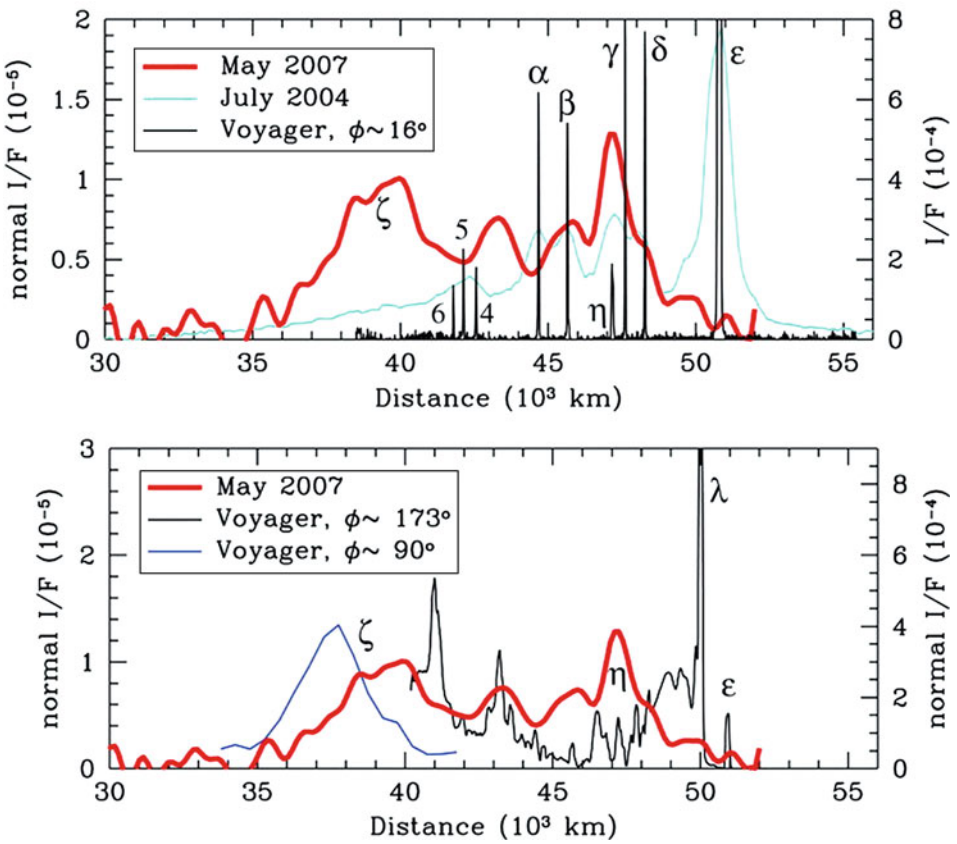
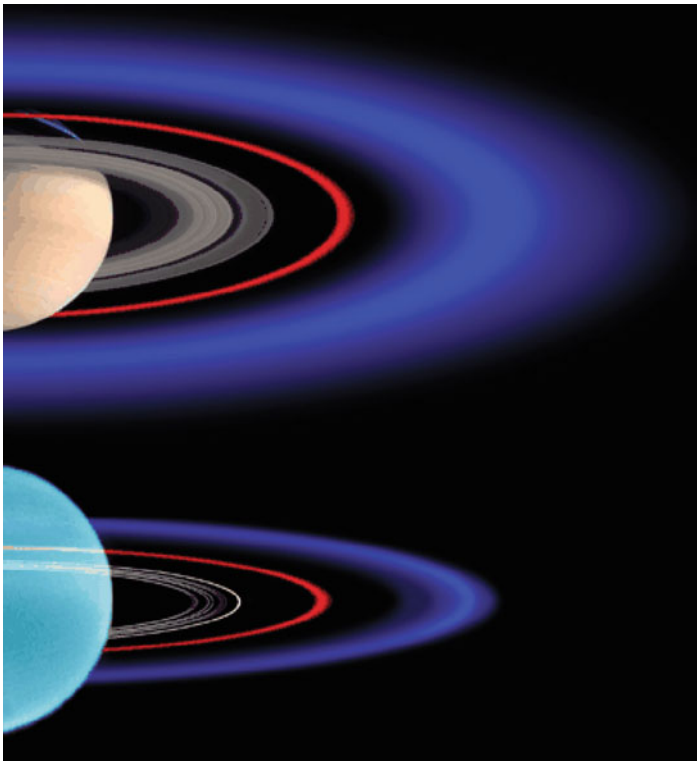


Fig. 7-28

Radial brightness profiles for Uranus rings from *Voyager* in 1986 and from the Hubble Space Telescope near (2004) and at (2007) the Uranian equinox (Figure from de Pater et al. 2007)

seen adjacent to that ring. Some of the dusty rings seen at high phase had bright counterparts in HST's unlit-side view, as did a similar "extension" to the δ ring, while others did not. Thus, the latter are likely pure dust rings, while the former contain a component of larger particles. A broad innermost ring, which had been given the provisional designation 1986 U2R based on *Voyager* images, was also seen in HST's unlit-side view and named the ζ ring. However, the core of the ζ ring as seen by HST is several thousand kilometers outward of its position in *Voyager* images. While acknowledging the possibility that overlapping particle populations with different optical properties would explain both the mid-phase *Voyager* observations and the HST data, de Pater et al. (2007) argue it is more likely that the ζ ring has changed dramatically in the intervening 20 years.

While most dusty rings are either gray or red in their spectral properties, two known rings have a prominent blue color: Saturn's E ring and Uranus' μ ring (◉ *Fig. 7-29*). Each of these rings is centered on the orbit of a moon – Enceladus and Mab, respectively (de Pater et al. 2006). For rings made of μm-sized particles, which are close in size to the wavelengths of visible and infrared light, a blue spectral trend is best explained by a particle-size distribution (regardless of



◘ Fig. 7-29
A schematic view of the outer rings of Saturn and Uranus, in which each system has been scaled to a common planetary radius. Highlighted are the red coloration of Saturn's G ring and Uranus' ν ring, both of which appear to be typical dusty rings, and the unusual blue coloration of Saturn's E ring and Uranus' μ ring (Figure from de Pater et al. 2006)

particle composition) that is either narrowly centered on a particular size or characterized by a very steep power law (Showalter et al. 1991; de Pater et al. 2004). For Enceladus and the E ring, the mechanism by which this happens seems well understood: ice particles are spewed from Enceladus south-polar geysers at a variety of speeds, with smaller particles being more easily accelerated by gas in the plume, leading to higher velocities that enable them to join the E ring (Schmidt et al. 2008; Hedman et al. 2009a; Kempf et al. 2010). However, it is harder to imagine this mechanism applying to the μ ring. Enceladus, with a diameter of ~500 km, is already so small that the source of sufficient internal heat to account for its plume is a matter of much debate (e.g., Meyer and Wisdom 2007, 2008a, b), and Mab has approximately 1/10 the diameter (and thus 1/1,000 the mass) of Enceladus. A satellite the size of Mab ought to be a good source of dust liberated by collisions (see above) but that should lead to shallower size distributions and a redder spectral trend like those of most dusty rings.

3.5 Ring Arcs and Azimuthal Clumps

Arcs	
Saturn:	G ring
	Methone ring arc
	Anthe ring arc
Neptune:	Galatea ring arc
	Adams ring

Azimuthal clumps	
Jupiter:	Main ring
Saturn:	Encke Gap ringlets
	F ring

Neptune's iconic ring arcs, known since *Voyager* and strongly suspected even earlier from occultation data, have now been joined by several Saturnian structures among the ranks of ring arcs. All share the common characteristic of an azimuthally confined region of enhanced brightness embedded within a fainter circumferential ring.[13] All known arcs orbit at the appropriate keplerian rate for their distance from planet center. But there are also significant differences among known arcs, as they cover a wide range of densities and are caused by at least two different mechanisms.

Left to itself, any clump of material orbiting a planet should spread out into a ring on a fairly short timescale. This is a direct result of keplerian shear (❯ 7.9), by which two objects with semimajor axes a and $a + \delta a$ will have mean motions n and $n + \delta n$, where $\delta n = -(3n/2a)\delta a$, and the time for one to "lap" the other by an entire orbit is

$$T_{\text{spread}} = \frac{2\pi}{\delta n} = -\frac{4\pi}{3(GM)^{1/2}} \frac{a^{5/2}}{\delta a}, \tag{7.13}$$

[13] For Saturn's Anthe and Methone arcs, as for Neptune's Galatea arc, the circumferential ring is too faint to have been detected as yet but likely consists of material recently escaped from the arc-confining mechanism.

where we have used the precise form of Kepler's Third Law, $n^2 a^3 = GM$. For the giant planets, $(GM)^{1/2}$ is of order 10^4 km$^{3/2}$ s^{-1}, and a typical distance from the planet is $a \sim 10^5$ km. Therefore, even for a compact initial clump $\delta a \sim 1$ km, the spreading time is only a few decades and is inversely proportional to δa.

This exercise demonstrates that azimuthal variations in a ring's mass are not intrinsically stable. Therefore, the several examples that exist of observed ring arcs must be dynamically generated or maintained. The two most prominent mechanisms for this are resonant confinement and asymmetric injection of mass into the ring.

Corotation resonances have resonance argument (see Sect. 3.1.1) of the form

$$\varphi = (m + k)\lambda' - m\lambda - kX', \tag{7.14}$$

where, as for (7.6), m and k are integers that label the resonance as $(m+k)$:m, primed quantities refer to the forcing moon, and X is either ϖ or Ω. For a corotation eccentricity resonance (CER), X is ϖ and the resonance strength is proportional to the perturbing moon's eccentricity, while for a corotation inclination resonance (CIR), X is Ω and the resonance strength is proportional to the perturbing moon's inclination.

Corotation resonances differ from their Lindblad and vertical cousins (see Sect. 3.1.1) in that the lowest-order resonances involve the eccentricity or inclination of the perturbing moon, rather than that of the particle being perturbed. Thus, rather than pumping up the eccentricities and inclinations of ring particles and driving waves, corotation resonances tend to azimuthally confine particles into an orbit commensurate with that of the perturbing moon. This mode of confinement was first proposed for Neptune's ring arcs by Goldreich et al. (1986).

3.5.1 Neptune's Adams Ring

The first-discovered and best-known set of ring arcs are found in Neptune's Adams ring. These are the densest components of Neptune's ring system (with $\tau_\perp \sim 0.1$) and were oftentimes the only component detectable in pre-*Voyager* Earth-based occultations, leading to much confusion as to whether the detected signatures were rings at all until *Voyager 2* settled the question. There are five arcs occupying ~20° out of a region extending over ~40° of longitude (*Figs. 7-9* and *7-30*). The three main arcs were named for the French revolutionary slogans Liberté, Egalité, and Fraternité, then closer inspection showed a bifurcation in Egalité and a dimmer fourth arc that was named Courage.

The first integrated dynamical explanation of the arcs was given by Porco (1991), who combined the theoretical framework of Goldreich et al. (1986) with observations showing that the Adams ring lies very close to a 43:42 resonance with the moon Galatea. Porco (1991) suggested that the azimuthal confinement mechanism was due to Galatea's 43:42 corotation inclination resonance (CIR) while radial spreading due to particle collisions was prevented by Galatea's nearby 43:42 outer Lindblad resonance (OLR). This explanation requires both Liberté and Egalité to stretch over multiple consecutive corotation sites (each of which is only[14] $360°/86 = 4.2°$ long) and predicted that the arcs would orbit at the pattern speed of the 43:42 CIR. A synthesis of all *Voyager* and Earth-based data (Nicholson et al. 1995) found two possible solutions for

[14] Because inclination resonances cannot be first-order (see, e.g., Sect. 10.3.3 of Murray and Dermott 1999), the CIR actually functions as an 86:84 resonance, which is why the number 86 appears here. For the CER discussed by Namouni and Porco (2002), the corotation sites would be twice as long.

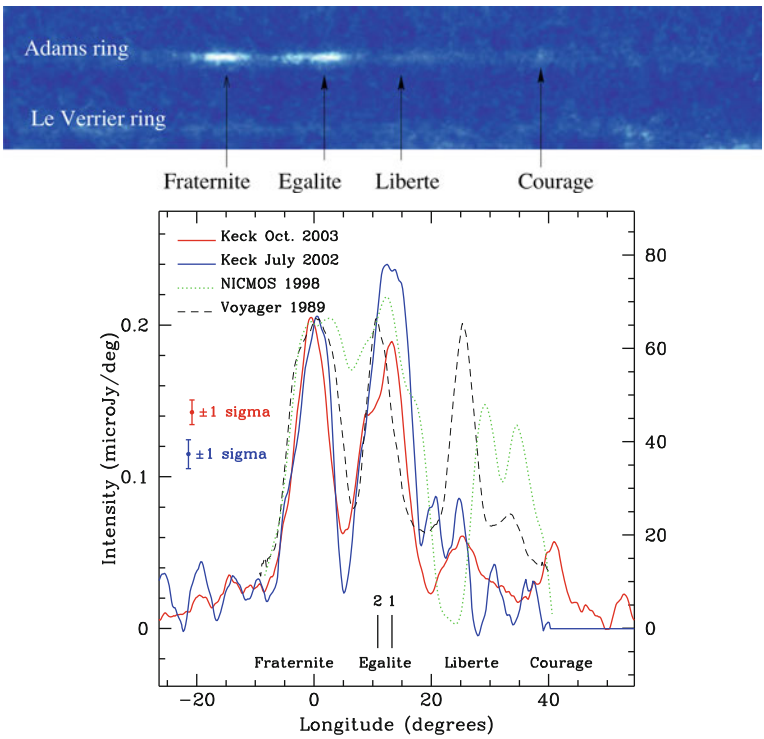


Fig. 7-30

(*upper*) Reprojected ground-based image of the Adams ring (with arcs) and the Le Verrier ring acquired in October 2003. (*lower*) Azimuthal profiles of the Adams arcs from four separate observations. The leading arc Liberté was seen by *Voyager* 2 (black dashed line) to be as bright as the other two main arcs, but subsequently dimmed (Figure from de Pater et al. (2005). See *Fig. 7-9* for schematic view)

the arcs' pattern speed, one of which was consistent with the Galatea 43:42 CIR, and detailed dynamical work (Foryta and Sicardy 1996; Hänninen and Porco 1997) supported the plausibility of resonant confinement by this mechanism. However, the reacquisition of the arcs with Earth-based imaging (Dumas et al. 1999; Sicardy et al. 1999) called the resonant pattern speed into question, favoring instead Nicholson et al.'s (1995) other solution. An attempt to salvage the Galatea model was made by Namouni and Porco (2002), who proposed an alternative model employing the 43:42 corotation eccentricity resonance (CER) rather than the CIR. They furthermore invoked the mass of the ring arcs themselves to adjust the resonant pattern speed to match the observations, thus deriving a value for the arcs' mass that is required for their model to work. Further Earth-based observations (de Pater et al. 2005) not only confirmed that the arcs are moving at the "wrong" pattern speed for the original Porco (1991) model but showed significant changes in the arcs' structure over the 20 years in which they have been observed in detail (*Fig. 7-30*). The brightness of Liberté, already diminished in the 1998 observations, had further declined until it was dimmer than Courage, while Courage had moved forward in longitude

relative to the other arcs. Egalité has also undergone less dramatic shifts in its morphology and longitude, while Fraternité appears largely stable.

The Adams arcs remain an enigma. While the model invoking resonant confinement by Galatea has appeared promising, the details have not come together as hoped. The original model by Porco (1991) was appealing because its predicted pattern speed appeared to match very closely with the observations; however, the reworking of the model by Namouni and Porco (2002) is less convincing because it requires the invocation of an additional free parameter in order to fit the data. Furthermore, now that other examples of resonantly confined ring arcs have come to light (see Sect. 3.5.3), all of which have the morphology predicted by Goldreich et al. (1986), with maxima at the center of a corotation site and brightness falling off long before the site's edges are reached, the objection can be raised anew that Egalité and Fraternité span multiple corotation sites while showing no clear evidence of internal minima at the expected spatial frequency. On the other hand, the azimuthal structure in Jupiter's Main ring (Sect. 3.5.2), though far less well observed, appears similar to the Adams arcs in that the spatial frequencies do not match those of nearby corotation resonances. Perhaps the corotation-resonance model will be vindicated in the end, or perhaps the answer will be more like the shepherding model of Lissauer (1985), which invokes yet-undiscovered moons embedded within the Adams ring, or perhaps the Adams arcs will be found not to be long-term stable structures.

3.5.2 Jupiter's Main Ring and Other Azimuthal Clumps

In the core of Jupiter's Main ring, *New Horizons* images found one close pair of azimuthal clumps and another family of three to five clumps (Showalter et al. 2007). The semimajor axes of the α and β clump families, measured from their orbital rates, fall less than 1 km from, respectively, the Metis 115:116 and 114:115 corotation inclination resonances (CIR), the same kind of resonance originally invoked by Porco (1991) for Neptune's Adams arcs. Given the distance between the two semimajor axes and the distance between the resonances, Showalter et al. (2007) estimate the probability of this happening by coincidence to be 4% . Finally, to complete the analogy with Neptune's Adams arcs, the 1.8° azimuthal spacing of the clumps does not correspond to the expected $360^\circ/230 = 1.56^\circ$ for these resonances (Showalter et al. 2007).

The Adams ring and the Main ring may turn out to be more like Saturn's F ring and the ringlets in the Encke Gap (Sect. 3.3), which have rich azimuthal structure that is clearly not associated with a resonant spatial frequency and is most likely due to embedded source moons.

3.5.3 Saturn's G Ring and Other Moon-Embedded Arcs

Saturn's G ring was, for a long time, the ring that should not be there. Saturn's main disk was clearly a long-term stable structure, the D ring clearly derived from it, the E ring associated with Enceladus, and the F ring confined by Prometheus and Pandora. Yet the G ring, composed of dust grains that cannot persist over long time periods, had no apparent mechanism for its needed continuous generation.

The first step toward addressing this question came with the discovery of a relatively bright arc within the G ring (Hedman et al. 2007d). The arc is brighter than the rest of the G ring by a factor of several (but still at $\tau_\perp \sim 10^{-6}$) and is radially much narrower (~250 km). It is very plausible that the rest of the G ring is composed of grains evolving away from the arc

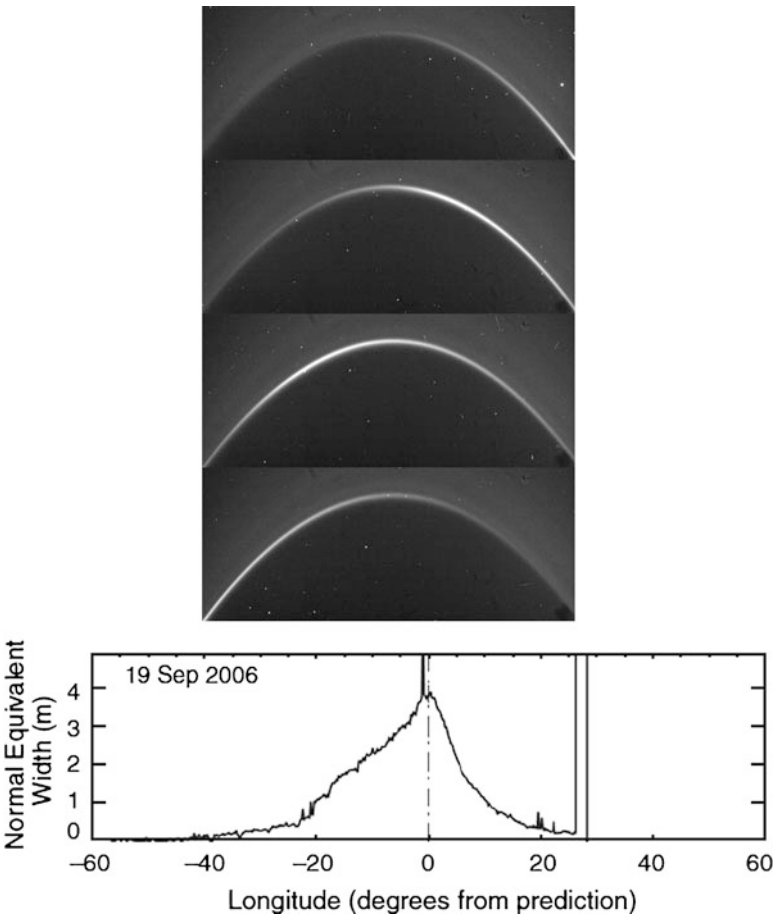


Fig. 7-31
(*top*) Images and (*bottom*) azimuthal profile of the G ring arc, centered on the Mimas 7:6 corotation site. The peaky shape in the azimuthal profile indicates a significant fraction of arc material is tightly bound to the resonance (i.e., has low libration amplitude) (Figure from Hedman et al. 2007d)

under the influence of nongravitational forces. Furthermore, the arc's orbital rate matches the 7:6 corotation eccentricity resonance (CER) with Mimas, again invoking the Goldreich et al. (1986) mechanism. The azimuthal profile of the arc is roughly triangle shaped (*Fig. 7-31*), consistent with a source body with only a small-amplitude libration about the exact resonance, but has broad wings that plausibly fill the corotation site's length of $360^\circ/7 = 51^\circ$ in longitude.

A further piece of the puzzle came with the discovery of Aegaeon (Hedman et al. 2010c), a 1-km moon orbiting within the G ring arc. Aegaeon is probably only the largest of a population of source objects within the arc, as electron-absorption measurements indicate massive objects spread over a distance much larger than Aegaeon itself (Hedman et al. 2007d).

Two other arcs, even more tenuous than the G ring's, surround the small moons Anthe and Methone, both situated between Mimas and Enceladus. A third moon in this vicinity, Pallene,

appears to have a very tenuous ring but not an arc. The Methone arc was first detected by charged-particle absorptions (Roussos et al. 2008), and all three structures were then seen in imaging data (Hedman et al. 2009b). Both Anthe and Methone are in resonance with Mimas, in the 11:10 and 15:14 CERs, respectively, and these resonances continue to confine the arc material once it has left the source moons (Agarwal et al. 2009). The azimuthal lengths of the Anthe and Methone arcs are both approximately half that of the available corotation site (Hedman et al. 2009b), but that may simply reflect the point at which the faint signal falls below detectability.

Because these ring arcs at Saturn are so much more tenuous than Neptune's, they are essentially collisionless. This allows them to be confined by the corotation resonance alone, which is necessary because the associated Lindblad resonances are radially farther away than they are for Neptune's case, both because of Saturn's higher J_2 compared to Neptune and because the resonances have lower azimuthal parameter m, and thus are not available to perform the radial confinement that was an essential part of the Goldreich et al. (1986) model for the collisional Adams ring.

Finally, we note that *Voyager 2* imaged a faint unnamed ring sharing the orbit of the moon Galatea, which may be an arc given its intermittent detectability (Showalter and Cuzzi 1992; Porco et al. 1995). Although little is known about this structure, it can now be placed in context with the Anthe and Methone arcs and may very well be driven by similar mechanisms.

3.6 Rings as Detectors

Planetary rings have shown themselves to be useful, in many cases, as detectors of planetary processes around them.

Spiral structures in the D ring, and in the similar tenuous dusty sheet in the Roche Division, are driven by Lindblad resonances with the rotation period of the planet's magnetic field. These structures are only seen in these tenuous sheets populated by tiny grains that are easily charged and thus subject to electromagnetic forces. The multiple pattern speeds required to explain the resonant structures are a major source of information for the complex rotation of Saturn's enigmatic magnetic field (Hedman et al. 2009c).

Evidences for vertical corrugations in Jupiter's Main ring were first seen in *Galileo* images (Ockert-Bell et al. 1999; Burns et al. 1999) but were not well understood. Next, *Cassini* images of Saturn's D ring showed evidence of a vertical corrugation whose radial wavelength is decreasing with time, a trend easily accounted for with a model of differential precession that begins with the ring as an inclined flat sheet ~20 years before *Cassini*'s arrival at Saturn (Hedman et al. 2007c). Images obtained during Saturn's 2009 equinox showed the vertical corrugation pattern extending far into the C ring, and the variation of the wavelength with radius confirms the previous result that the pattern originated in 1983 when an event of some kind caused the ring to be tilted by $\sim 10^{-7}$ radians (a few meters) with respect to the Laplace plane (Hedman et al. 2011d). Finally, similar analysis of the corrugations in Jupiter's Main ring yields a superposition of wavelengths implying two tilting events, one in July 1994 and one in 1990 (Showalter et al. 2011b). July 1994 is, of course, the date of the impact of comet Shoemaker-Levy 9 (SL9) into Jupiter, leading to the hypothesis that both corrugation patterns carry records of the impact of a spatially dispersed cloud, which is one mechanism for spreading the impulse over a large portion of the ring. The cloud of dust surrounding SL9 would likely fill the bill at Jupiter, and either a similar cometary system or a meteor stream could be the cause of the Saturn event.

Saturn's rings showed themselves capable of more direct detections of impactors during the 2009 equinox event. Bright markings, canted with respect to the azimuthal direction, were seen on both the A ring and the C ring during the few days before and after the Sun's passage through the ring midplane (Tiscareno et al. 2009b). Assuming that these are dust clouds evolving under keplerian shear (7.9), one can calculate the time elapsed since the cloud was radially aligned. In one case, the same cloud was seen 24 h apart, and the ages derived from keplerian shear differed by the same interval, confirming this interpretation of the observed structures. Also, in very high-phase high-resolution images of the C ring (not during equinox), small streaks appear that are probably transitory dust clouds produced by impacts (Tiscareno et al. 2009b). Both of these discoveries have the potential to use rings observations to constrain the influx of interplanetary impactors, though the derivation of impactor size from observed dust-cloud parameters has yet to be worked out.

Spiral waves and wavy gap edges are both the result of gravitational forcing due to moons. In the case of Janus and Epimetheus, the nature of the forcing changes with time, and the ring maintains a record of that change. Initial steps toward understanding the nature of the rings' record have been made for both density waves (Tiscareno et al. 2006b) and wavy edges (Spitale and Porco 2009). In the case of spiral density waves, the group velocity at which information propagates through the wave is slow enough that information can be gained about the state of the co-orbital moons as much as 10 years prior to the time of observation.

4 Experimental Rings Science

4.1 Numerical Simulations

Advances in simulation techniques, alongside advances in computing hardware and software, have allowed numerical simulations to become an increasingly important part of the study of planetary rings. Modern *n*-body dynamics exploded after Wisdom and Holman (1991) published a symplectic mapping that uses Hamiltonian methods to calculate perturbations as a deviation from keplerian orbit, rather than as deviations from motion in a straight line, and thus requires fewer correction events per unit of simulated time. This, along with ever-increasing computer power, allowed large *n*-body simulations to become routine; among the many solar system applications is the case of dusty rings evolving under gravity and other forces. For dense rings, though, the number of mutually interacting particles is too large for simulations that follow particles along their full orbits about the planet. Instead, for cases that do not focus on large-scale azimuthal structure, a very productive line of simulations follows a relatively small "patch" of the ring with sliding boundary conditions as devised by Wisdom and Tremaine (1988). The patch is surrounded by "mirror" patches (*Fig. 7-32*); those on either side in the azimuthal direction are stationary, so that a particle that leaves the patch on one side simply reappears on the opposite side, but the neighboring patches in the radial direction slide past according to keplerian shear (7.9). Even with their greatly reduced spatial dimensions, ring-patch simulations routinely include so many mutually interacting particles that they push the limits of the available computing power, and several approaches to efficiently accounting for their mutual interactions have been devised.

The first great success of ring-patch simulations, and a continuing area of their usefulness, is the characterization of self-gravity wakes (SGWs) (Sect. 3.1.4). First described by

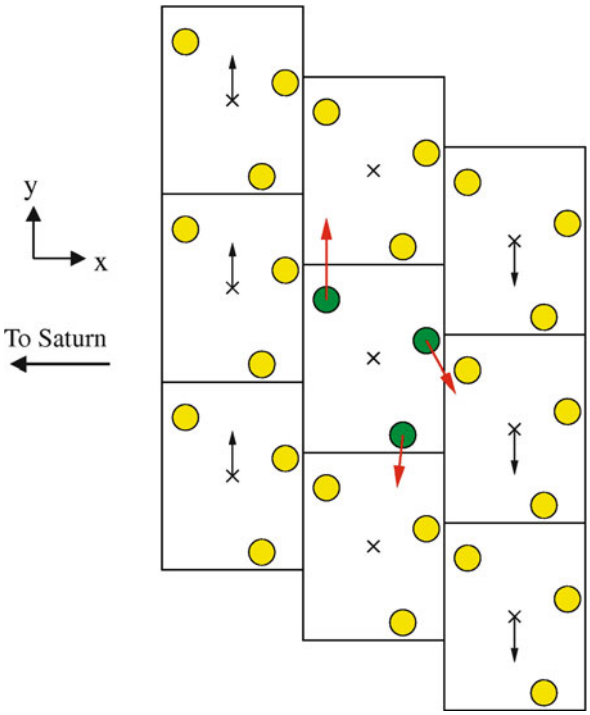


Fig. 7-32
Schematic representation of a ring-patch simulation with sliding boundary conditions. The simulation cell (*center*, with *green particles*) is replicated on all sides, with the replicant cells (with *yellow particles*) positioned according to relative keplerian velocities (7.9). In this representation, increasing radius ($+\hat{x}$) is to the right and keplerian orbital motion ($+\hat{y}$) is up (Figure from Perrine et al. 2011)

Julian and Toomre (1966) for the case of galaxies, it was understood by the 1980s that this non-axisymmetric structure due to a balance between accretion and disruption was likely present in Saturn's A and B rings and was likely responsible for the observer-centered quadrant azimuthal asymmetry observed in the A ring's brightness (Colombo et al. 1976; Franklin et al. 1987; Dones and Porco 1989). SGWs were successfully produced in ring-patch simulations by Salo (1992, 1995) and by Richardson (1994), and simulations continue to be a vital tool for understanding the mechanics and structure of SGWs and comparing their simulated photometric properties to observational data, as well as other ring properties including the rotational states and thermal properties of ring particles, and the mechanics of sharp edges and propellers (for details and references, see Schmidt et al. 2009).

An entirely different class of simulations are the semianalytical streamline models of Hahn (2007, 2008) and Hahn et al. (2009), which use the theoretical framework of Borderies-Rappaport and Longaretti (1994, and references therein) to probe the distribution of surface density, pressure, and viscosity of a ring in the vicinity of a sharp edge.

4.2 Physical Experiments and the Coefficient of Restitution

What happens when two ring particles collide? Gentle collisions, with velocities of order mm s^{-1}, occur constantly (frequencies of order the orbital frequency) within dense rings. A law describing the coefficient of restitution ε (the ratio between outgoing and incoming kinetic energies for two colliding particles in their center-of-mass reference frame) is an essential input for numerical simulations. A number of physical experiments have been conducted to measure the coefficient of restitution directly, but these must begin with assumptions as to the shape, porosity, and surface friction of ring particles. Consequently, the results of these experiments have been inconsistent. More recently, comparisons between simulations and data have attempted to arrive empirically at a favored coefficient of restitution law, from which ring-particle properties can then be inferred.

Bridges et al. (1984) conducted groundbreaking experiments for determining the collisional properties of icy objects, using a double-pendulum apparatus to achieve the exceedingly low collision velocities seen in ring systems. They found relations for ε_n, as a function of mutual incoming velocity v_n (the subscript denotes normal collisions), for frosty ice spheres at temperatures of ~200 K. Their result,

$$\varepsilon_n = \left(\frac{v_n}{v_c}\right)^b , \tag{7.15}$$

with exponent $b = -0.234$ and critical velocity $v_c = 0.0077\,\mathrm{cm\,s^{-1}}$, has been widely used in numerical simulations ever since. Further experiments broadened the range of particle properties tested. Hatzes et al. (1988, 1991) found that a frost layer on the surface can greatly increase the lossiness of a collision (i.e., depress ε), even leading to sticking at low collision velocities. Dilley and Crawford (1996) varied the mass of the incoming particles and found that smaller ice balls led to lossier collisions. Supulver et al. (1995) found less lossy collisions overall ((❯ 7.15) with $b = -0.14$ and $v_c = 0.01\,\mathrm{cm\,s^{-1}}$) for frost-free spheres at ~100 K and also found that glancing collisions are less lossy than normal collisions. Initial results from experiments in a microgravity environment were given by Colwell et al. (2008) and Heißelmann et al. (2010), while Durda et al. (2011) measured the coefficient of restitution for two 1-m granite spheres colliding at low speeds.

All collision experiments thus far have represented ring particles as hard spheres of H_2O ice. The actual shape of ring particles is not known; they are far too small to be sphericalized by hydrostatic equilibrium and might be expected to be roughly ellipsoidal if they take the shape of their Roche lobes. Their surfaces may be smoothed by the numerous gentle collisions they experience, after the manner of pebbles in a stream bed, but this also is not known. Finally, ring particles are almost certainly not as hard as solid ice, as their outer layers at least are probably quite porous (❯ Sect. 1.2).

Eventually, the role of physical experiments will ideally shift from determining a preferred restitution law based on an assumption of ring-particle properties, to using an empirically indicated restitution law to infer ring-particle properties. A step toward this goal was taken by Porco et al. (2008), who compared numerical simulations with *Voyager* imaging data of the azimuthal asymmetry to arrive at a preliminary conclusion favoring a lower coefficient of restitution (i.e., lossier collisions) than given by the prevailing Bridges et al. (1984) law. They favored instead a law with the form (Borderies et al. 1984)

$$\varepsilon = \left[-\frac{2}{3}\left(\frac{v^*}{v}\right)^2 + \left[\frac{10}{3}\left(\frac{v^*}{v}\right)^2 - \frac{5}{9}\left(\frac{v^*}{v}\right)^4 \right]^{1/2} \right]^{1/2} . \tag{7.16}$$

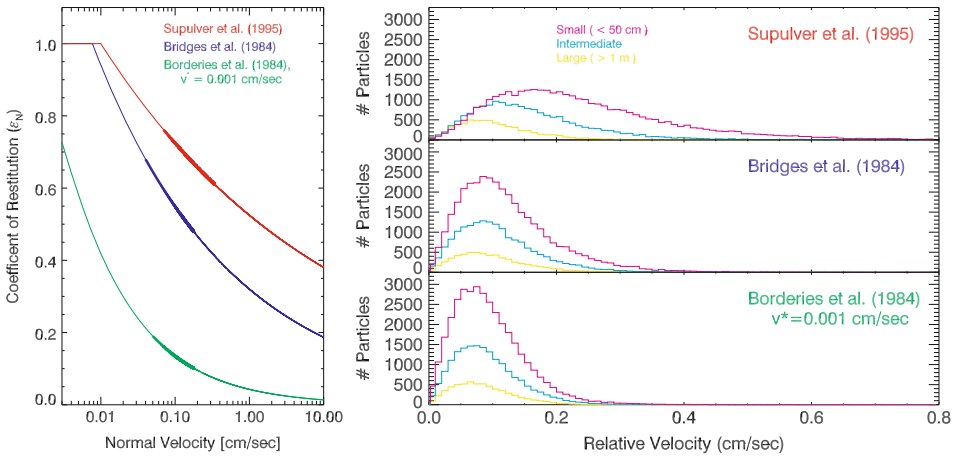


Fig. 7-33

(*left*) Coefficients of restitution, as a function of incoming velocity, according to different input laws. The bold portion of each line indicates the velocity range for most simulated small particles. (*right*) Velocity dispersion profiles for simulated ring patches using the same input laws (Figures modified from Porco et al. 2008)

This law is based on the Andrews (1930) theory of colliding spheres, but the more flexible formulation of Borderies et al. (1984) allows v^* to be a free parameter adjusting the law's lossiness. Porco et al. (2008) favored a value of $v^* = 0.001\,\mathrm{cm\,s^{-1}}$; however, a full treatment of the question, incorporating the substantial *Cassini* data set, has yet to be completed. A plot of coefficient of restitution distributions derived from a variety of laws is shown in ❯ *Fig. 7-33*.

In addition to the azimuthal photometric asymmetry, both the photometry of propellers (❯ Sect. 3.1.5) and the sharpness of gap edges (❯ Sect. 3.1.2) have the potential of being used to constrain the physical properties of ring particles by combining data with simulations, though work on these lines is just beginning. Propellers are large enough to appear in images with a reasonable amount of detail, but small enough that particle-particle interactions play a significant role in determining their structure (e.g., Lewis and Stewart 2009), although their poorly understood photometry (Tiscareno et al. 2010c), possibly due to vertical structure, may make it difficult to use them as a standard. Occultation data infer edges in Saturn's rings as sharp as 10–20 m (Colwell et al. 2010), significantly sharper than the edges obtained in simulations using the standard Bridges et al. (1984) restitution law (Weiss 2005), while preliminary results from simulations with lossier restitution laws yield sharper edges more in line with observations (J.E. Colwell 2010, personal communication).

4.3 Spectroscopic Ground Truth

The project of characterizing the optical properties of materials occurring in the outer solar system is ongoing. A large amount of spectroscopic data now exists for the rings of Jupiter and Saturn, in spectral ranges from the infrared to the ultraviolet, that contain numerous bands that are potentially diagnostic of ring-particle composition and/or particle size and state. Laboratory measurements of candidate ring-forming materials are compared to these

observations in an effort to constrain the chemical and physical composition of the rings. For details and references, see Cuzzi et al. (2009).

5 Age and Origin of Ring Systems

The question of the age of a ring system is actually two separate questions, one being the age of the material and the other the age of the structure currently in place. This distinction is particularly important for diffuse dusty rings (Sect. 3.4). For example, if Amalthea, Thebe, Adrastea, and Metis were to suddenly disappear, or to suddenly cease emitting dust, Jupiter's rings would dissipate on a timescale on the order of 10^5 years at the most (Burns et al. 2001). Thus, the age of individual particles in a diffuse ring is on the order of the mean residence time, while the age of the overall structure is related to how long the sources have been in place.

The distinction between material age and overall structural age also turns out to be useful for Saturn's main rings. Several aspects of the rings are difficult to reconcile with a ring age comparable to that of Saturn (see Charnoz et al. 2009a, and references therein), including (1) the $\gtrsim$95% water ice composition of the A and B rings is difficult to reconcile with the constant pollution of the rings by infall of interplanetary micrometeoroids; (2) the same infall should significantly erode ring particles, especially in regions of lower optical depth; and (3) exchange of angular momentum with moons currently confining ring edges can only be "rewound" for ~10^7 year. On the other hand, the disruption of a Sun-orbiting object containing sufficient mass in such a way as to form a ring system is an unlikely event given the recent state of the solar system, and Saturn is actually the least likely of the planets to capture an interloper because the balance of mass and solar distance gives it the smallest Hill sphere among the giant planets, leading to some doubts as to the viability of the young-rings scenario (Charnoz et al. 2009b).

One potentially viable theory suggests that the B ring core is ancient, dating from the first Gyr of the solar system in which collisions were much more frequent (Charnoz et al. 2009b), but that much of the specific organization of material in Saturn's rings is only ~10^7 year old. Recent indications that the B ring's mass has previously been underestimated (see Sect. 3.1.4 and Robbins et al. 2010) provide an increased buffer against interplanetary pollution (Elliott and Esposito 2011), as well as making the ring precursor body even larger and thus a recent ring origin even more unlikely. Erosion of ring particles can be counteracted by an accretion-driven recycling process (Durisen et al. 1989; Esposito 2006) and is also slowed if ring particles have spent most of their history in a denser structure than those where they are now found. The current ring-moons may also have formed within the rings and emerged only recently (Charnoz et al. 2010). Many questions still remain, however (Charnoz et al. 2009a). The mass of the B ring needs to be better known, as may well happen as a result of the planned close passes during *Cassini*'s end-of-mission maneuvers (Seal and Buffington 2009) both through direct sensing of the ring's gravitational pull and through its interaction with charged particles and gamma rays (the latter improving on the *Pioneer 11* measurement reported by Cooper et al. 1985). Also needed are improvements in knowledge of the interplanetary impactor flux throughout the history of the rings. The proposed recycling mechanism has yet to be carefully described and modeled. And any theory for the formation of Saturn's rings, whether the age be young or old, requires careful separation of water ice from any silicate and/or metal that must have been part of any body accreting directly from the protoplanetary nebula.

Canup (2010) recently suggested that Saturn's rings might have been formed by the disruption of a Titan-sized (~2,500-km radius) proto-moon near the end of the planetary formation period during which the circumplanetary gas disk is present. Due to its size, the proto-moon is differentiated, and as it spirals inward due to interaction with the disk, it first passes the Roche radius for ice (see ❯ Sect. 1.2) and its icy mantle is stripped away. Furthermore, because the incompletely consolidated Saturn would have been ~50% larger than at present, the proto-moon is engulfed by the planet before reaching the Roche radius for its rocky core, thus explaining the rings' icy composition. The initially resulting ring is ~1,000 times more massive than that of today, but Salmon et al. (2010) showed that viscous spreading of a ring over the age of the solar system can lead to a ring like that of today with relatively little sensitivity to the ring's initial mass (❯ *Fig. 7-34*). Supersizing the argument of Charnoz et al. (2010), Canup (2010) suggested that some of Saturn's midsize moons (with radii of hundreds of kilometers) may have been spawned by the viscous spreading of this massive ring, accounting for their relatively low densities (cf. ❯ Sect. 1.2), and Charnoz et al. (2011) in turn have explored this scenario in more detail. The details of the Canup (2010) simulations are conducted in the context of that author's theory for circumplanetary disks (e.g., Canup and Ward 2002, 2006), which has been criticized by some (e.g., Mosqueira and Estrada 2003). However, the basic outline of the Canup (2010) story appears to have plausible merit even beyond its specific theoretical context.

The rings of Uranus and Neptune, with their lower masses and much darker surfaces, have more plausibly remained little changed over the age of the solar system. The Uranian rings, as well as the nearby group of moons, may be part of a system that has oscillated between accretion and disruption for many Gyr (❯ Sect. 2.3). But why are the rings where they are? Is the dynamical environment such that a uniform supply of material at all distances from the planet would result in the rings as we see them today? Or was the supply of source material somehow confined to the locations at which we now see rings? On the planetary level, why does Saturn have its glorious broad dense disk while Uranus and Neptune have much more modest systems and Jupiter has only moon-generated dust? None of these questions has a clear answer as yet.

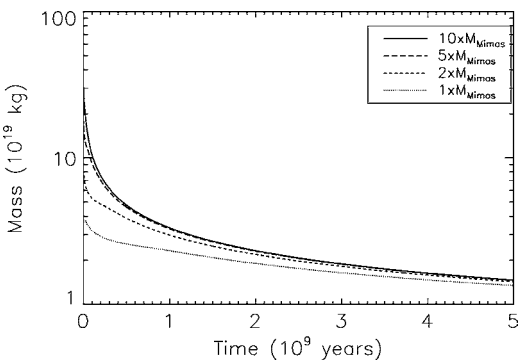


Fig. 7-34

Viscous spreading models predict that disk mass after 5 Gyr of evolution is relatively insensitive to the initial disk mass (Figure from Salmon et al. 2010)

6 Rings and Other Disks

Planetary rings are just one variety among many disk-shaped systems known to astronomers, but they are the only variety that is not exceedingly far away in time or space or both and thus that is available for close inspection (Burns and Cuzzi 2006). Other astrophysical disks include proto-planetary, protolunar, and protosatellite disks as theorized for the origins of our solar system and as presently observed as gas disks, dust disks, and debris disks at other stars. They also include accretion disks for binary stars, black holes, active galactic nuclei, etc.

Some examples of fruitful cross-pollination between planetary rings studies and other astrophysical disks follow. The interpretation of the inner edge of the Fomalhaut disk in terms of a resonant confinement as seen in planetary rings (Kalas et al. 2005) was validated by the discovery of Fomalhaut b (Kalas et al. 2008). The dynamics of eccentric rings like Uranus' ϵ ring have been extended and applied to astrophysical eccentric disks such as "superhump" binary star systems (Lubow 2010). Both spiral waves (Sect. 3.1.1) and self-gravity wakes (Sect. 3.1.4) were first proposed as physical processes likely to occur in galactic disks, with specific applications to planetary rings coming a decade or more later. But the shoe is now on the other foot, with direct observations of spiral waves and of SGWs in planetary rings having become so detailed that they can potentially inform understanding of similar processes occurring in galaxies. A similar process may soon occur with free unstable normal modes, which have long been seen in numerical simulations of protoplanetary disks (Laughlin et al. 1997) and have now likely been observed directly in Saturn's B ring (Spitale and Porco 2010). Finally, the orbital evolution of disk-embedded "propeller" moons and the nature of their interaction with the disk is only just beginning to be directly observed in Saturn's rings (Sect. 3.1.5), potentially shedding light on the evolution of planetesimals and other disk-embedded masses.

Acknowledgments

I thank Mark Showalter, Joe Burns, Josh Colwell, Jeff Cuzzi, Jonathan Fortney, Doug Hamilton, Matt Hedman, Doug Lin, Phil Nicholson, and John Weiss for helpful conversations. I additionally thank Robin Canup, John Cooper, Estelle Deau, Larry Esposito, and Rob French for valuable comments on the manuscript, and Hanno Rein for help in creating *Fig. 7-3*. I acknowledge funding from NASA Outer Planets Research (NNX10AP94G), NASA Cassini Data Analysis (NNX08AQ72G and NNX10AG67G), and the Cassini Project.

References

Acuna, M. H., & Ness, N. F. 1976, The main magnetic field of Jupiter. J. Geophys. Res., 81, 2917–2922

Agarwal, M., Tiscareno, M. S., Hedman, M. M., & Burns, J. A. 2008, Dynamics of faint rings associated with Methone, Anthe and Pallene. AAS Div. Planet. Sci. Meet. Abstr., 40, 30.02

Agarwal, M., Tiscareno, M. S., Hedman, M. M., & Burns, J. A. 2009, Dynamics of rings and arcs associated with three small moons of Saturn: Methone, Anthe and Pallene. AAS Div. Dyn. Astron. Meet. Abstr., 40, 3.05

Albers, N., Sremčević, M., Colwell, J. E., & Esposito, L. W. 2012, Saturn's F ring as seen by Cassini UVIS: kinematics and statistics. Icarus, 217, 367–388

Alexander, A. F. O. 1962, The Planet Saturn: A History of Observation, Theory and Discovery (London: Faber and Faber)

Andrews, J. P. 1930, Theory of collision of spheres of soft metals. Phil. Mag. Ser. 7, 9, 593–610

Barnes, J. W., & Fortney, J. J. 2004, Transit detectability of ring systems around extrasolar giant planets. Astrophys. J., 616, 1193–1203

Beurle, K., Murray, C. D., Williams, G. A., Evans, M. W., Cooper, N. J., & Agnor, C. B. 2010, Direct evidence for gravitational instability and moonlet formation in Saturn's rings. Astrophys. J. Lett., 718, L176–L180

Borderies, N., & Longaretti, P. Y. 1987, Description and behavior of streamlines in planetary rings. Icarus, 72, 593–603

Borderies-Rappaport, N., & Longaretti, P.-Y. 1994, Test particle motion around an oblate planet. Icarus, 107, 129–141

Borderies, N., Goldreich, P., & Tremaine, S. 1983, The dynamics of elliptical rings. Astron. J., 88, 1560–1568

Borderies, N., Goldreich, P., & Tremaine, S. 1984, Unsolved problems in planetary ring dynamics, in Planetary Rings, ed. R. Greenberg, & A. Brahic (Tucson: University of Arizona Press), 713–734

Bosh, A. S., Olkin, C. B., French, R. G., & Nicholson, P. D. 2002, Saturn's F ring: kinematics and particle sizes from stellar occultation studies, Icarus, 157, 57–75

Bridges, F. G., Hatzes, A., & Lin, D. N. C. 1984, Structure, stability and evolution of Saturn's rings. Nature, 309, 333–335

Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W., & Burrows, A. 2001, Hubble Space Telescope time-series photometry of the transiting planet of HD 209458. Astrophys. J., 552, 699–709

Burns, J. A., & Cuzzi, J. N. 2006, Our local astrophysical laboratory. Science, 312, 1753–1755

Burns, J. A., Lamy, P. L., & Soter, S. 1979, Radiation forces on small particles in the solar system. Icarus, 40, 1–48

Burns, J. A., Showalter, M. R., & Morfill, G. E. 1984, The ethereal rings of Jupiter and Saturn, in Planetary Rings, ed. R. Greenberg, & A. Brahic (Tucson: University of Arizona Press), 200–272

Burns, J. A., Schaffer, L. E., Greenberg, R. J., & Showalter, M. R. 1985, Lorentz resonances and the structure of the Jovian ring. Nature, 316, 115–119

Burns, J. A., Showalter, M. R., Hamilton, D. P., Nicholson, P. D., de Pater, I., Ockert-Bell, M. E., & Thomas, P. C. 1999, The formation of Jupiter's faint rings. Science, 284, 1146–1150

Burns, J. A., Hamilton, D. P., & Showalter, M. R. 2001, Dusty rings and circumplanetary dust: observations and simple physics, in Interplanetary Dust, ed. E. Grün, B. Å. S. Gustafson, S. Dermott, & H. Fechtig (Berlin: Springer), 641–725

Burns, J. A., Simonelli, D. P., Showalter, M. R., Hamilton, D. P., Porco, C. D., Throop, H., & Esposito, L. W. 2004, Jupiter's ring-moon system, in Jupiter: The Planet, Satellites and Magnetosphere, ed. F. Bagenal, T. E. Dowling, & W. B. McKinnon (Cambridge: Cambridge University Press), 241–262

Canup, R. M. 2010, Origin of Saturn's rings and inner moons by mass removal from a lost Titan-sized satellite. Nature, 468, 943–926

Canup, R. M., & Esposito, L. W. 1995, Accretion in the Roche zone: coexistence of rings and ring moons. Icarus, 113, 331–352

Canup, R. M., & Ward, W. R. 2002, Formation of the Galilean satellites: conditions of accretion. Astron. J., 124, 3404–3423

Canup, R. M., & Ward, W. R. 2006, A common mass scaling for satellite systems of gaseous planets. Nature, 441, 834–839

Chandrasekhar, S. 1969, Ellipsoidal Figures of Equilibrium (New Haven: Yale University Press)

Charnoz, S., Porco, C. C., Déau, E., Brahic, A., Spitale, J. N., Bacques, G., & Baillie, K. 2005, Cassini discovers a kinematic spiral ring around Saturn. Science, 310, 1300–1304

Charnoz, S., Brahic, A., Thomas, P. C., & Porco, C. C. 2007, The equatorial ridges of Pan and Atlas: terminal accretionary ornaments? Science, 318, 1622–1624

Charnoz, S., Dones, L., Esposito, L. W., Estrada, P. R., & Hedman, M. M. 2009a, Origin and evolution of Saturn's ring system, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 537–575

Charnoz, S., Morbidelli, A., Dones, L., & Salmon, J. 2009b, Did Saturn's rings form during the Late Heavy Bombardment? Icarus, 199, 413–428

Charnoz, S., Salmon, J., & Crida, A. 2010, The recent formation of Saturn's moonlets from viscous spreading of the main rings. Nature, 465, 752–754

Charnoz, S., et al. 2011, Accretion of Saturn's mid-sized moons during the viscous spreading of young massive rings: solving the paradox of silicate-poor rings versus silicate-rich moons. Icarus, 216, 535–550

Chavez, C. E. 2009, Appearance of Saturn's F ring azimuthal channels for the anti-alignment configuration between the ring and Prometheus. Icarus, 203, 233–237

Chiang, E. I., & Goldreich, P. 2000, Apse alignment of narrow eccentric planetary rings. Astrophys. J., 540, 1084–1090

Colombo, G., Goldreich, P., & Harris, A. W. 1976, Spiral structure as an explanation for the asymmetric brightness of Saturn's A ring. Nature, 264, 344–345

Colwell, J. E., Esposito, L. W., & Sremčević, M. 2006, Self-gravity wakes in Saturn's A ring measured by stellar occultations from Cassini. Geophys. Res. Lett., 33, L07201

Colwell, J. E., Esposito, L. W., Sremčević, M., Stewart, G. R., & McClintock, W. E. 2007, Self-gravity wakes and radial structure of Saturn's B ring. Icarus, 190, 127–144

Colwell, J. E., et al. 2008, Ejecta from impacts at 0.2–2.3 m/s in low gravity. Icarus, 195, 908–917

Colwell, J. E., Cooney, J. H., Esposito, L. W., & Sremčević, M. 2009a, Density waves in Cassini UVIS stellar occultations. 1. The Cassini division. Icarus, 200, 574–580

Colwell, J. E., Nicholson, P. D., Tiscareno, M. S., Murray, C. D., French, R. G., & Marouf, E. A. 2009b, The structure of Saturn's rings, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 375–412

Colwell, J. E., Jerousek, R. G., & Esposito, L. W. 2010, Sharp edges in Saturn's rings: radial structure and longitudinal variability. AAS Div. Planet. Sci. Meet. Abstr., 42, 50.01

Cooper, J. F., Eraker, J. H., & Simpson, J. A. 1985, The secondary radiation under Saturn's A-B-C rings produced by cosmic ray interactions. J. Geophys. Res., 90, 3415–3427

Crida, A., Papaloizou, J. C. B., Rein, H., Charnoz, S., & Salmon, J. 2010, Migration of a moonlet in a ring of solid particles: theory and application to Saturn's propellers. Astron. J., 140, 944–953

Cuzzi, J. N., & Burns, J. A. 1988, Charged particle depletion surrounding Saturn's F ring: evidence for a moonlet belt? Icarus, 74, 284–324

Cuzzi, J. N., & Scargle, J. D. 1985, Wavy edges suggest moonlet in Encke's gap. Astrophys. J., 292, 276–290

Cuzzi, J., Clark, R., Filacchione, G., French, R., Johnson, R., Marouf, E., & Spilker, L. 2009, Ring particle composition and size distribution, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 459–509

Cuzzi, J. N., et al. 2010, An evolving view of Saturn's dynamic rings. Science, 327, 1470–1475

Daubechies, I. 1992, Ten Lectures on Wavelets (Philadelphia: SIAM)

Dawson, R. I., French, R. G., & Showalter, M. R. 2010, Packed perturbers: short-term interactions among Uranus' inner moons. AAS Div. Dyn. Astron. Meet. Abstr., 41, 8.07

de Pater, I., Martin, S. C., & Showalter, M. R. 2004, Keck near-infrared observations of Saturn's E and G rings during Earth's ring plane crossing in August 1995. Icarus, 172, 446–454

de Pater, I., Gibbard, S. G., Chiang, E., Hammel, H. B., Macintosh, B., Marchis, F., Martin, S. C., Roe, H. G., & Showalter, M. 2005, The dynamic neptunian ring arcs: evidence for a gradual disappearance of Liberté and resonant jump of Courage. Icarus, 174, 263–272

de Pater, I., Hammel, H. B., Gibbard, S. G., & Showalter, M. R. 2006, New dust belts of Uranus: one ring, two ring, red ring, blue ring. Science, 312, 92–94

de Pater, I., Hammel, H. B., Showalter, M. R., & van Dam, M. A. 2007, The dark side of the rings of Uranus. Science, 317, 1888–1890

Denk, T., et al. 2010, Iapetus: unique surface properties and a global color dichotomy from Cassini imaging. Science, 327, 435–439

Dermott, S. F., & Murray, C. D. 1980, Origin of the eccentricity gradient and the apse alignment of the epsilon ring of Uranus. Icarus, 43, 338–349

Dilley, J., & Crawford, D. 1996, Mass dependence of energy loss in collisions of icy spheres: an experimental study. J. Geophys. Res., 101, 9267–9270

Dones, L., & Porco, C. C. 1989, Spiral density wakes in Saturn's A ring? Bull. Am. Astron. Soc. 21, 929

Dumas, C., Terrile, R. J., Smith, B. A., Schneider, G., & Becklin, E. E. 1999, Stability of Neptune's ring arcs in question. Nature, 400, 733–735

Duncan, M. J., & Lissauer, J. J. 1997, Orbital stability of the Uranian satellite system. Icarus, 125, 1–12

Durda, D. D., Movshovitz, N., Richardson, D. C., Asphaug, E., Morgan, A., Rawlings, A. R., & Vest, C. 2011, Experimental determination of the coefficient of restitution for meter-scale granite spheres. Icarus, 211, 849–855

Durisen, R. H., Cramer, N. L., Murphy, B. W., Cuzzi, J. N., Mullikin, T. L., & Cederbloom, S. E. 1989, Ballistic transport in planetary ring systems due to particle erosion mechanisms I. Theory, numerical methods, and illustrative examples. Icarus, 80, 136–166

Durisen, R. H., Bode, P. W., Cuzzi, J. N., Cederbloom, S. E., & Murphy, B. W. 1992, Ballistic transport in planetary ring systems due to particle erosion mechanisms II. Theoretical models for Saturn's A- and B-ring inner edges. Icarus, 100, 364–393

Elliot, J. L., Dunham, E., & Mink, D. 1977, The rings of Uranus. Nature, 267, 328–330

Elliott, J. P., & Esposito, L. W. 2011, Regolith depth growth on an icy body orbiting Saturn and evolution of bidirectional reflectance due to surface composition changes. Icarus, 212, 268–274

Esposito, L. W. 2006, Cassini observations and the history of Saturn's rings. AGU Fall Meeting Abstracts, P23E-0110

Esposito, L. W. 2010, Composition, structure, dynamics, and evolution of Saturn's rings. Ann. Rev. Earth Planet. Sci., 38, 383–410

Esposito, L. W., Brahic, A., Burns, J. A., & Marouf, E. A. 1991, Particle properties and processes in Uranus' rings, in Uranus, ed. J. T. Bergstralh, E. D. Miner, & M. S. Matthews (Tucson: University of Arizona Press), 410–465

Esposito, L. W., Meinke, B. K., Colwell, J. E., Nicholson, P. D., & Hedman, M. M. 2008, Moonlets and clumps in Saturn's F ring. Icarus, 194, 278–289

Farmer, A. J., & Goldreich, P. 2005, Spoke formation under moving plasma clouds. Icarus, 179, 535–538

Fillius, R. W., McIlwain, C. E., & Mogro-Campero, A. 1975, Radiation belts of Jupiter: a second look. Science 188, 465–467

Foryta, D. W., & Sicardy, B. 1996, The dynamics of the neptunian Adams ring's arcs. Icarus, 123, 129–167

Franklin, F. A., Cook, A. F., Barrey, R. T. F., Roff, C. A., Hunt, G. E., & de Rueda, H. B. 1987, Voyager observations of the azimuthal brightness variations in Saturn's rings. Icarus, 69, 280–296

French, R. S., & Showalter, M. R. 2011, Cupid is doomed: an analysis of the stability of the inner Uranian satellites. AAS Div. Dyn. Astron. Meet. Abstr., 42, 6.02

French, R. G., Nicholson, P. D., Porco, C. C., & Marouf, E. A. 1991, Dynamics and structure of the Uranian rings, in Uranus, ed. J. T. Bergstralh, E. D. Miner, & M. S. Matthews, (Tucson: University of Arizona Press), 327–409

French, R. G., et al. 1993, Geometry of the Saturn system from the 3 July 1989 occultation of 28 SGR and Voyager observations. Icarus, 103, 163–214

Gaudi, B. S., Chang, H., & Han, C. 2003, Probing structures of distant extrasolar planets with microlensing. Astrophys. J., 586, 527–539

Giese, B., Denk, T., Neukum, G., Roatsch, T., Helfenstein, P., Thomas, P. C., Turtle, E. P., McEwen, A., & Porco, C. C. 2008, The topography of Iapetus' leading side. Icarus, 193, 359–371

Goertz, C. K., & Morfill, G. 1983, A model for the formation of spokes in Saturn's rings. Icarus, 53, 219–229

Goldreich, P., & Tremaine, S. 1978a, The velocity dispersion in Saturn's rings. Icarus, 34, 227–239

Goldreich, P., & Tremaine, S. 1978b, The formation of the Cassini division in Saturn's rings. Icarus, 34, 240–253

Goldreich, P., & Tremaine, S. 1979a, Towards a theory for the Uranian rings. Nature, 277, 97–99

Goldreich, P., & Tremaine, S. 1979b, Precession of the epsilon ring of Uranus. Astron. J., 84, 1638–1641

Goldreich, P., & Tremaine, S. 1980, Disk-satellite interactions. Astrophys. J., 241, 425–441

Goldreich, P., & Tremaine, S. 1981, The origin of the eccentricities of the rings of Uranus. Astrophys. J., 243, 1062–1075

Goldreich, P., & Tremaine, S. 1982, The dynamics of planetary rings. Ann. Rev. Astron. Astrophys., 20, 249–283

Goldreich, P., Tremaine, S., & Borderies, N. 1986, Towards a theory for Neptune's arc rings. Astron. J., 92, 490–494

Goldreich, P., Murray, N., Longaretti, P. Y., & Banfield, D. 1989, Neptune's story. Science, 245, 500–504

Hahn, J. M. 2007, The secular evolution of a close ring-satellite system: the excitation of spiral bending waves at a nearby gap edge. Astrophys. J., 665, 856–865

Hahn, J. M. 2008, The secular evolution of a close ring-satellite system: the excitation of spiral density waves at a nearby gap edge. Astrophys. J., 680, 1569–1581

Hahn, J. M., Spitale, J. N., & Porco, C. C. 2009, Dynamics of the sharp edges of broad planetary rings. Astrophys. J., 699, 686–710

Halme, V.-P., Salo, H., Sremčević, M., Albers, N., Schmidt, J., Seiss, M., & Spahn, F. 2010, Dynamical and photometric simulations of propeller features in Saturn's A ring. AAS Div. Planet. Sci. Meet. Abstr., 42, 50.02

Hamilton, D. P. 1996, The asymmetric time-variable rings of Mars. Icarus, 119, 153–172

Hamilton, D. P. 2006, The collisional cascade model for Saturn's ring spokes. AAS Div. Planet. Sci. Meet. Abstr., 38, 51.04

Hamilton, D. P., & Krüger, H. 2008, The sculpting of Jupiter's gossamer rings by its shadow. Nature, 453, 72–75

Hänninen, J., & Porco, C. 1997, Collisional simulations of Neptune's ring arcs. Icarus, 126, 1–27

Hatzes, A. P., Bridges, F. G., & Lin, D. N. C. 1988, Collisional properties of ice spheres at low impact velocities. Mon. Not. Roy. Astron. Soc., 231, 1091–1115

Hatzes, A. P., Bridges, F., Lin, D. N. C., & Sachtjen, S. 1991, Coagulation of particles in Saturn's rings: measurements of the cohesive force of water frost. Icarus, 89, 113–121

Hedman, M. M., et al. 2005, Morphology, movements and models of ringlets in Saturn's Encke gap. AAS Div. Planet. Sci. Meet. Abstr., 37, 64.01

Hedman, M. M., Nicholson, P. D., Salo, H., Wallis, B. D., Buratti, B. J., Baines, K. H.,

Brown, R. H., & Clark, R. N. 2007a, Self-gravity wake structures in Saturn's A ring revealed by Cassini VIMS. Astron. J., 133, 2624–2629

Hedman, M. M., Burns, J. A., Tiscareno, M. S., & Porco, C. C. 2007b, The heliotropic rings of Saturn. AAS Div. Planet. Sci. Meet. Abstr., 39, 10.09

Hedman, M. M., et al. 2007c, Saturn's dynamic D ring. Icarus, 188, 89–107

Hedman, M. M., Burns, J. A., Tiscareno, M. S., Porco, C. C., Jones, G. H., Roussos, E., Krupp, N., Paranicas, C., & Kempf, S. 2007d, The source of Saturn's G ring. Science, 317, 653–656

Hedman, M. M., Nicholson, P. D., Showalter, M. R., Brown, R. H., Buratti, B. J., & Clark, R. N. 2009a, Spectral observations of the Enceladus plume with Cassini-VIMS. Astrophys. J., 693, 1749–1762

Hedman, M. M., Murray, C. D., Cooper, N. J., Tiscareno, M. S., Beurle, K., Evans, M. W., & Burns, J. A. 2009b, Three tenuous rings/arcs for three tiny moons. Icarus, 199, 378–386

Hedman, M. M., Burns, J. A., Tiscareno, M. S., & Porco, C. C. 2009c, Organizing some very tenuous things: resonant structures in Saturn's faint rings. Icarus, 202, 260–279

Hedman, M. M., Nicholson, P. D., Baines, K. H., Buratti, B. J., Sotin, C., Clark, R. N., Brown, R. H., French, R. G., & Marouf, E. A. 2010a, The architecture of the Cassini division. Astron. J., 139, 228–251

Hedman, M. M., Burt, J. A., Burns, J. A., & Tiscareno, M. S. 2010b, The shape and dynamics of a heliotropic dusty ringlet in the Cassini division. Icarus, 210, 284–297

Hedman, M. M., Cooper, N. J., Murray, C. D., Beurle, K., Evans, M. W., Tiscareno, M. S., & Burns, J. A. 2010c, Aegaeon (Saturn LIII), a G-ring object. Icarus, 207, 433–447

Hedman, M. M., Nicholson, P. D., Filacchione, G., Capaccioni, F., Ciarnello, M., & Clark, R. N. 2011a, Correlations between the spectra and structure of Saturn's main rings. AAS Div. Planet. Sci. Meet. Abstr., 43, 532

Hedman, M. M., Burns, J. A., & Tiscareno, M. S. 2011b, Of horseshoes and heliotropes: the dynamics of dust in the Encke gap. AAS Div. Dyn. Astron. Meet. Abstr., 42, 8.02

Hedman, M. M., Nicholson, P. D., Showalter, M. R., Brown, R. H., Buratti, B. J., Clark, R. N., Baines, K., & Sotin, C. 2011c, The Christiansen effect in Saturn's narrow dusty rings and the spectral identification of clumps in the F ring. Icarus, 215, 695–711

Hedman, M. M., Burns, J. A., Evans, M. W., Tiscareno, M. S., & Porco, C. C. 2011d, Saturn's curiously corrugated C ring. Science, 332, 708–711

Heißelmann, D., Blum, J., Fraser, H. J., & Wolling, K. 2010, Microgravity experiments on the collisional behavior of saturnian ring particles. Icarus, 206, 424–430

Hill, J. R., & Mendis, D. A. 1981, On the braids and spokes in Saturn's ring system. Moon Planet, 24, 431–436

Horányi, M., Burns, J. A., Hedman, M. M., Jones, G. H., & Kempf, S. 2009, Diffuse rings, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 511–536

Ip, W.-H. 2006, On a ring origin of the equatorial ridge of Iapetus. Geophys. Res. Lett., 33, L16203

Jacobson, R. A., Campbell, J. K., Taylor, A. H., & Synnott, S. P. 1992, The masses of Uranus and its major satellites from Voyager tracking data and Earth-based Uranian satellite data. Astron. J., 103, 2068–2078

Jacobson, R. A., Spitale, J., Porco, C. C., Beurle, K., Cooper, N. J., Evans, M. W., & Murray, C. D. 2008, Revised orbits of Saturn's small inner satellites. Astron. J., 135, 261–263

Jones, G. H., et al. 2006, Formation of Saturn's ring spokes by lightning-induced electron beams. Geophys. Res. Lett., 33, L21202.

Jones, G. H., et al. 2008, The dust halo of Saturn's largest icy moon, Rhea. Science, 319, 1380–1384

Julian, W. H., & Toomre, A. 1966, Non-axisymmetric responses of differentially rotating disks of stars. Astrophys. J., 146, 810–830

Kalas, P., Graham, J. R., & Clampin, M. 2005, A planetary system as the origin of structure in Fomalhaut's dust belt. Nature, 435, 1067–1070

Kalas, P., Graham, J. R., Chiang, E., Fitzgerald, M. P., Clampin, M., Kite, E. S., Stapelfeldt, K., Marois, C., & Krist, J. 2008, Optical images of an exosolar planet 25 light-years from Earth. Science, 322, 1345–1348

Kempf, S., Beckmann, U., & Schmidt, J. 2010, How the Enceladus dust plume feeds Saturn's E ring. Icarus, 206, 446–457

Krivov, A. V., & Hamilton, D. P. 1997, Martian dust belts: waiting for discovery. Icarus, 128, 335–353

Laughlin, G., Korchagin, V., & Adams, F. C. 1997, Spiral mode saturation in self-gravitating disks. Astrophys. J., 477, 410–423

Levison, H. F., Walsh, K. J., Barr, A. C., & Dones, L. 2011, Ridge formation and de-spinning of Iapetus via an impact-generated satellite. Icarus, 214, 773–778

Lewis, M. C., & Stewart, G. R. 2009, Features around embedded moonlets in Saturn's rings: the role of self-gravity and particle size distributions. Icarus, 199, 387–412

Lin, C. C., & Shu, F. H. 1964, On the spiral structure of disk galaxies. Astrophys. J., 140, 646–655

Lissauer, J. J. 1985, Shepherding model for Neptune's arc ring. Nature, 318, 544–545

Longaretti, P.-Y., & Borderies, N. 1991, Streamline formalism and ring orbit determination. Icarus, 94, 165–170

Lubow, S. H. 2010, Eccentricity growth rates of tidally distorted discs. Mon. Not. Roy. Astron. Soc., 406, 2777–2786

Mamajek, E. E., Quillen, A. C., Pecaut, M. J., Moolekamp, F., Scott, E. L., Kenworthy, M. A., Collier Cameron, A., & Parley, N. R. 2012, Planetary construction zones in occultation: discovery of an extrasolar ring system transiting a young Sun-like star and future prospects for detecting eclipses by circumsecondary and circumplanetary disks. Astron. J., 143, 72

Matson, D. L., Castillo-Rogez, J. C., Schubert, G., Sotin, C., & McKinnon, W. B. 2009, The thermal evolution and internal structure of Saturn's mid-sized icy satellites, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 577–612

McGhee, C. A., French, R. G., Dones, L., Cuzzi, J. N., Salo, H. J., & Danos, R. 2005, HST observations of spokes in Saturn's B ring. Icarus, 173, 508–521

Meyer, J., & Wisdom, J. 2007, Tidal heating in Enceladus. Icarus, 188, 535–539

Meyer, J., & Wisdom, J. 2008a, Tidal evolution of Mimas, Enceladus, & Dione. Icarus, 193, 213–223

Meyer, J., & Wisdom, J. 2008b, Episodic volcanism on Enceladus: application of the Ojakangas-Stevenson model. Icarus, 198, 178–180

Michikoshi, S., & Kokubo, E. 2011, Formation of a propeller structure by a moonlet in a dense planetary ring. Astrophys. J. Lett., 732, L23.

Millis, R. L., Wasserman, L. H., & Birch, P. V. 1977, Detection of rings around Uranus. Nature, 267, 330–331

Miner, E. D., Wessen, R. R., & Cuzzi, J. N. 2007, Planetary Ring Systems (Chichester: Springer Praxis)

Mitchell, C. J., Horányi, M., Havnes, O., & Porco, C. C. 2006, Saturn's spokes: lost and found. Science, 311, 1587–1589

Mitchell, C., Porco, C., Dones, L., & Spitale, J. 2012, The behavior of spokes in Saturn's B ring. Icarus, submitted

Morfill, G. E., & Thomas, H. M. 2005, Spoke formation under moving plasma clouds: The Goertz-Morfill model revisited. Icarus, 179, 539–542

Mosqueira, I., & Estrada, P. R. 2002, Apse alignment of the Uranian rings. Icarus, 158, 545–556

Mosqueira, I., & Estrada, P. R. 2003, Formation of the regular satellites of giant planets in an extended gaseous nebula I: subnebula model and accretion of satellites. Icarus, 163, 198–231

Murray, C. D., & Dermott, S. F. 1999, Solar System Dynamics (Cambridge: Cambridge University Press)

Murray, C. D., & Thompson, R. P. 1990, Orbits of shepherd satellites deduced from the structure of the rings of Uranus. Nature, 348, 499–502

Murray, C. D., & Thompson, R. P. 1991, Erratum: orbits of shepherd satellites deduced from the structure of the rings of Uranus. Nature, 350, 90

Murray, C. D., Chavez, C., Beurle, K., Cooper, N., Evans, M. W., Burns, J. A., & Porco, C. C. 2005, How Prometheus creates structure in Saturn's F ring. Nature, 437, 1326–1329

Murray, C. D., Beurle, K., Cooper, N. J., Evans, M. W., Williams, G. A., & Charnoz, S. 2008, The determination of the structure of Saturn's F ring by nearby moonlets. Nature, 453, 739–744

Namouni, F., & Porco, C., 2002, The confinement of Neptune's ring arcs by the moon Galatea. Nature, 417, 45–47

Nicholson, P. D., & Hedman, M. M. 2010, Self-gravity wake parameters in Saturn's A and B rings. Icarus, 206, 410–423

Nicholson, P. D., Mosqueira, I., & Matthews, K. 1995, Stellar occultation observations of Neptune's rings: 1984–1988. Icarus, 113, 295–330

Nicholson, P. D., et al. 2008, A close look at Saturn's rings with Cassini VIMS. Icarus, 193, 182–212

Ockert-Bell, M. E., Burns, J. A., Daubar, I. J., Thomas, P. C., Veverka, J., Belton, M. J. S., & Klaasen, K. P. 1999, The structure of Jupiter's ring system as revealed by the Galileo imaging experiment. Icarus, 138, 188–213

Ohta, Y., Taruya, A., & Suto, Y. 2009, Predicting photometric and spectroscopic signatures of rings around transiting extrasolar planets. Astrophys. J., 690, 1–12

Øieroset, M., Brain, D. A., Simpson, E., Mitchell, D. L., Phan, T. D., Halekas, J. S., Lin, R. P., & Acuña, M. H. 2010, Search for Phobos and Deimos gas/dust tori using in situ observations from Mars global surveyor MAG/ER. Icarus, 206, 189–198

Orton, G. S., Baines, K. H., Cruikshank, D., Cuzzi, J. N., Krimigis, S. M., Miller, S., & Lellouch, E. 2009, Review of knowledge prior to the Cassini-Huygens mission and concurrent research, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 9–54

Owen, T., Danielson, G. E., Cook, A. F., Hansen, C., Hall, V. L., & Duxbury, T. C. 1979, Jupiter's rings, Nature, 281, 442–446

Pan, M., & Chiang, E. 2010, The propeller and the frog. Astrophys. J. Lett., 722, L178–L182

Papaloizou, J. C. B., Nelson, R. P., Kley, W., Masset, F. S., & Artymowicz, P. 2007, Disk-planet interactions during planet formation, in Protostars and Planets V, ed. B. Reipurth, D. Jewitt, & K. Keil (Tucson: University of Arizona Press), 655–668

Perrine, R. P., Richardson, D. C., & Scheeres, D. J. 2011, A numerical model of cohesion in planetary rings. Icarus, 212, 719–735

Porco, C. C. 1991, An explanation for Neptune's ring arcs. Science, 253, 995–1001

Porco, C. C., Nicholson, P. D., Cuzzi, J. N., Lissauer, J. J., & Esposito, L. W. 1995, Neptune's ring system, in Neptune and Triton, ed. D. P. Cruikshank (Tucson: University of Arizona Press), 703–804

Porco, C. C., et al. 2005a, Cassini imaging science: initial results on Phoebe and Iapetus. Science, 307, 1237–1242

Porco, C. C., et al. 2005b, Cassini imaging science: initial results on Saturn's rings and small satellites. Science, 307, 1226–1236

Porco, C. C., Thomas, P. C., Weiss, J. W., & Richardson, D. C. 2007, Saturn's small satellites: clues to their origins. Science, 318, 1602–1607

Porco, C. C., Weiss, J. W., Richardson, D. C., Dones, L., Quinn, T., & Throop, H. 2008, Simulations of the dynamical and light-scattering behavior of Saturn's rings and the derivation of ring particle and disk properties. Astron. J., 136, 2172–2200

Rappaport, N. J., Longaretti, P., French, R. G., Marouf, E. A., & McGhee, C. A. 2009, A procedure to analyze nonlinear density waves in Saturn's rings using several occultation profiles. Icarus, 199, 154–173

Rein, H., & Papaloizou, J. C. B. 2010, Stochastic orbital migration of small bodies in Saturn's rings. Astron. Astrophys., 524, A22.

Renner, S., & Sicardy, B. 2006, Use of the geometric elements in numerical simulations. Cel. Mech. Dyn. Astron., 94, 237–248

Richardson, D. C. 1994, Tree code simulations of planetary rings. Mon. Not. Roy. Astron. Soc., 269, 493–511

Richardson, J. E., Melosh, H. J., Lisse, C. M., & Carcich, B. 2007, A ballistics analysis of the Deep Impact ejecta plume: determining Comet Tempel 1's gravity, mass, & density. Icarus, 190, 357–390

Robbins, S. J., Stewart, G. R., Lewis, M. C., Colwell, J. E., & Sremčević, M. 2010, Estimating the masses of Saturn's A and B rings from high-optical depth *N*-body simulations and stellar occultations. Icarus, 206, 431–445

Rosen, P. A., & Lissauer, J. J. 1988, The Titan -1:0 nodal bending wave in Saturn's ring C. Science, 241, 690–694

Roussos, E., Jones, G. H., Krupp, N., Paranicas, C., Mitchell, D. G., Krimigis, S. M., Woch, J., Lagg, A., & Khurana, K. 2008, Energetic electron signatures of Saturn's smaller moons: evidence of an arc of material at Methone. Icarus, 193, 455–464

Salmon, J., Charnoz, S., Crida, A., & Brahic, A. 2010, Long-term and large-scale viscous evolution of dense planetary rings. Icarus, 209, 771–785

Salo, H. 1992, Gravitational wakes in Saturn's rings. Nature, 359, 619–621

Salo, H. 1995, Simulations of dense planetary rings III. self-gravitating identical particles. Icarus, 117, 287–312

Salo, H., & Karjalainen, R. 2003, Photometric modeling of Saturn's rings I. Monte Carlo method and the effect of nonzero volume filling factor. Icarus, 164, 428–460

Salo, H., Karjalainen, R., & French, R. G. 2004, Photometric modeling of Saturn's rings. II. Azimuthal asymmetry in reflected and transmitted light. Icarus, 170, 70–90

Schenk, P. M., & McKinnon, W. B. 2009, Global color variations on Saturn's icy satellites, and new evidence for Rhea's ring. AAS Div. Planet. Sci. Meet. Abstr., 41, 3.03

Schenk, P., Hamilton, D. P., Johnson, R. E., McKinnon, W. B., Paranicas, C., Schmidt, J., & Showalter, M. R. 2011, Plasma, plumes and rings: saturn system dynamics as recorded in global color patterns on its midsize icy satellites. Icarus, 211, 740–757

Schlichting, H. E., & Chang, P. 2011, Warm Saturns: on the nature of rings around extrasolar planets that reside inside the ice line. Astrophys. J., 734, 117

Schmidt, J., Brilliantov, N., Spahn, F., & Kempf, S. 2008, Slow dust in Enceladus' plume from condensation and wall collisions in tiger stripe fractures. Nature, 451, 685–688

Schmidt, J., Ohtsuki, K., Rappaport, N., Salo, H., & Spahn, F. 2009, Dynamics of Saturn's dense rings, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer-Verlag), 413–458

Seager, S. (ed.), 2010, Exoplanets (Tucson: University of Arizona Press)

Seal, D. A., & Buffington, B. B. 2009, The Cassini extended mission, in Saturn from Cassini-Huygens, ed. M. Dougherty, L. Esposito, & S. M. Krimigis (Dordrecht: Springer), 725–744

Seiß, M., Spahn, F., Sremčević, M., & Salo, H. 2005, Structures induced by small moonlets in Saturn's rings: implications for the Cassini mission. Geophys. Res. Lett., 32, L11205.

Showalter, M. R. 1991, Visual detection of 1981S13, Saturn's eighteenth satellite, and its role in the Encke gap. Nature, 351, 709–713

Showalter, M. R. 2011, The rings of Uranus: shepherded or not? AAS Div. Planet. Sci. Meet. Abstr., 43, 1224

Showalter, M. R., & Cuzzi, J. N. 1992, Physical properties of Neptune's ring system. Bull. Am. Astron. Soc. 24, 1029

Showalter, M. R., & Lissauer, J. J. 2006, The second ring-moon system of Uranus: discovery and dynamics. Science, 311, 973–977

Showalter, M. R., Cuzzi, J. N., Marouf, E. A., & Esposito, L. W. 1986, Satellite "wakes" and the orbit of the Encke Gap moonlet. Icarus, 66, 297–323

Showalter, M. R., Cuzzi, J. N., & Larson, S. M. 1991, Structure and particle properties of Saturn's E ring. Icarus, 94, 451–473

Showalter, M. R., Hamilton, D. P., & Nicholson, P. D. 2006, A deep search for Martian dust rings and inner moons using the Hubble Space Telescope. Planet. Space Sci., 54, 844–854

Showalter, M. R., Cheng, A. F., Weaver, H. A., Stern, S. A., Spencer, J. R., Throop, H. B., Birath, E. M., Rose, D., & Moore, J. M. 2007, Clump detections and limits on moons in Jupiter's ring system. Science, 318, 232–234

Showalter, M. R., French, R., Sfair, R., Argüelles, C., Pajuelo, M., Becerra, P., Hedman, M., & Nicholson, P. 2009, The brightening of Saturn's F ring. AAS Div. Planet. Sci. Meet. Abstr., 41, 22.07

Showalter, M. R., Hamilton, D. P., Stern, S. A., Weaver, H. A., Steffl, A. J., & Young, L. A. 2011a, New satellite of (134340) Pluto: S/2011 (134340) 1. Cent. Bur. Electron. Telegr., 2769, 1

Showalter, M. R., Hedman, M. M., & Burns, J. A. 2011b, The impact of Comet Shoemaker-Levy 9 sends ripples through the rings of Jupiter. Science, 332, 711–713

Shu, F. H. 1984, Waves in planetary rings, in Planetary Rings, ed. R. Greenberg, & A. Brahic (Tucson: University of Arizona Press), 513–561

Sicardy, B., Roddier, F., Roddier, C., Perozzi, E., Graves, J. E., Guyon, O., & Northcott, M. J. 1999, Images of Neptune's ring arcs obtained by a ground-based telescope. Nature, 400, 731–733

Smith, B. A., et al. 1982, A new look at the Saturn system - The Voyager 2 images. Science, 215, 504–537

Soter, S. 1974, Remarks on the origin of Iapetus' photometric asymmetry. IAU Colloq., 28

Spahn, F., & Sremčević, M. 2000, Density patterns induced by small moonlets in Saturn's rings? Astron. Astrophys., 358, 368–372

Spencer, J. R., & Denk, T. 2010, Formation of Iapetus' extreme albedo dichotomy by exogenically triggered thermal ice migration. Science, 327, 432–435

Spitale, J., & Porco, C. C. 2006, Shapes and kinematics of eccentric features in Saturn's C ring and Cassini division. AAS Div. Dyn. Astron. Meet. Abstr., 37, 7.02

Spitale, J. N., & Porco, C. C. 2009, Time variability in the outer edge of Saturn's A-ring revealed by Cassini imaging. Astron. J., 138, 1520–1528

Spitale, J. N., & Porco, C. C. 2010, Detection of free unstable modes and massive bodies in Saturn's outer B ring. Astron. J., 140, 1747–1757

Spitale, J. N., Porco, C. C., & Colwell, J. 2008, An inclined saturnian ringlet at 1.954 Rs. AAS Div. Planet. Sci. Meet. Abstr., 40, 21.02

Sremčević, M., Spahn, F., & Duschl, W. J. 2002, Density structures in perturbed thin cold discs. Mon. Not. Roy. Astron. Soc., 337, 1139–1152

Sremčević, M., Schmidt, J., Salo, H., Seiß, M., Spahn, F., & Albers, N. 2007, A belt of moonlets in Saturn's A ring. Nature, 449, 1019–1021

Sremčević, M., Colwell, J. E., & Esposito, L. W. 2009, Small-scale ring structure observed in Cassini UVIS occultations. AGU Fall Meeting Abstracts, P54A–05

Steffl, A. J., & Stern, S. A. 2007, First constraints on rings in the Pluto system. Astron. J., 133, 1485–1489

Stern, S. A., Weaver, H. A., Steffl, A. J., Mutchler, M. J., Merline, W. J., Buie, M. W., Young, E. F., Young, L. A., & Spencer, J. R. 2006, A giant impact origin for Pluto's small moons and satellite multiplicity in the Kuiper belt. Nature, 439, 946–948

Supulver, K. D., Bridges, F. G., & Lin, D. N. C. 1995, The coefficient of restitution of ice particles in glancing collisions: experimental results for unfrosted surfaces. Icarus, 113, 188–199

Tamayo, D., Burns, J. A., Hamilton, D. P., & Hedman, M. M. 2011, Finding the trigger to Iapetus' odd global albedo pattern: dynamics of dust from Saturn's irregular satellites. Icarus, 215, 260–278

Thomson, F. S., Marouf, E. A., Tyler, G. L., French, R. G., & Rappoport, N. J. 2007, Periodic microstructure in Saturn's rings A and B. Geophys. Res. Lett., 34, L24203

Tiscareno, M. S. 2012, A modified "Type I migration" model for propeller moons in Saturn's rings. Planet. Space Sci., in press (arXiv:1206.4942)

Tiscareno, M. S., Burns, J. A., Hedman, M. M., Spitale, J. N., Porco, C. C., Murray, C. D., & Cassini

Imaging Team, 2005, Wavy edges and other disturbances in Saturn's Encke and Keeler gaps. AAS Div. Planet. Sci. Meet. Abstr., 37, 64.02

Tiscareno, M. S., Burns, J. A., Hedman, M. M., Porco, C. C., Weiss, J. W., Dones, L., Richardson, D. C., & Murray, C. D. 2006a, 100-metre-diameter moonlets in Saturn's A ring from observations of "propeller" structures. Nature, 440, 648–650

Tiscareno, M. S., Nicholson, P. D., Burns, J. A., Hedman, M. M., & Porco, C. C. 2006b, Unravelling temporal variability in Saturn's spiral density waves: results and predictions. Astrophys. J. Lett., 651, L65–L68

Tiscareno, M. S., Burns, J. A., Nicholson, P. D., Hedman, M. M., & Porco, C. C. 2007, Cassini imaging of Saturn's rings II. a wavelet technique for analysis of density waves and other radial structure in the rings. Icarus, 189, 14–34

Tiscareno, M. S., Burns, J. A., Hedman, M. M., & Porco, C. C. 2008, The population of propellers in Saturn's A ring. Astron. J., 135, 1083–1091

Tiscareno, M. S., Hedman, M. M., Burns, J. A., Weiss, J. W., & Porco, C. C. 2009a, Saturn's A ring has no inner edge. AAS Div. Planet. Sci. Meet. Abstr., 41, 25.04

Tiscareno, M. S., Burns, J. A., Hedman, M. M., DiNino, D., Porco, C. C., Beurle, K., & Evans, M. W. 2009b, Observations of ejecta clouds produced by impacts onto Saturn's rings. AGU Fall Meeting Abstracts, P54A–08

Tiscareno, M. S., Perrine, R. P., Richardson, D. C., Hedman, M. M., Weiss, J. W., Porco, C. C., & Burns, J. A. 2010a, An analytic parameterization of self-gravity wakes in Saturn's rings. Astron. J., 139, 492–503

Tiscareno, M. S., Burns, J. A., Cuzzi, J. N., & Hedman, M. M. 2010b, Cassini imaging search rules out rings around Rhea. Geophys. Res. Lett., 37, L14205

Tiscareno, M. S., et al. 2010c, Physical characteristics and non-keplerian orbital motion of "propeller" moons embedded in Saturn's rings. Astrophys. J. Lett., 718, L92–L96

Torrence, C., & Compo, G. P. 1998, A practical guide to wavelet analysis. Bull. Am. Meteorol. Soc., 79, 61–78 http://atoc.colorado.edu/research/wavelets/

Torrey, P. A., Tiscareno, M. S., Burns, J. A., & Porco, C. C. 2008, Mapping complexity: the wavy edges of the Encke and Keeler gaps in Saturn's rings. AAS Div. Dyn. Astron. Meet. Abstr., 39, 15.19

Van Helden, A. 1984, Saturn through the telescope: a brief historical survey, in Saturn, ed. T. Gehrels, & M. S. Matthews (Tucson: University of Arizona Press), 23–43

Verbiscer, A. J., Skrutskie, M. F., & Hamilton, D. P. 2009, Saturn's largest ring. Nature, 461, 1098–1100

Ward, W. R. 1986, Density waves in the solar nebula: differential Lindblad torque. Icarus, 67, 164–180

Ward, W. R. 1997, Survival of planetary systems. Astrophys. J. Lett., 482, L211–L214

Weiss, J. W. 2005, The physics of unconstrained edges in planetary rings. Ph.D. Thesis, University of Colorado

Weiss, J. W., Porco, C. C., & Tiscareno, M. S. 2009, Ring edge waves and the masses of nearby satellites. Astron. J., 138, 272–286

Winter, O. C., Mourão, D. C., Giuliatti Winter, S. M., Spahn, F., & da Cruz, C. 2007, Moonlets wandering on a leash-ring. Mon. Not. Roy. Astron. Soc., 380, L54–L57

Wisdom, J., & Holman, M. 1991, Symplectic maps for the *n*-body problem. Astron. J., 102, 1528–1538

Wisdom, J., & Tremaine, S. 1988, Local simulations of planetary rings. Astron. J., 95, 925–940

Zebker, H. A., Marouf, E. A., & Tyler, G. L. 1985, Saturn's rings: particle size distributions for thin layer model. Icarus, 64, 531–548

8 An Overview of the Asteroids and Meteorites

Andrew Rivkin
Johns Hopkins University/Applied Physics Laboratory, Laurel, MD, USA

1 Introduction ***378***

2 Dynamics of Asteroids ***383***
2.1 Gravitational and Nongravitational Forces 383
2.2 Orbital Stability and Lifetime of NEOs 386
2.3 Collisional Evolution and Families 386
2.4 Binary and Multiple Objects 388
2.5 The NEO Hazard 389
2.6 The Nice Model 390

3 Geology and Surfaces of Asteroids ***391***
3.1 Cratering 391
3.2 Regolith 394
3.3 Processes 396
3.4 Regolith Movement and Mass Wasting 398
3.5 Outgassing 400
3.6 Cohesive Forces 400

4 Asteroidal Interiors and Geophysics ***401***
4.1 Asteroid Sizes and Densities 401
4.2 Monoliths and Rubble Piles 404
4.3 Rotation Rates and Interior Structure 405
4.4 Strength 406
4.5 Meteoritical Data 406

5 Asteroid and Meteorite Compositions ***409***
5.1 Isotopic Studies 412
5.2 Asteroidal Compositions from Remote Sensing 413
5.3 Compositions of Specific Objects and Classes: Current Interpretations 417

References ***422***

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_8, © Springer Science+Business Media Dordrecht 2013

Abstract: The asteroids are a population of millions of small bodies found throughout the inner solar system but concentrated between Mars and Jupiter and ranging in size from hundreds of kilometers to meter-sized or smaller. Samples of asteroids are available in the form of meteorites, affording the opportunity for a unique synergy between geochemical and astronomical study. From laboratory and remote sensing studies, it is apparent that the asteroids harbor a range of compositions from metallic cores of once-molten bodies to primitive, ice-rich mixtures. This range of compositions is used to place constraints on the heating and cooling histories of asteroids. While igneous processes ceased on the asteroids very early on in solar system history, they have been heavily impacted since they formed. These impacts have generated regolith on asteroids of even small size, as well as formed craters, and in some cases entirely disrupted the target asteroids. As a result, most asteroids appear to be rubble piles, collections of material held together by little more than gravity and small-scale cohesive forces. Dynamical families are also created via impacts, with nongravitational forces affecting their orbits since their formation. The large number of asteroids, both in and out of families, has made them useful as test masses, allowing dynamicists to probe conditions early in solar system history, just as geochemists do with meteorites, and to model and constrain the motions of the larger planets. This chapter will provide a broad overview of asteroidal studies from the compositional to geophysical, from surfaces to cores.

1 Introduction

The understanding of the asteroids and meteorites is critical to understanding the history and evolution of the solar system. However, there is more than one way to define what an asteroid is, usually in opposition to the comets, their sibling small bodies. These definitions can be seen as compositional, dynamical, and observational.

The longest-standing definition for asteroids is that they are solar system bodies that are point sources even in large telescopes at high magnification. Herschel found Ceres and Pallas to be starlike shortly after their discovery and dubbed them "asteroids" for their appearance. Over the intervening centuries, the Minor Planet Center and other workers distinguished comets from asteroids by the presence or absence of a coma and/or tail. This remains the main discriminator between comets and asteroids today, as large telescopes, space telescopes, radar experiments, and adaptive optics have rendered Herschel's original name inaccurate for at least some objects.

Most asteroids are found in-between Mars and Jupiter, with significant populations leading and trailing Jupiter near the Lagrange points and within the orbit of Mars. As discussed in Sect. 2. The Tisserand parameter is conserved after encounters with Jupiter and can be used to judge the likelihood of an object originating in the inner or outer solar system. This has given rise to another means of distinguishing asteroids, as small bodies that formed in the inner solar system regardless of their appearance. As an outgrowth of the two previous classification schemes, a third composition-based definition has arisen with asteroids generally considered objects made of rock or rock and metal, while comets are ice-rock mixtures.

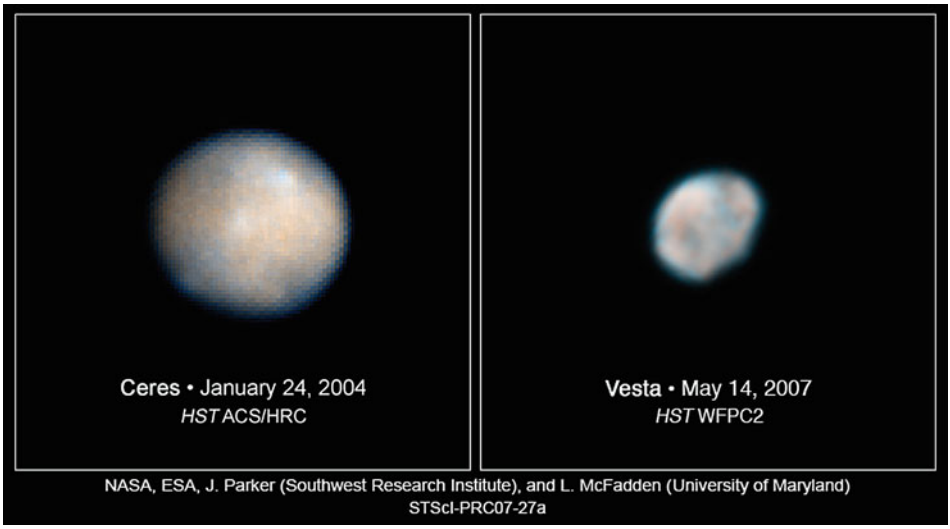
In practice, the vast majority of objects can be classified unambiguously either as asteroids or not. However, there are important populations that seem to straddle the comet/asteroid line

or qualify as asteroids in some but not all definitions. Extinct comets, which have exhausted their supply of near-surface volatiles, lack comae and tails and appear asteroidal. The origin of the Trojan asteroids near 5 AU remains uncertain, with recent models suggesting they may have formed in the transneptunian region and later been transported to their current orbits (Morbidelli et al. 2005). The "main-belt comets" or "activated asteroids" are observed to have typical main-belt orbits but also harbor tails and are interpreted to have icy compositions. Recent work suggests some of the largest asteroids may have substantial complements of ice (Sect. 5) and some objects have been observed to have ice on their surfaces (Sect. 2).

The IAU's adoption of a definition for planets has also affected the asteroids. The IAU scheme created a category of "dwarf planets," which are massive enough to achieve hydrostatic equilibrium, but did not have a mass high enough to "clear their neighborhood" early in solar system history (though "neighborhood clearing" involved factors in addition to mass). Ceres, the largest object in the asteroid belt, is formally classified as a dwarf planet by the IAU, the only non-TNO so classified at present. All other objects in the asteroid belt (as well as the NEO and Trojan populations) are classified as "small solar system bodies" by the IAU, a catch-all term that includes both comets and asteroids. Vesta seems likely to have achieved hydrostatic equilibrium early in its history (probably during differentiation and/or shortly thereafter), but a large impact subsequently changed its shape, and it has not returned to hydrostatic equilibrium. While Vesta is not currently classed as a dwarf planet, it is possible that future refinements to the IAU classification system will include it in that category. Pallas is slightly larger than Vesta, but slightly less massive. Observations of Pallas have not provided us with the detail available for Ceres or Vesta, but similarly, it appears to not be in hydrostatic equilibrium (*Figs. 8-1* and *8-2*).

Other populations not discussed in detail in this chapter may also have begun their existence as asteroids before removal from the asteroid belt. Phobos and Deimos, the satellites of Mars, have the physical appearance of asteroids and share spectral characteristics with outer-belt asteroids. Indeed, before the Galileo encounters with Gaspra and Ida, Phobos and Deimos were used as asteroid proxies when considering main-belt objects. However, the origin of Phobos and Deimos is still not clear, and each of the leading theories has strengths and drawbacks: capture from the asteroid belt is difficult to explain from a dynamical point of view given the current orbits of the two bodies, while formation in their current locations (or as Mars ejecta) is difficult to reconcile with their spectral properties (Burns 1992). Our knowledge of the Martian satellites has been gained largely through observation by spacecraft whose first priority was Mars, and the loss of Phobos-Grunt, a planned Russian sample return from Phobos, will postpone our understanding further. A hoped for reflight of Phobos-Grunt would be able to place its relationship to the asteroids in sharper focus (Galimov 2010).

Beyond Mars, the outer planets host over 100 irregular satellites, with 55 orbiting Jupiter alone. It is thought that all of these objects were captured into their current orbits from the small body population, most likely from the TNO population (Nicholson et al. 2008). However, the Jovian irregular satellites have spectra consistent with the Trojan asteroids, outer-belt asteroids, or in some cases mid-belt asteroids. Furthermore, while the "Nice model" predicts irregular satellite populations captured from TNOs for Saturn, Uranus, and Neptune that are consistent with observations, it cannot reproduce the Jovian population (Nesvorný et al. 2007). This leaves open the possibility that the irregular satellites of Jupiter were captured from the asteroid belt, in contrast to a Jupiter Trojan asteroid population that could be captured from TNOs (Morbidelli et al. 2005).



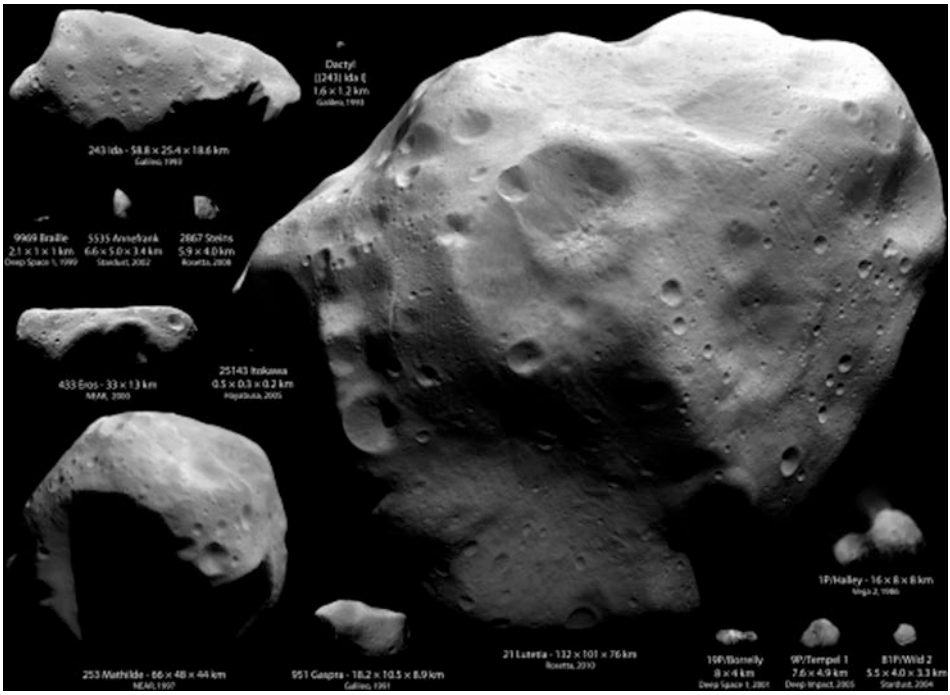
http://hubblesite.org/newscenter/archive/releases/2007/27/image/a/

Fig. 8-1

The two most massive asteroids, 1 Ceres and 4 Vesta, give only tantalizing hints of their surfaces even in the most powerful telescopes. However, Earth-based observations have been able to determine the shapes of these objects, allowing Ceres to be classified as a "dwarf planet," while the large crater near Vesta's south pole (seen in profile along the *lower right* edge of Vesta in the *right panel*) takes it out of consideration in the current IAU scheme. The arrival of the Dawn spacecraft at each of these bodies during this decade will produce a flood of high-resolution data (Image taken by the Hubble Space Telescope, reproduced courtesy of NASA)

While the large end of the asteroid size range is addressed by the proposed IAU classification, the small end is less well-defined. Dust bands in the asteroid belt are derived from asteroid disruptions, but the dust grains themselves are not generally considered asteroids. An informal distinction is sometimes made with objects smaller than a given size (sometimes 50 m) termed "meteoroids," though this term is often reserved for objects that will hit Earth. Recent NEO surveys have been able to discover meter-scale objects, which are usually termed (and always classified as) asteroids rather than meteoroids. The difficulty of observing, tracking, and characterizing centimeter-scale bodies in space probably makes their classification of little practical use, but they could be considered to be a qualitatively intermediate group between asteroids (large enough to survive atmospheric entry) and dust (small enough to be affected by radiation pressure and/or Poynting-Robertson drag).

While dust is not covered in this chapter, its study is an important, if often underappreciated, part of planetary studies (Dermott et al. 2002). Interplanetary dust particles (IDPs) are collected and studied as the tiny meteorites they are, and zodiacal clouds of such dust are studied in both our solar system and also in other stellar systems. It is generally accepted that a large



http://www.planetary.org/blog/article/00002585/

Fig. 8-2

A handful of asteroids have been encountered by spacecraft at this writing. This compilation, by Emily Lakdawalla and Ted Stryk, shows these asteroids (as well as targets of cometary encounters) to scale, from the ~100 km-sized 21 Lutetia to the ~300 m-sized 25143 Itokawa. These asteroids all show evidence of widespread cratering marking their surfaces and sculpting their irregular shapes. Vesta, the next asteroid to be encountered, is nearly five times larger than Lutetia (Image courtesy of Emily Lakdawalla, Ted Stryk/Planetary Society)

fraction, perhaps the dominant fraction, of IDPs originated on comets rather than asteroids (see Nesvorný et al. 2010b, for instance).

In the following sections, the study of asteroids and meteorites is divided into three disciplines. First, the dynamics of asteroids, including a discussion of dynamical families and binary systems, are considered. Next, the surfaces of asteroids and the processes that affect those surfaces are addressed. In the third section, asteroid interiors and geophysics are studied. Finally, the compositions of asteroids, including the wealth of information gained from meteorite studies, are addressed. Naturally, it is impossible to cover material in great depth in a single chapter, so references are provided for further study including both primary and review papers.

A list of particularly interesting asteroids is included as *Table 8-1*.

Table 8-1
Numbered asteroids of note

Name	Diameter (km)	Semimajor axis (AU)	Albedo	Spectral type	Note
1 Ceres	952	2.77	0.09	G	Largest asteroid, dwarf planet, to be visited by Dawn
2 Pallas	544	2.77	0.16	B	Second largest asteroid by diameter
4 Vesta	529	2.36	0.42	V	Second largest asteroid by mass, parent of HED meteorites, Dawn target asteroid
6 Hebe	186	2.43	0.27	S	Proposed parent body of H chondrites
8 Flora	128	2.20	0.24	S	Namesake of Flora dynamical family
10 Hygiea	431	3.14	0.07	C	Fourth largest asteroid
15 Eunomia	268	2.64	0.21	S	Largest S-class asteroid
16 Psyche	186	2.92	0.12	M	Largest M-class asteroid, thought metallic
21 Lutetia	100	2.44	0.21	M	Rosetta target in 2010
24 Themis	198	3.13	0.068	C	Namesake of Themis dynamical family, thought icy
65 Cybele	273	3.44	0.07	P	Namesake of Cybele group, thought icy
87 Sylvia	286	3.49	0.044	P	First known triple asteroid system
153 Hilda	171	3.97	0.062	P	Namesake of Hilda group
158 Koronis	35	2.87	0.28	S	Namesake of Koronis dynamical family
216 Kleopatra	118	2.79	0.12	M	Triple asteroid system, thought metallic
221 Eos	104	3.01	0.14	K	Namesake of Eos dynamical family
243 Ida	32	2.86	0.24	S	First known binary asteroid system, Galileo target in 1993
253 Mathilde	53	2.65	0.044	C	NEAR Shoemaker target in 1997
323 Brucia	36	2.38	0.18	S	First asteroid discovered via photography
349 Dembowska	140	2.92	0.38	R	Unique surface composition

(continued)

Table 8-1
(Continued)

Name	Diameter (km)	Semimajor axis (AU)	Albedo	Spectral type	Note
433 Eros	17	1.46	0.25	S	First NEO discovered, NEAR Shoemaker target in 2000–2001
588 Achilles	136	5.19	0.033	D	First Trojan asteroid discovered
624 Hektor	241	5.24	0.025	D	Largest Trojan asteroid
951 Gaspra	12	2.21	0.22	S	First asteroid encounter, Galileo target in 1991
1036 Ganymed	32	2.662	0.17	S	Largest known NEO
2867 Šteins	5	2.36	0.34	E	Rosetta target in 2008
3200 Phaethon	5	1.27	~0.1	B	Only asteroid associated with meteor shower
5261 Eureka	~3	1.52	0.35	A	Largest and first-known Mars Trojan
25143 Itokawa	0.32	1.32	0.53	S	Hayabusa target in 2005, first asteroid sample return
99942 Apophis	~0.3	0.922	0.33	S	Makes very close approach to Earth in 2029

Summary table of asteroids notable for orbital or physical characteristics, or which have been spacecraft targets. The spectral class is derived from the Tholen taxonomy

2 Dynamics of Asteroids

2.1 Gravitational and Nongravitational Forces

Asteroids can be found in practically every region of the inner solar system where their orbits are stable on long timescales and in some regions where they are not. The vast majority of them are found in the main asteroid belt, stretching roughly from 1.9 to 3.4 AU. Eighty seven percent of known objects have inclinations between 0° and 15°. However, sizeable populations can be found outside the main belt near 5.2 AU (the Trojans) and 4.0 AU (the Hildas) as well as on orbits inside the main belt both in-between and crossing the orbits of the inner planets.

The shape of the asteroid belt in orbital element space is controlled by resonances. Mean-motion resonances occur where the orbital period is an integral fraction of a planet, for the asteroids almost always Jupiter. Secular resonances occur where the precession of an orbit matches that of a planet. The most important mean-motion resonance for sculpting the asteroid belt is the 3:1 with Jupiter, occurring near 2.5 AU.

An object in a resonant orbit finds its orbit modified by the unbalanced tug of a planet's gravity. The orbital eccentricity will increase until its orbit begins to cross that of Mars, where it can undergo encounters or impacts with that planet. Until an encounter (or impact) removes the body from the resonance, eccentricity continues to increase until the orbit crosses

the Earth's, Venus', and Mercury's in turn. If it manages to remain in resonance without impacting a planet, the body will eventually have its eccentricity raised high enough to impact the Sun (Gladman et al. 1997). During orbital evolution, the quantity $(a(1-e^2))^{1/2}\cos i$(see below) is conserved, and thus objects can move from highly eccentric to highly inclined orbits and back again through time.

Figure 8-3 (Morbidelli et al. 2002) shows the orbital parameters of asteroids on planet crossing orbits in a–e space.

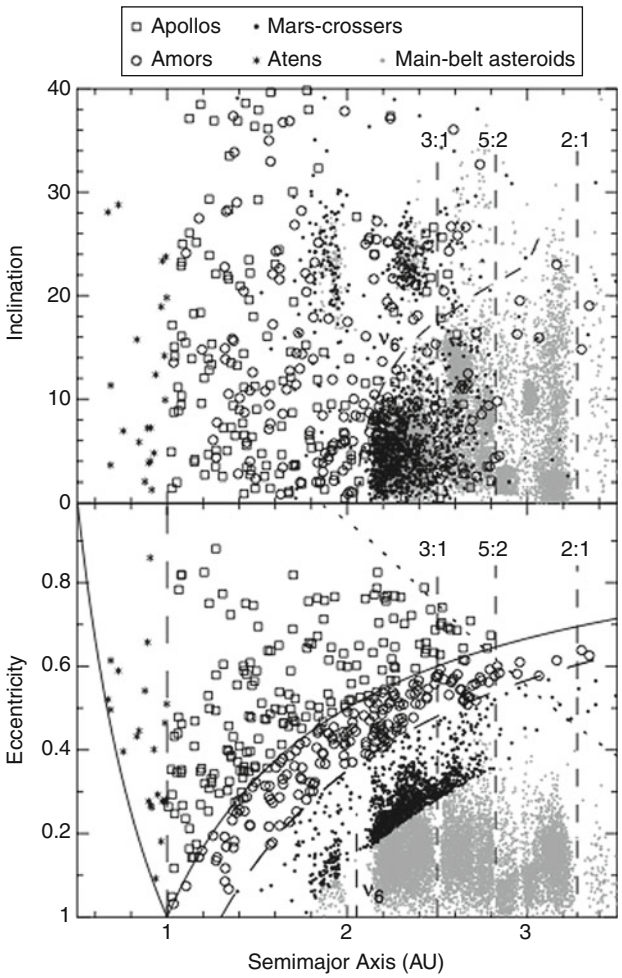


Fig. 8-3

Asteroids in the inner solar system are here plotted in semi-major axis/inclination (*upper* panel) and semi-major axis/eccentricity space (*lower* panel). Dashed lines show the positions of major resonances in the asteroid belt, solid lines in the lower panel show the boundary within which asteroids are Earth-crossing (the Apollos and Atens). (Morbidelli et al. 2002). This figure originally appeared in Asteroids III, U. of Arizona Press

The objects interior to the asteroid belt are of particular interest because of their accessibility and the hazard they pose. An asteroid whose orbit passes inside that of Mars but remains outside 1.3 AU is classified as a Mars-crosser. Objects with perihelia less than 1.3 AU are classified as near-Earth objects (NEOs). This population contains both asteroidal and cometary objects, though is thought to be dominated by the former: Stuart and Binzel (2004) used the differing spectral properties of cometary nuclei and inner-belt asteroids (see ❯ Sect. 2) and the frequency with which comet-like spectra appear in the NEO population, corrected for biases due to differing albedos, to put an upper limit of 17% on the cometary contribution (in the >1 km size range). Dynamically, the likelihood of an NEO originating beyond Jupiter's orbit (and thus being "cometary") is often assessed using the Tisserand parameter T_j:

$$T_j = a_j/a + 2(a/a_j(1-e^2))^{1/2}\cos i,$$

where a_j denotes the semimajor axis of Jupiter. This value is quasi-conserved, remaining close to the same value before and after encounters with Jupiter. For asteroids, T_j is typically greater than 3, while for comets, it is less than 3. NEOs are often further classified based on their orbits: Amors orbit entirely outside Earth's orbit. Apollos have semimajor axes greater than the Earth's, but perihelia inside Earth's orbit, while Atens are the reverse, mostly orbiting interior to Earth but crossing its orbit for aphelion.

In addition to the effect of resonances, nongravitational forces also can move asteroids. The most important is the Yarkovsky force (Bottke et al. 2006). The Yarkovsky force results from the nonalignment between the average insolation vector for an object (pointing from the Sun) and the average emission vector (pointing from the hottest part of the body). Objects with finite thermal inertia are hottest in the afternoon, which for prograde rotators is in the direction away from orbital motion. The extra emission in that direction causes a small momentum transfer that serves to speed up the object and increase its orbit size. For retrograde rotators, the opposite is true and the orbit decays. This is the "diurnal Yarkovsky effect." The "seasonal Yarkovsky effect" is similar, except its concern is the hottest time of the year rather than the day. For the seasonal case, the component of extra emission is always in the direction of orbital motion and therefore it always serves to decrease orbit size. The YORP force, discussed further below, is related to the Yarkovsky force but results in a torque on the body rather than a momentum change.

The Yarkovsky force is critically dependent upon the thermal properties of an object. Measurement of Yarkovsky-induced deflection has been achieved for the NEO 6489 Golevka (Chesley et al. 2003). This measurement has been used to constrain Golevka's density, assuming typical thermal properties of rock. Given the magnitude of the Yarkovsky force and typical thermal properties of asteroids, it is most effective on objects in the ~10 m size range, with appreciable effects still possible up to 1–10 km diameters given long enough timescales. Larger objects are simply too massive for the Yarkovsky effect to be effective, while smaller objects become isothermal and the force goes to zero.

Bottke et al. (2002) numerically modeled the orbital evolution of asteroidal population as a whole, including gravitational and nongravitational effects, estimating the fraction of the NEO population derived from six pathways: those that came to near-Earth space via the ν_6 resonance, via the 3:1 resonance, from the outer asteroid belt, the Jupiter family comets (JFCs), the ecliptic comets, and via low-inclination, Mars-crossing orbits with semimajor axes within the main asteroid belt (IMCs). These pathways ultimately connect the NEO population with the major reservoirs of small bodies in the solar system: the Hungaria, main-belt, and Trojan asteroids, the transneptunian region, and the Oort cloud.

Bottke et al. found the ν_6 resonance to be the most common pathway for NEOs, with roughly 37% of objects entering near-Earth space via that route. The 3:1 and IMCs each accounted for 20–25% of NEOs, with the outer-belt and JFC pathways splitting the remaining 15%.

2.2 Orbital Stability and Lifetime of NEOs

As mentioned at the start of this section, and unlike objects in the main belt, near-Earth objects experience rapid orbital evolution and have relatively short lifetimes. Close encounters with planets alter orbital elements and in combination with resonances result in complicated trajectories through orbital element phase space. Those NEOs with $a > \sim 2.5$ AU can encounter Jupiter during periods of high eccentricity and are preferentially ejected from the solar system by that planet as a final fate unless they enter a resonance and/or encounter the inner planets and lower their semimajor axis. For NEOs with smaller semimajor axes ($a < \sim 2.5$), the most likely fate is impact into the Sun when the eccentricity is high enough. Other possibilities include impacts into the terrestrial planets or the Moon (Morbidelli et al. 2002).

The amount of time spent in the near-Earth region by an object (before it is removed dynamically) is affected by the pathway it used to enter near-Earth space. Those objects that enter via resonances have shorter lifetimes than those that diffuse from the inner asteroid belt. The mean lifetime in all asteroidal cases is of order 1–10 Myr, with very few surviving as long as 100 Myr (Bottke et al. 2002). In order to keep a steady-state population, Bottke et al. predict a flux of 790 ± 200 objects of $H < 18$ (or larger than roughly 1 km in diameter) per Myr from the asteroid belt into the NEO population. They also calculate that this rate implies a mass loss of roughly 5% from the main belt over the last 3 Gyr.

2.3 Collisional Evolution and Families

Dohnanyi (1969) showed that given certain simplifying assumptions, the population of asteroids would reach an equilibrium size-frequency distribution (SFD) with a cumulative diameter power law slope of −2.5, that is, $N(>D) \sim D^{-2.5}$ where N(>D) is the number of objects larger than D km. While later work included several additional important effects, including self-gravitation, the Dohnanyi power law slope is still often used as a point of reference. The observed SFD is consistent with a population in collisional equilibrium.

It is thought that all of the objects in the main belt larger than roughly 10 km have been discovered, and it is seen that the SFD departs from power law behavior, with a "wavy" appearance. Several workers using different techniques have estimated roughly 1 million objects of 1 km or larger in the asteroid belt, though the exact number varies by ±30% in the literature (Morbidelli and Vokrouhlický 2003; Tedesco and Desert 2002; Ivezić et al. 2002; Bottke et al. 2005).

Durda et al. (1998) looked at the SFD in detail to extract the strength properties of the population, also concluding its "wavy" nature reflected the original SFD of the asteroids. The nature of the original "hump" near sizes of ~100 km has remained somewhat mysterious, though very recent work by Morbidelli et al. (2009) suggest centimeter- to meter-sized bodies coalesced due to turbulence in the solar nebula to form objects of roughly this size, which would then represent the original size of planetesimals.

The collisional lifetime of asteroids in the asteroid belt is dependent on diameter; as objects become larger, fewer objects are capable of overcoming the target's gravitational binding energy

to lead to disruption. O'Brien and Greenberg (2005) modeled the main-belt and NEO populations and found a fairly flat collisional lifetime of ~7 Myr for meter-scale objects, varying little for the 10 cm–100 m diameter range. However, it rapidly increases between 100 m and 10 km, reaching 1 Gyr at roughly 1 km, and the age of the solar system by a few kilometers. Given the short lifetime of NEOs versus removal via impact into the Sun or a planet (or for some, ejection by Jupiter), collisional evolution is not important for this population.

The mean impact speed for objects in the asteroid belt is roughly 5 km/s, though individual objects will experience different values and relatively wider or narrower spreads of impact speeds depending upon their eccentricity and inclination (Bottke et al. 1994). These impact speeds result in ejecta that typically moves at m/s to several tens of m/s. This is enough to overcome the escape speed of most objects, but not enough to lead to large immediate dispersion between the orbit of the impacted object and that of its ejecta. As a result, dynamical families of asteroids with similar orbital elements are created after impacts. The largest of these were recognized in the early twentieth century by Hirayama (1918): Eos, Koronis, Themis, Maria, and Flora. By convention, families are named after their lowest-numbered member.

Dynamical families are typically recognized based on similarity among orbital elements. Proper elements are used rather than osculating elements, which have potentially changed significantly with time. However, proper elements require orbit simulations to be performed. The dramatic increase in asteroid discoveries in the 1990s along with increased computing power led to more refined and reliable family identification techniques, as described in Bendjoya and Zappalà (2002).

A combined dynamical and spectroscopic approach is often used to distinguish family members from unrelated objects with coincidentally similar orbital elements (usually called "interlopers" or members of the "background population"). This technique is most useful in situations where the family members have significant spectral differences from the typical asteroid in that part of the asteroid belt: for instance, the S-class Koronis family in the C-dominated outer belt, or the spectrally unusual K-class Eos family or V-class Vesta family. Early workers anticipated the presence of spectrally diverse families as an outcome of the breakup of differentiated objects, but spectral studies have not found any convincing evidence for such families. Instead, those families that have been studied in detail show remarkable spectral similarity (Bus 1999). This is suggestive of either a great rarity of families derived from differentiated objects or else the mixing and homogenizing of material in catastrophic impacts to a scale that results in family members emerging with very similar bulk compositions (Michel et al. 2002) or a combination of both.

Durda et al. (2007) modeled the formation of prominent dynamical families by analyzing the size-frequency distribution of current-day fragments, constraining the properties of the original parent bodies for the largest families. They found that roughly 20 families were created by the catastrophic disruption of objects greater than 100 km in diameter, with the largest including the Themis family (original $D \sim 450$ km) and Eos family (~380 km). The differing SFDs of various families also provide information about their formation. Durda et al., following earlier workers, defined cratering versus catastrophic collisions as those where the largest remaining fragment was greater than 50% of the original mass versus those where it was 50% or less. Cratering collisions result from less-energetic collisions than catastrophic ones. Most of the families associated with the largest asteroids are modeled as cratering impacts by Durda et al., like Juno, Vesta, and Hygiea.

With the (re)discovery of the Yarkovsky effect and its inclusion in dynamical studies, it became possible to estimate the ages of some families based on the spread of their orbits.

Jedicke et al. (2004) used spectral data to argue that the largest families broke up billions of years ago. However, it was also discovered using dynamical models that some smaller clusters of objects broke up much more recently: the Karin cluster only 6 million years ago (Nesvorný et al. 2002) and the Veritas cluster 8 million (Nesvorný et al. 2003). In addition to providing constraints on regolith processes and impact rates, these young families have been linked to dust bands seen in IRAS data. Ongoing work seeks to determine the fraction of interplanetary dust particles (IDPs) falling to Earth from these different sources, as well as from cometary sources.

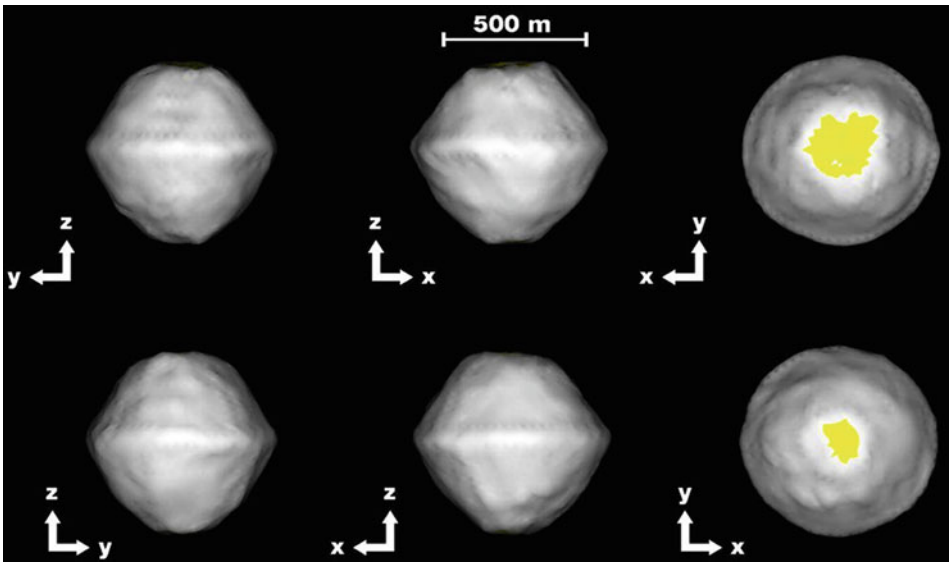
Reflectance spectroscopy, discussed in more detail in Sect. 5, has been used to try and identify "true" families from groups of objects with coincidentally close orbits. Bus (1999) found families tended to be homogeneous in spectral properties, making the identification of interlopers relatively straightforward, especially in cases where the family is of an unusual spectral type either when compared to its neighbors in that part of the asteroid belt or in general. Interestingly, astronomers have yet to conclusively find a family of the type that workers in the 1980s thought would be commonplace (Chapman et al. 1989): a large iron core as the central body with fragments of mantle and crust as the other family members. Burbine et al. (1996) suggested that the difference in strength between mantle and core (or crust) material would lead to the rapid collisional erosion of the former.

2.4 Binary and Multiple Objects

The discovery of 243 Ida's satellite Dactyl by the Galileo spacecraft in 1993 was a surprise (Chapman et al. 1995). Surveys in the discovery's aftermath found ~2% of large asteroids had satellites. This is far less than radar and light curve surveys of NEOs, which find roughly 15% of those objects are part of multiple systems (Richardson and Walsh 2006). This is thought to reflect different formation styles for large and small objects. The satellite systems of larger asteroids are thought to form via collisions, either as large fragments of ejecta fortuitously escaping in the same direction and capturing each other or as ejecta getting captured around the target object through N-body interactions.

NEO systems are thought to be created through YORP, which is more effective on smaller, less massive bodies. YORP acts to put a torque on irregular bodies and increases (or decreases) rotation rates and can move rotation poles (Bottke et al. 2006). For loosely consolidated objects, increased rotation rate causes material to move to the equator and can eventually cause fission and creation of a satellite. Studies of 1999 KW4 (Ostro et al. 2006) show the clearest example of this satellite creation process, which has also been seen in objects like 1994 CC (*Fig. 8-4*). YORP should also be acting on small main-belt asteroids to create satellites, but it has been difficult to detect such systems. Nevertheless, those data that do exist suggest a binary fraction more in line with the NEO values than with the large main-belt objects, consistent with YORP acting on the small MBAs.

Binary systems continue to evolve after formation, and those that are unstable will eventually give rise to either "contact binaries" where the satellite orbit evolves such that it re-impacts the primary at low speed or else the orbit can evolve such that the satellite escapes into an independent heliocentric orbit. Such asteroid pairs have been found and analyzed by Pravec et al. (2010) and are seen to be consistent with fission and escape. Cuk (2007) found that a YORP-like process he called "binary YORP" leads to rapid evolution of the orbits of asteroid binary systems, suggesting that such systems are quite young in geological terms ($<10^5$ years) and also



http://www.sciencedirect.com/science/article/pii/S0019103511003472
(Figure 11)

Fig. 8-4

The YORP process is thought to create many of the binary systems found with kilometer-scale primaries. In this process, the YORP torque increases the spin rate of an object, which causes loose material to move to the equator. At high enough spin rates, fission can eventually occur with a satellite forming. This radar-derived shape model of 1994 CC, a small binary system, shows the characteristics expected for such a process, including a prominent equatorial ridge where material collects. This shape is commonly seen in small binary systems (Figure from Brozovic et al. (2011))

suggesting that a typical kilometer-sized NEO may go through several cycles of satellite formation and destruction before it is itself destroyed, with main-belt objects undergoing many cycles since their formation.

2.5 The NEO Hazard

The resonances that alter NEO orbits and cause them to become planet crossing act quickly. As mentioned above, a typical NEO only lasts for a few tens of millions of years before it is removed via impact. In addition, the cosmic-ray exposure ages of meteorites, which measure the time spent in the main belt or NEO space as a meter-scale object before Earth impact, are relatively short compared to the age of the solar system (<120 Myr for stony meteorites, <1.5 Gyr for irons: Eugster et al. 2006). As a result, a constant resupply is necessary to explain the population seen today. Collisional ejecta is very unlikely to find itself on a resonant orbit. The Yarkovsky force, however, leads to a steady input of material into resonances far beyond what collisions can directly input (Bottke et al. 2005). While some objects near resonances may contribute an increased share of material, the influence of the Yarkovsky force leads to a more well-mixed NEO population.

The recognition that asteroidal impacts play a role in terrestrial extinctions has led to increased emphasis on understanding the threat from, and trying to avoid, future impacts. This effort not only involved engineering discussions of mitigation and models of how large impacts affect the biosphere, not covered here, but also discovery and characterization of NEOs. It was found that impacts with objects of 1 km are sufficient to have global consequences. Roughly 90% of NEOs 1 km diameter or larger have been discovered according to recent models, with an extrapolated population of ~1,000–1,100 such bodies (Stuart and Binzel 2004), though that number is under regular revision. Discussion of the hazard has led to the concept of "potentially hazardous asteroids" (PHAs), which can approach within 0.05 AU of the Earth's orbit and have absolute magnitudes of 22.0 or brighter, implying sizes of 120–150 m or larger, depending on assumed albedo. This set of objects is of sufficient size to have at least regional effects and orbits that can change on the timescale of thousands of years to become Earth-impacting. Actuarial studies suggest that the largest current threat to Earth, in a statistical sense, is the undiscovered population of PHAs, rather than any known and cataloged object.

The asteroid 25143 Itokawa is the best-known and best-studied PHA, though it does not approach closer to the Earth than 2 million km. Other than Itokawa, 99942 Apophis is likely the most famous PHA, with an upcoming approach to the surface of the Earth of under 40,000 km in 2029. The only object ever discovered on a collision course with Earth, 2008 TC3, was only 2–3 m in diameter and thus too small to be a PHA or have any serious effect on Earth or its life.

The threat of larger impacts has led to interest both inside and outside the scientific community. The US Congress has charged NASA with conducting asteroid surveys, and in addition, some researchers have studied how best to mitigate any hazards that are found. A variety of methods have been proposed, depending on the amount of time available before impact and the size of the impactor. However, the effectiveness of these mitigation techniques is difficult to confidently forecast without additional information about asteroid interiors in general. Ideally, detailed information about the physical properties of any impactor would be available prior to any mitigation is attempted, but effort is being placed into determining average properties, and the variation of those properties for the NEO population at large in case such mitigation timescales are short. Beyond the physics of handling a possible impact, the policy and legal aspects of such a situation are far from settled. While a detailed discussion of these issues is beyond the scope of this chapter, recent NASA-sponsored studies by the community are available (NASA's NEO office hosts links to those studies at http://neo.jpl.nasa.gov/links/).

2.6 The Nice Model

In addition to current-day dynamics, recent work has looked at the conditions in the early solar system. The discovery of "hot Jupiters" around other stars has led to models of planetary migration, where those planets were formed far from their primaries and moved inward due to scattering of planetesimals. While any single interaction between a planetesimal and a gas giant has very little effect on the giant planet, the cumulative effect of countless interactions can rob its orbit of energy and cause the orbit to decay. Models of our own solar system found that the cumulative effects were to move Jupiter inward (as it scattered planetesimals outward) but to move Saturn, Uranus, and Neptune outward (as they scattered objects inward). According to the models, Jupiter and Saturn enter resonance with one another after several hundred million years, with immediate and strong effects on the remaining small body populations as all four gas giants have rapid orbital changes, resonances sweep the asteroid belt, and Neptune

is thrust out to its current orbit into what was then a Kuiper belt that extended closer to the Sun. This model, named the "Nice model" after the French city where the first calculations were made, was published in a set of papers in the first decade of this century (Gomes et al. 2005; Morbidelli et al. 2005; Tsiganis et al. 2005).

Neptune's motion into its new orbit is thought to have resulted in a massive shower of KBOs into the inner solar system. It also may have allowed large numbers of KBOs to be trapped in a 1:1 resonance with Jupiter, resulting in today's Trojan asteroid population. Furthermore, some models show that some KBOs can become implanted in the asteroid belt, where they can still lurk today. The first reconnaissance of KBO spectral properties is still underway, but from the study of cometary nuclei, it has been suggested that the D-class asteroids (see Sect. 2), which is a main spectral type for Trojan asteroids, may be good candidates for implanted KBOs in the inner solar system. Intriguingly, Phobos and Deimos also have D-class spectra over most or all of their surfaces. The origin of the Martian moons is still a matter of fierce debate, but an outer solar system origin is potentially consistent with the Nice model.

This shower of KBOs (and likely, asteroids) may have left evidence in the radiometric ages in meteorites. An apparent peak in impacts near 3.9 billion years ago was noted in lunar samples, dubbed the "terminal lunar cataclysm" by some lunar researchers, though the evidence remains controversial (Cohen et al. 2000). Swindle et al. (2009) studied a set of H-chondrite samples and reviewed other meteorite work, concluding that their argon-argon ages were consistent with disturbances at roughly 3.9 Gyr ago, interpreted as due to a large impact flux.

3 Geology and Surfaces of Asteroids

The great range of sizes present in the asteroidal population leads to a large range of surface properties. The largest objects are thought to have surfaces like that of the Moon: heavily cratered and covered in surface layer of regolith. This is borne out by spacecraft visits. The smaller objects, those only a few tens of meter in size or smaller, are thought to be too small and easy to disrupt to have experienced much in the way of processing at their current sizes and are expected to have the properties of similar-sized rocks on Earth. Sullivan et al. (2002) reviewed our knowledge of asteroidal surface geology at the time between the NEAR and Hayabusa missions, with the geology of Itokawa presented in a number of papers listed below, and Steins in Keller et al. (2010). The most recent asteroid encounter that of 21 Lutetia by the Rosetta spacecraft has not yet been analyzed fully at this writing.

3.1 Cratering

The most pervasive process occurring on asteroids is impact cratering, which continues to this day. While structures interpreted as craters have been detected on some radar-observed NEOs (Busch et al. 2008, for instance) and a very large crater has been detected in HST observations of Vesta (Thomas et al. 1997), the vast majority of our knowledge of asteroidal cratering at this writing comes from spacecraft imagery of Ida, Gaspra, Eros, Mathilde, and Itokawa.

Crater counts can be used to determine the ages of surfaces, given a knowledge of the impactor population (or production function), and assuming cratering has not been so intense as to saturate the surface. Extrapolations from the lunar cratering record can be used, provided

adjustments are made for the different gravity, composition, and impact speeds experienced by the asteroids (and indeed, from one asteroid to another).

Chapman et al. (1996a, b) looked at the crater populations of Gaspra and Ida from the Galileo encounters with those objects, finding Gaspra to have a relatively sparse crater density and an age ~200 Myr, while Ida was found to have a surface saturated with craters in the 100 m–1 km size range and an age roughly ten times older than Gaspra. The record on Eros is similar to that of Ida for craters larger than 100 m, though with a distinct depletion of smaller-sized craters (Chapman et al. 2002). Eros' exit from the main-belt population to the NEO population changed the nature of its impactor population, but this depletion is now ascribed to erasure via seismic shaking of the regolith during subsequent impacts (see below).

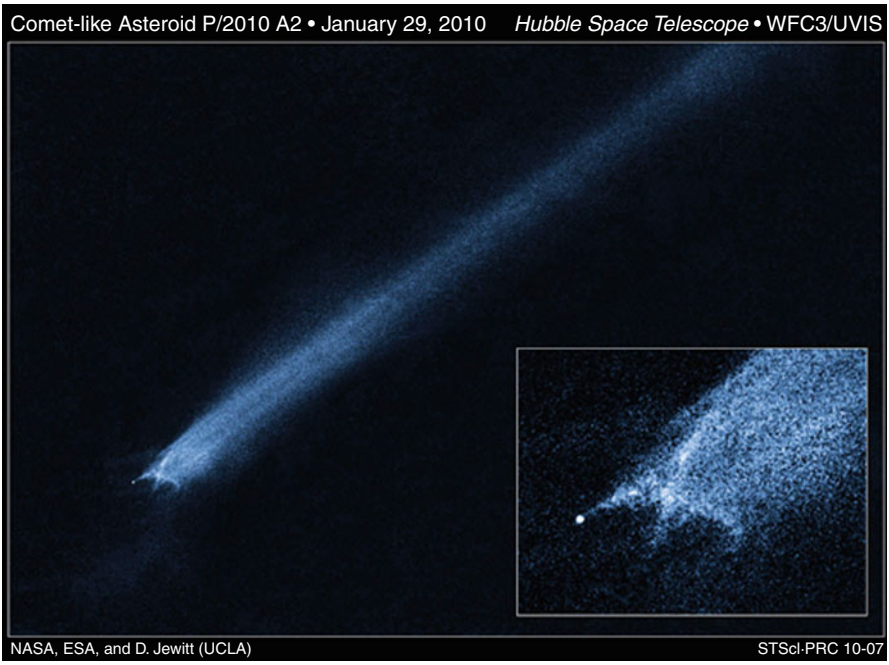
In contrast to these S asteroids, the crater population of the C-asteroid Mathilde is dominated by several very large features, which play a large part in defining Mathilde's shape. While Mathilde's slow rotation hid roughly half of its surface from NEAR Shoemaker during its flyby, four craters were observed with diameters larger than Mathilde's radius (Chapman et al. 1999), which had been generally thought of as the limit beyond which disruption would occur. There was, however, little evidence these cratering events had large-scale effects on the object. Davis (1999) showed that these craters and Mathilde's non-disruption could be explained if Mathilde was exceedingly porous, a result consistent with its low density. Davis estimated Mathilde's surface age at roughly 4 billion years old, though uncertainties in the production function and collisional physics allow an age as young as 2 billion years.

O'Brien et al. (2006) used a model production function created from various astronomical and meteoritic constraints to fit the crater populations of Gaspra, Ida, Mathilde, and Eros, including processes responsible for crater degradation and erasure. They found Gaspra to be best fit by roughly 1 billion years of cratering, with an event roughly 65 Myr ago erasing craters <3 km, which have been reaccumulating since. Ida's surface is equally well fit by ages beyond 250 Myr, though they used additional constraints to argue a likely age of 0.5–1 Byr for that object. For Eros, the best fit age is 120 Myr, while for Mathilde, a 4 Gyr age was found, consistent with Davis (1999).

Michel et al. (2009) applied the O'Brien et al. production function to Itokawa, finding its surface age to be at least 75 Myr and possibly as long as 1 Byr. However, there is a pronounced deficiency of craters 10 m and smaller on Itokawa's surface, beyond what the seismic shaking process can explain. Michel et al. noted that models of Itokawa's structure by Scheeres et al. (2007) suggest it is a possible contact binary, formed via a low-velocity impact under 1 Myr ago, and that such an impact could explain the deficiency in small craters. Alternately, armoring of the surface by Itokawa's gravely regolith could also play a part by making small crater creation more difficult, rather than invoking selective erasure.

The most recent asteroid encounter for which full results are available at this writing is the Rosetta encounter with the E-class asteroid 2867 Steins (Marchi et al. 2010), which is intermediate in size between Eros and Itokawa but unlike those objects resides in the main asteroid belt. As with Itokawa, Marchi et al. found the modeled cratering age of Steins to depend strongly on model parameters, ranging from 154 Myr for one scaling law to 0.49–1.6 Byr for a different scaling law. Another similarity with Itokawa is a lack of small craters, less than ~500 m on Steins. This was interpreted as due to erasure and resetting at the time of the largest impact on Steins (provisionally named Ruby crater). Under that interpretation, the Ruby impact occurred ~2–10 Myr ago, with smaller craters accumulating since.

Very preliminary results for the July 2010 encounter with 21 Lutetia, also by Rosetta, have been presented at scientific conferences. While details have not yet been published at the time of



http://www.nasa.gov/mission_pages/hubble/science/asteroid-20100202.html

Fig. 8-5

This Hubble Space Telescope image shows the asteroid 2010 A2. Given its cometary appearance, initial speculation centered on it being a volatile-rich object with a sublimation-driven tail. However, detailed imaging over time showed instead the 3-D structure of the tail and its evolution much more consistent with dust and debris generated from an impact event on the asteroid. This is thought to be the first time the aftermath of an impact has been seen in the small bodies population (Image taken by the Hubble Space Telescope, reproduced courtesy of NASA)

this writing, the conference abstracts suggest Lutetia has regions of different ages on its surface, with some "near craterless terrains" suggesting recent impacts (Keller et al. 2010). The same imagery also has found large craters, with diameters comparable to Lutetia's radius (up to about 60 km: Marchi et al. 2010).

The recent discovery of the "comet" P/2010 A2 (*Fig. 8-5*) has allowed the aftereffects of an asteroid-asteroid impact in the main belt to be studied. This object had a tail at its discovery and has been classified as a comet as a result. However, its orbit lies wholly within the inner asteroid belt, and it was suspected early on of not being a "true" comet (i.e., one whose tail and coma are generated by sublimation of volatiles). A pair of observing campaigns including the HST and Rosetta spacecraft allowed a 3-D reconstruction of the tail morphology and led to the conclusion that P/2010 A2 was the target of a collision in February 2009, less than a year before its discovery (Snodgrass et al. 2010; Jewitt et al. 2010). Tail evolution was also best fit when modeled with millimeter- to centimeter-sized particles, rather than the smaller ones expected from typical cometary dust. It is not yet clear whether this indicates the impact was not capable of creating smaller particles or if it reflects the typical particle size at the surface of

P/2010 A2. P/2010 A2 (at ~120 m) is smaller than any object visited by spacecraft, with Itokawa the nearest in size. Snodgrass et al. concluded the dust mass represented roughly 16% of the original parent body mass, and with an initial target size of ~120 m further concluded that the impactor was likely ~6–9 m in diameter. P/2010 A2 is expected to experience impacts of this size every 1.1 billion years, but such impacts should be occurring every 12 years somewhere in the asteroid belt.

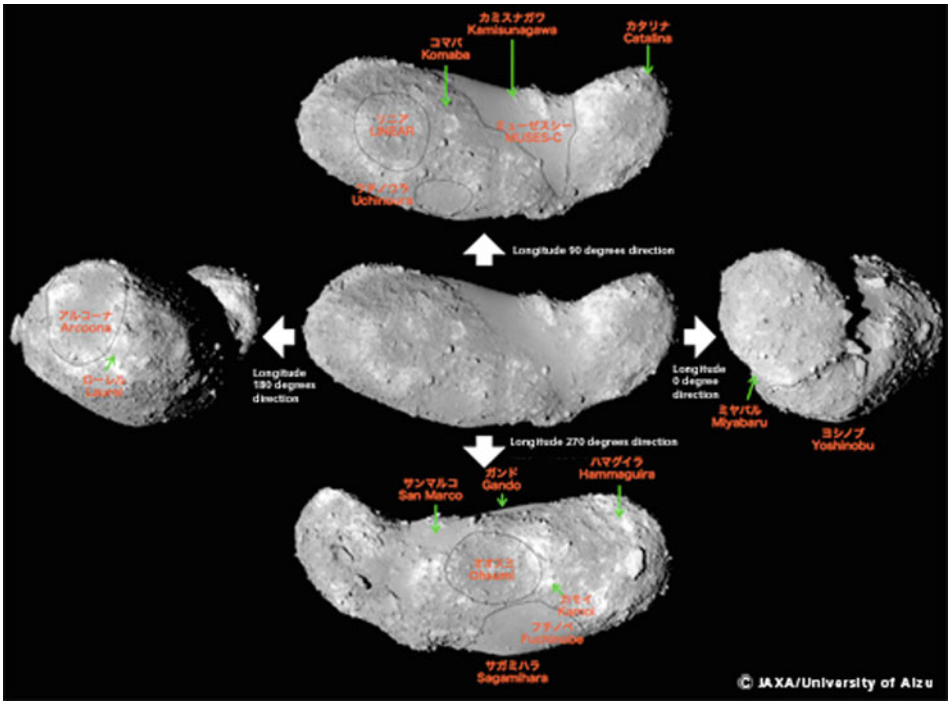
At this writing, the asteroid 596 Scheila has very recently been observed with a tail and coma. It is not yet clear whether this comet-like activity is dust lofted after an impact, like the interpretation of 2010 A2, or whether it is volatile-driven (perhaps prompted by impact excavation), similar to what is seen in the main-belt comets.

3.2 Regolith

In addition to cratering, impacts on asteroids serve to pulverize the surface. Because the mean impact speed in the asteroid belt is roughly 5 km/s, even very small impactors can have an effect: a 1 mm diameter impactor at that speed carries the same kinetic energy as a typical person swinging a baseball bat. Over time, a layer of broken up material, the regolith, is left covering the surface. The regolith in a given location is generated both from smaller impacts breaking up local rocks and from the ejecta of more distant impacts arriving on ballistic trajectories. In a quantitative sense, large impacts on the Moon serve to fracture bedrock and deepen the regolith layer, while smaller impacts further pulverize and mix the regolith (Hartmann 2003). The lower gravities and irregular shapes found among typical asteroids compared to the Moon complicate application of lunar models to the asteroids. Dombard et al. (2010) studied NEAR Shoemaker imagery of boulders on Eros, concluding that some boulders erode in place into regolith. They proposed that thermal cycling of boulders could cause their disaggregation, with potential applicability to other asteroids, particularly those with minerals of differing thermal properties (like the ordinary chondrites, with both metallic and silicate components). Quantitative modeling of regolith formation on asteroids is only in its early stages, hampered by a lack of data other than Eros and Itokawa (*Fig. 8-6*).

What data does exist comes largely from spacecraft imagery. Studies of depth/diameter ratios of craters compared to theoretical values for newly formed craters suggest that Ida, Gaspra, and Eros have tens of meters of impact-generated regolith on their surfaces (summarized in Sullivan et al. 2002 and references therein). Similar values are derived by studies of other geological features. Cheng and Barnouin (2010) suggested that asteroids generate regolith depths of roughly 0.2% of their diameter, consistent with previous estimates and implying a regolith depth of ~200 m for Lutetia and ~1–2 km for Vesta and Ceres if these objects behave the same way smaller objects do. These regolith depths can be higher than what is estimated for the lunar regolith (again, tens of meters: Wilcox et al. 2005), but given the possible differences in formation, it is again not clear how to best apply lunar constraints to the asteroids.

When dealing with small rubble piles, the situation becomes still murkier. Cheng argued that Itokawa may have no bedrock at all and could be an accumulation of boulders and finer material formed on an earlier parent body: effectively regolith the whole way through. The satellites of smaller asteroids could also be entirely composed of regolith if they formed via the YORP mechanism as current theories predict (see Sect. 2), as would be any escaped satellites now on independent orbits.



http://www.jaxa.jp/press/2009/03/20090303_itokawa_e.html

Fig. 8-6

The asteroid Itokawa was explored by the Hayabusa spacecraft, built by the Japanese space agency JAXA. This diagram shows Itokawa from all sides, with major named areas labeled. The appearance of Itokawa has led to a general consensus that it is a rubble pile, a loose collection of boulders and smaller debris held together by gravity and small-scale cohesion rather than tensile strength (Image courtesy of JAXA)

The differing thermal properties of bare rock and regolith (and the varying thermal properties of a dusty regolith versus a cobbly regolith) provide another means of understanding asteroid surfaces. Delbo' et al. (2007) showed that thermal models of NEOs as a group were consistent with an average thermal inertia of 200 ± 40 J m^{-2} $s^{-0.5}$ K^{-1} (compared to values near 10–20 in those same units for the largest main-belt asteroids), which suggests the presence of regolith rather than bare rock (bare rock thermal inertias are over 1,000 in those units). However, they also suggested the thermal inertia of asteroids steadily increases with decreasing size, consistent with increasing particle size with decreasing diameter. Data are relatively sparse, however, and as objects become smaller, their uncertainties become larger. It is not yet certain that the thermal inertia does not reach a plateau at ~10 km and remain at that value for smaller objects in general.

Impacts into regolith-covered objects have different results depending upon whether or not the impact is large enough to reach the bedrock at the base of the regolith layer. Large enough impacts will pulverize bedrock to create more regolith. Small impacts contained entirely within the regolith layer will serve instead to mix, or garden that layer. In this way, surfaces that

experience space weathering or similar maturation processes (as discussed below) can be "reset," and fresh material brought to the optical surface. Because the porosity of the regolith may differ from that of the target as a whole, these smaller impacts may also be qualitatively different in terms of the amount of impact melt produced, ejecta speeds, etc.

3.3 Processes

The surfaces of the asteroids have been exposed to the solar wind and micrometeorite impacts for billions of years. The extent to which this exposure affects those surfaces has been a matter of intense debate, although there does appear to be a growing consensus. The term "space weathering" is often used to describe these processes, or more broadly any process that changes the surface of an airless body, whether the Moon, Mercury, asteroids, or icy outer solar system objects.

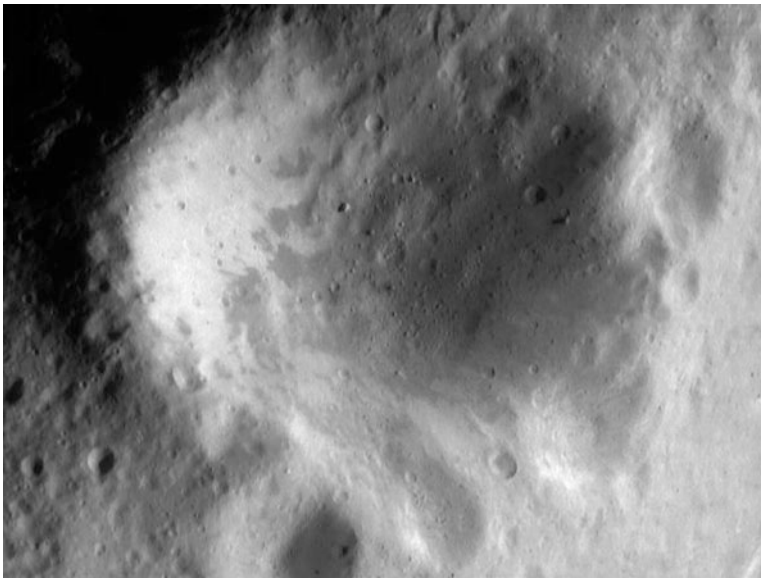
Studies of the Apollo samples determined lunar soils exhibited a range of correlated properties: solar wind-derived gasses, metallic iron content, and amount of agglutinitic glasses were found in greater quantities in soils that spent more time on the lunar surface. A series of papers at the turn of this century deduced that many of these markers of increased maturity were concentrated in the finest particle sizes and were associated with coatings of reduced iron, termed "nanophase iron" or npFe (Pieters et al. 2000; Noble et al. 2001; Hapke 2001). These coatings are thought to be due to the reduction of iron in silicates during micrometeorite impacts and its subsequent deposition from vapor phase on other grains. Sputtering and chemical reduction of surfaces caused by solar wind interactions has also been invoked as a driver of space weathering, which would act independently of micrometeorite-driven processes (Loeffler et al. 2009). Destruction of a mature, space-weathered regolith (also called resetting) can also occur, either through dilution through mixing with fresh, previously buried material during the course of subsequent impacts, or acceleration off of the object entirely if those subsequent impacts are sufficiently large (or alternately removal through YORP, see Sect. 2).

The NEAR Shoemaker mission provided conclusive evidence for space-weathered material on Eros' surface (*Fig. 8-7*). Low-albedo regolith sits atop high-albedo regolith, and the low-albedo material can also be found collecting in lower areas after mass wasting.

The amount of space weathering experienced by a given body is likely to be a complex combination of its collisional and orbital history. This has led to efforts to study the processes collectively by observing the statistical properties of large samples. These studies will often use size as a proxy for age: in general, smaller objects will have shorter lifetimes against collisional disruption than larger objects on equivalent orbits, as discussed in Sect. 2. They also will potentially retain a less-developed regolith (due to lower gravity). While a specific 2-km object may or may not have an older surface than a specific 1-km object, the aggregate population of 2-km objects is expected to have older surfaces than the comparable population of 1-km objects.

Binzel et al. (2004) observed the spectral slopes of the S-complex NEAs (specifically those falling in the S, Q, or Sq classes) and found the average spectral slope increased from a value like the OC meteorites for smaller objects (~1 km diameter and less) to a value typical of the main-belt S asteroids for objects 5 km diameter and larger.

The timescale on which space weathering acts has been the subject of ongoing research. The study of very recent families along with some laboratory studies suggests a very rapid timescale for the process: Vernazza et al. (2009) proposed weathering occurs within 1 Myr, while Nesvorný et al. (2010a) concluded that the onset occurs at roughly 1 Myr for NEOs, with



http://near.jhuapl.edu/iod/20001110/index.html

Fig. 8-7

This crater on 433 Eros shows evidence of space weathering. Regolith darkens with exposure to micrometeorites and the solar wind, while material just beneath the surface retains its original character. This is seen at the *left side* of the crater, where dark material has moved down the slope to the center, exposing fresher, brighter, and unaltered material (NEAR Shoemaker image, courtesy of JHU/APL and NASA)

a different balance of processes in the main asteroid belt likely leading to a different but roughly comparable timescale. The short dynamical lifetime of the NEAs and the common presence of apparently weathered S asteroids in that population seem to require a timescale shorter than 10 Myr. However, there is evidence for another much longer-acting weathering process as well. The multifamily study of Willman and Jedicke (2011) led to an estimate of ~2,050 Myr for the weathering timescale, roughly consistent with the work of Grier et al. (2001) who found lunar materials did not fully weather until roughly 1 Gyr. Vernazza et al. (2009) also found evidence for a second, slower process along with a faster process.

The presence of OH/H_2O in the lunar regolith (Pieters et al. 2009; Sunshine et al. 2009; Clark 2009) has been attributed to the interaction of solar wind protons with the lunar surface in a form of space weathering. While still a very recent result, the implications for asteroidal regolith are obvious. Taken at the simplest level, the asteroids should be more susceptible to solar wind-created OH/H_2O than the Moon since they are colder and have had ample time to more than make up for their greater distance from the Sun. However, a more detailed look suggests a more complicated picture: objects are commonly found lacking the deep absorption bands seen on the Moon, including on relatively lunar-like objects like Vesta (Rivkin et al. 2006). NEOs like Eros and Toutatis are also lacking the bands in available data (Rivkin and Clark 2001; Howell et al. 1994). Further work will likely clarify the extent to which the processes seen on the Moon are acting on the asteroids and why differences may exist.

3.4 Regolith Movement and Mass Wasting

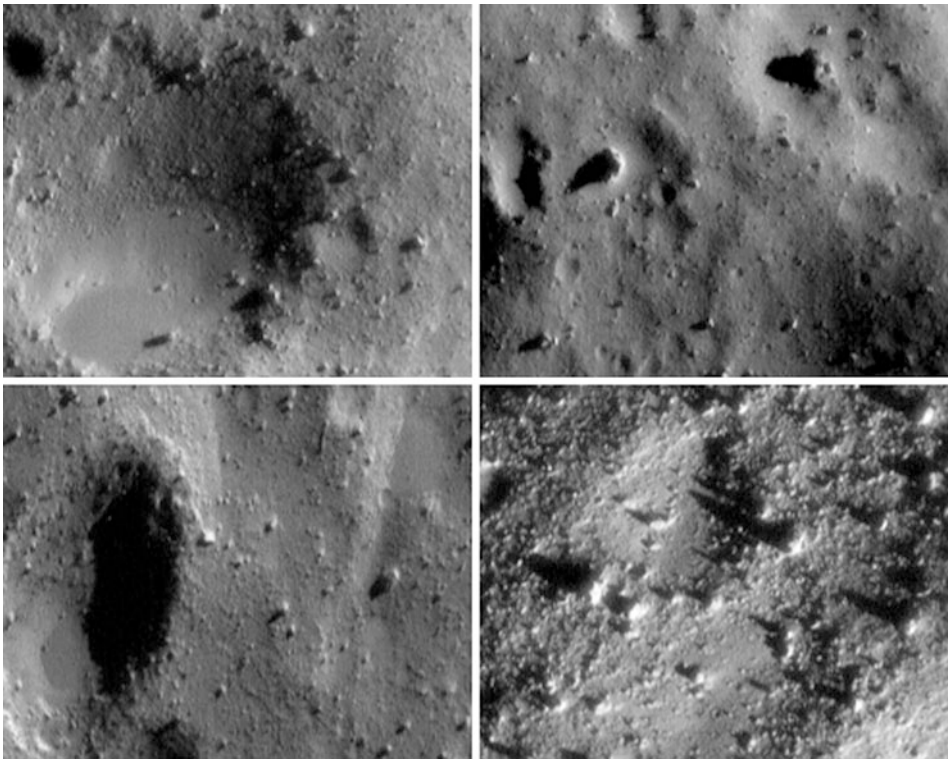
Mass movement occurs on asteroidal surfaces via several processes, which serve to soften (or erase) craters and move material from high to low elevations. Although not typically considered in this category, the accumulation of impact ejecta is one such process. The low gravities of the asteroids visited by spacecraft are not conducive to the formation of continuous and discontinuous ejecta blankets, as is seen on the Moon, but models have shown that ejecta and boulders from impacts can have a blanketing effect far from the impact point (Geissler et al. 1996).

Two similar effects, jolt and seismic shaking, are thought to be responsible for the majority of mass movement on larger asteroids. Jolt is due to seismic waves from a large impact globally destabilizing slopes and moving mass (Nolan et al. 2001). Seismic shaking is a smaller-scale process, which when integrated over long timescales and over an entire object leads to similar results as jolt. Thomas and Robinson (2005) concluded the formation of a single 7.6 km crater on Eros was responsible for the removal of 500-m craters over nearly 40% of its surface via seismic "degradation of topography." Richardson et al. (2005) demonstrated that the collisional evolution of an Eros-like object, including the effects of seismic shaking, resulted in a good fit to the Eros crater size-frequency distribution. They found that depending on the seismic propagation properties, impactors as small as a few meters or less could potentially destabilize all of the slopes on an Eros-sized object. Conversely, they predict that there is an upper limit to the size of affected bodies, and objects larger than ~70–100 km will not experience global effects from a single impact. This size is roughly the size of Lutetia, which should provide an interesting constraint for seismic shaking models.

More recently, it has been realized that the YORP force, in addition to affecting the direction of asteroid spin vectors, can also play a role in regolith transport. In cases where YORP increases spin rate, a potential gradient is created from pole to equator, preferentially causing regolith to move to lower latitudes. As the process continues, material may be lifted from the surface to create a satellite (see ❯ Sect. 2). Evidence for this has been seen on two objects, 1999 KW4 (diameter ~1.2 km) and 1996 RQ36 (diameter ~560 m), both seen through radar-derived shape models to have dramatic equatorial belts consistent with YORP-driven regolith movement (Ostro et al. 2006; Nolan et al. 2007). Although data are still scarce, it seems possible that the shape of KW4 will come to be understood as a typical shape for asteroids in its size range.

The rapid weathering timescale seen by Vernazza et al. and other workers suggests that NEO surfaces should be fully weathered through most of their lifetimes in near-Earth space. The relatively large fraction of Q-class asteroids in the NEO population requires a means of "resetting" their surfaces to explain the fresh material seen spectroscopically. A series of papers looked at the distribution of orbits for S- and Q-class NEOs, interpreted as fresh and mature surfaces, reaching the conclusion that tidal forces during close passes to the Earth and/or Venus are responsible for the resetting (Binzel et al. 2010; Nesvorný et al. 2010a). This could be seen as similar to comet disruptions during close passes to Jupiter or the Sun (Asphaug and Benz 1996), but obviously much less energetic than those cases. The close pass of Apophis to Earth in 2036 has been suggested as a test case for theories of tidal resetting of regolith (Holsapple and Michel 2006).

The most striking and unexpected surface features seen on Eros were the so-called ponds: smooth areas apparently composed of fine-grained material, or at least free of larger cobbles and rocks typically seen elsewhere on the body (see ❯ *Fig. 8-8*, Robinson et al. 2001;



http://near.jhuapl.edu/iod/20010131/index.html

Fig. 8-8

Regolith on Eros is seen close-up in these four panels. The two *left panels* show the phenomenon known as "ponds," very smooth areas where fine-grained regolith collects. It is not yet clear whether this is primarily due to effects like seismic shaking, electrostatic levitation, or a combination of those and other processes (NEAR Shoemaker image, courtesy of JHU/APL and NASA)

Sullivan et al. 2002). Often, ponds were seen on crater floors. These deposits, including NEAR Shoemaker's landing spot, are mostly found near the equator (over 90% of larger ponds are within 30° of the equator), which also include regions with relatively low gravity and areas that are preferentially near the terminator due to Eros' very high obliquity. In addition to images of ponds available from orbit and NEAR Shoemaker's landing sequence, spectral data shows the ponds to be less red and have a deeper silicate absorption band than non-pond regions on Eros, suggesting pond material is less altered than the typical background material.

Pond-like morphology can also be seen on Itokawa, despite its small size. Fujiwara et al. (2006) divided Itokawa into "rough" and "smooth" terrain and noted the similarity between the latter and ponds on Eros in low-resolution imagery, though the particle size in the Itokawa smooth terrain is coarser than what is expected in the ponds on Eros. Barnouin-Jha et al. (2008) concluded that the smooth terrain on Itokawa (which they called the "lowlands") is flatter than any area on Eros except for perhaps its ponds and also noted that at least 1 pond could be seen on a crater floor on Itokawa, echoing what is seen on Eros.

Two theories have been invoked for explaining these ponds. The first is electrostatic levitation. Regolith illuminated by sunlight will become charged via the photoelectric effect, with solar wind electron flux providing a competing effect. If the electrostatic repulsion force exceeds the gravitational force, the charged regolith may levitate and follow field lines to other parts of the body or else escape entirely. Scheeres et al. (2010, following Colwell et al. 2005) calculated asteroids at a variety of solar distances would obtain a surface potential of 4.4 V near the subsolar point, thought to be insufficient for motion of even the smallest regolith particles. However, near-terminator regions may have a greatly enhanced electric field, and scaling from lunar models suggests particles up to centimeter in size could be transported on an Itokawa-sized object. Hughes et al. (2008) modeled dust levitation on Eros and found it provided a plausible scenario for pond formation there, although they were unable to explain their geographic distribution and considered the model incomplete.

Seismic shaking has also been proposed as the explanation for the ponds, as well as for the lack of small craters. In this case, mobilized regolith would collect at low areas on an asteroidal surface, and additional impacts would serve to further add to the pond (Cheng et al. 2002).

3.5 Outgassing

The asteroid belt is thought to be near the position of the "snow line" or "ice line": the distance in the solar nebula at which local conditions allowed the condensation and accretion of water ice into solid bodies. There is ample evidence from the meteorite collection that water was important in the evolution of at least some carbonaceous chondrites, and as is discussed in Sect. 4, there is evidence for surviving reservoirs of internal ice in Ceres and other objects.

However, the surface temperatures experienced today for most main-belt asteroids are generally too high to allow ice to remain stably across their surfaces. Schorghofer (2008) modeled the temperature distribution for asteroids as a function of rotation rate, obliquity, semimajor axis, and thermal inertia, finding that beyond roughly 3.1–3.2 AU ice could remain stable in the shallow subsurface at low latitudes for realistic asteroidal objects, where it could plausibly be liberated by impacts. This modeling is consistent with the observations of "main-belt comets," (MBCs) whose activity is thought to be driven by sublimation of ice newly excavated from their interiors. As sublimation continues, an ice-free lag deposit forms, like what is found on typical comets, and eventually, ice availability and activity ceases.

It is not known whether outgassing is occurring today on objects other than the MBCs. Observations of Ceres by A'Hearn and Feldman (1992) found evidence of OH emission from its sunlit pole, with none from the pole in darkness. Ceres shows no evidence of ice in spectra of its surface, and it is too warm to retain ice in its near-surface save very close to its poles. However, the possibility that Ceres' surface is a relatively thin lag deposit over relatively pure ice (Rivkin et al. 2011b) provides reasonably easy resupply of ice to the surface, and while the A'Hearn and Feldman observations have not been confirmed, the possibility remains that Ceres hosts a local outgassing-induced atmosphere near its summer pole.

3.6 Cohesive Forces

It has recently been recognized that the cohesive forces that act between particles in the microgravity conditions found on the surfaces of small asteroids can be much stronger than seen in more familiar conditions (Asphaug 2008; Scheeres et al. 2010). Scheeres et al. (2010) compared

the importance of several forces experienced by asteroidal regolith, including electrostatic, solar radiation, and surface cohesive forces. They found that for objects the size and gravity of Itokawa, van der Waals cohesive forces dominate over other forces, including gravity. They propose that clumping of small particles could mimic larger particles, for instance, what appears to be centimeter-sized gravel on Itokawa could instead be centimeter-sized clumps of much finer particles. This leads to the possibility that some properties seen on asteroidal surfaces and attributed to strength, such as cliffs or ability to withstand rapid rotation, may instead be consistent with cohesion in a rubble pile of centimeter-sized or smaller particles.

4 Asteroidal Interiors and Geophysics

The study of asteroidal interiors is still in its infancy, though it has made great strides in the past decade. Values for asteroidal masses and densities are still scarce, but becoming more common through precise astrometry and discoveries of binary objects. Furthermore, a precise shape model has helped constrain the interior of Ceres, with applications to other objects likely to soon follow.

4.1 Asteroid Sizes and Densities

Knowing an object's density provides a relatively easy qualitative insight into its interior structure. Determining the densities for asteroids has been challenging, however, and well-constrained densities are only available for a relative handful of objects. Indeed, for the vast majority of objects, even the size is poorly constrained, with only a brightness available. As a result, the absolute magnitude (H) of asteroids is often reported in place of a size. The H magnitude represents the brightness of an asteroid if it were at a distance of 1 AU from the Sun, 1 AU from the Earth, and at 0° phase angle. While that geometry is impossible to attain in reality, the H magnitude can be calculated from observations at other geometries. The diameter (D) of an asteroid is related to the absolute magnitude through the albedo (p), or fraction of incident light reflected from the surface:

$$D = 1,329(p^{-1/2})10^{-0.2H}$$

Albedos are best measured by simultaneous observations of emitted and reflected flux, although often these must be measured separately due to instrumental limitations. Harris and Lagerros (2002) review thermal models for asteroids.

If no albedo information is available or inferable, the range of typical albedos for asteroids (~0.04–0.4) leads to an uncertainty of a factor of ~3 in size, leading to a factor of ~27 uncertainty in volume. Reducing the albedo uncertainty to a factor of 2 (by constraining a spectral type or by other means) reduces the size uncertainty to a factor of 50%, reducing the volume uncertainty accordingly. A measurement of size via albedo measurement, occultation, direct imaging (via AO or spaceborne techniques), or otherwise leads to much improved volume estimates, and typically volumes are much more well-constrained for asteroids than mass. However, precise volume estimates are very much dependent upon good shape models, which require extended light-curve campaigns or close enough passes for radar experiments if objects are not large enough to be directly imaged. Spacecraft flybys can reverse the typical situation, providing a secure mass but leaving the volume relatively less well-constrained, particularly for slow rotators

or objects observed near a solstice, where a large fraction of the object may remain in shadow through the spacecraft observing period.

Masses can be quite difficult to obtain for all but the largest asteroids, which themselves have only been given precise masses relatively recently. Most of the mass of the asteroid belt is concentrated in a very small number of bodies, with Ceres alone accounting for roughly one-third of its mass. Nearer the small end, the meter-scale bodies that are the proximate parent bodies of most meteorites are no more massive than large cars or light trucks. Itokawa, visited by the Hayabusa spacecraft, is of roughly the same mass as the Three Gorges Dam in China, the most massive human-made object.

The most accurate asteroidal masses are those made during spacecraft rendezvous, like Hayabusa's with Itokawa and NEAR Shoemaker's with Eros. These missions were successful in measuring object masses to 1–3% via spacecraft tracking (Yeomans et al. 2000; Abe et al. 2006), and Dawn is expected to provide yet more precise mass measurements for Ceres and Vesta. It has proven more difficult to measure masses during flybys, particularly when the encounters occur as "add-ons" to missions with a non-asteroidal focus and cannot be optimized for asteroidal science. Even in these cases, however, the constraints that can be placed on target mass (~±15–20% for the Galileo flyby of Ida: Petit et al. 1997) are better than could be generated by other means.

Ida, as a binary, provided another means of mass measurement. The ability to determine system masses relatively easily for binary systems (and the importance of mass measurements) has been a major driver for binary search programs. Roughly 50 asteroids have been confirmed to have satellites, and the data suggests that satellites are probable for roughly 50 more (Richardson and Walsh 2006; Johnston 2012). However, fewer than half of the confirmed systems have been observed well enough to calculate their mass. In some cases, particularly NEO systems, the recurrence time between good apparitions is very long, hampering efforts to determine their mass. The preponderance of singleton asteroids also limits the usefulness of this technique. However, it remains the most promising avenue for generating high-quality mass measurements for the largest number of targets.

The final technique for measuring asteroid masses is through observations of close encounters between objects or the masses necessary to account for orbital perturbations for a population. This method requires very precise astrometric measurements, particularly as objects become smaller. Roughly two dozen asteroids have had their masses measured by this technique, including Ceres and Vesta. Precision measurements of the position of Mars by spacecraft in its orbit are sensitive to effects of asteroids that are both more massive and/or pass relatively close to Mars, and masses of those objects have also become available (Fienga et al. 2009; Konopliv et al. 2011). While anticipated data from Dawn will improve our knowledge of the masses of Ceres and Vesta, this type of astrometric technique will likely remain the best hope for constraining the mass of Pallas and other large single asteroids (Hilton 2002).

Studies of meteorite densities were greatly spurred by the first spacecraft encounters with asteroids during the 1990s, with a nondestructive technique developed using helium to accurately measure the cracks and voids present in a sample (Consolmagno et al. 2008). Densities among the meteorites range from roughly 2,000–8,000 kg/m^3, with the CI chondrites on the low end and iron meteorites at the high end of that range, and the most important controlling factors for the meteorite densities are fractions of low-density hydrated minerals and of iron metal, as one might expect. However, before these densities can be compared to asteroidal values, the additional factor of porosity must be taken into account.

Porosity occurs in two forms: the intrinsic porosity found between grains in a coherent rock (for instance, the interior cracks and voids mentioned above) called microporosity and larger-scale porosity between rocks due to voids or joints called macroporosity. For comparison, the porosity in a pile of sand is roughly 40% (Abe et al. 2006), and the porosity of a coherent slab of granite can be less than 1% (Nur and Simmons 1969). Some very low densities among asteroidal bodies, notably Phobos, Deimos, 253 Mathilde, and 45 Eugenia, can only be explained by significant macroporosity (30–50% or higher) or very large amounts of interior ice, which are geochemically unlikely.

While the low density of C asteroids and the general acceptance of the link between them and carbonaceous chondrites makes the argument for macroporosity fairly clear, the densities of other asteroids can be more difficult to interpret due to less well-defined linkages and more intermediate densities. The density of 243 Ida, with no other information, can be interpreted as a low-porosity ordinary chondrite or as a higher-porosity stony iron. Similarly, the interpretations of the internal structures of Eros, Ida, and Itokawa are influenced by the general consensus that these are ordinary chondrite-like bodies and that deviations from typical OC densities are indicative of varying amounts of macroporosity (*Fig. 8-9*).

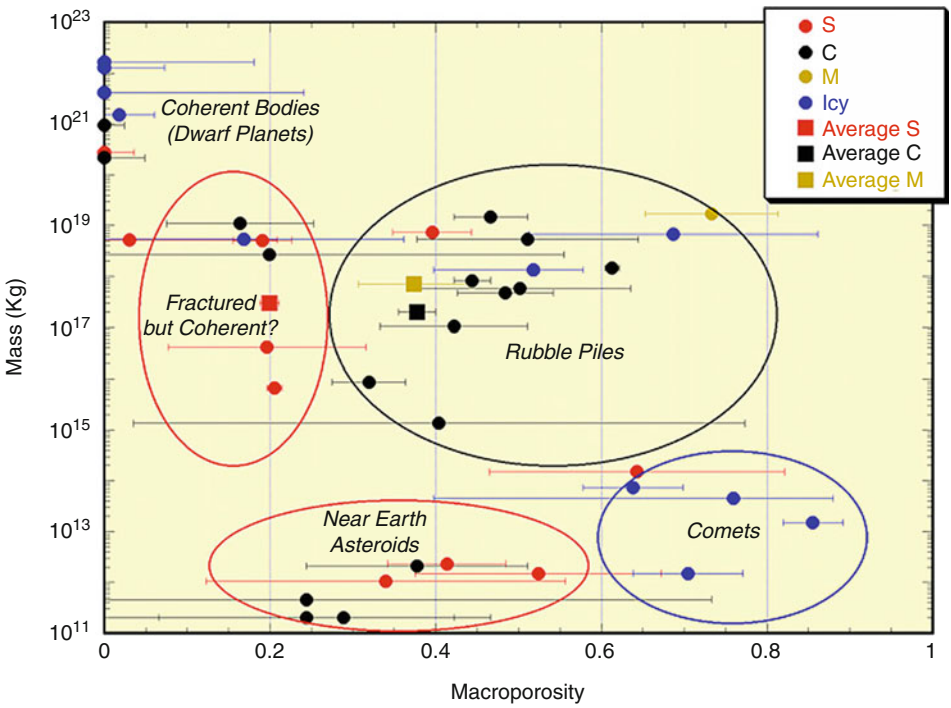


Fig. 8-9

Using likely compositional analogs for asteroids and comets, the macroporosities can be calculated for objects with known densities. The distribution of macroporosities suggests that many objects are rubble piles, with large fractions of void space. However, objects above ~10^{20} kg have macroporosities consistent with zero, as seen in this figure originally from Consolmagno et al. (2008)

4.2 Monoliths and Rubble Piles

There has been no universally accepted consensus on the terminology to be used for asteroid interiors. However, Asphaug et al. (2002) suggested "monolith" for coherent objects, "shattered" for objects that are thoroughly faulted and jointed but whose interior pieces have not appreciably moved or rotated with respect to one another, and "rubble pile" for those objects composed of pieces that have moved and/or rotated. Rubble piles can also be generated by the reaccumulation of material after a catastrophic disruption. Stress waves propagate much more poorly in rubble piles than monoliths, leading to different results upon impact, as described below. The lack of a firm consensus on terminology leads some to consider any object to be either monolith or rubble pile, with no third category.

The masses and gravities of asteroids do not generate large central pressures, important for considering phase changes and mechanical properties in asteroidal interiors. The most massive body in the asteroid belt, Ceres, has a central pressure of only 1.5–2 kbar, still appreciably larger than that of Vesta (roughly 500–700 bars). Smaller objects like Mathilde have central pressures less than 5 bars, with Phobos size and smaller objects under 1 bar. For comparison, the central pressure in the Earth is roughly 3.6 Mbar, over 1,000 times that of Ceres.

Asphaug et al. (2002) considered these central pressures by calculating the equivalent depth of burial here on Earth where the same pressures are found. For the largest asteroids, this is on the order of kilometers, while for objects of sizes like those visited by spacecraft, the equivalent depth is ~10 m or less (in some cases, like Itokawa, *much* less, of order 1 mm, though of course porosity in the uppermost portions of the Earth's surface means there is much more pressure from the ambient atmosphere than overlying rock). These central pressures, even for the largest asteroids, are much smaller than the pressures required to remove all of the porosity from chondritic material, and therefore the density of undifferentiated objects is expected to be a lower bound on the allowable density of any meteorite analogs.

Consolmagno et al. (2008) compiled the available densities of asteroids and, comparing them to the microporosities of potential analog meteorites, extracted estimates of the macroporosity present in those bodies. They reported that those asteroids with masses larger than 10^{20} kg have macroporosities consistent with zero. These include the three most massive asteroids (Ceres, Pallas, and Vesta). The mass of Hygiea is slightly below the 10^{20} kg threshold, but its density (2.55 g/cm^3) and likely carbonaceous chondrite-like composition makes a lack of significant macroporosity also consistent for that body. There may be additional objects with masses similar to Hygiea that also have little or no macroporosity, but the relatively few asteroids with well-constrained masses prevents a more thorough census at this point. Consolmagno et al. suggest the low (or absent) macroporosities of the largest asteroids could be due to avoidance of catastrophic collisions and reaccretion or gravitational removal of macroporosity via interior pressure, though that is unlikely as seen above. However, objects can potentially reduce their porosities via differentiation, which we suspect has occurred in Vesta and Ceres, and could plausibly also have occurred in Hygiea and Pallas.

In contrast to the larger objects, Consolmagno et al. found that "virtually all" objects smaller than 10^{20} kg have at least 20% macroporosity if not much larger values. Some larger objects can sustain surprisingly large macroporosities, like the ~50% macroporosity inferred for 45 Eugenia (~200 km diameter), ~45% for the components of the 90 Antiope system (each ~100 km diameter), and an astonishingly high ~60% macroporosity for the 150-km object 283 Emma. When separating the objects by spectral class, they also find that the smaller macroporosity values are preferentially found in the S-class asteroids, with the C-class objects carrying larger

macroporosities in general. Abe et al. (2006) reported a porosity of ~40% for Itokawa, similar to a pile of angular sand, and interpreted as evidence of a rubble pile structure. Given the 40%+ porosities seen in rocky bodies over the entire 200 m–200 km size range (if not beyond in either direction), it seems safe to say that rubble piles are the typical state of affairs for asteroids.

4.3 Rotation Rates and Interior Structure

Reliable rotation rates are known for over 3,000 asteroids (Warner et al. 2009) and also provide information about asteroid interiors. Harris (1996) considered the balance of centrifugal force and gravity, specifically the equation

$$\rho_c = (3.3/P)^2(1 + \Delta m)$$

where ρ_c is the critical density at which an object would fly apart for a period of P hours and a light-curve amplitude of Δm magnitudes (for a spherical object, $\Delta m = 0$). For a sample of 688 asteroids, Harris found $\rho_c = 2.7\ g/cm^3$ was consistent with the distribution of periods, consistent with the idea that all of the objects in the sample were rubble piles with no tensile strength. The larger database now available still shows evidence of the so-called spin barrier, where no large objects have rotation periods less than roughly 2.2 h (*Fig. 8-10*). Conversely, below a size

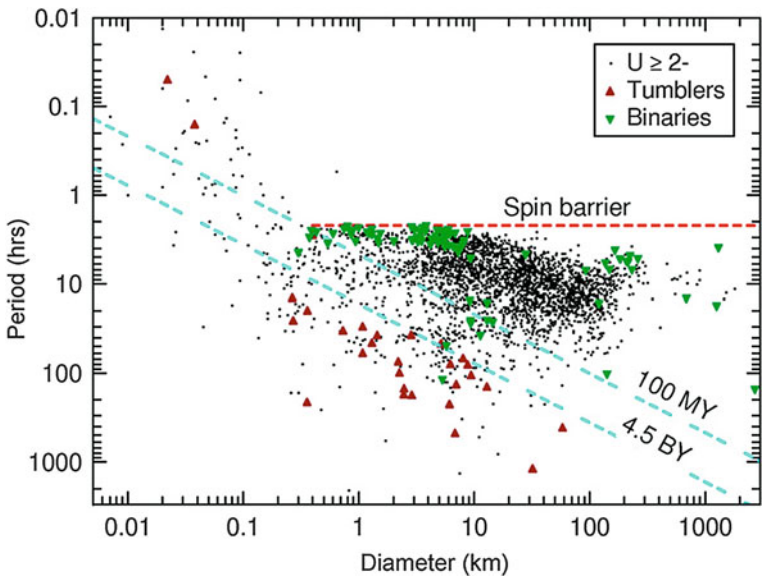


Fig. 8-10

This figure, originally published by Warner et al. (2009), shows known rotational periods for asteroids compared to their sizes. Objects larger than ~150 m have periods longer than roughly 2 h, the so-called spin barrier that is suggestive that these objects are rubble piles, or at least that tensile strength is not required for those bodies. Below ~150 m, rotation periods can be much faster, but it is not clear whether tensile strength is necessary to explain them or if cohesive forces in a low-gravity environment alone are sufficient

of roughly 200 m, rapid rotators become very common, including some with periods as short as a few minutes.

While these results are very suggestive that the rapid rotators can be single intact rocks (or "monoliths"), there are additional factors that complicate matters. An object seen to have slow rotation is not proven to be a rubble pile, although the weight of numbers is, again, suggestive. More seriously, calculations including cohesion (Holsapple 2007) indicate that smaller objects could have rotation periods far shorter than the spin barrier but still be rubble piles. Cohesive forces on asteroid surfaces are further discussed in ❯ Sect. 3.

4.4 Strength

The response of an object (asteroidal or otherwise) to stresses like impacts and disruption is critically dependent upon its strength. The greater gravitational binding energy of large objects leads to increasing strength as size increases. Laboratory experiments have probed small sizes (of order meter size and smaller) and find that strength in that size region increases with smaller size, as there is less likelihood of smaller objects having a crack or flaw that aligns with the direction of stress (Durda et al. 1998). This leads to behavior where the strength (here defined as energy per unit mass required for disruption), called Q^*, first decreases with object sizes before reaching a minimum value, then increasing again. Objects larger than the minimum Q^* are considered to be in the "gravity regime," where effects are relatively independent of composition while those below the minimum are in the "strength regime," where material properties and composition play a leading role. Material properties should have an important influence on the asteroidal size-frequency distribution.

Determining the transition between strength and gravity regimes has been a matter of continuing work, though it appears to be roughly near 100 m in size. At this writing, there are no firm plans for spacecraft visits to objects suspected to be in the strength regime, though one may be visited by astronauts in the mid-2020s.

4.5 Meteoritical Data

Some direct information about asteroidal interiors can be obtained from meteorite samples, many of which came from differentiated bodies (see ❯ Sects. 3 and ❯ 5). It is thought the main heat source driving differentiation was decay of short-lived radionuclides, mainly ^{26}Al (Grimm and McSween 1993). Other heat sources may have contributed, however. Magnetic induction heating, in which the changing magnetic fields in the solar nebula led to Joule heating of materials, has also been investigated and is considered a possible contributor. Accretional energy and energy brought via larger impacts may also have played a role, though such energy is deposited near an object's surface and is more readily radiated away. McCoy et al. (2006) includes a review of possible processes.

The asteroids began as undifferentiated, unequilibrated mixtures of material stable at main-belt temperatures: rock-forming minerals, metal, and in some cases ice. As interior temperatures increased in (some) asteroidal parent bodies, the differing melting temperatures and densities of these components led to differentiation into rocky mantles and crusts over metallic cores, or in some water-rich bodies like Ceres an icy mantle over a rock-metal core.

Roughly a dozen iron meteorite groups show evidence of having formed in the core of a differentiated object, which would also have had a rocky mantle. There are over 80 ungrouped iron meteorites, from which parent body data is difficult to extract, but some estimates suggest that as many as 60 objects may have differentiated into a core, mantle, and crust in the early asteroid belt (Chabot and Haack 2006). Many of the achondrite groups may also have come from differentiated objects. The HED meteorites have been linked to Vesta, generally thought to be a differentiated object, but the HEDs appear to harbor at least two distinct parent bodies, suggesting a second Vesta-like object once was in existence. Gaffey et al. (1993) summarized the general properties of fully differentiated objects, with core radii 40–50%, mantle thicknesses of 30–45%, and crusts ~20% of the total radius for OC compositions (H chondrites having the largest cores and thinnest mantles, with the opposite for LL chondrites). This provides a rough rule of thumb that any existing meteorite parent bodies that were cores of differentiated bodies require disruption of an object at least twice the current parent body size.

Partially differentiated objects may also exist. While the gravity of larger objects is too large to maintain significant density inversions in layers of rock and metal, in smaller objects, the inherent material strength of the layers may be sufficient to maintain relatively small forces from overlying (or underlying buoyant) layers, especially at colder temperatures or relatively small amounts of melting.

Modeling of Vesta's interior using both meteoritical and astronomical evidence is summarized in Keil (2002) and McCoy et al. (2006). Mass balance arguments by Ruzicka et al. (1997) suggest Vesta's core is most likely near 5 wt% of the body's entire mass, though with large uncertainties. They further suggest an upper limit of 130 km for core radius, a ~65–220 km thick mantle, and overlying diogenite and eucrite layers of ~12–43 and ~23–42 km thicknesses, respectively. An independent calculation by Dreibus et al. (1997) based on Vesta's bulk silicates leads to a larger estimated core mass of 21.7%, with a radius of 123 km. These values lead to a good match with the measured density of Vesta for reasonable estimates of mantle and core densities.

It is clear from the HED meteorites that differentiation and melting on Vesta occurred very early in solar system history, within its first few million years. This is also the timescale on which Vesta's basaltic crust was formed, with the crust forming after core separation and volcanism lasting for at least 10 million years or so (McSween et al. 2010 and references therein). Study of the sizes of "vestoids," asteroids dynamically and spectrally related to Vesta, leads to an estimate that its crust is greater than roughly 10 km in thickness, consistent with calculations based on the cooling rates of eucrites and the mass balance studies mentioned above. The geochemical details of the HED meteorites also are consistent with a magma ocean phase on Vesta, although consensus has not yet been reached. It is also not clear what Vesta's undifferentiated composition was most like, with different lines of evidence supporting ordinary chondritic, carbonaceous chondritic, and nonchondritic precursors (McSween et al. 2010).

The interior structure of Ceres has been the subject of several recent studies. McCord and Sotin (2005) argued that Ceres was sufficiently large that its macroporosity should be near zero and that its density should represent the density of its starting materials. They modeled several cases fitting that constraint and found a model of an icy mantle over a rocky core to be the most likely case. This was independently borne out by analysis of HST data by Thomas et al. (2005), who used improved shape information for Ceres to argue for the same interior structure. Castillo-Rogez and McCord (2010) found that depending upon initial conditions, liquid water may remain in Ceres' deep interior to the present. Ceres is differentiated into a rocky core and a volatile shell in all cases considered by Castillo-Rogez and McCord. They also found that the

shape of Ceres is very sensitive to the amount of hydrated silicates present in the core, though measurements of Ceres' shape is not of sufficient precision for further constraints. The Castillo-Rogez and McCord models do allow for the differentiation not only of ice from rock but of an iron core, though the most likely cases do not differentiate to that extent.

In contrast to the geophysical models, Zolotov (2009) has argued that many of Ceres' properties are also consistent with a porous, undifferentiated body made of hydrated minerals. However, such a composition and structure cannot explain Ceres' shape as seen by Thomas et al. The Dawn encounter with Ceres in 2015 should provide data to further constrain models of Ceres' interior.

These types of thermal models are also being turned to other large asteroids. Castillo-Rogez and Schmidt (2010) modeled the original parent body of the Themis family, with starting conditions varying composition and the time of formation. They find that for starting compositions of homogeneous ice/rock mixtures, rapid formation (~3 Myr after CAI formation) leads to differentiation of the parent body similar to what is seen in the models of Ceres. Longer formation times (5 Myr) lead to only partial differentiation or for a homogeneous 100% hydrated silicate composition, with no geophysical evolution at all. They suggest that astronomical observations of 24 Themis and 90 Antiope, the two largest members of the current Themis family, could be explained by a differentiated or partially differentiated case, where Themis and Antiope reaccreted a large fraction of ice after parent-body disruption.

Interior models of chondritic parent bodies are constrained largely from meteoritic rather than astronomical evidence. Uncertainties are further compounded by the unknown number of parent bodies represented by the meteorites: while it is generally considered that each of the OC groups (H, L, LL, etc.) comes from a different parent body, it is still not known whether more than one parent body could contribute to each class. Conversely, some of the carbonaceous chondrite classes may have originated on the same parent body, experiencing different levels of alteration, and some of the achondrite groups are associated with one another (e.g., winonaites and IAB irons, for example: Weisberg et al. 2006).

The differing metamorphic grades of OC meteorites and the various temperatures required to cause the observed metamorphism have led to a general conception of an "onion skin" model for the interiors of OC parents (or at least their predisruption interiors, as appropriate). In this model, the highest temperatures and highest metamorphic grades are found at the center of the parent body, with both decreasing with decreasing depth. The unequilibrated (type 3) chondrites would be found near the surface. However, differing peak temperatures and cooling rates are implied by different techniques, complicating detailed modeling (Huss et al. 2006). Besides these direct constraints, factors such as the depth of regolith and megaregolith (and its time history) and the time and pattern of parent body accretion also can have significant influence on the interior temperature and evolution of asteroids (Ghosh et al. 2003).

Differentiation and formation of iron cores is at least superficially straightforward. Chabot and Haack (2006) review geochemical studies of the iron meteorites, with their cooling rates suggesting original parent-body diameters on 10–100 km scales. Unlike the formation of the Earth's core, asteroidal cores most likely formed via dendritic growth downward from the core-mantle boundary. However, other possible modes of crystallization cannot currently be ruled out. Because of the small sizes of asteroids, any liquid cores must have solidified very early in solar system history, perhaps within the first few million years (Roberts et al. 2011).

Liberation of iron meteorites from asteroidal interiors is, however, proving to be more difficult than originally envisioned. Collisional models suggest that disruption of asteroids and dispersal of fragments is exceedingly difficult today. Furthermore, the survival of only a single

Vesta-like object in the asteroid belt is consistent with few having ever been present, even in early solar system history (Bottke et al. 2005). It is not currently obvious how best to reconcile these constraints, though Bottke et al. (2006) argued that the data are consistent with iron meteorites forming in parent bodies' interior to today's asteroid belt, being liberated during disruptions, and experiencing dynamical evolution to reach their current orbits all during the first few 10^8 years of solar system history.

Despite the lack of liquid cores today, and the narrow window during which meteorite parent bodies could have sustained a magnetic field, remnant magnetism can be found in some meteorites. Weiss et al. (2010) review the field of meteorite paleomagnetic studies.

5 Asteroid and Meteorite Compositions

Our knowledge of the asteroids is immeasurably enhanced by the meteorite collection. The combination of impacts in the asteroid belt and orbital evolution due to the Yarkovsky force and planetary resonances leads to a constant influx of asteroidal material to the Earth. Roughly $5,000 \pm 2,000$ kg of meteorites fall to the Earth in a typical year (Bland et al. 1996; Zolensky et al. 2006), although due to the fraction of the Earth's surface covered by the oceans and other inaccessible areas, most are not recovered. However, meteorites can be recognized decades or longer after their fall because of distinctive features such as their fusion crust (a thin surficial layer caused by frictional heating during atmospheric passage) and high metal fraction compared to terrestrial rocks. In addition, some terrains and climates result in concentration of and/or easier recognition of meteorites, for instance, the Sahara desert and regions of Antarctica. Indeed, most of the known meteorites were recovered in Antarctica by dedicated search teams like ANSMET. Meteorites collected long after their fall, or with no knowledge of their fall date, like the Antarctic meteorites, are termed "finds," while those seen to fall and collected soon after are called "falls." Falls are considered more valuable to the community than finds since they have less terrestrial alteration and contamination, but often for rare meteorite types, only finds are available.

Meteorites are formally classified into different groups based on their compositions, textures, and alteration histories as shown in ❯ *Fig. 8-11*. At the highest level, meteorites are typically divided into the chondrites and achondrites.

Chondrites have elemental abundances in the same relative concentrations seen in the Sun, save for those elements that are not easily incorporated into rocks such as hydrogen and helium. These "chondritic abundances" are evidence that these materials are largely unprocessed since formation. Chondrites also are distinguished by chondrules, millimeter-scale glassy spherules or fragments of spherules. Chondrules formed within 2–3 million years of the condensation of the first solar system solids (typically considered the "time zero" for the solar system), melting and recrystallizing as free-floating objects within the solar nebula before being incorporated into the chondrites (Morris and Desch 2010). When discussing chondrites, a distinction is usually made between the chondrules and the "matrix," comprising the material in the rock found between chondrules. Chondrules can comprise up to 80% of the volume of some chondrite types. Despite their importance, determining the process by which chondrules formed has been a long-standing problem in meteoritics. Early suggestions that they melted via nebular lightning, solar flares, and impacts have been replaced by preference for models in which

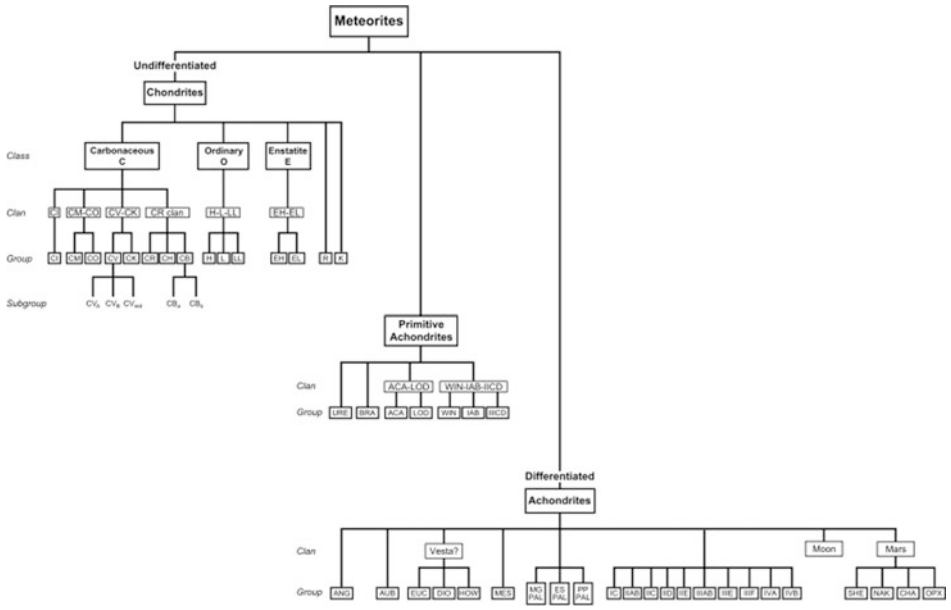


Fig. 8-11

The meteorites are typically divided into the achondrites, which have experienced some form of melting, and the chondrites which have not. This figure, from Weisberg et al. (2006), includes a distinction between those achondrites that have experienced melting and differentiation from those that have only experienced a small amount of melting (the primitive achondrites). These large-scale distinctions are further subdivided into the groups and classes discussed in the text

the chondrule precursor material melted after passing through shock waves in the solar nebula (Morris and Desch 2010).

Chondrites are thought to represent the starting material for rocky bodies and the rockier portion of icy bodies. This gives an important constraint for chemical evolution models, for instance, allowing the composition of the Earth's core (or the lunar or Martian core) to be calculated from the near-surface samples available to us. They also provide textural and compositional information directly related to early solar system history. The calcium-aluminum-rich inclusions (CAIs) present in many chondrites are the oldest solids in the solar system, comprised of refractory material that was the first to condense out of the solar nebula.

The chondrites are further divided into several groups and subgroups based on details of elemental patterns and isotopic ratios. The largest group is the ordinary chondrites (OCs), responsible for the vast majority of falls (roughly 90% of chondrites and 80% of all meteorites) and presumably comprising the majority of asteroidal material, at least in near-Earth space. The OCs are primarily composed of olivine and pyroxene, with plagioclase, other silicates, iron-nickel metal, sulfides, and oxides in smaller proportions (Rubin 1996). Typically, they contain less matrix than the carbonaceous chondrites, the other major chondrite grouping. The abundance of iron and its fraction in metallic versus oxidized forms leads to a further division in the OCs between the H (high), L (low), and LL (very low) groups.

The carbonaceous chondrites (CCs) are distinguished from the OC meteorites by higher amounts of refractory lithophile elements and oxygen isotopic ratios near or below the terrestrial fractionation values. As further discussed below, many CC meteorites show evidence of parent-body aqueous alteration early in solar system history, and clays and other alteration minerals are commonly found in their matrices. They need not be carbon-rich, despite their name, though they can contain organic material. The CI group of CC is thought to be the closest match to the Sun's nonvolatile composition that we have in the meteorite collection.

The organic material in carbonaceous chondrites has been the subject of obvious interest in the community. There is great variability in the composition of this organic material and in the relative concentrations of soluble and insoluble material from one CC group to another. The soluble material can vary greatly in abundance even in different parts of the same meteorite (Pizzarello et al. 2006). Some of the molecules seen in the CC meteorites are also commonly found in life on Earth, like amino acids. However, the variety found in these molecules and the randomness of the mixtures in which they are found argues for their abiotic origin. Interestingly, while both right-handed and left-handed versions of organic compounds are found in the meteorites, there appears to be a slight excess of left-handed (Pizarello et al.). While the role of meteoritic/asteroidal organic material in the origin of life on Earth is not settled, it has been proposed that this excess could have formed the basis of the exclusively left-handed molecules used by all terrestrial life.

The remaining chondrites (E, R, and ungrouped) cover a wide variety of properties and mineralogies, but comprise only ~2% of the chondrite falls, the vast majority of them in the E (enstatite) chondrite group.

Although the chondrites did not suffer melting since the time they were assembled from chondrules and matrix, many of them did experience some degree of thermal and aqueous metamorphism. A numerical metamorphic grade can be assigned based on criteria like equilibration of iron and calcium, presence or absence of hydrated minerals, mineral textures, and other factors, discussed in Huss et al. (2006). In the typical scheme, unmetamorphosed material has a grade of 3, aqueous alteration leads to a grade between 1 and 3 (with 1 the most altered), and thermal metamorphism to a grade between 3 and 6 (with 6 the most metamorphosed). A heavily metamorphosed LL chondrite may be classified as LL6, then, while a heavily aqueously altered carbonaceous chondrite may be a CI1. Some workers have subdivided the grades from 3 to 4 into decimal tenths, leading to an H3.1, for instance.

Achondrites comprise all of those meteorites which are not chondrites, including the small fraction of meteorites that come from Mars and the Moon. All achondrites have been heated to the point that their minerals melted, though the amount of melting varies widely. Some of the achondrite parent bodies were heated to the point of differentiating into a rocky crust and mantle over a metallic core. The iron meteorites represent not only fragments of such a core, but also material that suffered local melting and metal separation.

The primitive achondrites include several groups that have experienced some melting, but did not themselves crystallize from a melt or are not from differentiated parent bodies, retaining close geochemical similarities to the chondrites (Weisberg et al. 2006). Much work on primitive achondrites is ongoing, and their place in an overall meteorite classification scheme is still unsettled.

The most abundant achondrites (roughly 60% of achondrite falls) are the HED meteorites, igneous material that has been linked to Vesta through spectral, dynamical, and geochemical arguments (Keil 2002 and references therein). The HED meteorites are named for three subgroups: the howardites, eucrites, and diogenites. Eucrites are basaltic rocks that formed near

Vesta's surface, while diogenites are orthopyroxene cumulates that originated deeper in its crust. Howardites are brecciated mixtures of eucritic and diogenitic material. Interestingly, while the case for Vesta as a HED parent body is robust, recent work suggests that at least two distinct parent bodies are required to explain the difference in oxygen isotopes found among members of the group (Yamaguchi et al. 2002). It is generally thought that the main HED group is the one associated with Vesta rather than the less common Ibitira group, but work to conclusively demonstrate this is still underway.

The angrite and aubrite groups are also basaltic rocks derived from the crust of differentiated asteroidal parent bodies. The aubrites appear to be related to the E chondrites, perhaps with an E chondrite precursor material, while the precursor material for angrites appears to be absent from the meteorite collection.

The iron meteorites comprise 13 geochemically distinct groups, classified based on geochemical and textural differences. It is thought that ten of these groups formed in the cores of separate differentiated parent bodies, though other groups may be more consistent with formation in melt pools rather than requiring full differentiation. In addition to these groups, roughly 15% of iron meteorites are ungrouped, perhaps representing as many as 50 additional parent bodies. Without additional samples to characterize trends within a group, it is not certain whether these ungrouped samples formed in the core of a differentiated object versus in melt pools.

The final asteroidal achondrite groups are the mesosiderites and pallasites, both partially silicate and partially metallic. The mesosiderites appear to have originated in a large impact, which created a silicate-metal breccia (Scott et al. 2001). Pallasites are olivine-metal mixtures of uncertain origin, with earlier theories suggesting they formed near a core-mantle boundary and more recent ones proposing mixing of mantle and core material after a large impact (Yang et al. 2010).

5.1 Isotopic Studies

The different geochemical behavior of various elements and isotopic mixes provides a means of tracing asteroidal histories. Oxygen may be the most important of the stable isotopes. The relative amounts of the three main oxygen isotopes (^{16}O, ^{17}O, ^{18}O) serve as a fingerprint, homogeneous within a single parent body but differing from one body to the next, allowing linkages to be recognized or rejected more easily (Weisberg et al. 2006). This, along with other isotopic and mineralogic analysis, allows different meteorite groups from the same parent body to be interpreted together, or alternately demonstrates that some meteorite types (e.g., the eucrites) are derived from more than one parent body.

Radioactive and radiogenic elements are used to date geochemical events on asteroids using the same techniques as in terrestrial studies. The different isotopic systems typically are used to study different processes: the production of ^{40}Ar by ^{40}K in silicates is used to study the timing of impacts, which release built-up Ar and "reset" the system.

In addition to the ongoing creation of radiogenic elements, the products of extinct radionuclides can also be studied in meteoritical materials, which have experienced little disturbance. The isotope ^{182}Hf decays into ^{182}W with a half-life of 9 Myr. Hf is a strongly lithophile element, while W is strongly siderophile. As a result, core formation in differentiated objects would have effectively sequestered all W into the core, and any found today was created by the decay of ^{182}Hf. Studies using this dating system show that core formation occurred very early on in solar

system history, within the first few Myr. The extinct radionuclide ^{26}Al is thought to have been a primary heat source in the early solar system and decays into ^{26}Mg. Determining the initial concentration of ^{26}Al in the solar nebula is critical to understanding the thermal history and evolution of objects, but it is not straightforward since ^{26}Mg is itself a commonly found isotope and because relatively small differences in accretion rates can lead to large differences in the amount of ^{26}Al (with a half-life of ~700,000 year) accreted in a body. Current estimates of the initial ^{26}Al /^{27}Al ratio are converging near 6×10^{-5} (Halliday and Kleine 2006).

5.2 Asteroidal Compositions from Remote Sensing

The vast amount of available information about meteorite compositions has been leveraged to help interpret asteroidal data. However, progress has been uneven in conclusively associating asteroids with meteorite types and making quantitative measurements of asteroidal compositions. The rate of asteroid discoveries far outpaces the rate at which those objects can be characterized, even at a cursory level. *Figure 8-12*, from a 2006 NASA study on hazardous asteroid mitigation, was drafted in the context of near-Earth asteroid observations but also applies well to the population as a whole: the number of objects that we can characterize in great depth is relatively small, and what is learned from well-characterized objects

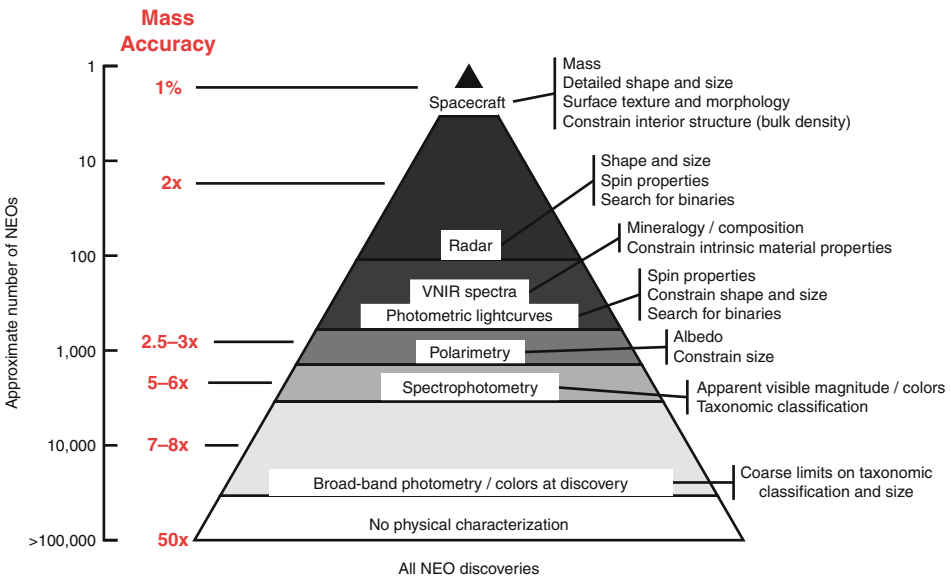


Fig. 8-12

While many thousands of asteroids are known, very few are completely characterized. This figure provides a schematic measure of the number of objects characterized to a particular level, with associated mass uncertainties present for each level of characterization. While generated for near-Earth objects specifically, the general form of this pyramid and its levels are applicable to the main-belt asteroids as well (NASA)

must be applied to make educated estimates about the great numbers of poorly characterized objects.

Visible and near-IR observations (wavelengths roughly 0.4–1.0 μm) were the first to be made of the asteroids because of available technology and remain the easiest because of that technology's maturity and the relative transparency of the atmosphere in that spectral region. Broadband photometry gave way to observations with specially selected filters by the late 1970s, and the first asteroid taxonomies were constructed at that time. Vesta, the brightest asteroid, was the first to be studied, and its spectral similarity to the HED meteorites was quickly noted (McCord et al. 1970). This led to hopes that other asteroids would easily be associated with meteorite types, and the first major taxonomic classes were somewhat optimistically named S, C, and M for "stony," "carbonaceous," and "metal": the three meteorite types they at least superficially resembled. While later work has shown these mnemonics to be oversimplified at best, later taxonomies are based on this early work.

Currently, there are two visible near-IR classification schemes in general use, the Tholen and Bus taxonomies (Tholen 1984; Bus and Binzel 2002). Both use principal component analysis techniques to find clusters of objects which are grouped together into classes and subclasses. The Tholen taxonomy was created from eight-color spectrophotometric data, but also included albedo information. The Bus taxonomy uses spectroscopic data of much higher spectral resolution, but has more restricted wavelength coverage compared to what is used in the Tholen taxonomy and does not include albedo. Both schemes use letter names for their classes, and each has tried to stay consistent with predecessor taxonomies where possible. That has led to good overall agreement with each other, but some level of confusion in detail as objects can be classified differently in the different schemes, and an informal hybrid taxonomy has evolved in some cases. ❯ *Table 8-2* shows the main classes in each taxonomic system, along with typical examples.

The Tholen taxonomy (1984) separates the asteroids into three main classes: S, C, and X. The X class is further divided based on albedo: high-albedo (>0.3) X asteroids belong to the E class, medium (0.08–0.3) to the M, and low (<0.08) to the P. X asteroids with no albedo information remain in the X class. C asteroids may be further classified into the B, F, or G subclasses based on their albedos and UV reflectances. The A, Q, R, and V asteroids are close to the S class in principal component space, but are not technically subclasses of it. The Q class is of particular note because it has been associated with the OC meteorites and the search for Q-class asteroids in the main asteroid belt has long occupied asteroid scientists. At this writing, excellent candidates for such objects have been identified (Mothé-Diniz et al. 2010; Rivkin et al. 2011b), though consensus will likely require near-IR data to augment the available visible-wavelength spectra. Finally, the D and T classes are transitional between the C and X classes.

The Bus taxonomy (Bus and Binzel 2002) uses the term "complex" to denote the largest groupings: again S, C, and X. Subclasses are defined based on the presence or absence of specific absorptions and inflections. The most notable subclass for our purposes is the Ch subclass, which feature absorption bands near 0.7 μm associated with phyllosilicates (see below). The lack of albedo information means that the E, M, and P classes are not defined and not used. The other major Tholen classes are also found in the Bus taxonomy.

Gradie and Tedesco (1982) found that different regions in the asteroid belt are dominated by different asteroid spectral classes (❯ *Fig. 8-13*). Subsequent work, including the very large catalog built using SDSS data (>50,000 asteroids, >10,000 linked to known objects: Ivezić et al. 2001), has reinforced this result, although mixing between groups is certainly present and members of each complex can be found in all regions.

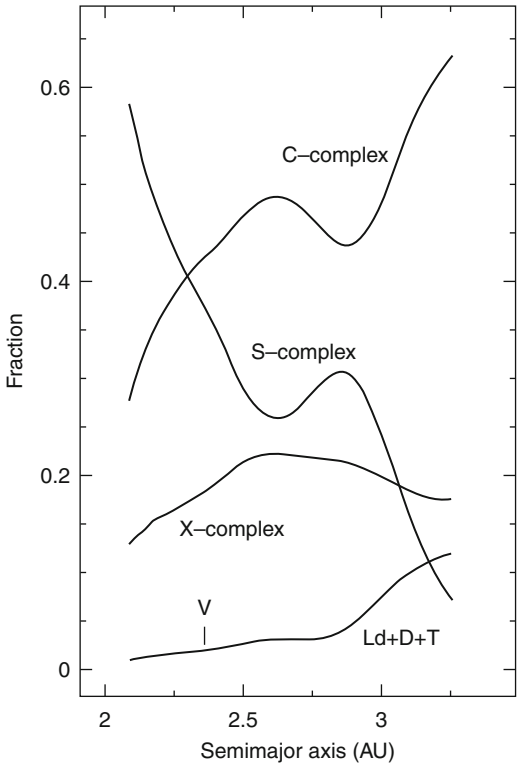
Table 8-2

Asteroid-derived meteorite groups and postulated parent bodies or populations. Table adapted from Burbine et al. in Reviews in Mineralogy and Geochemistry: http://rimg.geoscienceworld.org/cgi/content/abstract/68/1/273

Group	Fall percentage (%)	Postulated parent or source bodies
L	38.0	S(IV) asteroids (Gaffey et al. 1993)
H	34.1	6 Hebe [S(IV)] (Gaffey and Gilbert 1998)
LL	7.9	S(IV) asteroids (Gaffey et al. 1993)
Irons	4.2	M asteroids (Cloutis et al. 1990; Magri et al. 1999)
Eucrites	2.7	4 Vesta (V) (McCord et al. 1970)
Howardites	2.1	4 Vesta (V) (McCord et al. 1970)
CM	1.7	19 Fortuna (G, Ch) (Burbine 1998)
Diogenites	1.2	4 Vesta (V) (McCord et al. 1970)
Aubrites	1.0	3103 Eger (Gaffey et al. 1992)
EH	0.8	M asteroids (Gaffey and McCord 1978)
EL	0.7	M asteroids (Gaffey and McCord 1978)
Mesosiderites	0.7	M asteroids (Gaffey et al. 1993)
CV	0.6	K asteroids (Bell 1988)
CI	0.5	C asteroids (Gaffey and McCord 1978)
CO	0.5	221 Eos (K) (Bell 1988)
Pallasites	0.5	A asteroids (Cruikshank and Hartmann 1984; Lucey et al. 1998)
Ureilites	0.5	S asteroids (Gaffey et al. 1993)
CR	0.3	C asteroids (Hiroi et al. 1996)
CK	0.3	C asteroids (Gaffey and McCord 1978)
CB	0.3	M asteroids (Hardersen et al. 2005)
Acapulcoites	0.1	S asteroids (McCoy 2000)
Angrites	0.1	3628 Božnemcová (Cloutis et al. 2006)
Lodranites	0.1	S asteroids (Gaffey et al. 1993; McCoy 2000)
R	0.1	A or S asteroids
Winonaites	0.1	S asteroids (Gaffey et al. 1993)
(Tagish Lake)	0.1	D or T asteroids (Hiroi et al. 2001; Hiroi and Hasegawa 2003)
Brachinites	Only finds	A asteroids (Cruikshank and Hartmann 1984; Sunshine et al. 1998)
CH	Only finds	C or M asteroids

References have a discussion of the postulated linkage and may not be the first to propose the linkage

The E-class asteroids (Tholen taxonomy) are the main population in the Hungaria region at the inner edge of the main asteroid belt. The S asteroids dominate the inner belt, with the C asteroids dominant through much of the rest of the main belt (and thought to be the most abundant asteroids when considering the main belt as a whole). The outer-belt and Trojan asteroids are more poorly sampled, but appear to be mostly P-class (Tholen taxonomy) and D-class asteroids.



Gradie/Tedesco plot

Fig. 8-13

This figure from Bus and Binzel (2002) demonstrates the changing distribution of asteroid spectral types in the main belt. While the C complex is a major component throughout the region, it dominates the middle and outer belt. The low-albedo D and T classes become more important in the outer belt, while the S complex dominates the inner belt but becomes relatively unimportant as semimajor axis increases (note the bump near 2.9 AU, due to the large S-class Koronis dynamical family)

While visible observations were the first to be made for asteroids, the near-IR wavelengths extending to 2.5 μm contain more diagnostic information for silicate minerals than the wavelengths shortward of 1 μm. As a consequence, and as detector technology matured, the number of observations in the 0.8–2.5 μm region has grown greatly. An extension of the Bus taxonomic system using data reaching 2.5 μm has been published by DeMeo et al. (2009) and is likely to serve as the successor to the Bus system where the full span of wavelengths are available.

The most important absorption bands in this wavelength region are due to olivine (band center near 1 μm) and pyroxene (bands near 1 and 2 μm). They are the dominant spectral features in the S-complex asteroids as well as other higher albedo classes. Because of their prominence in the spectra of the brightest asteroids, as well as importance for terrestrial rocks, quantitative analyses have been developed to use these absorption bands to measure or constrain silicate compositions (Gaffey et al. 2002; Sunshine et al. 2004; Dunn et al. 2010).

Compositional information is also obtained at longer wavelengths. The 3-μm region has diagnostic absorptions due to water and hydroxyl, either bound into minerals or as free water (or ice). It is also where organic material (containing C–H bonds) have strong absorptions. The 8–13 μm region is currently experiencing great interest, thanks in part to a large number of Spitzer Space Telescope (SST) observations of asteroids. Both of these spectral regions are more difficult to utilize than the 0.4–2.5 μm region because of competing absorptions in the Earth's atmosphere, which render certain wavelength ranges unobservable from the ground. However, many absorption bands of interest are still easily detectable, and the anticipated launch of the James Webb Space Telescope (JWST) will cover spectral ranges unobservable from the surface of the Earth, as did the SST.

The light analyzed in spectral studies typically penetrates a few wavelengths into the material. For the visible-NIR studies most commonly done for asteroids, this is only a few micrometers. However, because of regolith gardening and homogenization, it is thought that the top several centimeters (or more) are well-mixed on asteroids and the very surface is representative of much more than the top few micrometers. In addition, analysis and interpretation of surface reflectance spectra can lead to insight about the entire body; for instance, the detection of igneous minerals on a body leads naturally to inferences about the history of that object, as would a close to match to a particular meteorite group.

5.3 Compositions of Specific Objects and Classes: Current Interpretations

Because of the long-standing "S-asteroid problem" (see above), those objects have been intensively investigated using these techniques. Binzel et al. (2001) concluded the NEO 25143 Itokawa (then still known by its provisional designation 1998 SF36) was a space-weathered LL chondrite, a conclusion borne out by the vis-NIR and X-ray spectrometers onboard the Hayabusa spacecraft (Okada et al. 2006). Hayabusa's return to Earth in June 2010 brought back thousands of particles from this object, allowing the remote sensing datasets to be further cross-checked, work that is ongoing at this writing. Further remote spectroscopic studies of NEOs found a higher percentage of LL-like objects than expected from meteorite statistics, a mismatch attributed by Vernazza et al. (2008) to delivery of different chondrite types from different main-belt source regions via size-dependent mechanisms.

Gaffey et al. (1993) divided the S asteroids into seven subclasses based on their visible-NIR spectra and interpretations of their olivine/pyroxene compositions and ratios. These subclasses included the S(IV) group, thought to be consistent with the OCs, while the other groups corresponded to achondrites and hypothetical but yet-unseen mineralogies. Gaffey and Gilbert (1998) studied the asteroid 6 Hebe in further detail, concluding it was akin to the H chondrites and a "probable" parent body of those meteorites. Studies of 15 Eunomia, the largest S asteroid, point to a differentiated body, though the details of its surface metal content and exact mineralogy vary among workers (Reed et al. 1997; Nathues et al. 2005; Moskovitz et al. 2008).

Spacecraft have encountered four S asteroids: 25143 Itokawa (mentioned above), 243 Ida, 951 Gaspra, and 433 Eros. While the Gaspra and Ida encounters were only flybys, the Galileo spacecraft was able to return some spectral data for both asteroids. Gaspra was found to be more olivine-rich than a typical S asteroid, while Ida had a spectrum consistent with a weathered OC, and indeed the spatially resolved data available for Ida, and the spectral difference seen

between fresh craters and older background areas, was one of the first demonstrations that space weathering was occurring on asteroids (Chapman 1996).

Decades of study due to its status as one of the largest NEOs, along with the NEAR Shoemaker mission, have left 433 Eros as one of the best-understood asteroids. A data set using both visible-NIR imagery and NIR spectroscopy from NEAR Shoemaker led Izenberg et al. (2003) to interpret Eros' composition as consistent with L6 meteorites, with an opx/(opx+ol) ratio of 0.25–0.29. Lucey et al. (2002) looked at Eros' spectral changes with temperature and concluded it was most consistent with LL-type OCs, with H chondrites excluded. Lim and Nittler (2009) performed a recalibration of the NEAR Shoemaker X-ray spectrometer (XRS) data, finding an elemental composition like the OC meteorites, save a strong depletion of sulfur, which they interpreted as due to space weathering.

While many of the recent studies are pointing toward space-weathered OC mineralogies for many S asteroids, it is worth noting that this is not universally accepted. Abell et al. (2007) have argued that Itokawa has experienced some partial melting and that its spectrum is not consistent with OC compositions per se. Similarly, Gaffey (2010) asserts that space weathering does not change important spectral parameters and ought not be invoked to explain some mismatches. Conversely, Mothé-Diniz et al. (2010) argued that OC spectra can be found that match non-Q asteroids and that the Sq and Sk groups in the Bus taxonomy might also be viable OC analogs.

Observations of Vesta from the HST and ground-based telescopes equipped with adaptive optics (AO) have provided spatially resolved spectral information, augmented by point-source observations obtained over a full rotation. These studies have found Vesta to have hemispheric-level differences in spectral properties, with one hemisphere more akin to the diogenite meteorites and the other analogous to the eucrites (Binzel et al. 1997). The eucritic regions have been interpreted as the original basaltic surface of Vesta, with a lower albedo than the other regions due to particle size differences, regolith maturation, or other effects (Gaffey 1997). Imagery also shows evidence for a giant crater near Vesta's south pole, but there have not been spectral observations with viewing geometry favorable to studying the crater. The 3-μm region shows evidence of a roughly 1–2% reduction in reflectance compared to the 2.5 μm region, but interpretations have ranged from the influence of carbonaceous impactors in the regolith to simple variations in continuum behavior (Hasegawa et al. 2003; Rivkin et al. 2006). The discovery of water/OH in the lunar regolith, interpreted as formed through solar wind interactions with the regolith, has created further possibilities for Vesta.

Investigations of additional V-class asteroids, both in the NEO population and in the main belt, have demonstrated spectral similarities, though there is some evidence that the smaller objects (often termed "vestoids" if they are dynamically linked to Vesta) may have been space weathered differently than Vesta and may only represent a fraction of the materials present on Vesta itself (Shestopalov et al. 2010). Those V asteroids that are not dynamically linked to the Vesta family show some spectral differences compared to Vesta, though it is not yet clear to what extent this is due to different mineralogy and to what extent it is due to the different temperatures experienced by these objects. A more thorough understanding of Vesta and V types awaits the arrival of the Dawn spacecraft.

Understanding of low-albedo objects has lagged compared to the S asteroids and Vesta due to the lack of strong, diagnostic absorptions in the 0.4–2.5 μm region where most asteroidal observations have been made. A broad absorption band near 0.7 μm has been interpreted as due to Fe2+–Fe3+ intervalence charge transfers and is associated with (though not diagnostic of) phyllosilicate minerals (Vilas 1994). This band is an integral part of the Bus taxonomy,

is the discriminator of the Ch and Cgh taxonomic classes, and is also seen in some carbonaceous chondrites. A strong UV absorption is also seen in the Tholen C-class asteroids (the spectroscopic data used by Bus does not consistently extend to short enough wavelengths to see this band) and is due to oxidized iron. Because of the strong absorptions due to water and OH in the 3-μm region, observations in this region have provided more insight into the C-asteroid population.

Ceres is the best-understood of the low-albedo asteroids. Its 0.4–2.5 μm spectrum is largely featureless, other than a broad band centered near 1.2 μm that is likely due to Fe3+, possibly in magnetite. A series of absorptions in the 3.0–3.3 μm region have been interpreted separately and as a group. The most recent interpretation for the bands, by Milliken and Rivkin (2009), is brucite (responsible for the band near 3.05 μm) and carbonates (responsible for the bands at 3.3–3.4 μm as well as additional bands near 3.7–3.8 μm). These minerals would have formed during intense aqueous alteration of an olivine-rich assemblage early in Ceres' history. Interpretations of mid-IR observations are consistent with the 3-μm data. Still unidentified in Ceres' spectrum is an absorption near 0.25 μm, though anticipated advances in UV spectral libraries and additional UV asteroidal observations should aid its interpretation.

The asteroid 24 Themis, largest member of the Themis asteroid family, is observed to have some water ice on its surface, probably as thin coatings on silicate grains, as well as organic material consistent with ice tholins and material seen in carbonaceous chondrites (Rivkin and Emery 2010; Campins et al. 2010). Hydrated minerals are not seen on Themis, consistent with any interior ice remaining unmelted since accretion. The cometary activity seen in the "main-belt comets," many of which are in the Themis dynamical family, is most likely driven by ice sublimation, which is consistent with what is seen on Themis. The uncertainty in measurements of Themis' obliquity combined with the extreme sensitivity of ice stability to small changes in temperature near Themis' expected temperature leaves current models undecided on the required resupply rate to Themis' surface from its interior. Indeed, for some obliquities surface, ice remains stable over a range of latitudes sufficient to account for the existence of ice (Rivkin and Emery 2010).

While Ceres and Themis are the best-studied of the C asteroids at 3 μm, dozens of other objects have also been observed. These seem to have band shapes that separate into a three to four different classes, though a formal taxonomy has not been established at this writing. Rivkin (2007) informally named these groups after their most prominent members: Ceres type, Pallas type, Cybele type (or Themis type), and anhydrous. Members of the same group might be expected to have similar mineralogies, but modeling of spectra as well as a more formal taxonomy will be necessary to strictly reach that conclusion. The Pallas types, the most common of these groupings, appear to be at least qualitatively consistent with the carbonaceous chondrite meteorites. The Ceres types are presumed to have surfaces with compositions similar to Ceres' surface as mentioned above: brucite and carbonates along with a low-albedo component and possible other aqueous alteration products. The Cybele types, which include Themis, may include a wider range of mineralogies and may be split into several groups. However, other ice and organic-bearing objects might be in this group. The "anhydrous" objects need to be considered with some caution: in some cases, the available data have large uncertainties and are classified here as a default. In all cases, this group technically only has upper limits on band depth, and higher-quality data and/or spectral coverage now unavailable could plausibly find evidence for a band currently not detected or undetectable from Earth's surface.

The asteroid 2008 TC3 had spectral properties that would place it in the C complex, and pieces of it were recovered after its impact near the Egypt-Sudan border. These pieces, referred to

as the Almahata Sitta meteorite, are mostly classified as ureilites, one of the primitive achondrite groups mentioned above (Jenniskens et al. 2009). However, small clasts from other meteorite groups including ordinary chondrites were also found, suggesting meter-class objects could have significant heterogeneity, presumably via collision-induced mixing while still in a regolith on a larger parent body (Zolensky et al. 2010).

The D-class asteroids found in the outer asteroid belt and Trojan clouds, as well as some planetary satellite populations, are featureless throughout the 0.4–4 μm region. This is somewhat surprising, as their steep spectral slopes are very similar to what is seen in cometary nuclei, which has led to their interpretation as organic-rich objects by analogy. However, the absorptions due to C–H bands, like those found on 24 Themis, are absent from similar spectra of the D asteroids. Their faintness, due to their low albedo and residences further from the Sun, may be in part to blame. Cruikshank et al. (2001) analyzed the spectrum of the large Trojan asteroid 624 Hektor, finding its surface composition to be fit by a combination of magnesium-rich pyroxene and low-albedo neutral material like elemental carbon, with the steep spectral slopes provided by the pyroxene. Cruikshank et al. note, however, that up to a few percent of volatile material could still be present in the form of ice or hydrous silicates but masked by the low-albedo material. Emery and Brown (2004) modeled the spectra of several Trojan asteroids through the visible and near IR (out to 4 μm), extending the findings of Cruikshank et al. and concluding that organic-rich material was very unlikely to be responsible for the steep spectral slopes of the Trojans in general. Further work by Emery et al. (2006) using mid-IR (5.2–38 μm) observations from the Spitzer Space Telescope found evidence for fine-grained silicates on Trojan surfaces and emission spectra more typical to what is seen in cometary comae than asteroidal regoliths. This is suggestive of a severely underdense surface, or perhaps grains embedded in a transparent matrix.

The last major complex is the X complex, whose members are usually studied in the context of their Tholen classes (E, M, and P). The P class has not been the recipient of intense study, as its properties are similar to and sometimes overlap the C class and related groups, and indeed, some Tholen P asteroids are classified in the C complex in the Bus taxonomy. Hiroi et al. (2004) interpreted the P asteroids as akin to the carbonaceous chondrites, intermediate between the CI/CM and Tagish Lake meteorites and perhaps having experienced thermal and aqueous alteration.

The M-class asteroids have long been associated with iron meteorites and enstatite chondrites, both of which are relatively featureless in the 0.4–2.5 μm region. However, observations in the 3-μm region by Jones et al. (1990) and Rivkin et al. (1995, 2000) found absorptions in roughly a third of the M asteroids surveyed, with a strong size correlation. These 3-μm bands were interpreted as evidence that the M asteroids could not all be iron meteorite parents. In addition, higher spectral resolution in the visible and near IR has shown the presence of silicate absorptions in a large number of M asteroids (Hardersen et al. 2005), and it is recognized that a relatively wide variety of mineralogies are consistent with what is seen in the class.

While metal is featureless in visible-near IR wavelengths and thus spectral bands diagnostic for metal do not exist, thermal and radar data can provide more insights. Radar studies summarized by Shepard et al. (2010) show a wide range in the radar albedos of M asteroids, ranging from what is expected of pure metal surfaces to values more typical of rocky bodies. The radar albedo of an object is a function of both surface porosity and metal content, and only a handful of objects have radar albedos high enough to unambiguously require a high metal content. Radar-shape modeling of 216 Kleopatra shows a remarkable "dog bone" shape along

with a very high metal content. The asteroid 16 Psyche also has a high-radar albedo and is generally interpreted as a metallic core of a disrupted, differentiated body. Given Psyche's size, the original parent body must have been roughly the size of Vesta before it was disrupted. Shepard et al. found seven M asteroids had radar albedos indicative of high metal content, but suggested that most M-class asteroids are neither remnant cores or enstatite chondrites, but are instead composites more akin to stony iron meteorites or high-iron carbonaceous chondrites. Shepard et al. also noted a positive correlation between visual and radar albedos for the M asteroids that had moderate metal content, though they were uncertain as to the cause. Observations in the mid-IR by Lim and Emery (2006) also show a range of thermal properties for M asteroids, again with some objects consistent with the S and C asteroids and others with behavior more consistent with high metal contents. With only one possible exception, 129 Antigone, those objects seen to have a 3-μm band by Rivkin et al. are also the objects more consistent with lower metal abundances. Ockert-Bell studied 16 M and X asteroids, 9 of which had radar albedos available, and found a positive correlation between the 1.7–2.45 μm spectral slope and radar albedo.

The Rosetta spacecraft encountered two X asteroids: the E-class asteroid 2867 Šteins and the M-class asteroid 21 Lutetia. Data from Lutetia were collected over a very wide range of wavelengths, from the far UV (~0.1 μm) through the visible and infrared to 5 μm, with some millimeter-wave data also collected. Data from the encounter is still under analysis at this writing, but it is already well established from ground-based data that Lutetia does not have a high metal content at its surface, as reviewed by Belskaya et al. (2010).

E-class asteroids are distinguished by their high-albedos and featureless spectra and have been associated with the aubrite meteorites. As with the M asteroids (and D asteroids), the lack of interpretable absorption bands has hampered a full understanding of this class. Some E asteroids, notably 64 Angelina and 317 Roxane, possess an absorption band near 0.5 μm. This has been interpreted as due to sulfides, particularly oldhamite (CaS), though oldhamite has a significantly lower albedo than the E asteroids. Clark et al. (2004) performed mixing models based on Hapke theory on 9 E asteroids and divided them into 3 groups, "Nysa-like," "Angelina-like," and "Hungaria-like." In addition to a high-albedo mineral like enstatite in each group, the Nysa-like group was interpreted as bearing a low-iron orthopyroxene, the Angelina-like group possibly containing oldhamite and forsterite, and the Hungaria-like group "not inconsistent" with aubrites. Clark et al. concluded the aubrites were only likely to have come from the Hungaria-like group with the others unrepresented among the meteorites. Gaffey and Kelley (2004) independently but similarly divided the E asteroids into three groups. The 0.5-μm band is the basis for the Xe class in the Bus taxonomy, but it is cautioned that there is not a one-to-one correspondence between the E and Xe classes in the two taxonomies, as evidenced by the E asteroids in Clark et al. and Kelley et al. that do not have the 0.5-μm band.

Benner et al. (2008) found a correlation between the radar polarization ratio (usually interpreted as a function of surface roughness) and NEO spectral class, with the E-class asteroids exhibiting the highest SC/OC (same sense circular polarization to opposite sense circular polarization) ratios. This is suggestive of very high roughness on these objects on centimeter to decimeter scales. Benner et al. also note a surprisingly high fraction of E asteroids in their sample, roughly six times higher than the fraction of aubrites seen in the meteorite collection. It is not clear why this is the case, though possibilities they considered include extreme fragility for aubrites or additional analogs for the E asteroids unrecognized in the meteorite collection. They also identified several candidate E-class objects based on radar polarization alone, with confirmation via spectroscopy still pending.

References

Abe, S., Mukai, T., Hirata, N., Barnouin-Jha, O. S., Cheng, A. F., Demura, H., Gaskell, R. W., Hashimoto, T., Hiraoka, K., Honda, T., Kubota, T., Matsuoka, M., Mizuno, T., Nakamura, R., Scheeres, D. J., & Yoshikawa, M. 2006, Mass and local topography measurements of Itokawa by Hayabusa. Science, 312, 1344–1349

Abell, P. A., Vilas, F., Jarvis, K. S., Gaffey, M. J., & Kelley, M. S. 2007, Mineralogical composition of (25143) Itokawa 1998 SF36 from visible and near-infrared reflectance spectroscopy: evidence for partial melting. Meteorit. Planet. Sci., 42, 2165–2177

A'Hearn, M. F., & Feldman, P. D. 1992, Water vaporization on Ceres. Icarus, 98, 54–60

Asphaug, E. 2008, Critical crater diameter and asteroid impact seismology. Meteorit. Planet. Sci., 43, 1075–1084

Asphaug, E., & Benz, W. 1996, Size, density, and structure of comet Shoemaker-Levy 9 inferred from the physics of tidal breakup. Icarus, 121, 225–248

Asphaug, E., Ryan, E. V., & Zuber, M. T. 2002, Asteroid interiors, in Asteroids III, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 463–484

Barnouin-Jha, O. S., Mukai, T., Abe, S., Hirata, N., Nakamura, R., Gaskell, R. W., Saito, J., & Clark, B. E. 2008, Small-scale topography of 25143 Itokawa from the Hayabusa laser altimeter. Icarus, 198, 108–124

Bell, J. F. 1988, A probable asteroidal parent body for the CV or CO chondrites. Meteoritics, 23, 256–257

Belskaya, I. N., Fornasier, S., Krugly, Y. N., Shevchenko, V. G., Gaftonyuk, N. M., Barucci, M. A., Fulchignoni, M., & Gil-Hutton, R. 2010, Puzzling asteroid 21 Lutetia: our knowledge prior to the Rosetta fly-by. A&A, 515, A29+

Bendjoya, P., & Zappalà, V. 2002, Asteroid family identification, in Asteroids III, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 613–618

Benner, L. A. M., Ostro, S. J., Magri, C., Nolan, M. C., Howell, E. S., Giorgini, J. D., Jurgens, R. F., Margot, J., Taylor, P. A., Busch, M. W., & Shepard, M. K. 2008, Near-Earth asteroid surface roughness depends on compositional class. Icarus, 198, 294–304

Binzel, R. P., Gaffey, M. J., Thomas, P. C., Zellner, B. H., Storrs, A. D., & Wells, E. N. 1997, Geologic mapping of Vesta from 1994 Hubble Space Telescope images. Icarus, 128, 95–103

Binzel, R. P., Rivkin, A. S., Bus, S. J., Sunshine, J. M., & Burbine, T. H. 2001, MUSES-C target asteroid (25143) 1998 SF36: a reddened ordinary chondrite. Meteorit. Planet. Sci., 36, 1167–1172

Binzel, R. P., Rivkin, A. S., Stuart, J. S., Harris, A. W., Bus, S. J., & Burbine, T. H. 2004, Observed spectral properties of near-Earth objects: results for population distribution, source regions, and space weathering processes. Icarus, 170, 259–294

Binzel, R. P., Morbidelli, A., Merouane, S., DeMeo, F. E., Birlan, M., Vernazza, P., Thomas, C. A., Rivkin, A. S., Bus, S. J., & Tokunaga, A. T. 2010, Earth encounters as the origin of fresh surfaces on near-Earth asteroids. Nature, 463, 331–334

Bland, P. A., Smith, T. B., Jull, A. J. T., Berry, F. J., Bevan, A. W. R., Cloudt, S., & Pillinger, C. T. 1996, The flux of meteorites to the Earth over the last 50,000 years. MNRAS, 283, 551–+

Bottke, W. F. Jr., Nolan, M. C., Greenberg, R., & Kolvoord, R. A. 1994, Velocity distributions among colliding asteroids. Icarus, 107, 255–268

Bottke, W. F., Morbidelli, A., Jedicke, R., Petit, J.-M., Levison, H. F., Michel, P., & Metcalfe, T. S. 2002, Debiased orbital and absolute magnitude distribution of the near-Earth objects. Icarus, 156, 399–433

Bottke, W. F., Durda, D. D., Nesvorný, D., Jedicke, R., Morbidelli, A., Vokrouhlický, D., & Levison, H. F. 2005, Linking the collisional history of the main asteroid belt to its dynamical excitation and depletion. Icarus, 179, 63–94

Bottke, Jr., W. F., Vokrouhlický, D., Rubincam, D. P., & Nesvorný, D. 2006, The Yarkovsky and Yorp effects: implications for asteroid dynamics. Annu. Rev.Earth Planet. Sci., 34, 157–191

Brozovic, M., Benner, L. A. M., Taylor, P. A., Nolan, M. C., Howell, E. S., Magri, C., Scheeres, D. J., Giorgini, J. D., Pollock, J. T., Pravec, P., Galád, A., Fang, J., Margot, J.-L., Busch, M. W., Shepard, M. K., Reichart, D. E., Ivarsen, K. M., Haislip, J. B., Lacluyze, A. P., Jao, J., Slade, M. A., Lawrence, K. J., & Hicks, M. D. 2011, Radar and optical observations and physical modeling of triple near-Earth Asteroid (136617) 1994 CC. Icarus, 216(1), 241–256

Burbine, T. H. 1998, Could G-class asteroids be the parent bodies of the CM chondrites? Meteorit. Planet. Sci., 33, 253–258

Burbine, T. H., Meibom, A., & Binzel, R. P. 1996, Mantle material in the main belt: battered to bits? Meteorit. Planet. Sci., 31, 607–620

Burns, J. A. 1992, Contradictory clues as to the origin of the martian moons. In Mars, ed. H. H. Kieffer,

B. M. Jakosky, C. W. Snyder, & M. S. Matthews (Tucson: University of Arizona Press) 1283–1302

Bus, S. J. 1999, Compositional structure in the asteroid belt: results of a spectroscopic survey. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA

Busch, M. W., Benner, L. A. M., Ostro, S. J., Giorgini, J. D., Jurgens, R. F., Rose, R., Scheeres, D. J., Magri, C., Margot, J.-L., Nolan, M. C., & Hine, A. A. 2008, Physical properties of near-Earth Asteroid (33342) 1998 WT24. Icarus, 195(2), 614–621

Bus, S. J., & Binzel, R. P. 2002, Phase II of the small main-belt asteroid spectroscopic survey: a feature-based taxonomy. Icarus, 158, 146–177

Campins, H., Hargrove, K., Pinilla-Alonso, N., Howell, E. S., Kelley, M. S., Licandro, J., Mothé-Diniz, T., Fernández, Y., & Ziffer, J. 2010, Water ice and organics on the surface of the asteroid 24 Themis., 464, 1320–1321

Castillo-Rogez, J. C., & McCord, T. B. 2010, Ceres evolution and present state constrained by shape data. Icarus, 205, 443–459

Castillo-Rogez, J. C., & Schmidt, B. E. 2010, Geophysical evolution of the Themis family parent body. Geophys. Res. Lett., 37, 10202–+

Chabot, N. L., & Haack, H. 2006, Evolution of asteroidal cores, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press) 747–771

Chapman, C. R. 1996, S-type asteroids, ordinary chondrites, and space weathering: the evidence from Galileo's fly-bys of Gaspra and Ida. Meteorit. Planet. Sci., 31, 699–725

Chapman, C. R., Paolicchi, P., Zappala, V., Binzel, R. P., & Bell, J. F. 1989, Asteroid Families: Physical Properties and Evolution (Tucson: University of Arizona Press), 386–415

Chapman, C. R., Veverka, J., Thomas, P. C., Klaasen, K., Belton, M. J. S., Harch, A., McEwen, A., Johnson, T. V., Helfenstein, P., Davies, M. E., Merline, W. J., & Denk, T. 1995, Discovery and physical properties of Dactyl, a satellite of asteroid 243 Ida. Nature, 374, 783–785

Chapman, C. R., Ryan, E. V., Merline, W. J., Neukum, G., Wagner, R., Thomas, P. C., Veverka, J., & Sullivan, R. J. 1996a, Cratering on Ida. Icarus, 120, 77–86

Chapman, C. R., Veverka, J., Belton, M. J. S., Neukum, G., & Morrison, D. 1996b, Cratering on Gaspra. Icarus, 120, 231–245

Chapman, C. R., Merline, W. J., & Thomas, P. 1999, Cratering on Mathilde. Icarus, 140, 28–33

Chapman, C. R., Merline, W. J., Thomas, P. C., Joseph, J., Cheng, A. F., & Izenberg, N. 2002, Impact history of Eros: craters and boulders. Icarus, 155, 104–118

Cheng, A. F., & Barnouin, O. S. 2010, Eros and Itokawa comparisons: NEAR Shoemaker and Hayabusa, in volume 41 of Lunar and Planetary Institute Science Conference Abstracts, Lunar and Planetary Institute Science Conference Abstracts (Houston : Lunar and Planetary Institute), pages 2747–+

Cheng, A. F., Izenberg, N., Chapman, C. R., & Zuber, M. T. 2002, Ponded deposits on asteroid 433 Eros. Meteorit. Planet. Sci., 37, 1095–1105

Chesley, S. R., Ostro, S. J., Vokrouhlický, D., Apek, D., Giorgini, J. D., Nolan, M. C., Margot, J., Hine, A. A., Benner, L. A. M., & Chamberlin, A. B. 2003, Direct detection of the Yarkovsky effect by Radar Ranging to Asteroid 6489 Golevka. Science, 302, 1739–1742

Clark, R. N. 2009, Detection of adsorbed water and hydroxyl on the Moon. Science, 326, 562–

Clark, B. E., Bus, S. J., Rivkin, A. S., McConnochie, T., Sanders, J., Shah, S., Hiroi, T., & Shepard, M. 2004, E-type asteroid spectroscopy and compositional modeling. J. Geophys. Res., 109. doi:10.1029/2003JE002200

Cloutis, E. A., Gaffey, M. J., Smith, D. G. W., & St. Lambert, R. J. 1990, Reflectance spectra of "featureless" materials and the surface mineralogies of M- and E-class asteroids. J. Geophys. Res., 95, 281–293

Cohen, B. A., Swindle, T. D., & Kring, D. A. 2000, Support for the lunar cataclysm hypothesis from lunar meteorite impact melt ages. Science, 290(5497), 1754–1756

Colwell, J. E., Gulbis, A. A. S., Horányi, M., & Robertson, S. 2005, Dust transport in photoelectron layers and the formation of dust ponds on Eros. Icarus, 175, 159–169

Consolmagno, G., Britt, D., & Macke, R. 2008, The significance of meteorite density and porosity. Chemie der Erde/Geochemistry, 68, 1–29

Cruikshank, D. P., & Hartmann, W. K. 1984, The meteorite-asteroid connection – two olivine-rich asteroids. Science (ISSN 0036-8075), 223, 281–283

Cruikshank, D. P., Ore, C. M. D., Roush, T. L., Geballe, T. R., Owen, T. C., de Bergh, C., Cash, M. D., & Hartmann, W. K. 2001, Constraints on the composition of Trojan asteroid 624 Hektor. Icarus, 153, 348–360

Ćuk, M. 2007, Formation and destruction of small binary asteroids. ApJ, 659, L57–L60

Davis, D. R. 1999, The collisional history of Asteroid 253 Mathilde. Icarus, 140, 49–52

Delbo', M., Dell'Oro, A., Harris, A. W., Mottola, S., & Mueller, M. 2007, Thermal inertia of near-Earth

asteroids and implications for the magnitude of the Yarkovsky effect. Icarus, 190, 236–249

DeMeo, F. E., Binzel, R. P., Slivan, S. M., & Bus, S. J. 2009, An extension of the Bus asteroid taxonomy into the near-infrared. Icarus, 202, 160–180

Dermott, S. F., Durda, D. D., Grogan, K., & Keohoe, T. J. J. 2002, Asteroidal dust, in Asteroids III, ed. W. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 423–442

Dohnanyi, J. S. 1969, Collisional model of asteroids and their debris. J. Geophys. Res., 74, 2531–2554

Dombard, A. J., Barnouin, O. S., Prockter, L. M., & Thomas, P. C. 2010, Boulders and ponds on the Asteroid 433 Eros. Icarus, 210, 713–721

Dreibus, G., Bruckner, J., & Wanke, H. 1997, On the core mass of the Asteroid Vesta. Meteorit. Planet. Sci. Suppl., 32, 36–+

Dunn, T. L., McCoy, T. J., Sunshine, J. M., & McSween, H. Y. 2010, A coordinated spectral, mineralogical, and compositional study of ordinary chondrites. Icarus, 208, 789–797

Durda, D. D., Greenberg, R., & Jedicke, R. 1998, Collisional models and scaling laws: a new interpretation of the shape of the main-belt asteroid size distribution. Icarus 135, 431–440

Durda, D. D., Bottke, W. F., Nesvorný, D., Enke, B. L., Merline, W. J., Asphaug, E., & Richardson, D. C. 2007, Size frequency distributions of fragments from SPH/N-body simulations of asteroid impacts: comparison with observed asteroid families. Icarus, 186, 498–516

Emery, J. P., & Brown, R. H. 2004, The surface composition of Trojan asteroids: constraints set by scattering theory. Icarus, 170, 131–152

Emery, J. P., Cruikshank, D. P., & van Cleve, J. 2006, Thermal emission spectroscopy (5.2 38 μm) of three Trojan asteroids with the Spitzer Space Telescope: detection of fine-grained silicates. Icarus, 182, 496–512

Eugster, O., Herzog, G. F., Marti, K., & Caffee, M. W. 2006, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press) Irradiation records, cosmic-ray exposure ages, and transfer times of meteorites, 829–851

Fienga, A., Laskar, J., Morley, T., Manche, H., Kuchynka, P., Le Poncin-Lafitte, C., Budnik, F., Gastineau, M., & Somenzi, L. 2009, INPOP08, a 4-D planetary ephemeris: from asteroid and time-scale computations to ESA Mars Express and Venus Express contributions. A&A, 507, 1675–1686

Fujiwara, A., Kawaguchi, J., Yeomans, D. K., Abe, M., Mukai, T., Okada, T., Saito, J., Yano, H., Yoshikawa, M., Scheeres, D. J., Barnouin-Jha, O., Cheng, A. F., Demura, H., Gaskell, R. W., Hirata, N., Ikeda, H., Kominato, T., Miyamoto, H., Nakamura, A. M., Nakamura, R., Sasaki, S., & Uesugi, K. 2006, The Rubble-pile Asteroid Itokawa as observed by Hayabusa. Science, 312, 1330–1334

Gaffey, M. J., & Gilbert, S. L. 1998, Asteroid 6 Hebe: the probable parent body of the H-type ordinary chondrites and the IIE iron meteorites. Meteorit. Planet. Sci., 33, 1281–1295

Gaffey, M. J., & McCord, T. B. 1978, Asteroid surface materials – mineralogical characterizations from reflectance spectra. Space Sci. Rev., 21, 555–628

Gaffey, M. J., Reed, K. L., & Kelley, M. S. 1992, Relationship of E-type asteroid 3103 (1982 BB) to the enstatitechondrite meteorites and the Hungaria asteroids. Icarus, 100, 95–109

Gaffey, M. J. 1997, Surface Lithologic Heterogeneity of Asteroid 4 Vesta. Icarus, 127, 130–157

Gaffey, M. J. 2010, Space weathering and the interpretation of asteroid reflectance spectra. Icarus, 209, 564–574

Gaffey, M. J., & Kelley, M. S. 2004, Mineralogical variations among high Albedo E-type Asteroids: implications for asteroid igneous processes, in volume 35 of Lunar and Planetary Institute Science Conference Abstracts, Lunar and Planetary Institute Science Conference Abstracts, ed. S. Mackwell & E. Stansbery (Houston : Lunar and Planetary Institute), 1812–+

Gaffey, M. J., Bell, J. F., Brown, R. H., Burbine, T. H., Piatek, J. L., Reed, K. L., & Chaky, D. A. 1993, Mineralogical variations within the S-type asteroid class. Icarus, 106, 573–602

Galimov, E. M. 2010, Phobos sample return mission: scientific substantiation. Solar Syst. Res., 44, 5–14

Geissler, P., Petit, J., Durda, D. D., Greenberg, R., Bottke, W., Nolan, M., & Moore, J. 1996, Erosion and Ejecta Reaccretion on 243 Ida and its Moon. Icarus, 120, 140–157

Ghosh, A., Weidenschilling, S. J., & McSween, H. Y., Jr. 2003, Importance of the accretion process in asteroid thermal evolution: 6 Hebe as an example. Meteorit. Planet. Sci. , 38, 711–724

Gladman, B. J., Migliorini, F., Morbidelli, A., Zappala, V., Michel, P., Cellino, A., Froeschle, C., Levison, H. F., Bailey, M., & Duncan, M. 1997, Dynamical lifetimes of objects injected into asteroid belt resonances. Science, 277, 197–201

Gomes, R., Levison, H. F., Tsiganis, K., & Morbidelli, A. 2005, Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Nature, 435, 466–469

Gradie, J. C., & Tedesco, E. F. 1982, Compositional structure of the asteroid belt. Science, 216, 1405–1407

Grier, J. A., McEwen, A. S., Lucey, P. G., Milazzo, M., & Strom, R. G. 2001, Optical maturity of ejecta from large rayed lunar craters. J. Geophys. Res., 106, 32847–32862

Grimm, R. E., & McSween, H. Y. 1993, Heliocentric zoning of the asteroid belt by aluminum-26 heating. Science, 259, 653–655

Halliday, A. N., & Kleine, T. 2006, Meteorites and the timing, mechanisms, and conditions of terrestrial planet accretion and early differentiation, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press), 775–801

Hapke, B. 2001, Space weathering from Mercury to the asteroid belt. J. Geophys. Res., 106, 10039–10074

Hardersen, P. S., Gaffey, M. J., & Abell, P. A. 2005, Near-IR spectral evidence for the presence of iron-poor orthopyroxenes on the surfaces of six M-type asteroids. Icarus, 175, 141–158

Harris, A. W. 1996, The rotation rates of very small asteroids: evidence for 'Rubble Pile' structure, in volume 27 of Lunar and Planetary Institute Science Conference Abstracts, Lunar and Planetary Institute Science Conference Abstracts (Houston : Lunar and Planetary Institute), 493–+

Harris, A. W., & Lagerros, J. S. V. 2002, Asteroids in the thermal infrared, in Asteroids III, ed. W. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 205–218

Hartmann, W. K. 2003, Megaregolith evolution and cratering cataclysm models–Lunar cataclysm as a misconception (28 years) later. Meteorit. Planet. Sci., 38, 579–593

Hasegawa, S., Murakawa, K., Ishiguro, M., Nonaka, H., Takato, N., Davis, C. J., Ueno, M., & Hiroi, T. 2003, Evidence of hydrated and/or hydroxylated minerals on the surface of asteroid 4 Vesta. GRL, 30. doi:10.1029/2003GL01862

Hilton, J. L. 2002, Asteroid masses and densities., in Asteroids III, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 103–112

Hirayama, K. 1918, Groups of asteroids probably of common origin. AJ, 31, 185–188

Hiroi, T., & Hasegawa, S. 2003, Revisiting the search for the parent body of the Tagish Lake meteorite: case of aT/D asteroid 308 Polyxo. Antarct. Meteor. Res., 16, 176–184

Hiroi, T., Pieters, C. M., Rutherford, M. J., Zolensky, M. E., Sasaki, S., Ueda, Y., & Miyamoto, M. 2004, What are the P-type Asteroids made of? in volume 35 of Lunar and Planetary Institute Science Conference Abstracts, Lunar and Planetary Institute Science Conference Abstracts, ed. S. Mackwell & E. Stansbery (Houston : Lunar and Planetary Institute), 1616–+

Hiroi, T., Zolensky, M. E., Pieters, C. M., & Lipschutz, M. E. 1996, Thermal metamorphism of the C, G, B, and F asteroids seen from the 0.7 μm, 3 μm and UV absorption strengths in comparison with carbonaceous chondrites. Meteorit. Planet. Sci., 31, 321–327

Hiroi, T., Zolensky, M. E., & Pieters, C. M. 2001, The Tagish lake meteorite: a possible sample from a D- type asteroid. Science, 293, 2234–2236

Holsapple, K. A. 2007, Spin limits of Solar system bodies: from the small fast-rotators to 2003 EL61. Icarus, 187, 500–509

Holsapple, K. A., & Michel, P. 2006, Tidal disruptions: a continuum theory for solid bodies., 183, 331–348

Howell, E. S., Britt, D. T., Bell, J. F., Binzel, R. P., & Lebofsky, L. A. 1994, Visible and near-infrared spectral observations of 4179 Toutatis. Icarus, 111, 468–474

Hughes, A. L. H., Colwell, J. E., & Dewolfe, A. W. 2008, Electrostatic dust transport on Eros: 3-D simulations of pond formation. Icarus, 195, 630–648

Huss, G. R., Rubin, A. E., & Grossman, J. N. 2006, Thermal metamorphism in chondrites, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press), 567–586

Ivezić, Ž., Tabachnik, S., Rafikov, R., Lupton, R. H., Quinn, T., Hammergren, M., Eyer, L., Chu, J., Armstrong, J. C., Fan, X., Finlator, K., Geballe, T. R., Gunn, J. E., Hennessy, G. S., Knapp, G. R., Leggett, S. K., Munn, J. A., Pier, J. R., Rockosi, C. M., Schneider, D. P., Strauss, M. A., Yanny, B., Brinkmann, J., Csabai, I., Hindsley, R. B., Kent, S., Lamb, D. Q., Margon, B., McKay, T. A., Smith, J. A., Waddel, P., York, D. G., & the SDSS Collaboration, 2001, Solar system objects observed in the sloan digital sky survey commissioning data. AJ, 122, 2749–2784

Ivezić, Ž., Lupton, R. H., Juriæ, M., Tabachnik, S., Quinn, T., Gunn, J. E., Knapp, G. R., Rockosi, C. M., & Brinkmann, J. 2002, Color confirmation of asteroid families. AJ, 124, 2943–2948

Izenberg, N. R., Murchie, S. L., Bell, III, J. F., McFadden, L. A., Wellnitz, D. D., Clark, B. E., & Gaffey, M. J. 2003, Spectral properties and geologic processes on Eros from combined NEAR NIS and MSI data sets. Meteorit. Planet. Sci., 38, 1053–1077

Jedicke, R., Nesvorný, D., Whiteley, R., Ivezić, Ž., & Juriæ, M. 2004, An age-colour relationship for main-belt S-complex asteroids. Nature, 429, 275–277

Jenniskens, P., Shaddad, M. H., Numan, D., Elsir, S., Kudoda, A. M., Zolensky, M. E., Le, L., Robinson, G. A., Friedrich, J. M., Rumble, D., Steele, A., Chesley, S. R., Fitzsimmons, A., Duddy, S., Hsieh, H. H., Ramsay, G., Brown, P. G., Edwards, W. N., Tagliaferri, E., Boslough, M. B., Spalding, R. E., Dantowitz, R., Kozubal, M., Pravec, P., Borovicka, J., Charvat, Z., Vaubaillon, J., Kuiper, J., Albers, J., Bishop, J. L., Mancinelli, R. L., Sandford, S. A., Milam, S. N., Nuevo, M., & Worden, S. P. 2009, The impact and recovery of asteroid 2008 TC3. Nature, 458, 485–488

Jewitt, D., Weaver, H., Agarwal, J., Mutchler, M., & Drahus, M. 2010, A recent disruption of the main-belt asteroid P/2010A2. Nature, 467, 817–819

Johnston, W. R. 2012. Asteroids with satellites. Retrieved from http://www.johnstonsarchive.net/astro/asteroidmoons.html

Jones, T. D., Lebofsky, L. A., Lewis, J. S., & Marley, M. S. 1990, The composition and origin of the C, P, and D asteroids: water as a tracer of thermal evolution in the outer belt. Icarus, 88, 172–192

Keil, K. 2002, Geological history of asteroid 4 Vesta: the "smallest terrestrial planet," in Asteroids III, ed. W. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 573–584

Keller, H. U., Barbieri, C., Koschny, D., Lamy, P., Rickman, H., Rodrigo, R., Sierks, H., A'Hearn, M. F., Angrilli, F., Barucci, M. A., Bertaux, J., Cremonese, G., Da Deppo, V., Davidsson, B., De Cecco, M., Debei, S., Fornasier, S., Fulle, M., Groussin, O., Gutierrez, P. J., Hviid, S. F., Ip, W., Jorda, L., Knollenberg, J., Kramm, J. R., Kührt, E., Küppers, M., Lara, L., Lazzarin, M., Moreno, J. L., Marzari, F., Michalik, H., Naletto, G., Sabau, L., Thomas, N., Wenzel, K., Bertini, I., Besse, S., Ferri, F., Kaasalainen, M., Lowry, S., Marchi, S., Mottola, S., Sabolo, W., Schröder, S. E., Spjuth, S., & Vernazza, P. 2010, E-Type Asteroid (2867) steins as imaged by OSIRIS on board Rosetta. Science, 327, 190–

Konopliv, A. S., Asmar, S. W., Folkner, W. M., Karatekin, Ö., Nunes, D. C., Smrekar, S. E., Yoder, C. F., & Zuber, M. T. 2011, Mars high resolution gravity fields from MRO, Mars seasonal gravity, and other dynamical parameters. Icarus, 211, 401–428

Lim, L. F., & Emery, J. P. 2006, A Spitzer IRS survey of hydrated and non-hydrated M asteroids: preliminary 5-13 Micron results, in volume 38 of Bulletin of the American Astronomical Society, Bulletin of the American Astronomical Society, (Washington: American Astronomical Society) 626–+

Lim, L. F., & Nittler, L. R. 2009, Elemental composition of 433 Eros: new calibration of the NEAR-Shoemaker XRS data. Icarus, 200, 129–146

Loeffler, M. J., Dukes, C. A., & Baragiola, R. A. 2009, Irradiation of olivine by 4 keV He: simulation of space weathering by the solar wind. J. Geophys. Res., 114, 3003–+

Lucey, P. G., Hinrichs, J., Kelly, M., Wellnitz, D., Izenberg, N., Murchie, S., Robinson, M., Clark, B. E., & Bell, J. F. 2002, Detection of temperature-dependent spectral variation on the asteroid Eros and new evidence for the presence of an olivine-rich silicate assemblage. Icarus, 155, 181–188

Lucey, P. G., Keil, K., & Whitely, R. 1998, The influence of temperature on the spectra of the A-asteroids and implications for their silicate chemistry. J. Geophys. Res., 103(E3), 5865–5871

Magri, C., Ostro, S. J., Rosema, K. D., Thomas, M. L., Mitchell, D. L., Campbell, D. B., Chandler, J. F., Shapiro, I. I., Giorgini, J. D., & Yeomans, D. K. 1999, Mainbelt asteroids: results of Arecibo and Goldstone radar observations of 37objects during 1980–1985. Icarus, 140, 379–407

Marchi, S., Barbieri, C., Küppers, M., Marzari, F., Davidsson, B., Keller, H. U., Besse, S., Lamy, P., Mottola, S., Massironi, M., & Cremonese, G. 2010, The cratering history of asteroid (2867) Steins. Planet. Space Sci., 58, 1116–1123

McCord, T. B., & Sotin, C. 2005, Ceres: evolution and current state. J. Geophys. Res., 110, 5009–+

McCord, T. B., Adams, J. B., & Johnson, T. B. 1970, Asteroid Vesta: spectral reflectivity and compositional information. Science, 168, 1445–1447

McCoy, T. J., Nittler, L. R., Burbine, T. H., Trombka, J. I., Clark, P. E., & Murphy, M. E. 2000, Anatomy of a partially differentiated asteroid: a "NEAR"-sighted view of acapulcoites and lodranites. Icarus, 148, 29–36

McCoy, T. J., Mittlefehldt, D. W., & Wilson, L. 2006, Asteroid differentiation, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press), 733–745

McSween, H. Y., Mittlefehldt, D. W., Beck, A. W., Mayne, R. G., & McCoy, T. J. 2010, HED meteorites and their relationship to the geology of Vesta and the Dawn mission. Space Science Reviews, 163(1–4), 141–174

Michel, P., Tanga, P., Benz, W., & Richardson, D. C. 2002, Formation of asteroid families by catastrophic disruption: simulations with fragmentation and gravitational reaccumulation. Icarus, 160, 10–23

Michel, P., O'Brien, D. P., Abe, S., & Hirata, N. 2009, Itokawa's cratering record as observed by

Hayabusa: implications for its age and collisional history. Icarus, 200, 503–513

Milliken, R. E., & Rivkin, A. S. 2009, Brucite and carbonate assemblages from altered olivine-rich materials on Ceres. Nat. Geosci., 2, 258–261

Morbidelli, A., & Vokrouhlický, D. 2003, The Yarkovsky-driven origin of near-Earth asteroids. Icarus, 163, 120–134

Morbidelli, A., Bottke, W. F., Jr., Froeschlé, C., & Michel, P. 2002, Origin and evolution of near-earth objects, in Asteroids III, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 409–422

Morbidelli, A., Levison, H. F., Tsiganis, K., & Gomes, R. 2005, Chaotic capture of Jupiter's Trojan asteroids in the early Solar System. Nature, 435, 462–465

Morbidelli, A., Bottke, W. F., Nesvorný, D., & Levison, H. F. 2009, Asteroids were born big. Icarus, 204, 558–573

Morris, M. A., & Desch, S. J. 2010, Thermal histories of Chondrules in Solar Nebula shocks. AJ, 722, 1474–1494

Moskovitz, N. A., Jedicke, R., Gaidos, E., Willman, M., Nesvorný, D., Fevig, R., & Ivezić, Ž. 2008, The distribution of basaltic asteroids in the Main Belt. Icarus, 198, 77–90

Mothé-Diniz, T., Jasmin, F. L., Carvano, J. M., Lazzaro, D., Nesvorný, D., & Ramirez, A. C. 2010, Re-assessing the ordinary chondrites paradox. A&A, 514, A86+

Nathues, A., Mottola, S., Kaasalainen, M., & Neukum, G. 2005, Spectral study of the Eunomia asteroid family. I. Eunomia. Icarus, 175, 452–463

Nesvorný, D., Bottke, W. F., Jr., Dones, L., & Levison, H. F. 2002, The recent breakup of an asteroid in the main-belt region. Nature, 417, 720–771

Nesvorný, D., Alvarellos, J. L. A., Dones, L., & Levison, H. F. 2003, Orbital and collisional evolution of the irregular satellites. AJ, 126, 398–429

Nesvorný, D., Bottke, W. F., Vokrouhlický, D., Chapman, C. R., & Rafkin, S. 2010a, Do planetary encounters reset surfaces of near Earth asteroids? Icarus, 209, 510–519

Nesvorný, D., Jenniskens, P., Levison, H. F., Bottke, W. F., Vokrouhlický, D., & Gounelle, M. 2010b, Cometary origin of the zodiacal cloud and carbonaceous micrometeorites. Implications for hot debris disks. AJ, 713, 816–836

Nesvorný, D., Vokrouhlický, D., & Morbidelli, A. 2007, Capture of irregular satellites during planetary encounters. AJ, 133(5), 1962–1976

Nicholson, P. D., Cuk, M., Sheppard, S. S., Nesvorny, D., & Johnson, T. V. 2008, Irregular satellites of the giant planets. In The Solar System Beyond Neptune, eds. M. A. Barucci, H. Boehnhardt, D. P. Cruikshank, and A. Morbidelli, (Tucson: University of Arizona Press), 411–424

Noble, S. K., Pieters, C., Taylor, L. A., Morris, R. V., Allen, C. C., McKay, D. S., & Keller, L. P. 2001, The optical properties of the finest fraction of lunar soil: implications for space weathering. Meteorit. Planet. Sci., 36, 31–42

Nolan, M. C., Asphaug, E., Greenberg, R., & Melosh, H. J. 2001, Impacts on asteroids: fragmentation, regolith transport, and disruption. Icarus, 153, 1–15

Nolan, M. C., Magri, C., Ostro, S. J., Benner, L. A., Giorgini, J. D., Howell, E. S., & Hudson, R. S. 2007, The Shape and Spin of 101955 (1999 RQ36) from Arecibo and Goldstone Radar Imaging, in volume 38 of Bulletin of the American Astronomical Society, AAS/Division for Planetary Sciences Meeting Abstracts #39, 433–+ (Washington: American Astronomical Society)

Nur, A., & Simmons, G. 1969, The effect of saturation on velocity in low porosity rocks. Earth Planet. Sci. Lett., 7, 183–+

O'Brien, D. P., & Greenberg, R. 2005, The collisional and dynamical evolution of the main-belt and NEA size distributions. Icarus, 178, 179–212

O'Brien, D. P., Greenberg, R., & Richardson, J. E. 2006, Craters on asteroids: reconciling diverse impact records with a common impacting population. Icarus, 183(1), 79–92

Okada, T., Shirai, K., Yamamoto, Y., Arai, T., Ogawa, K., Hosono, K., & Kato, M. 2006, X-ray fluorescence spectrometry of asteroid Itokawa by Hayabusa. Science, 312, 1338–1341

Ostro, S. J., Margot, J., Benner, L. A. M., Giorgini, J. D., Scheeres, D. J., Fahnestock, E. G., Broschart, S. B., Bellerose, J., Nolan, M. C., Magri, C., Pravec, P., Scheirich, P., Rose, R., Jurgens, R. F., De Jong, E. M., & Suzuki, S. 2006, Radar imaging of binary near-Earth Asteroid (66391) 1999 KW4. Science, 314, 1276–1280

Petit, J., Durda, D. D., Greenberg, R., Hurford, T. A., & Geissler, P. E. 1997, The Long-term dynamics of Dactyl's orbit. Icarus, 130, 177–197

Pieters, C., Taylor, L. A., Noble, S. K., Keller, L. P., Hapke, B., Morris, R. V., Allen, C. C., McKay, D. S., & Wentworth, S. 2000, Space weathering on airless bodies: resolving a mystery with lunar samples. Meteorit. Planet. Sci., 35, 1101–1107

Pieters, C. M., Goswami, J. N., Clark, R. N., Annadurai, M., Boardman, J., Buratti, B., Combe, J., Dyar, M. D., Green, R., Head, J. W., Hibbitts, C., Hicks, M., Isaacson, P., Klima, R., Kramer, G., Kumar, S., Livo, E., Lundeen, S., Malaret, E., McCord, T., Mustard, J., Nettles, J., Petro, N., Runyon, C., Staid, M., Sunshine, J., Taylor, L. A., Tompkins, S., & Varanasi, P. 2009, Character and

spatial distribution of OH/H_2O on the surface of the Moon seen by M on Chandrayaan-1. Science, 326, 568–

Pizzarello, S., Cooper, G. W., & Flynn, G. J. 2006, The nature and distribution of the organic material in carbonaceous chondrites and interplanetary dust particles, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press) 625–651

Pravec, P., Vokrouhlický, D., Polishook, D., Scheeres, D. J., Harris, A. W., Galád, A., Vaduvescu, O., Pozo, F., Barr, A., Longa, P., Vachier, F., Colas, F., Pray, D. P., Pollock, J., Reichart, D., Ivarsen, K., Haislip, J., Lacluyze, A., Kušnirák, P., Henych, T., Marchis, F., Macomber, B., Jacobson, S. A., Krugly, Y. N., Sergeev, A. V., & Leroy, A. 2010, Formation of asteroid pairs by rotational fission. Nature, 466, 1085–1088

Reed, K. L., Gaffey, M. J., & Lebofsky, L. A. 1997, Shape and albedo variations of asteroid 15 Eunomia. Icarus, 125, 446–454

Richardson, D. C., & Walsh, K. J. 2006, Binary minor planets. Annu. Rev. Earth Planet. Sci., 34, 47–81

Richardson, J. E., Melosh, H. J., Greenberg, R. J., & O'Brien, D. P. 2005, The global effects of impact-induced seismic activity on fractured asteroid surface morphology. Icarus, 179, 325–349

Rivkin, A. S. 2007, Diversity of hydrated minerals on C-Class asteroids, in volume 38 of Bulletin of the American Astronomical Society, AAS/Division for Planetary Sciences Meeting Abstracts #39, 476–+ (Washington: American Astronomical Society)

Rivkin, A. S., & Clark, B. E. 2001, Observations of 433 Eros from 1.25–3.35 μm. Meteorit. Planet. Sci., 36, 1729–1729

Rivkin, A. S., & Emery, J. P. 2010, Detection of ice and organics on an asteroidal surface. Nature, 464, 1322–1323

Rivkin, A. S., Howell, E. S., Britt, D. T., Lebofsky, L. A., Nolan, M. C., & Branston, D. D. 1995, 3-μm spectrophotometric survey of M and E-class asteroids. Icarus, 117, 90–100

Rivkin, A. S., Lebofsky, L. A., Clark, B. E., Howell, E. S., & Britt, D. T. 2000, The nature of M-class asteroids in the 3-μm region. Icarus, 145, 351–368

Rivkin, A. S., McFadden, L. A., Binzel, R. P., & Sykes, M. 2006, Rotationally-resolved spectroscopy of Vesta I: 2 4 μm region. Icarus, 180, 464–472

Rivkin, A. S., Thomas, C. A., Trilling, D. E., Enga, M., & Grier, J. A. 2011b, Ordinary chondrite-like colors in small Koronis family members. Icarus, 211, 1294–1297

Roberts, J. H., Rivkin, A. S., & Chabot, N. L. 2011, A transient dynamo on vesta? in The 42nd Lunar Planet. Sci. Conf., The Woodlands, TX, 7–11 Mar. 2011. LPI Contribution No. 1608, 2242

Robinson, M. S., Thomas, P. C., Veverka, J., Murchie, S., & Carcich, B. 2001, The nature of ponded deposits on Eros. Nature, 413, 396–400

Rubin, A. E. 1996, Mineralogy of meteorite groups. Meteorit. Planet. Sci., 32, 231–247

Ruzicka, A., Snyder, G. A., & Taylor, L. A. 1997, Vesta as the HED parent body: implications for the size of a core and for large-scale differentiation. Meteorit. Planet. Sci., 32, 825–840

Scheeres, D. J., Abe, M., Yoshikawa, M., Nakamura, R., Gaskell, R. W., & Abell, P. A. 2007, The effect of YORP on Itokawa. Icarus, 188, 425–429

Scheeres, D. J., Hartzell, C. M., Sánchez, P., & Swift, M. 2010, Scaling forces to asteroid surfaces: the role of cohesion. Icarus, 210, 968–984

Schorghofer, N. 2008, The lifetime of ice on main belt asteroids. ApJ, 682, 697–705

Scott, E. R. D., Haack, H., & Love, S. G. 2001, Formation of mesosiderites by fragmentation and reaccretion of a large differentiated asteroid. Meteorit. Planet. Sci., 36, 869–891

Shepard, M. K., Clark, B. E., Ockert-Bell, M., Nolan, M. C., Howell, E. S., Magri, C., Giorgini, J. D., Benner, L. A. M., Ostro, S. J., Harris, A. W., Warner, B. D., Stephens, R. D., & Mueller, M. 2010, A radar survey of M- and X-class asteroids II. Summary and synthesis. Icarus, 208, 221–237

Shestopalov, D. I., McFadden, L. A., Golubeva, L. F., & Orujova, L. O. 2010, About mineral composition of geologic units in the northern hemisphere of Vesta. Icarus, 209, 575–585

Snodgrass, C., Tubiana, C., Vincent, J., Sierks, H., Hviid, S., Moissi, R., Boehnhardt, H., Barbieri, C., Koschny, D., Lamy, P., Rickman, H., Rodrigo, R., Carry, B., Lowry, S. C., Laird, R. J. M., Weissman, P. R., Fitzsimmons, A., Marchi, S., & OSIRIS Team 2010, A collision in 2009 as the origin of the debris trail of asteroid P/2010A2. Nature, 467, 814–816

Stuart, J. S., & Binzel, R. P. 2004, Bias-corrected population, size distribution, and impact hazard for the near-Earth objects. Icarus, 170(2), 295–311

Sullivan, R. J., Thomas, P. C., Murchie, S. L., & Robinson, M. S. 2002, Asteroid geology from Galileo and NEAR Shoemaker data, in Asteroids III, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 331–350

Sunshine, J. M., Binzel, R. P., Burbine, T. H., & Bus, S. J. 1998, Is asteroid 289 Nenetta compositionally analogous to the Brachinite meteorites? Lunar Planet. Sci., XXIX, 1430

Sunshine, J. M., Bus, S. J., McCoy, T. J., Corrigan, C. M., Burbine, T. H., & Binzel, R. P. 2004, High-Calcium pyroxene as an indicator of igneous differentiation in asteroids and meteorites. Meteorit. Planet. Sci., 39, 1343–1357

Sunshine, J. M., Farnham, T. L., Feaga, L. M., Groussin, O., Merlin, F., Milliken, R. E., & A'Hearn, M. F. 2009, Temporal and spatial variability of lunar hydration as observed by the deep impact spacecraft. Science, 326, 565–

Swindle, T. D., Isachsen, C. E., Weirich, J. R., & Kring, D. A. 2009, 40Ar-39Ar ages of H-chondrite impact melt breccias. Meteorit. Planet. Sci., 44(5), 747–762

Tedesco, E. F., & Desert, F. 2002, The infrared space observatory deep asteroid search. AJ, 123, 2070–2082

Tholen, D. J. 1984, Asteroid taxonomy from cluster analysis of photometry. Ph.D. dissertation, University of Arizona, Tucson

Thomas, P. C., & Robinson, M. S. 2005, Seismic resurfacing by a single impact on the asteroid 433 Eros. Nature, 436, 366–369

Thomas, P. C., Binzel, R. P., Gaffey, M. J., Zellner, B. H., Storrs, A. D., & Wells, E. 1997, Vesta: spin pole, size, and shape from HST images. Icarus, 128, 88–94

Thomas, P. C., Parker, J. W., McFadden, L. A., Russell, C. T., Stern, S. A., Sykes, M. V., & Young, E. F. 2005, Differentiation of the asteroid Ceres as revealed by its shape. Nature, 437, 224–226

Tsiganis, K., Gomes, R., Morbidelli, A., & Levison, H. F. 2005, Origin of the orbital architecture of the giant planets of the Solar System. Nature, 435(7041), 459–461

Vernazza, P., Binzel, R. P., Thomas, C. A., DeMeo, F. E., Bus, S. J., Rivkin, A. S., & Tokunaga, A. T. 2008, Compositional differences between meteorites and near-Earth asteroids. Nature, 454, 858–860

Vernazza, P., Binzel, R. P., Rossi, A., Fulchignoni, M., & Birlan, M. 2009, Solar wind as the origin of rapid reddening of asteroid surfaces. Nature, 458, 993–995

Vilas, F. 1994, A cheaper, faster, better way to detect water of hydration on solar system bodies. Icarus, 111, 456–467

Warner, B. D., Harris, A. W., & Pravec, P. 2009, The asteroid lightcurve database. Icarus, 202, 134–146

Weisberg, M. K., McCoy, T. J., & Krot, A. N. 2006, Systematics and evaluation of meteorite classification, in ed. H. Y. Lauretta & H. Y., Jr. McSween, Meteorites and the Early Solar System II (Tucson: University of Arizona Press), 19–52

Weiss, B. P., Gattacceca, J., Stanley, S., Rochette, P., & Christensen, U. R. 2010, Paleomagnetic records of meteorites and early planetesimal differentiation. Space Sci. Rev., 152, 341–390

Wilcox, B. B., Robinson, M. S., Thomas, P. C., & Hawke, B. R. 2005, Constraints on the depth and variability of the lunar regolith. Meteorit. Planet. Sci., 40, 695–+

Willman, M., & Jedicke, R. 2011, Asteroid age distributions determined by space weathering and collisional evolution models. Icarus, 211, 504–510

Yamaguchi, A., Clayton, R. N., Mayeda, T. K., Ebihara, M., Oura, Y., Miura, Y. N., Haramura, H., Misawa, K., Kojima, H., & Nagao, K. 2002, A new source of basaltic meteorites inferred from Northwest Africa 011. Science, 296, 334–36

Yang, J., Goldstein, J. I., & Scott, E. R. D. 2010, Main-group pallasites: thermal history, relationship to IIIAB irons, and origin. Geochim. Cosmochim. Acta, 74, 4471–4492

Yeomans, D. K., Antreasian, P. G., Barriot, J.-P., Chesley, S. R., Dunham, D. W., Farquhar, R. W., Giorgini, J. D. Helfrich, C. E. Konopliv, A. S., McAdams, J. V., Miller, J. K., Owen, W. M., Scheeres, D. J., Thomas, P. C., Veverka, J., & Williams, B. G. 2000, Radio science results during the NEAR-Shoemaker spacecraft Rendezvous with eros. Science, 289(5487), 2085–2088

Zolensky, M., Bland, P., Brown, P., & Halliday, I. 2006, Flux of extraterrestrial materials, In Meteorites and the Early Solar System II, ed. D. S. Lauretta and H. Y. McSween Jr. (Tucson: University of Arizona Press), 869–888

Zolensky, M., Herrin, J., Mikouchi, T., Ohsumi, K., Friedrich, J., Steele, A., Rumble, D., Fries, M., Sandford, S., Milam, S., Hagiya, K., Takeda, H., Satake, W., Kurihara, T., Colbert, M., Hanna, R., Maisano, J., Ketcham, R., Goodrich, C., Le, L., Robinson, G., Martinez, J., Ross, K., Jenniskens, P., & Shaddad, M. H. 2010, Mineralogy and petrography of the Almahata Sitta ureilite. Meteorit. Planet. Sci., 45, 1618–1637

Zolotov, M. Y. 2009, On the composition and differentiation of Ceres. Icarus, 204, 183–193

9 Dusty Planetary Systems

Amaya Moro-Martín[1, 2]
[1]Department of Astrophysics, Centro de Astrobiología, INTA-CSIC, Instituto de Tecnica Aeroespacial, Madrid, Spain
[2]Department of Astrophysical Sciences, Princeton University , Princeton, NJ, USA

1 Part I: Solar and Extrasolar Debris Disks 433
1.1 Debris Disks Are Evidence of the Presence of Extrasolar Planetesimals 433
1.2 The Solar System Debris Disk 435
1.2.1 Debris Dust in the Inner Solar System 435
1.2.2 Debris Dust in the Outer Solar System 440
1.2.3 Evolution of the Dust Production Rate in the Solar System 442
1.3 Extrasolar Debris Disks 444
1.3.1 Debris Disk Frequency 444
1.3.2 Debris Disk Evolution 447
1.3.3 Debris Disk Structure and Inferred Planetesimal Location 449
1.3.4 Planet-Debris Disk Relation 456
1.3.5 Debris Disk Composition 458
1.4 Future Prospects in Debris Disks Studies 461

2 Part II: Physical Processes Acting on Dust 462
2.1 Radiation and Stellar Wind Forces 463
2.1.1 Radiation Pressure 463
2.1.2 Poynting-Robertson Drag 464
2.1.3 Stellar Wind Forces 465
2.1.4 Effect of Radiation Forces on the Dust Dynamics 465
2.1.5 Effect of Radiation Forces on the Dust Spatial Distribution 467
2.2 Gravitational Forces in the Presence of Planets 467
2.2.1 Resonant Perturbations 469
2.2.2 Gravitational Scattering 470
2.2.3 Secular Perturbations 470
2.2.4 Effect of Gravitational Forces on the Dust Spatial Distribution 471
2.3 Collisions 474
2.3.1 Collisional Lifetimes 474
2.3.2 Effect of Collisions on the Dust Size Distribution 475
2.3.3 Effect of Collisions on the Dust Spatial Distribution 478
2.3.4 Effect of Collisions on the Dust Disk Evolution 479
2.4 Other Physical Processes 480
2.4.1 Dust Sublimation 480

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_9, © Springer Science+Business Media Dordrecht 2013

2.4.2 Lorentz Force 481
2.4.3 Sputtering 481
2.5 Open Questions 482

References ***483***

Extensive photometric stellar surveys show that many main-sequence stars show emission at infrared and longer wavelengths that is in excess of the stellar photosphere; this emission is thought to arise from circumstellar dust. The presence of dust disks is confirmed by spatially resolved imaging at infrared to millimeter wavelengths (tracing the dust thermal emission) and at optical to near-infrared wavelengths (tracing the dust scattered light). Because the expected lifetime of these dust particles is much shorter than the age of the stars ($>10^7$ year), it is inferred that this solid material not primordial, i.e., the remaining from the placental cloud of gas and dust where the star was born, but instead is replenished by dust-producing planetesimals. These planetesimals are analogous to the asteroids, comets, and Kuiper belt objects (KBOs) in our solar system that produce the interplanetary dust that gives rise to the zodiacal light (tracing the inner component of the solar system debris disk). The presence of these "debris disks" around stars with a wide range of masses, luminosities, and metallicities, with and without binary companions, is evidence that planetesimal formation is a robust process that can take place under a wide range of conditions.

This chapter is divided in two parts. *Part I* discusses how the study of the solar system debris disk and the study of debris disks around other stars can help us learn about the formation, evolution, and diversity of planetary systems by shedding light on the frequency and timing of planetesimal formation, the location and physical properties of the planetesimals, the presence of long-period planets, and the dynamical and collisional evolution of the system. It first describes the interplanetary dust in the inner and outer solar system and the evolution of dust production though the solar system's history, followed by a summary of the properties of debris disks around other stars (their frequency, evolution, spatial structure, and composition). *Part II* reviews the physical processes that affect dust particles in the gas-free environment of a debris disk, like the solar system's interplanetary space, and their effect on the dust particle size and spatial distribution; the discussion focuses on radiation and stellar wind forces, gravitational forces in the presence of planets and grain collisions.

1 Part I: Solar and Extrasolar Debris Disks

1.1 Debris Disks Are Evidence of the Presence of Extrasolar Planetesimals

Circumstellar disks play a fundamental role in the formation of stars and planets and the subsequent evolution of planetary systems. Pre-stellar disks form from the contraction and conservation of angular momentum of the densest regions of molecular clouds (consisting on gas and dust in a 100:1 mass ratio). The accretion of mass onto the forming star is regulated by mass and angular momentum transfer mechanisms within the disk and requires the ejection of material and angular momentum by bipolar outflows. With time, the mass reservoir of the cloud gets depleted, the disk begins to dissipate from the inside out, and the transfer of mass onto the star weaken. Observations indicate that by 3 Myr, half of the disks show inner cavities 10s of AU in size, and the signatures of mass accretion onto the star weakens. Simultaneously, planet formation takes place by accretion of dust particles into larger and larger bodies, resulting in a few massive cores and a swarm of embryos. Because the formation of giant planets requires a gas disk to provide material for the formation of a gaseous envelope around a massive core, it needs to happen before the protoplanetary gas disk dissipates in approximately 6 Myr.

These massive planets might be responsible for carving the large inner cavities inferred to be present in young disks. *Spitzer* Space Telescope surveys indicate that the majority of young stars (with stellar types later than B) are surrounded by gas-rich protoplanetary disks, indicating that most stars harbor the raw material required to form planetary systems. The formation of terrestrial planets and massive planets beyond the ice line is not limited by the presence of gas in the disk and continues for approximately 100 Myr; a critical step in this process is the formation of planetesimals.

Observations with *Spitzer* show that there is evidence that *at least* 15% of mature stars (10 Myr–10 Gyr) of a wide range of masses (0.5–3 M_{Sun}) harbor planetesimal belts with sizes of 10s–100s AU. This evidence comes from the presence of an infrared emission in excess of that expected from the stellar photosphere, thought to arise from a circumstellar dust disk. The reason why these dust disks are evidence of the presence of planetesimals is because the lifetime of the dust grains is of the order of 0.01–1 Myr (see ❯ Sects. 2.1.4 and ❯ 2.3.1), much shorter than the age of the star (>10 Myr); therefore, the origin of these dust grains cannot be primordial, i.e., from the cloud of gas and dust where the star was born, but must be the result of ongoing dust production generated by planetesimals, like the asteroids, comets, and Kuiper belt objects (KBOs) in our solar system (see ❯ *Figs. 9-1* and ❯ *9-2*). This is why these dust disks are known as *debris disks*.

Debris disks are therefore evidence of the formation of planetesimals around other stars. The study of this population of extrasolar planetesimals can give us a more complete picture of the diversity of planetary systems, shedding light on their formation and dynamical histories. Even though these planetesimals will remain undetected in the foreseeable future, the study of their debris dust can help us learn about some of the planetesimals' characteristics (e.g., location, composition, and evolution). Debris disks can also help us learn about the planet population.

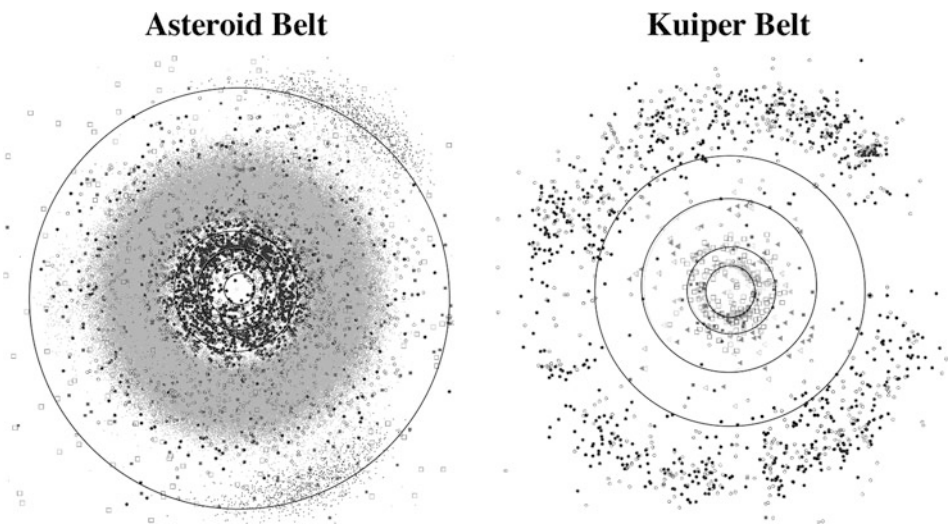


Fig. 9-1

Distribution of planetesimals in the solar system. The *circles* represent the orbits of the eight planets; the outermost circles correspond to the orbits of Jupiter (*left*) and Neptune (*right*) (Courtesy of G. Williams at the Minor Planet Center)

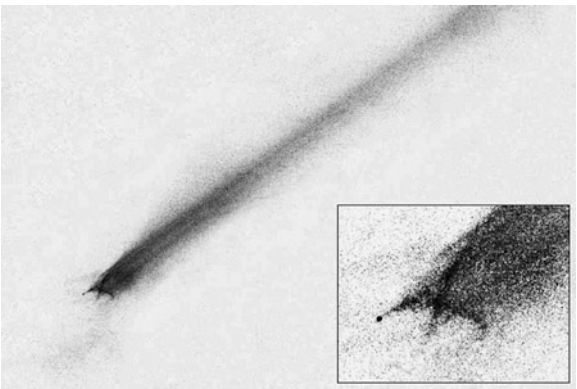


Fig. 9-2
Dust production by the inner belt asteroid P/2010A2 as seen by the *Hubble* Space Telescope Wide Field Camera 3. Its comet-like morphology is due to a tail of millimeter-sized dust particles emanating from a 120 m nucleus (Figure from Jewitt et al. (2010))

As will be discussed in Sects. 1.3.3 and 2.2, the structure of the debris disk is sensitive to planets with a wide range of masses and semimajor axes and is independent of the system's age (see *Fig. 9-3*). Therefore, the study of debris disk structure can serve as a planet-detection method, covering a parameter space complementary to that of radial velocity, transit, and direct imaging techniques. The goal of this chapter is to describe how debris disks can shed light on the formation, evolution, and diversity of planetary systems, helping us place our solar system into context.

1.2 The Solar System Debris Disk

The Sun also harbors a disk of dust produced in the inner and outer solar system by its population of minor bodies, like the asteroids, Kuiper belt objects (KBOs), and comets (*Fig. 9-1*).

1.2.1 Debris Dust in the Inner Solar System

Figure 9-2 shows an spectacular example of the production of dust by the asteroid P/2010A2 (from Jewitt et al. 2010). In this case, the dust is produced either by the rotational disruption of the asteroid (due to the spin-up produced by radiation torques) or by the collision with a meter-sized projectile, an event that must have taken place around February or March 2009. Long before the technological development allowed detailed images like this one to be obtained, there were some phenomena, visible to the naked eye and that must have been noticed since the beginning of humankind, that hinted that our planetary system was a dusty one: the zodiacal light, the "shooting stars," and the comet tails. Their analysis with current instrumentation sheds light on the origin and properties of the solar system's debris dust.

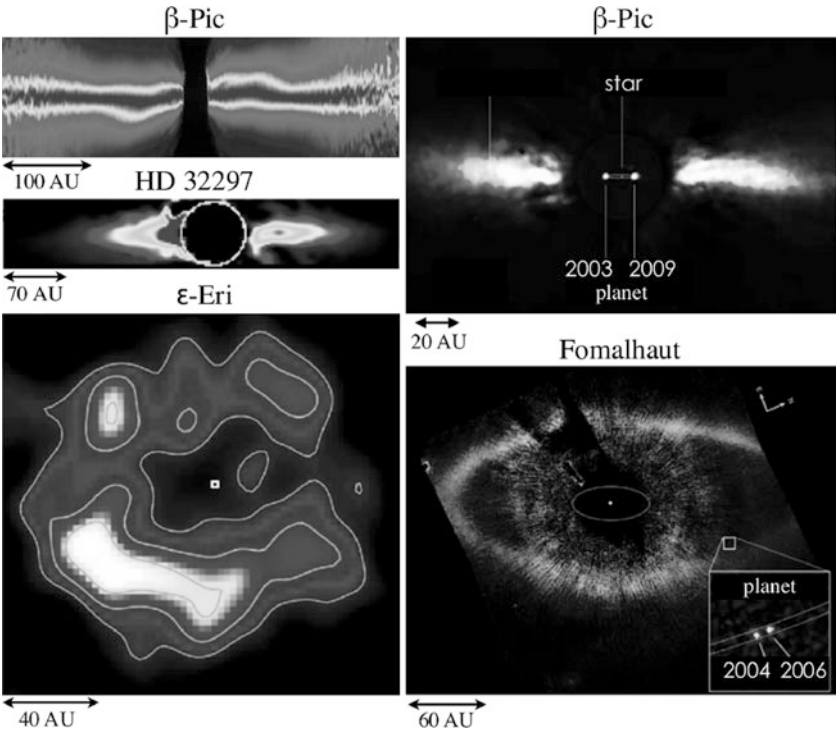


Fig. 9-3

(*Left*): Spatially resolved images of nearby debris disks showing dust emission from 10s to 100s of AU with a wide diversity of complex features including inner gaps, warps, brightness asymmetries, offsets, and clumply rings, some of which may be due to the presence of massive planets; from *top* to *bottom*: β-Pic (HST/STIS CCD coronography at 0.2–1 μm; Heap et al. 2000), HD 32297 (HST/NICMOS coronography at 1.1 μm; Schneider et al. 2005) and ϵ-Eri (JCMT/SCUBA at 850 μm; Greaves et al. 1998, 2005). (*Right*): Direct detection of planets in debris disk systems predicted to exist from the disk morphology; from *top* to *bottom*: β-Pic (VLT/NACO at 3.78 μm; Lagrange et al. 2010) and Fomalhaut (HST/ACS at 0.69–0.97 μm; Kalas et al. 2008)

Zodiacal Dust

The zodiacal light is a glow that appears along the ecliptic and can be seen about an hour before and after sunset (in the Eastern and Western horizons, respectively). The term "zodiacal" refers to its location along the ecliptic. It is caused by the scattering of solar photons by the dust grains in the inner solar system. It appeared in ancient Egyptian art represented by a triangle inclined with respect to the horizon (worshiped as the war god Sopdu), and it was also known to the Greek and the Romans, the Chaldeans, and the Aztecs (*Codex Telleriano-Remensis*). The first explicit description in Europe appeared in 1661 (Childrey's *Britannia Raconica*) and the first scientific observations were done by Cassini in 1683, and correctly interpreted by de Duiliers a year later as produced by sunlight reflected from small particles orbiting the Sun. Scattered light observations of the zodiacal light can help determine the properties of the dust particles, revealing nonspherical, irregular, or fluffy aggregates, 10–100 μm in size grain (see *Fig. 9-4*),

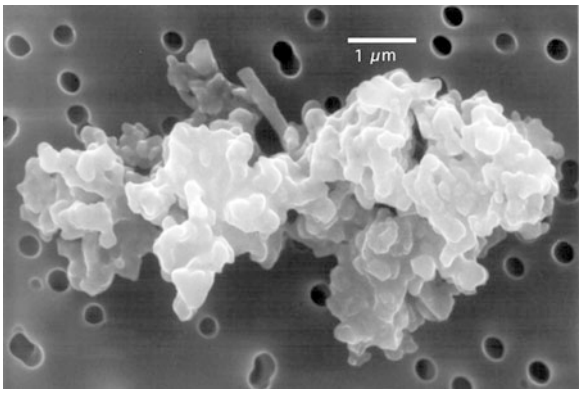


Fig. 9-4
Interplanetary dust particle (IDP) collected in the Earth's stratosphere (Figure from D. E. Brownlee (University of Washington) and E. Jessberger (Institut für Planetologie, Münster))

composed of a mixture of silicates and organic material with a low albedo of ~0.08. Regarding their spatial distribution, the observed brightness profile of the scattered light follows $r^{-2.3}$ from 0.3 to 1 AU, $r^{-2.5}$ out to 2.3 AU, and $r^{-2.37}$ out to the asteroid belt, corresponding to a number density distribution not too different from the r^{-1} dependency expected in the case of dust dynamics dominated by Poynting-Robertson drag (see Sect. 2.1.5). It is also observed that the plane of symmetry of the zodiacal light is warped, as a result of gravitational perturbations by the planets (see Sects. 2.2.3 and 2.2.4). The motion of the dust particles can be studied from the absorption lines in the scattered stellar light, revealing particles in elliptical prograde orbits that belong to two dust populations: one with a spherical distribution and a r^{-2} number density radial profile, likely due to dust particles produced by long-period comets, and another with a flattened low inclination distribution and a number density r^{-1} dependency, likely due to asteroids or short-period comets (see review in Levasseur-Regourd et al. 2001 and references therein).

The heating of the dust particles by the Sun makes them emit at infrared and submillimeter wavelengths; this thermal emission dominates the night sky between 5 and 500 μm and has been mapped by the *IRAS*, *COBE*, *ISO*, and *Spitzer* Space Telescopes. These observations indicate that the ratio of the dust-to-stellar luminosities, known as the fractional luminosity, is $f = L_{\rm dust}/L_{\odot} \sim 10^{-8}$–$10^{-7}$ (Dermott et al. 2002); this is more than two orders of magnitude fainter than the extrasolar debris disks observed with *Spitzer* (see *Fig. 9-8* – the difference being due to the limited sensitivity of the *Spitzer* observations). It is inferred that the particles dominating the thermal emission from the zodiacal cloud are rapidly rotating grains 10–100 μm in size, with low albedos, located near 1 AU, and with an amorphous forsterite/olivine composition; there is also a population of smaller 1 μm-sized grains made of crystalline olivine and hydrous silicate that accounts for a weak silicate emission feature at 10 μm (from the vibration of stretching Si-O bonds – Reach et al. 2003). The laboratory analysis of interplanetary dust particles (IPDs) collected in the Earth's stratosphere also reveals this mixture of amorphous and crystalline olivine and pyroxene that has also been identified in the mid-infrared spectra of extrasolar debris disks (see Sect. 1.3.5). Regarding the spatial distribution, the thermal emission from the zodiacal cloud shows long, narrow arcs that coincide with the perihelion passage

of some short-period comets (produced by low albedo, porous, millimeter-sized dust particles), and broader dust bands at low ecliptic latitudes, thought to originate from the breakup of the asteroids that gave rise to the Themis, Koronis, and Eos asteroidal families (Sykes and Greenberg 1986 and Dermott et al. 2002 argued that the formation of the Veritas family 8.3 Myr ago accounts for ~25% of the zodiacal thermal emission today). The gravitational perturbations exerted by the planets are also evident in the spatial distribution of the thermal emission: a ring of asteroidal dust particles is trapped in the exterior mean motion resonances (MMRs) with the Earth at around 1 AU, forming a ring-like structure with a 10% number density enhancement on the Earth's wake that results from the resonance geometry of the 1:1 MMR (Dermott et al. 1994b; Reach et al. 1995; see discussion in ❯ Sect. 2.2.4).

Dust Particles Falling on Earth

"Shooting stars" are evidence that dust particles fall on Earth. This phenomenon, that must have been noticed since the beginning of humankind, is produced by the ionization of atmospheric atoms along the path of the incoming high-velocity dust grain. The grain generally gets destroyed before it reaches the surface of the Earth because the impact of atmospheric molecules increases its temperature to ~1,000–2,000 K, at which point the grain's atoms and molecules begin to ablate; another destruction process is the loss of the volatile glue that maintains the aggregate together. The trails can be used to trace back the orbits of the incoming dust particles, revealing that most of the sporadic meteors are generated by particles on prograde, low eccentric orbits, with small relative velocities with respect to the Earth of a few km/s. Meteors also occur in showers when the Earth crosses the dusty trail of an asteroid or a comet, the latter produced by the gradual released of dust during ice sublimation or by the breakup of inactive comets. The origin of the dust can be traced back to the parent body because their orbital elements remain similar during the first 10^4 years (see review in McDonnell et al. 2001 and references therein).

Some of the dust grains survive atmospheric entry and can be collected on the Earth's surface (from deep sea sediments and Greenland and Antarctic ice) in the form of micrometeorites with sizes in the 20 μm–1 mm range (Maurette et al. 1994). Because of the effect of atmospheric entry, particles larger than 50 μm are strongly affected, and there is a selection effect against particles with high entry velocities, high densities, and/or fragile structure. Micrometeorites can be collected in the Antarctic ice, e.g., at the bottom of the water well of the Amundsen-Scott South Pole Station, where the particles accumulate as the ice melts during the drilling of the well; a ton of this ice contains approximately 100 cosmic spherules larger than 50 μm and about 500 micrometeorites in the 50–400 μm range (Maurette et al. 1994, 1996); the slope of the cumulative size distribution for particles larger than 200 μm is −5.2. The estimated particle flux, compared to that found in the upper atmosphere in a similar size range, indicates that only about 4% of the incoming dust grains survive atmospheric entry. Micrometeorites have also been collected in deep sea sediments (see ❯ Sect. 1.2.3). The fate of the organic material on the dust particles that enter the Earth's atmosphere is particularly interesting for astrobiology; this matter may survive entry if it is in the form of complex compounds. (For a review, see Jessberger et al. (2001) and references therein.)

Dust particles can also be collected from the Earth's stratosphere using aerogel collecting plates on high-altitude flying aircraft. These interplanetary dust particles (IDPs) are distinguished from high-altitude terrestrial dust (e.g., the dust produced by solid rockets) by their elemental and isotopic composition, and by the effect of their long journey in interplanetary space, the latter revealed by He atoms implanted by the solar wind in the bubbles, voids,

and crystal defects of the grains, and by the radiation damage due to the impacts with high-energy cosmic rays and stellar wind particles. The collected IDPs are on average 15 μm in size (with a size range of 5–25 μm determined by the collecting method and the terrestrial contaminants). *Figure 9-4* shows a typical IDP, formed by aggregates of thousands to millions of submicron grains, with about 40% porosity and a bulk density of about 2 g/cm^{-3} (this is consistent with the formation of the IDP as a random aggregate of similar-sized components). About 85% of the IDPs are formed by aggregates of different minerals (chondritic composition), while the rest are aggregates of just a few minerals (nonchondritic composition). Chondritic IDPs tend to be dark (with low albedos of 0.05–0.15, due to the presence of carbon, sulfides, and nanometer-sized metal grains), and their mineralogy is similar to that of the matrix of some carbonaceous chondrites and can be anhydrous (dominated by olivine and pyroxene) or hydrous; their structure ranges from porous (generally anhydrous silicates) to compact (generally hydrated silicates that have been subject to aqueous alteration). Because the chondritic IDPs are aggregates of thousands to millions of mineral grains, their overall elemental composition is similar to solar. The entry velocities of the IDPs can be estimated from the amount of solar wind-implanted He that was lost during atmospheric entry (due to the heating of the dust particles). There are two populations of IDPs: (1) Low-velocity grains (<14 km/s), likely associated with the low eccentricity and low inclination orbits of asteroidal dust; they tend to be compact, with average bulk densities of about 3 g/cm^3, and they commonly have a chondritic composition of low crystalline Fe–Mg hydrous silicates similar to that of carbonaceous chondrite meteorites. (2) High-velocity grains (>18 km/s), with velocities similar to those expected from cometary dust; they have low bulk densities of about 1 g/cm^3, and a composition is similar to that found in comets, e.g., Hale-Bopp, with both crystalline and noncrystalline Fe–Mg anhydrous silicate minerals (pyroxenes and olivines) and GEMS. (For a review, see Jessberger et al. (2001) and references therein.)

In Situ Dust Detections in the Inner Solar System

Another of the earliest observations of debris dust in the solar system are the sightings of comet tails. Records of comets appear as early as in Chinese oracle bones, and in many cultures, they have been considered bad omens. Comets are now treasured as the most pristine planetesimals in the solar system, harboring invaluable information about our planetary system's formation and evolution. Comets are one of the main sources of dust; because sublimation drives the cometary activity, it is assumed that their dust production generally takes place in the inner solar system. However, there are isolated flare-ups that also produce dust at large heliocentric distances. The dust production rate of comets is difficult to estimate because the cometary activity is not steady (e.g., comet Holmes had a massive mass loss in 2007 that made it become 10^6 times brighter in a day – Li et al. 2010). A study by Nesvorný et al. (2010) concluded that about 85% of the dust in the inner solar system is produced by Jupiter-family comets and <10% by long-period comets (a result that is model dependent).

Cometary dust particles have been studied in situ in the case of comets Halley, Tempel 1, and Wild 2 using instruments on-board the *VeGa* 1, *VeGa* 2, *Giotto*, *Deep Impact*, and *Stardust* spacecrafts. *Deep Impact* sent a copper-core impactor into comet Tempel 1 to study the ejected cloud of subsurface material (observed from ground-based telescopes and from the flyby section of the spacecraft). It found that the surface of the comet was more dusty than expected, depleted of ice that is present at depths of about 1 m (A'Hearn 2008); ground-based observations of the dust cloud released by the impact revealed the presence of forsterite (Mg_2SiO_4) and enstatite ($MgSiO_3$). Comet Wild 2 was the target of the *Stardust* mission that was able to return

cometary dust particles to Earth for detailed laboratory analysis. Their study reveals that even though Wild 2 is thought to have originated in the cold environment of the Kuiper belt, it contains aggregates where highly refractory and volatile components are found close together in the micron-scale, indicating that the comet has aggregated dust particles that have formed at a very wide range of heliocentric distances: there is pre-solar material identified from the isotopic evidence and that has suffered little alteration since it was accreted; there are large 5 μm Fe-poor forsterite crystals that are likely formed from the vaporization, melting, or heating of material above 1,000 K; there are calcium-aluminium inclusions (CAIs – the oldest solids of the Solar system) thought to have condensed at very high temperatures during a brief period of time and that must have been transported from the inner solar system (where CAIs formed) to the region where the comet formed, and there are also amorphous glassy grains, some of which have an interstellar origin. The returned samples contained a broad range of minerals, each of which has been found in primitive meteorites but never such a rich diversity in the same object. This heterogeneous mixture is far from being in chemical or mineralogical equilibrium (Brownlee et al. 2006; McKeegan et al. 2006; Zolensky et al. 2006a; Sandford et al. 2006).

Dust particles have also been detected in situ at different heliocentric distances by the spacecrafts *HEOS* (1 AU), *Hitten* (1 AU), *Helios* (0.3–1 AU), *Galileo* (0.7–5 AU), *Pioneer* 8 and 9 (0.75–1.08 AU), *Ulysses* (1.3–2.3 AU), *Cassini*, and *Pioneer* 10 and 11. Impact velocities >1 km/s result in the vaporization and ionization of material from the target and the impactor that is separated by an applied electric field and produces an electronic signal that depends on the mass, velocity, and chemical composition of the impacting grain. One of the challenges is to calibrate the detectors in the laboratory using impactors with characteristics similar to that of interplanetary particles. Another challenge is that these in situ measurements are limited to the trajectory of the spacecraft, so the interpretation of the impact data depends on assumptions regarding the dust spatial distribution. These measurements are also biased against particles at high inclinations because they spend little time near the ecliptic where most of the experiments take place (see review by Grün et al. 2001).

The current dust production rate in the inner solar system is of the order of 10^4 kg/s; the relative contribution of the different sources is still under debate and has likely changed with time. As was mentioned above, Nesvorný et al. (2010) argued that the splitting of Jupiter-family comets accounts for 85% of the dust in the inner solar system with <10% produced by Oort cloud long-period comets, while a previous study by Dermott et al. (1994b) concluded that the asteroids contribute to at least 33%; other contributors of dust to the inner solar system could include Halley-type comets and KBOs.

1.2.2 Debris Dust in the Outer Solar System

As discussed in Sect. 1.2.1, the dust in the inner solar system ("zodiacal dust") has been studied via remote observations of its scattered light and thermal emission. The situation is very different for the dust in the outer solar system, for which no remote detections have been achieved so far: at optical wavelengths, its emission is dwarfed by the much brighter zodiacal light in the foreground, while at infrared and longer wavelengths, its emission is also confounded by the thermal emission of the zodiacal dust in the foreground and emission from the galactic dust (known as the "cirrus") in the background. The cosmic microwave background radiation provides a very uniform source against which emission from the Kuiper belt (KB) might be detected in the future. The derived upper limit on the total mass of dust in the

outer solar system is about 10^{-5} $M_\oplus$ or 10^{20} kg, a thousand times the mass inferred from the *Voyager* 1 detection experiments (Backman et al. 1995; Jewitt and Luu 2000; Moro-Martín and Malhotra 2003). Dynamical modeling of the dust produced in the Kuiper belt, together with recent in situ detections by the *New Horizons* spacecraft, indicates that the fractional luminosity of the Kuiper belt dust, $L_{\rm dust}/L_{\rm Sun}$, is ~10^{-7}; this is below the detection limit *Herschel*/PACS, meaning that Kuiper belt dust disk analogs around other stars could not be detected with present-day instrumentation in space (Vitense et al. 2012).

Pioneer 10 and 11 detected dust out to 18 and 13 AU, respectively, with detectors that consisted on pressurized gas cells with metallic walls 25–50 μm thick that would lose pressure when penetrated by a projectile (with the problem that the usable area of the detector diminished with time). For the detectable mass range (10^{-12}–10^{-11} kg for impact velocities of 20 km/s), it was found that the flux was nearly constant between Jupiter and 18 AU (Humes 1980). Even though Pioneer 10 and 11 did not sample the KB region directly (because the former failed at 18 AU and the latter was turned off at 13 AU), dynamical models indicate and that the KB was likely the source of the dust detected beyond 10 AU (Landgraf et al. 2002). Therefore, the detection of Kuiper belt dust preceded the discovery of the Kuiper belt itself by Jewitt and Luu (1993), but at the time the origin of the dust went unrecognized. The dynamical models also require a significant contribution from comets (including short-period Oort cloud and Jupiter-family comets) to account for the flat number density distribution of the dust.

Voyager detected dust in the 30–60 AU Kuiper belt region with a number density of $n \sim 2 \times 10^{-8}$ m^{-3}. The data remains poorly calibrated because the detections were done indirectly by measuring the pulses in the conductivity of the medium adjacent to the spacecraft, caused by plasma generated from the vaporization of the impactor; it is thought that smallest dust particles detected were $s \sim 2$ μm in size (Gurnett et al. 1997). Because the size of the dust particles follows a power-law distribution (see Sect. 2.3.2), a reasonable approximation is that most impactors were of this minimum size. Adopting a KB vertical height of $H \sim 10$ AU, this would correspond to an optical depth of $\tau \sim \pi s^2 H n \sim 4 \times 10^{-7}$ (Jewitt and Luu 2000), about two orders of magnitude smaller than that of the extrasolar debris disks detected by *Spitzer* (*Fig. 9-8*, where the fractional luminosity can be approximated by the optical depth, $f = L_{\rm dust}/L_{\rm star} \sim \tau$ – see discussion in Sect. 2.3.1); this is comparable to the normal optical depth of the zodiacal dust, $\tau \sim 10^{-8}$–10^{-7} (see Sect. 1.2.1). Additional evidence of dust-producing collisions in the KB is the impact craters imaged by the *Deep Impact* mission on the surface of the comet Tempe1, thought to be created during the time the comet belonged to the KB.

The dust production rate estimates in the outer solar system are in the range $(0.2–5) \times 10^4$ kg/s (from Voyager and Pioneer data, respectively; Jewitt and Luu 2000; Landgraf et al. 2002). Dust production rates from theoretical models are in the range of $(0.1–1) \times 10^4$ kg/s, from the erosion of KBO surfaces by the flux of interstellar grains (Yamamoto and Mukai 1998), to $(1–300) \times 10^6$, from mutual grain-grain collisions (Stern 1996). For comparison, the dust production rate in the inner solar system is of the order of 10^4 kg/s (see Sect. 1.2.1).

Cassini also detected 17 dust impact events between the orbits of Jupiter and Saturn (however, the orientation of the spacecraft was not optimized to maximize the number of detections). The particles are inferred to have sizes in the submicron to micron range and are found on bound and unbound orbits. Particles on bound orbits have low eccentricities and low inclinations; the shape of the impact signals indicated that the grains are irregular and have high porosity, as expected from a cometary origin; particles on unbound orbits are inferred to have sizes of ~0.4 μm, in agreement with an interstellar origin (Altobelli et al. 2007).

1.2.3 Evolution of the Dust Production Rate in the Solar System

The dust production rate in the solar system has changed significantly with time. It is thought that the solar system was significantly more dusty in the past because the asteroid and the Kuiper belts were more densely populated. Evidence for a massive primordial KB is the existence of KBOs larger than 200 km, which formation by pairwise accretion must have required a number density of objects about two orders of magnitude higher than today. Evidence for a massive primordial asteroid belt (AB) comes from the minimum mass solar nebula, showing a strong depletion in the AB region unlikely to be primordial. The solar system then became progressively less dusty as the planetesimal belts eroded away by mutual planetesimal collisions. Evidence of collisional evolution comes from the modeling and observation of the size distribution of the asteroids and KBOs (see discussion in Sect. 2.3.2 and *Fig. 9-25*). This collisional evolution likely resulted in the production of large quantities of dust, as it can be seen in the left panel of *Fig. 9-5* from a model by Kenyon and Bromley (2005). They found that in a planetesimal belt, Pluto-sized bodies ~1,000 km in size excite the eccentricities of the more abundant 1–10 km sized planetesimals, triggering a collisional cascade that produces dust and changes the planetesimal size distribution. Because the dust production rate is proportional to the number of collisions, and this is proportional to the square of the number of planetesimals, as the planetesimals erode and grind down to dust, the dust production rate decreases and the expected thermal emission from the dust slowly decays with time as $1/t$ (see discussion Sect. 2.3.4). This decay is punctuated by large spikes that are due to large collisions happening stochastically

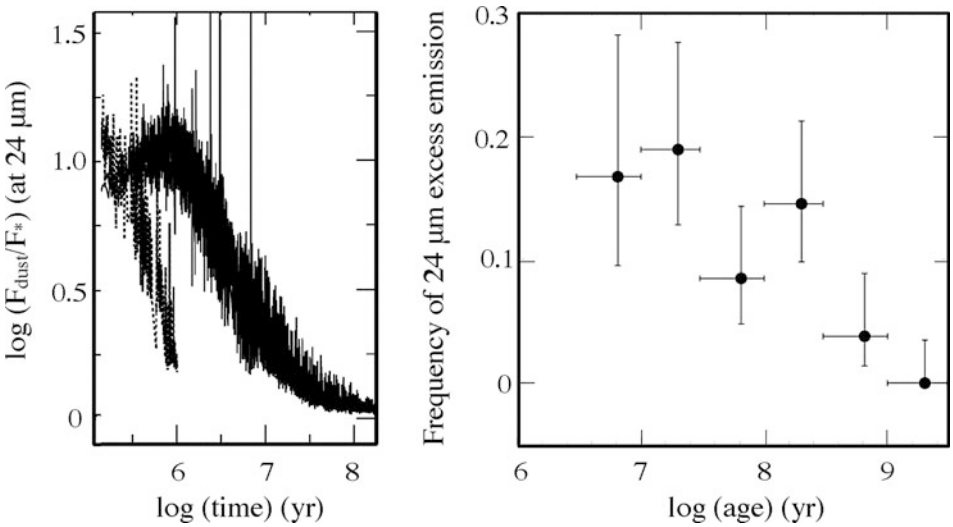


Fig. 9-5

(*Left*): Evolution of the 24 μm excess as a function of time for two planetesimal disks extending from 0.68 to 1.32 AU (*dashed line*) and 0.4 to 2 AU (*solid line*). The central star is solar type. Excess emission decreases as planetesimals grow into Mars-sized or larger objects and collisions become increasingly rare (From Kenyon and Bromley 2005). (*Right*): Fraction of stars in a sample of 309 FGK stars with detectable 24 μm excess plotted as a function of age. Each bin spans a factor of 3 in age. The vertical bars are Poisson errors (From Meyer et al. 2008)

(left panel of Fig. 9-5). Examples of stochastic events in the recent history of the solar system are the fragmentation of the asteroids giving rise to the asteroidal families and the dust bands.

A major change in the dust production rate is expected to have occurred in the early solar system at the time of the Late Heavy Bombardment (LHB), a period in which a large number of impact craters in the Moon and the terrestrial planets were created (with an impact rate at Earth approximately 20,000 × the current value). Because this heavy bombardment deleted all records of previous impacts, it is not clear whether the LHB was a single event, the tail of a very heavy bombardment process, or the last of a series of multiple cataclysm (but see Chapman et al. 2007). This event, dated from lunar samples of impact melt rocks, happened during a very narrow interval of time 3.8–4.1 Gyr ago (~600 Myr after the formation of the terrestrial planets). Thereafter, the impact rate decreased exponentially with a time constant ranging from 10 to 100 Myr (Chyba 1990). Strom et al. (2005) compared the impact cratering record and inferred crater size distribution on the Moon, Mars, Venus, and Mercury to the size distribution of different asteroidal populations and found that the LHB lasted ~20–200 Myr, that the source of the impactors was the main AB, and that the mechanism was size independent. The most likely scenario is that the orbital migration of the giant planets caused a resonance sweeping of the AB and as a result many of the asteroidal orbits became unstable, causing a large-scale ejection of bodies into planet-crossing orbits (explaining the observed cratering record); the orbital migration of the planets also caused a major depletion of the KB as Neptune migrated outward; it is estimated that ~90% of the KBOs were lost (Gomes et al. 2005). The LHB was probably a single event in the history of the solar system that would have been accompanied by a high rate of collisions and dust production (see Fig. 9-6). After the LHB, there must have been a sharp decrease in the dust production rate due to the drastic depletion of planetesimal (Booth et al. 2009).

Some record of the interplanetary dust flux falling on the Earth can be found in the sedimentary rocks at the sea floor. As mentioned in Sect. 1.2.1, stellar wind He atoms are implanted in the voids, bubbles, and crystal defects of interplanetary dust particles. A significant fraction

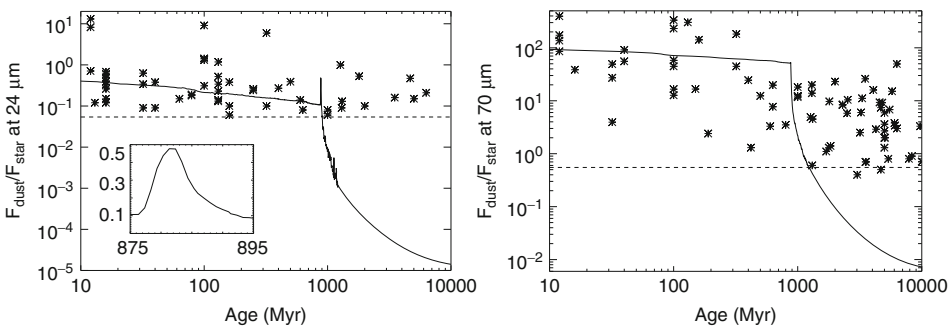


Fig. 9-6

Excess ratio (F_{dust}/F_{star}) versus time at 24 μm (*left*) and at 70 μm (*right*). The *asterisks* correspond to *Spitzer*/MIPS observations of FGK stars and the *dashed lines* are the observational limits. The *solid lines* correspond to a model of the dust production in the solar system, assuming blackbody grains with a power-law size distribution of $n(s)ds \propto s^{-3.5}ds$; the sharp decrease is due to the drastic planetesimal depletion that took place at the time of the Late Heavy Bombardment (From Booth et al. 2009)

of the He is lost during atmospheric entry as the particle heats up, but some He survives and under special conditions can remained trapped for millions of years in sedimentary rocks at the sea floor (Farley 1995). Extraterrestrial material is associated with high $^3He/^4He$ ratios, as opposed to that of terrestrial origin; the enhanced isotope ratio observed in some of the cores extracted from the sedimentary rock at the sea floor can be accounted for if approximately 0.5% of the total mass in interplanetary dust suffered little He loss upon atmospheric entry. The concentration of extraterrestrial material in the sedimentary rock depends on the flux of interplanetary dust particles and on the sedimentation rate; for pelagic clays, the concentration is large because they accumulate slowly. The record of interplanetary dust flux analyzed this way spans a timescale of 100 Myr (Farley 1995; see summary by Zolensky et al. 2006b), and because of the effects of atmospheric entry, it favors asteroidal dust in low eccentric orbits. It is found that: (1) The flux increased by a factor of 5 between 36.5 and 34 Myr ago, coinciding with deposition of Ir, shocked quatz and spinnel (associated with major impacts); the sedimentation rate at Earth is well known during that time, but the origin of this enhanced interplanetary dust flux, whether due to asteroidal collisions or to increased cometary activity, is still uncertain. (2) The flux increased by a factor of 4 between 8.2 ± 0.1 and 6.7 Myr ago; this is likely associated with the collisional cascade that resulted from the most recent asteroid breakup, the one that gave rise to the formation of the Veritas asteroidal family 8.3 ± 0.5 Myr ago, thought to be caused by the breakup of a 150 km size C-type asteroid rich in hydrated minerals; at that time, it probably constituted the dominant source of (water rich) interplanetary dust at Earth (Dermott et al. (2002) argued that still contributes to about 25% of the zodiacal dust today, but see Nesvorný et al. 2010.)

1.3 Extrasolar Debris Disks

The extrapolation of radial velocity studies indicate that ~17–19% of solar-type stars harbor giant planets within 20 AU (Marcy et al. 2005). A natural question arises whether stars also harbor planetesimals, thought to be the building blocks of planets. Long before extrasolar planets were discovered, it was inferred that the answer to this question was yes: dust-producing planetesimals had to be responsible for the infrared excesses observed around many mature stars (see discussion in ❯ Sect. 1.1). These dust disks, first discovered by *IRAS* (Aumann et al. 1984) and later studied by *ISO* (e.g., Habing et al. 2001; Decin et al. 2003), were extensively surveyed by *Spitzer*. The goal of the *Sptizer* surveys was to characterize the frequency and properties of debris disks around stars of different spectral types, ages, and environment; taking advantage of the unprecedented sensitivity of the *Spitzer/MIPS* (24 and 70 μm) and *Spitzer/IRS* (5–35 μm) instruments (Rieke et al. 2004; Houck et al. 2004), more than 700 stars were surveyed and hundreds of debris disks were identified. The following is a brief summary of the results from the *Spitzer* debris disk surveys.

1.3.1 Debris Disk Frequency

Most of the debris disks detected by *Spitzer* are found around mature main-sequence stars A to K2 type, with stellar luminosities ranging from 0.3 to 3 $L_\odot$. ❯ *Figure 9-7* summarizes the frequency of debris disks at two different wavelengths in a combined sample of 350 AFGK stars older than 600 Myr (from Trilling et al. 2008). From Wien's law, emission peaks at 24 and 70 μm

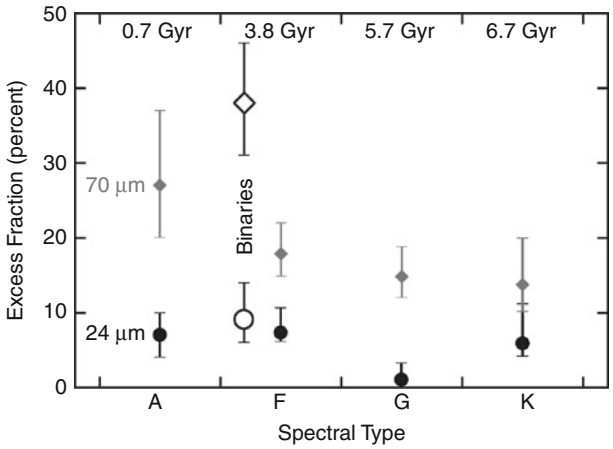


Fig. 9-7

The percentage of stars showing excess dust emission as a function of stellar type for ages >600 Myr (the mean ages within each type are shown at the *top*); the *vertical bars* correspond to 1-σ Gaussian errors; *solid black symbols* correspond to single stars with excess emission at 24 μm (tracing warmer dust), while the *solid grey diamonds* correspond to excesses at 70 μm (tracing colder dust); *empty symbols* are for binary systems (Figure from Trilling et al. (2008)). A different survey of 328 solar-type FGK stars (30 Myr–3 Gyr) found that the frequency of 24 μm excess is 14.7% at <300 Myr and 2% at >300 Myr, while at 70 μm, the excess rates are 6–10% and are fairly independent of age (Meyer et al. 2008; Hillenbrand et al. 2008; Carpenter et al. 2009)

would correspond to characteristic dust temperatures of 153 and 52 K, respectively; assuming 1 $L_{\odot}$, blackbody grains would adopt that temperature if located at 3 AU (24 μm) and 28 AU (70 μm), while a 10 μm grain in the intermediate size regime would adopt that temperature if located at 5 AU (24 μm) and 75 AU (70 μm – see (9.15) and (9.16)). *Figure 9-7* shows that debris disks are more common around A-type stars than around solar-type FGK stars (the younger age of the A-star sample may bias this result because, as discussed in Sect. 1.3.2, debris disks are common around young stars). It is also found that the frequency of debris disks is significantly higher for solar-type stars than for old M stars. There is a prominent debris disk around the M star AU Mic, but this star is relatively young (~12 Myr). In a survey of 64 old M dwarfs, none of the stars were found to have excesses (Gautier et al. 2007). This may be an observational bias because the peak emission of these colder disks would be at $\lambda > 70$ μm, i.e., beyond the wavelength where *Spitzer/MIPS* was most sensitive. Ongoing debris disk surveys with *Herschel* are increasing the number of disk detection rates made by *Spitzer*, and most of the new detected debris disks are found around cold late-type stars (Eiroa et al. 2011; Kennedy et al. in preparation; preliminary results from the *Herschel*/DEBRIS survey show debris disk frequencies of 26%, 24%, 19%, 9.5%, and 1.3% around spectra types A, F, G, K, and M, respectively – Kennedy et al. in preparation).

There is also evidence of the presence of planetesimals around white dwarfs. Some of these evolved stars show infrared excesses and high levels of pollutants (elements other than the expected pure H and He), thought to arise from tidally disrupted planetesimals. Section 1.3.5 discusses how the atmospheric abundances of these white dwarfs provide a unique opportunity

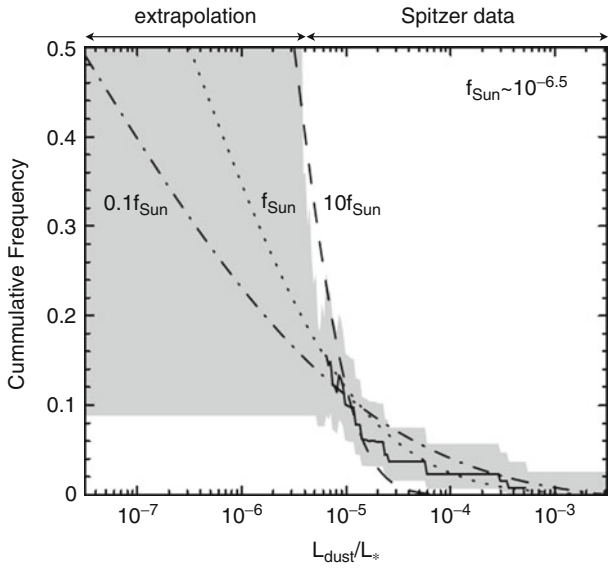


Fig. 9-8

Disk detection cumulative frequency as a function of dust fractional luminosity ($f = L_{dust}/L_*$). The *grey region* corresponds to Spitzer data (*right*) and an estimate based on extrapolation of these observations (*left*). The *lines* correspond to theoretical debris disk distributions assuming a Gaussian distribution of debris disk fractional luminosities (f) and considering an average fractional luminosity of $f_{ave} = f_{Sun} \sim 10^{-6.5}$, i.e., similar to that of the solar system's debris disk (*dotted line*); $f_{ave} = 10\, f_{Sun}$ (*dashed line*); $f_{ave} = 0.1\, f_{Sun}$ (*dot-dashed line*) (Figure from Bryden et al. (2006))

to study the elemental composition of the disrupted planetesimals. The presence of planetesimals around stars with a very wide range of spectral types, from M-type to the progenitors of white dwarfs – with several orders of magnitude difference in stellar luminosities – implies that planetesimal formation is a robust process that can take place under a wide range of conditions.

Figure 9-8 shows the disk detection frequency from *Spitzer* surveys, indicating that there is a steep increase with decreasing fractional luminosity ($f = L_{dust}/L_*$; from Bryden et al. 2006). Due to the limited sensitivity of the *Spitzer* debris disk surveys, the detected fractional luminosities are generally $f \gtrsim 10^{-5}$; this is larger than those inferred for the solar system's debris disk today : $f \sim 10^{-8} - 10^{-7}$ for the inner solar system and $f \sim 10^{-7} - 10^{-6}$ for the outer solar system (although the latter is only an estimate because its emission is overwhelmed by the zodiacal dust foreground – see Sects. 1.2.1 and 1.2.2). *Figure 9-8* compares the observations to theoretical debris disk distributions: assuming a Gaussian distribution of debris disk luminosities and extrapolating from *Spitzer* observations, Bryden et al. (2006) concluded the fractional luminosity of an average debris disk around a solar-type stars could be between 0.1 and 10 times that of the solar system debris disk. In other words, the observations are consistent with debris disks at the solar system level being common (but they would have been too faint to be detected by *Spitzer*). Ongoing debris disk surveys with *Herschel* (e.g., DUNES and DEBRIS – Eiroa et al. 2011; Kennedy et al. in preparation) are now probing the frequency of disk detections for fainter disks. However, recent estimates of the KB dust disk emission by

Vitense et al. (2012) indicate that, with a fractional luminosity of $f \sim 10^{-7}$ peaking at 40–50 μm, the dust emission of a KB dust disk analog would be less than 1% the stellar photosphere, still below the *Herschel*/PACS detection limits. This means that the detection of KB dust disk analogs still awaits more sensitive far-infrared space instrumentation (e.g., *SPICA*/SAFARI).

1.3.2 Debris Disk Evolution

The study of the frequency and properties of debris disks around stars of different ages can shed light on the evolution of debris disks with time. The main challenge in this case is that the ages of main-sequence stars, in particular those that are not in clusters, are difficult to determine. As mentioned in Sect. 1.2.3 and the left panel of *Fig. 9-5* shows, collisional models predict that the steady erosion of planetesimals will naturally lead to a decrease in the dust production rate; this slow decay will be punctuated by short-term episodes of increased activity triggered by large collisional events that can make the disk look an order of magnitude brighter (see the more detailed discussion in Sect. 2.3.4). These models agree broadly with the observations derived from the *Spitzer* surveys, as the ones shown in *Fig. 9-9* for A-type and FGK (solar-type) stars. It is found that the frequency and fractional luminosities ($f = L_{\rm dust}/L_*$) of debris disks around FGK stars with ages in the range 0.01–1 Gyr decline in a timescale of 100–400 Myr, but there is no clear evidence of a decline in the 1–10 Gyr age range (Trilling et al. 2008). This indicates that different physical processes might be dominating the evolution of the dust around the younger and the older systems. A possible scenario is that, at young ages, stochastic dust production due to individual collisions is more prominent, while at older ages, the steady grinding down of planetesimals dominates. The relative importance of these two processes is still under discussion.

The *Spizer* surveys also showed that the evolution of dust around both A-type and FGK stars proceeds differently in the inner and outer regions, with the warmer dust (dominating the emission at 24 μm) declines faster than the colder dust (seen at 70 μm). This indicates that the clearing of the disk in the inner regions is more efficient, as would be expected from the shorter dynamical timescales.

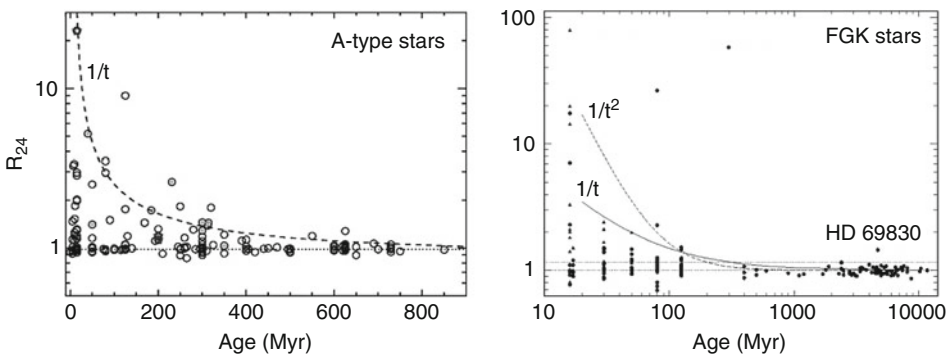


Fig. 9-9

Dust emission divided by the expected stellar emission at 24 μm as a function of stellar age, for A-type stars (*left*) and FGK stars (*right*) (From Su et al. (2006) and Siegler et al. (2007), respectively). The main features are a 1/t overall decay and a large variability for a given stellar age

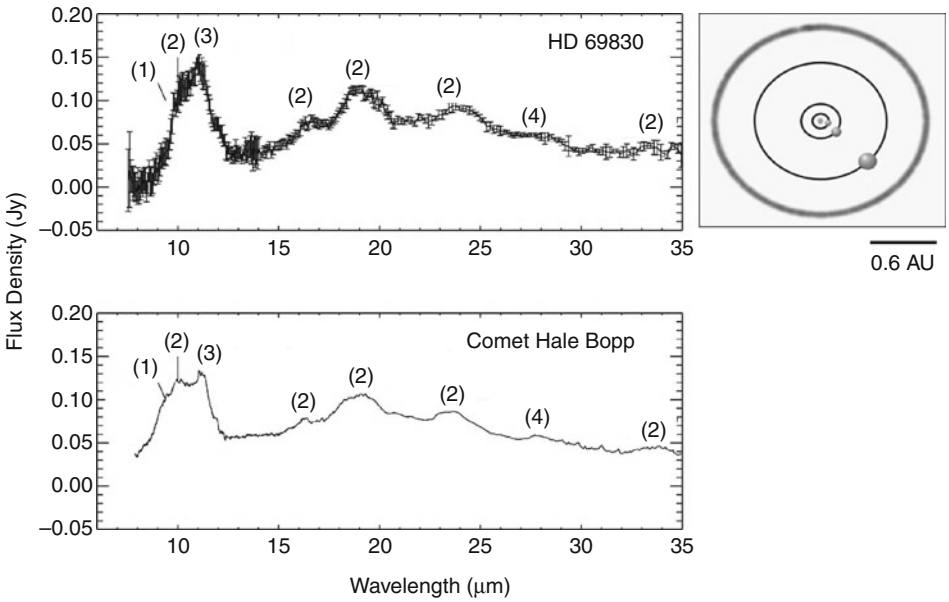


Fig. 9-10

(*Top left*): Spectrum of the dust emission around HD 69830. (*Bottom left*): Spectrum of comet Hale-Bopp normalized to a blackbody temperature of 400 K. The spectral features labeled are (*1*) amorphous olivine, (*2*) crystalline olivine, (*3*) crystalline olivine (forsterite), and (*4*) crystalline pyroxene (From Beichman et al. 2005). (*Right*): Further analysis showed that the best fit is to a highly processed low carbon P- or D-type asteroid belt located at ~1 AU, outside the orbit of the three Neptune-like planets in the system (Lisse et al. 2007); the planets are located at 0.0785 AU (⩾10.2 $M_\oplus$), 0.186 AU (⩾11.8 $M_\oplus$), and 0.63 AU (⩾18.1 $M_\oplus$)

Regarding the issue of steady state versus stochastic dust production, some systems show evidence that transient events dominate the dust production. This is the case of HD 69830, a star that shows no excess emission at 70 μm, but shows strong excess at 24 μm (HD 69830 is one of the outliers in the right panel of *Fig. 9-9*), with prominent spectral features in the *Spizter/IRS* wavelength range (see *Fig. 9-10* – Beichman et al. 2005). The spectral features are indicative of the presence of large quantities of small warm grains. Because these small grains have very short lifetimes (see discussion in Sect. 2.3.1), it is inferred that the level of dust production is too high to be sustained for the age of the system (because the planetesimals would have not survived the inferred erosion rate). These led to Wyatt et al. (2007) conclude that the high rate of dust production in HD 69830, as well as in a few other systems, is transient. HD 69830 is particularly interesting because it harbors three Neptune-like planets inside 1 AU (Lovis et al. 2006) and the dust is inferred to be located near the 2:1 and 5:2 MMRs of the outermost planet, so there is the possibility that this transient event is triggered by gravitational perturbations from the planets. The planet-debris disk connection will be discussed in Sect. 1.3.4.

The duration of the dust production events (expected to be short if stochastic collisions dominate, and long otherwise) is critical to estimate what percentage of stars show evidence of dust production throughout their lives. And this is an important question to address because

terrestrial planet formation is expected to result in the production of large quantities of dust in these regions (observable at 24 μm – see the left panel of *Fig. 9-5*), so the percentage of stars showing excess emission at these wavelengths can shed light on the frequency of planet formation. The right panel of *Fig. 9-5* shows the evolution of the 24 μm emission as a function of stellar age. If the dust-producing events are very long-lived, the stars that show dust excesses in one age bin will also show dust excesses at later times, and this may result in that <20% of the FGK stars in this survey show evidence of planetesimal formation near the terrestrial planet region. On the other hand, if the dust-producing events are shorter than the age bins in the figure, the stars showing excesses in one age bin are not the same as the stars showing excesses at other age bins, and this might result in that >60% of these stars show evidence of planetesimal formation (assuming that each star only has one epoch of high dust production). An additional caveat is that most of these observations are spatially unresolved, and therefore it is not evident where the 24 μm emission is coming from in the disk; there is the possibility that the steady erosion of planetesimals in the KB-like region could be contributing to the 24 μm excess emission, in which case the interpretation in terms of the percentage of stars showing evidence of planetesimal formation near the terrestrial planet region would change. Surveys of spatially resolved disks will help clarify this issue.

As it was discussed in Sect. 1.2.3, there is evidence that the migration of the giant planets in the early solar system had an important effect on the evolution of its debris disk: the drastic planetesimal clearing that took place at the time of the LHB, thought to be triggered by the migration of the planets, would have been associated with a sharp decreased in the dust production rate in both the AB and the KB (see solid line in *Fig. 9-6*). Because the extrapolation of radial velocity studies indicate that ~17–19% of solar-type stars harbor giant planets within 20 AU (Marcy et al. 2005), and the presence of hot Jupiters and planets locked in resonances are evidence that planet migration has taken place in some of these planetary systems, a natural question to ask is whether these drastic planetesimal clearing events are common in other system. The interest is that frequency and the timing of these LHB-type of events can have important implications for the habitability of these systems. *Figure 9-6* shows together the expected evolution of dust in the solar system at two different wavelengths, and the *Spitzer* debris disks observations. A statistical study by Booth et al. (2009) concluded that less than 12% of solar-type star suffered drastic planetesimal clearing events similar to that during the LHB. This is a preliminary result because, as explained above, the *Spitzer* surveys were limited in sensitivity, so this issue needs to be revisited in the future using deeper surveys. In any case, this result might not be surprising because there is no evidence of a positive correlation between the presence of debris disks and the presence of giant planets (Greaves et al. 2004; Moro-Martín et al. 2007a; Bryden et al. 2009; Kóspál et al. 2009), so there is no reason to expect that the debris disks observed so far should show evidence of planetesimal depletion due to planet migration. The planet-disk correlation will be discussed in Sect. 1.3.4.

1.3.3 Debris Disk Structure and Inferred Planetesimal Location

Because most of the debris disks observed so far are located at distances >10 pc from the Sun, the limited spatial resolution of the instruments leaves them spatially unresolved. Before the launch of the *Herschel* infrared telescope (3.5 m in diameter, compared to 0.85 cm for *Spitzer*), only a few dozen of the closest and the largest debris disks (out of the hundreds known) were spatially resolved. The sizes of these disks range from 10s of AU up to 1,000 AU and imply that

the dust-producing planetesimals are located on spatial scales similar to the KBOs in our solar system (extending out to ~50 AU for the classical KB and out to ~1,000 AU for the scattered KB). The location of the dust in the disks that are spatially unresolved can still be constrained or inferred. Because different wavelengths trace different dust temperatures and distances from the central star, the study of the debris disk SED can shed light on the radial distribution of the dust: for example, a disk with a central cavity where the warmer dust is absent would have an SED showing a depletion in the mid-IR wavelength region. In fact, the SEDs of the unresolved debris disks show evidence of central cavities, with characteristic dust temperatures in the range of 50–150 K, corresponding to dust located in the 10s–100 AU range. This result may be biased because of the limited sensitivity and wavelength coverage of the *Spitzer* observations. Ongoing debris disk surveys with *Herschel* are quickly increasing the number of resolved disks and/or providing better constraints for their outer radii with the help of more sensitive longer wavelength observations.

Inner Gaps

From a *Spitzer* survey of 328 FGK stars at 24 and 70 μm, it was found that about two third of the debris disk SEDs could be fitted with a single temperature blackbody ($T < 45$–85 K) consistent with a ring-like configuration, while the rest would require either multiple rings or a continuous distribution of dust out to tens of AU (Hillenbrand et al. 2008). Detailed analysis of the excess spectra from 12 to 35 μm of 44 of these stars showed that the characteristic dust temperature in these disks ranges from 60 to 180 K, that a cold component is needed to account for the 70 μm excess, and that inner disk cavities are common; the inner radii of these cavities are ~40 AU for the disks with 70 μm excess and ~10 AU for the disks without 70 μm excess (see *Fig. 9-11* from Carpenter et al. 2009). Because these dust disks are in a regime where the dynamics of the dust particles are mostly controlled by collisions (see discussion in Sect. 2.3.3), and therefore the dust traces the location of the planetesimals, these results indicate that most of the planetesimals inferred to exist around mature FGK (solar-type) stars are KB like (in the sense that they have large inner cavities – the inner radius of the KB is ~35 AU).

Inner cavities are also common around more massive stars: a *Spitzer* survey of 52 A-type and late B-type stars known to have debris disks showed that the majority of the disks (39/52) can be fitted with a single-temperature blackbody with a median temperature of 190 K, corresponding to a characteristic distance of 10 AU, while the rest (13/52) are better fitted by extended disks without cavities (Morales et al. 2009).

The presence of inner cavities has been confirmed by spatially resolved observations of nearby debris disks in both scattered light, as, e.g., in HR 4796A, Fomalhaut and HD 139664, and in thermal millimeter and submillimeter emission, as, e.g., in ϵ-Eri, Vega, and η Corvi (with inner cavities of 50, 80, and 100 AU, respectively).

Degeneracy of the SED Analysis

The analysis of the debris disk SEDs in terms of the dust location depends on assumptions on how efficiently the grains absorb and reemit the stellar radiation, because it is this balance that determines their equilibrium temperature. This in turn depends on the grain size and composition, which ideally can be constrained through the modeling of the solid-state features in the debris disk spectra. The issue is that most debris disks observed so far do not have any features in the 5–35 μm wavelength range covered by *Spitzer/IRS* (see in Sect. 1.3.5) where they show a smooth blackbody continuum (Chen et al. 2006; Beichman et al. 2006; Carpenter et al. 2009; Morales et al. 2009). It is generally assumed that this is because the grains have sizes

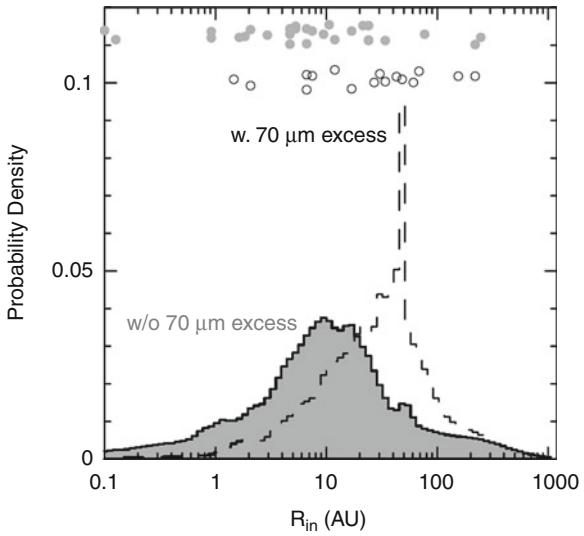


Fig. 9-11

Probability distribution for disk inner radii based on the analysis of the *Spitzer*/IRS spectra (12–35 μm) of 44 debris disks around FGK stars. The *dashed* and *grey histograms* correspond to sources with and without 70 μm excess, respectively (the best fit parameters are the *open* and *grey circles* on *top*). Typical disk inner radii are ~40 AU and ~10 AU for disks with and without 70 μm excess, respectively, indicating that most of the debris disks observed are KB like (From Carpenter et al. 2009)

≳10 μm (a grain size commonly adopted in the SED analysis). Regarding the grain composition, a common assumption is that they are made of "astronomical silicates" (i.e., silicates with optical constants from Weingartner and Draine 2001). However, the laboratory analysis of dust particles from the solar system (IDPs and *Stardust* returned samples – Sect. 1.2.1) and the analysis of debris disks with spectral features, like HD 69830 (*Fig. 9-10* and Sect. 1.3.5), indicate that this assumption might be too simplistic. Another caveat is that the debris disk SEDs are generally not constrained at the long wavelength range and most of them have only upper limits beyond 70 μm. As a result, the cold dust remains undetected and the outer disk radii unconstrained. Ongoing observations with *Herschel*, with increased sensitivity at longer wavelengths, are now contributing to characterize the cold dust. In fact, *Herschel* has detected a "new class" of very cold (~20 K) and faint ($f \sim 10^{-6}$) debris disks that only show dust emission beyond 100 or 160 μm (Eiroa et al. 2011).

Figure 9-12 illustrates the degeneracy in the SED analysis of the debris disk around the planet-bearing star HD 82943. The top left and center panels show the observed and modeled SEDs, while the allowed parameter space is shown right below. If the system is known to harbor planets (like in this case), an additional constraint on the dust location can be obtained from dynamical simulations that study the effect of the planetary perturbations on the stability of the planetesimals' orbits and that can identify the regions where the planetesimals could be stable and long-lived (see right panels of *Fig. 9-12*). *Figure 9-21* shows a similar dynamical analysis used to constrain the dust location in the HD 38529 planetary system. In this

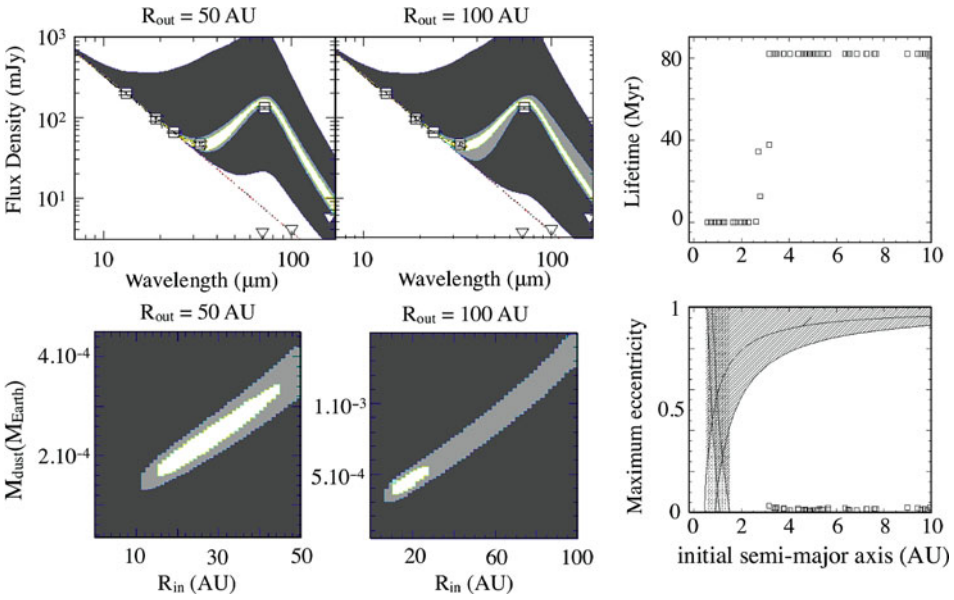


Fig. 9-12

(*Top left and center*): Observed and modeled SEDs for HD 82943. The *dotted line* is the stellar photosphere. The *Spitzer* observations are represented by *squares* with 1-σ error bars. The *triangles* are Herschel 5-σ/1 h sensitivity limits. The *colored regions* are formed by SED models of the star + disk emission, where the disk is composed of 10 μm size particles with optical properties typical of astronomical silicates, a total dust mass M_{dust} and extends from R_{in} to R_{out} with a constant surface density. R_{in} and M_{dust} are the free parameters; R_{out} is kept fixed at 50 AU (*left*) and 100 AU (*center*). The *colors* represent the goodness of the fit: for the *white inner region*, χ^2 probability $P(\chi^2 \mid \nu) < 0.683$; *light grey* for $P(\chi^2 \mid \nu) > 0.683$; and *dark grey* for $P(\chi^2 \mid \nu) > 0.9973$, i.e., models that are excluded with 3-σ certainty. (*Bottom left and center*): Parameter space of the modeled SEDs in the *top panels* showing the degeneracy of the SED analysis. In this case, the SED can rule out the presence of small grains, so these models assume a single grain radius of 10 μm. Adopting a disk outer radius of 50 AU, the best SED fits require the inner disk radius to be 16 AU ≤ Rin ≤ 44 AU, while if adopting a disk outer radius of 100 AU, the best fit will be for 12 AU ≤ Rin ≤ 26 AU. (*Right*): Results from the dynamical simulation (lasting 82 Myr) of 500 test particles in the HD 82943 planetary system, where the planets b and c have masses of 1.46 and 1.73 M_{Jup}, semimajor axes of 0.75 and 1.19 AU, and eccentricities of 0.45 and 0.27, respectively. (*Top right*): test particle's lifetimes. (*Bottom right*): allowed parameter space for the planetesimals' orbital elements, where the *shaded areas* indicate regions where test particle's orbits are unstable due to planet-crossing (*striped area*) or overlapping first-order mean motion resonances (*dotted area*); the *squared symbols* show the maximum eccentricity attained by test particles on initially circular orbits. The test particle orbits are stable beyond ~3 AU, with maximum eccentricities always <0.1. Long-lived, dust-producing planetesimals could therefore be located anywhere beyond 3 AU (From Moro-Martín et al. 2010)

case, it is found that the effect of the secular perturbations is very long-ranged (extending to 55 AU compared to the 3.74 AU semimajor axis of the outermost planet) and constrains the dust-producing planetesimals to the 20–50 AU region (resembling the KB).

Figure 9-13 illustrates how the study of debris disks can help us learn about the diversity of planetary systems. The figure shows the possible planet and planetesimal configurations of

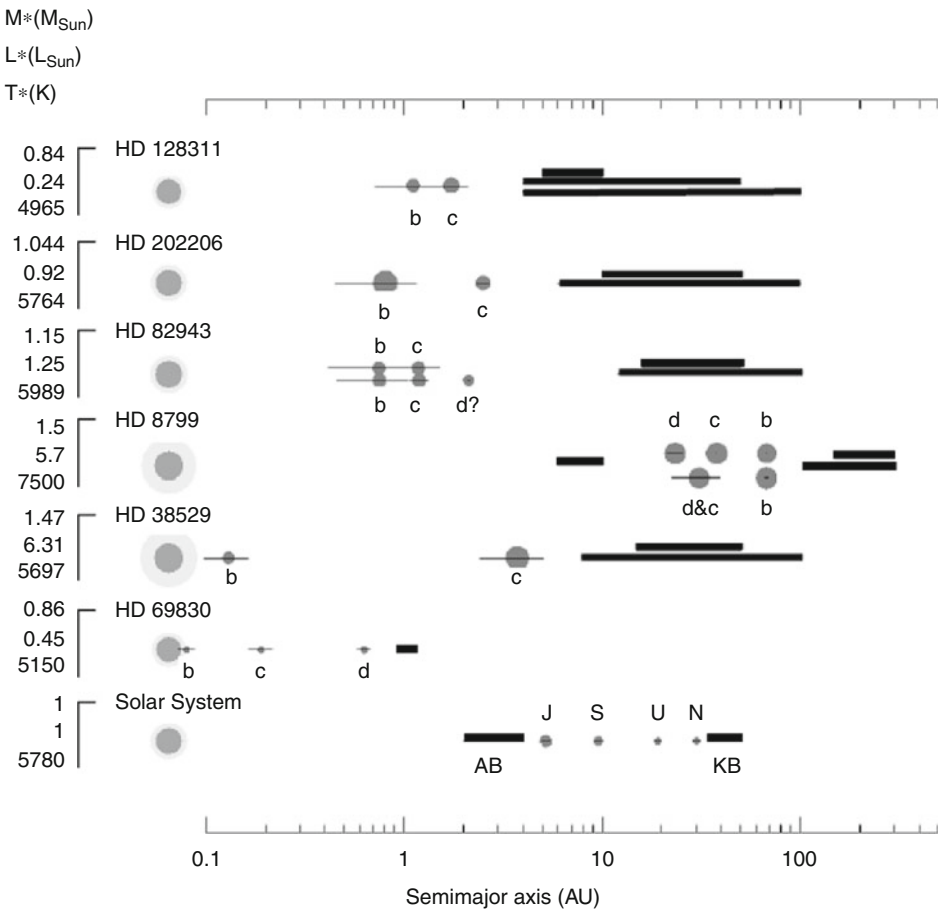


Fig. 9-13

Schematic representation of seven planetary systems known to harbor multiple planets and dust-producing planetesimals. The stellar mass, luminosity, and effective temperature are labeled to the *left*. The sizes of the *dark grey circles* are proportional to the cube root of the stellar and planetary masses, while the sizes of the *light grey circles* are proportional to the stellar luminosities. The *thin lines* extend from periastron to apoastron. For a given system, there is a range of planetary configurations that can fit the observations . The inferred location of the dust-producing planetesimals is represented by the *thick black lines*. Each line corresponds to a possible solution, showing the degeneracy of the problem and the need for spatially resolved observations (From Moro-Martín et al. 2010)

the systems known to date to harbor both multiple planets and debris dust (from Moro-Martín et al. 2010). For most of the stars, the study is based on a combined SED and dynamical analysis (similar to that in ❯ *Fig. 9-12*). The solutions are degenerate; to set tighter constraints on the planetesimal location, there is the need to obtain spatially resolved images and/or accurate photometric points from the mid-infrared to the submillimeter, so the inner and outer radius of the disk can be better determined. Observations with *Herschel*, *JWST*, and *ALMA* will be very valuable for this purpose.

Other Structural Features Revealed by Spatially Resolved Observations

The few dozen debris disks that have been spatially resolved so far show a rich diversity of structural features (❯ *Fig. 9-3*); Kalas et al. (2006) identified two basic architectures, either narrow belts about 20–30 AU wide and with well-defined outer boundaries (e.g., HR 4796A, Fomalhaut, and HD 139664), or wide belts with sensitivity limited edges implying widths >50 AU (HD 32297, β-Pic, AU Mic, HD 107146, and HD 53143). Structural features include clumpy rings (like in AU Mic, β-Pic, ε-Eri, and Fomalhaut), sharp inner edges (Fomalhaut), brightness asymmetries (β-Pic, AU Mic, HR 4796, HD 32297, Fomalhaut, and Vega), offsets of the dust disk center with respect to the central star (Fomalhaut and ε-Eri, with offsets of 15 AU and 6.6–16.6 AU, respectively), warps of the disk plane (like β-Pic and AU Mic), and spirals (HD 141569 – Wyatt et al. 1999; Heap et al. 2000; Clampin et al. 2003; Holland et al. 2003; Stapelfeldt et al. 2004; Kalas et al. 2005; Greaves et al. 2005; Schneider et al. 2005; Krist et al. 2005). Some of these features have also been observed in the solar system debris disk: the zodiacal cloud shows a warp in its plane of symmetry and an asymmetric ring near the Earth's orbit, and dynamical models of the KB dust disk predict an inner cavity around 10 AU, and an asymmetric ring outside the orbit of Neptune (see ❯ *Fig. 9-19* and discussion in ❯ Sects. 1.2.1 and ❯ 2.2).

These spatially resolved observations show that the disks look different at different wavelengths because at a given wavelength, the thermal emission is dominated by a particular grain size, and grains of different sizes have different dynamical evolutions that result in different structural features. As it is discussed in more detail in *Part II*, large particles that dominate the emission at longer wavelengths show more structural features because they interact more weakly with the stellar radiation field and therefore their dynamical evolution is slow; in this case, the particles trace the location of their parent planetesimals, or if planets are present, they can also be subject to resonant trapping. On the other hand, small grains that dominate the emission at shorter wavelengths interact more strongly with radiation, which results in a more uniform and extended disk. This can clearly be observed in the case of Vega (❯ *Fig. 9-14*). ❯ Section 1.3.3 discussed the need for spatially resolved observations to break the degeneracy in the debris disk SED analysis; the above discussion indicates that these resolved observations need to be taken at multiwavelengths.

Debris Disk Structure Can Unveil the Presence of Planets

Secular perturbations by Saturn are responsible for the creation of the inner edge of the asteroid belt around 2 AU, while other secular perturbations are thought to account for the offset of the zodiacal cloud center with respect to the Sun, the inclination of the cloud with respect to the ecliptic and the cloud warp; in addition, the trapping of dust particles in MMRs with the Earth is responsible for the asymmetric ring of dust along the Earth's orbit (❯ Sect. 1.2.1). In light of these observations, a natural question arises: are the structural features observed in debris disks the result of gravitational perturbations with unseen planets? Clumpy rings have been explained as dust and/or dust-producing planetesimals trapped in MMRs with a planet;

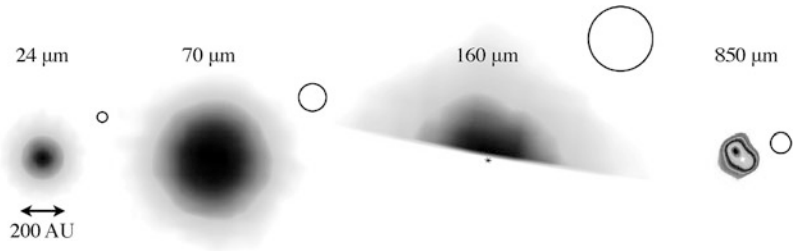


Fig. 9-14

Spatially resolved images of Vega from Spitzer/MIPS at 24, 70, and 160 μm (Su et al. 2005) and from JCMT/SCUBA at 850 μm (Holland et al. 1998). All images are in the same scale. The instrument beam sizes (shown in *white circles*) indicate that the wide radial extent of the MIPS disk images compared to the SCUBA disk image is not a consequence of the instrumental PSF but due to a different spatial location of the particles traced by the two instruments. This indicates the need for spatially resolved observations at different wavelengths. The submillimeter emission is thought to arise from large dust particles on bound orbits that originate from a planetesimal belt analogous to the KB, while the MIPS emission is thought to correspond to smaller particles on unbound orbits produced by collisions in the planetesimal belt traced by the submillimeter observations; these particles are blown away by radiation pressure to distances much larger than the location of the parent bodies; this scenario would explain not only the wider extent of the MIPS disk but also its uniform distribution, in contrast with the clumply and more compact submillimeter disk (Su et al. 2005; cf. to Müller et al. 2010 that interpret the observations with a collisional cascade in steady state)

warps can be the result of secular perturbations of a planet in an orbit inclined with respect to the planetesimal/dust disk; and spirals, offsets, and brightness asymmetries might also be the result of secular perturbations, in this case of an eccentric planet that forces an eccentricity on the planetesimals and the dust. Even though the origin of individual features is still under discussion,[1] and the models require further refinements – e.g., in the inclusion of dust collisional processes (Stark and Kuchner 2009) and the effects of gas drag – the complexity of these features, in particular the azimuthal asymmetries, indicates that planets likely play a role in the creation of structure in the debris disks.

The Fomalhaut case illustrates this idea: a massive planet was predicted to exist to account for both the sharp inner edge observed in the dust disk (created by gravitational scattering) and the offset of the dust disk center (created by secular perturbations – Wyatt et al. 1999; Kalas et al. 2005; Quillen 2006). Later on, follow-up observations were able to directly image a planet candidate around the predicted semimajor axis, where the planet eccentricity could be constrained

[1]It is possible that some of these disk features are created by mechanism other than planets. For example, clumps could trace the location of a recent planetesimal collision, instead of the location of dust-producing planetesimals or dust particles trapped in MMRs with a planet. Azimuthal asymmetries and spirals could be created by binary companions or close stellar flybys, but the later involves fine tuning to create the perturbations without destroying the disk. Brightness asymmetries on the outermost edge of the disk could also be created by sandblasting from interstellar grains. The interaction of the dust with remnant gas, stellar wind, or magnetic fields could also be responsible for some of the structure. Even though it is possible that some of the disk features are created by mechanism other than planets, it seems unlikely that non-planet mechanism can account for all the debris disk structures observed.

from the offset of the dust disk (Kalas et al. 2008). Because the eccentricity of the dust particles at the edge of the chaotic region where the MMRs overlap depends on the planet mass, the radial distribution of the dust near the inner edge of the dust disk can set limits on the mass of the planet (Chiang et al. 2009). This is particularly interesting because in some systems, a dynamical constraint for the planet mass can be compared to the mass estimate based on the observed luminosity, with the advantage that the former is independent of the planet age (which in most cases is difficult to estimate) and the initial conditions of the planet evolutionary model. In these systems, a dynamical constraint can be used to test and calibrate current evolutionary models of giant planets; this is important because, in most cases, dynamical constrains are not available and the planet masses need to be derived from evolutionary models alone.

Another example of a planet successfully predicted to exist based on the debris disk structure is β-Pic b (Mouillet et al. 1997; Lagrange et al. 2010): with an estimated mass of ~9 $M_{\rm Jup}$, this planet located at 8–14 AU can account for some of the asymmetries observed in the debris disk, including its inner warp. This system is particularly relevant because the planet has a relative short orbit that will allow to constrain the mass with the radial velocity technique, enabling the much needed calibration of the planet evolutionary models at young ages (~10 Myr).

The connection between planets and inner cavities is under debate. The gravitational scattering of dust particles by planets is a very efficient process that can create a dust-depleted region inside the orbit of the planet (see discussion in ❯ Sects. 2.2.2 and ❯ 2.2.4 and ❯ *Figs. 9-19*, ❯ *9-20*, ❯ *9-22*). But some of the observed inner cavities might be the result of grain-grain collisions rather than gravitational scattering with a planet; however, this latter scenario assumes that the parent bodies have an inner edge to their spatial distribution which may require planets to be present to keep the planetesimals confined. Core accretion models of planet formation predict the formation of inner cavities because the planets form faster closer to the star, depleting planetesimals from the inner disk regions.

Most of the structural features discussed in ❯ Sect. 2.2 depend on the mass and orbit of the planet. The case of the solar system illustrates that the structure is sensitive to small planets (like the Earth) and to planets located far from the star (like Neptune). This opens the possibility of using the study of the dust disk structure as a detection technique for planets of a wide range of masses and semimajor axes. This method is complementary to radial velocity and transit surveys (limited to planets relatively close to the star) and to direct imaging (limited to young and massive planets).

1.3.4 Planet-Debris Disk Relation

In the core accretion models of planet formation, planetesimals are the building blocks of planets; in these models, giant planet formation requires the presence of a protoplanetary disk rich in planetesimals, which would favor a positive correlation between the presence of giant planets and debris disks. However, as it was discussed in ❯ Sect. 1.2.3, the migration of giant planets can lead to drastic planetesimal clearing events, as the one expected to have occurred during the LHB in the early solar system; this would favor an anticorrelation between giant planets and debris disks.

Spitzer debris disk surveys do not find any sign of positive or negative correlation between the presence of giant planets identified in radial velocity surveys and the presence of debris disks

(Moro-Martín et al. 2007a; Bryden et al. 2009; Kóspál et al. 2009), i.e., there is no apparent difference between the incidence rate of debris disks around stars with and without known planetary companions.

A scenario that could account for this lack of correlation is the following. *Spitzer* observations are sensitivity limited and can only detect the brighter disks; based on the observed distribution of fractional luminosities, it seems likely that debris disks at the solar system level are very common, i.e., that many stars harbor planetesimals (see discussion in Sect. 1.3.1). Debris disks are found around stars with a wide range of spectra types, indicating that the planetesimal formation process is very robust and can take place under a wide range of conditions. This is also in agreement with the observation that the presence of debris disks is not correlated with high stellar metallicities (Greaves et al. 2006). Giant planets, on the other hand, are strongly correlated with high stellar metallicities (because their formation may require the presence of a large surface density of solids in the disk, so that the planet can grow a core sufficiently large to accrete an atmosphere before the gas disk disappears – Fischer and Valenti 2005). All this indicates that the conditions required to form planetesimals are more easily met than those to form giant planets, planetesimals are more common, and massive planets may not be required to produce the debris dust. In a possible scenario for the production of debris at Gyr ages, even in a disk that is too low in solids to form a giant planet, a large 1,000 km size planetesimals can stir up smaller planetesimals (0.1–10 km in size) along their orbits, starting a collisional cascade that can produce dust excess emission over the relevant range of ages. Results from numerical models exploring this scenario are shown in *Fig. 9-15*.

The planet-debris disk relation will be revisited with the *Herschel* DEBRIS and DUNES surveys, sensitive to lower-mass debris disks and to disks around colder stars.

The increasing population of super-Earths discovered by the radial velocity surveys will soon allow to test whether there is a correlation between the presence of low-mass planets ($<10\ M_{Earth}$) and the presence of debris disks. Contrarily to high-mass planets, recent results from the radial velocity surveys indicate that there is no correlation between high metallicities and low-mass planets (Mayor et al. 2011). This might indicate that the conditions to form

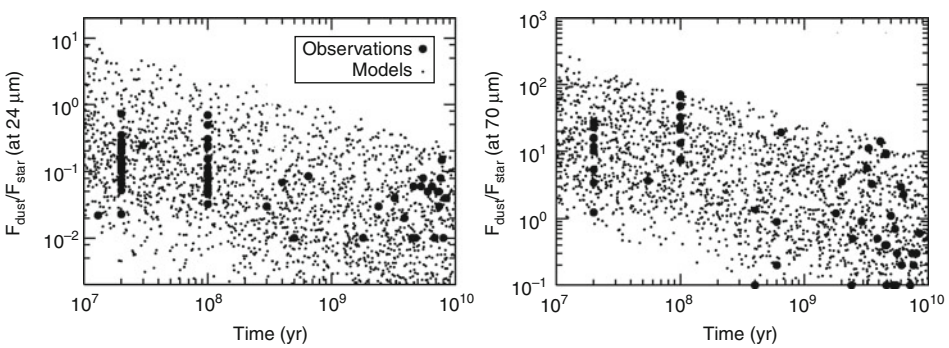


Fig. 9-15

Ratio of the excess dust flux to the stellar flux at 24 μm (*left*) and at 70 μm (*right*) as a function of stellar age. The collisional cascade models (*small dots*) can reproduce the range of ratios observed by *Spitzer* (*thick dots*) by varying the disk location and disk masses. The models predict an overall decay of the excess ratio with time (Figure from Löhne et al. (2008))

low-mass planets are more easily met than those to form high-mass planets and, therefore, there could be a correlation between low-mass planets and debris disks. At the time of the *Spitzer* planet-debris disk correlation studies, little was known about the frequency of low-mass planets and only a handful of these objects were known. Now that many more low-mass planet systems have been identified, and given that the *Herschel* surveys include low-mass planet host stars, it has became possible to explore whether a low-mass planet-debris correlation exists, as hinted by the metallicity studies.

Of particular interest is whether the presence of terrestrial planets might be correlated with debris dust. It is well known that the future detection and characterization of terrestrial planets in extrasolar planetary systems might be compromised by the emission from their inner debris disks (the equivalent to the zodiacal cloud). This means that stars with evidence of warm dust are not good candidates for terrestrial planet searches. On the other hand, a study by Raymond et al. (2011, 2012) indicates that stars with evidence of cold dust might turn out to be good targets. This study is based on numerical simulations of the dynamical evolution of about 400 different planetary systems, each consisting on three giant planets, two belts of planetesimals (inner and outer), and a population of Mars-sized embryos in the inner belt (that can grow into terrestrial planets). The simulations are able to reproduce the observed distribution of giant planet eccentricities. The ensemble of models by Raymond et al. (2011, 2012) show that (1) 40–70% of the systems that become unstable destroy their terrestrial planets; in these cases, the giant planet instability is too strong and the embryos are thrown into the star or ejected from the system. (2) The terrestrial planet outcome correlates with the eccentricity of the surviving giant planets; higher eccentricities imply that the instability was more violent and therefore the terrestrial planets were likely destroyed, while lower eccentricities indicate that more terrestrial planets survive. And (3) there is a strong correlation between the presence of bright cold dust and the occurrence of terrestrial planets (*Fig. 9-16*; Raymond et al. 2011, 2012); cold debris disks trace terrestrial planets because dusty systems mean a calm dynamical evolution where terrestrial planets are able to grow and survive.[2]

Finally, the recent release of the all-sky near- to mid-infrared survey carried out by the *WISE* space telescope, together with ground-based observations, may soon allow to assess whether there is a correlation between the presence of planets and the presence of warm dust (that may hint dynamical activity in the terrestrial planet region).

1.3.5 Debris Disk Composition

Spitzer/IRS carried out spectral surveys of debris disks around star from A to K type in the 5–35 μm wavelength range. Even though silicate emission features should be prominent around 10–20 μm, the spectra of most debris disks do not show any features, so little is known about the composition of the dust. In total, about two dozen disks show spectral features in this wavelength range; they are generally young stars (<50 Myr), and the majority of them show evidence of warm crystalline silicates and multiple planetesimal belts (with a warm component responsible for the spectral features, and a cold component that accounts for the emission at longer

[2] In this context, the solar system is an outlier because it harbors four terrestrial planets but very little dust. The explanation might be that the instability in the solar system was sufficiently strong to clear out the planetesimals, but sufficiently weak not to affect the stability of the terrestrial planets.

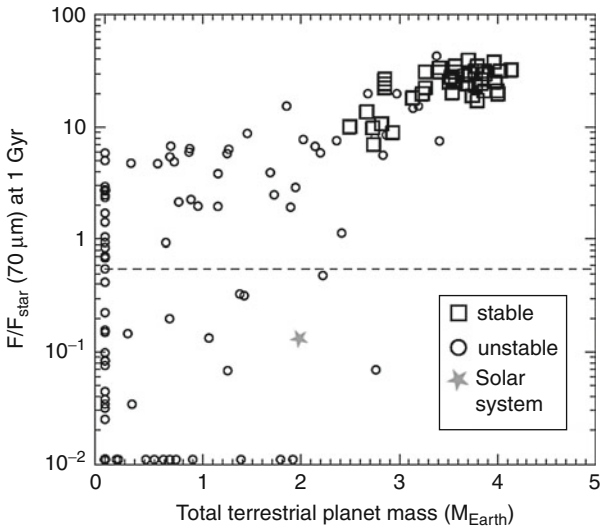


Fig. 9-16

The dust-to-stellar flux ratio at 70 μm versus the total mass in terrestrial planets from 400 dynamical simulations that include three giant planets, two belts of planetesimals (inner and outer), and a population of Mars-sized embryos in the inner belt (that can grow into terrestrial planets). The *circles* correspond to unstable planetary systems, the *squares* to stable systems, and the *grey star* is an estimate of the solar system's debris disk flux approximately 900 Myr after the LHB. The *dashed line* is the *Spitzer* detection limit. The figure shows the correlation between bright cold dust and the efficiency of terrestrial planet formation and survival: debris disks trace terrestrial planets because dusty systems mean a calm dynamical evolution where terrestrial planets are able to grow and survive (From Raymond et al. 2011)

wavelengths – Chen et al. 2006). It is found that for both FGK and A-type stars, there is a diversity of compositions and degree of crystallinity (related to the processing history) for stars of similar spectral types, even those of similar ages.

The lack of spectral features in the majority of the disks indicates that the dust grains are either large and/or cold ($T < 110$ K). Grains with sizes $s > \lambda/2\pi$ tend to absorb and emit energy efficiently with a constant emissivity (i.e., with no spectral features), so the lack of features at $\lambda < 35$ μm implies grain sizes $s > 5.5$ μm. For debris disks around A-type stars, radiation pressure can account for the lack of grain smaller than this size because the median blowout size for the stars in the A star sample is ~4.7 μm (see discussion on the blowout size in Sect. 2.1.4). For solar-type stars, the blowout size is ~1 μm, so small grains are expected to be present and should produce spectral features if they are warm. In fact, the two systems found with excess emission at $\lambda < 25$ μm, i.e., with warm dust, show spectral features at 7–20 μm.

One of these two systems is HD 69830 (*Fig. 9-10*). This mature K0 star, several Gyr old, harbors three close-in Neptune-like planets and shows no excess emission at 70 μm (i.e., no evidence of cold dust), a strong excess at 24 μm (corresponding to 1,000 × the emission of the zodiacal cloud), and its spectra are dominated by strong silicate features that can be fitted with 80% Mg-rich olivines (25% of which is amorphous) and 20% crystalline pyroxenes

(Beichman et al. 2005). At first sight, the spectrum is very similar to that of comet Hale-Bopp, but further analysis reveals it differs from cometary spectra (because there is no evidence for water gas, amorphous carbon, amorphous pyroxene, PAHs, phyllosilicates, and metallic sulfides); the best fit is to the dust produced by a highly processed low carbon P- or D-type asteroid (Lisse et al. 2007). These asteroids are the most common in the outer asteroid belt, and the breakup of these type of objects gave rise to the Veritas and Karin asteroidal families in events that should have been important sources of dust (see discussion in Sect. 1.2.1). The dust production in HD 69830 is also thought to be transient because the rate that would be needed to account for the observed amount of dust is too high to be sustain for the age of the star (Wyatt et al. 2007 – see discussion in Sect. 1.3.2). It is likely that the strong silicate features observed are associated with collision/disruption events, and in the case of HD 69830, the spectra reveal that the composition of the parent body was similar to that of an asteroid. It is inferred that the dust around this planet-bearing star is located at 0.93–1.16 AU, near the 2:1 and 5:2 MMRs of the outermost Neptune-like planet, raising the question whether the increased level of dust production is the result of gravitational perturbations by the planets.

The "colors" of the debris disks (obtained from scattered-light images taken at different wavelengths) can also provide some information about their composition. Debris disks tend to be red or neutral, with their redness commonly explained by the presence of small (~0.4 μm) silicate grains. However, spatially resolved spectra have shown that debris disks do not generally contain large amounts of small silicate grains, in which case it must be the composition of the grains that makes them look intrinsically red. The red color of the debris disks around HR 4796A, an A0V star 8 Myr old, has been identified as a signature of tholins, the complex organic materials found in the surface of icy bodies and in the atmosphere of Titan; however, other fits to the data consist in porous grains made of amorphous silicates, amorphous carbon, and water ice (Debes et al. 2008). Higher resolution spectroscopy (spatially resolved) is required to further constrain the models.

The study of the atmospheric composition of some white dwarfs provides a unique opportunity to probe the elemental composition of planetesimals in these systems. About one fifth of white dwarfs expected to have pure hydrogen or pure helium atmospheres (due to quick sedimentation of the heavy elements under the effect of the strong gravitational field) show evidence for heavier elements, likely due to pollution from external sources. In the case of white dwarfs that show atmospheric pollutants and near-infrared excesses, where the latter is inferred to be produced by dust located inside the tidal radius of the star, it is likely that the source of the pollutants is this circumstellar dust. This is favored over an interstellar origin based on several observations: the spectra of some of these infrared excesses show a strong 10 μm silicate feature with a shape that resembles that observed in the zodiacal cloud and differs from the characteristic interstellar dust emission, and an observed depletion of carbon relative to iron in the atmospheric composition of the white dwarf also suggests that the infalling material is asteroid like rather than interstellar (Jura 2006). Detailed observations of the atmospheric composition of one of the white dwarfs (GD 362) reveal that the relative enhancement of refractory elements and the relative depletion of volatiles are similar to that found in the Earth (Jura et al. 2007). All this favors a scenario in which tidally disrupted[3] asteroid-like bodies are responsible for the near-infrared excess emission and the atmospheric pollution. The advantage of using

[3]During the stellar evolution leading to the white dwarf stage, there is a significant mass loss from the star. As a result, the semimajor axes of orbiting bodies increase, and this may lead to dynamical instabilities caused by resonance sweeping and other effects that may send asteroid-like bodies into the tidal radius of the star.

this method to assess the composition of these asteroid-like bodies is that, since the material has already been broken into its elemental composition, the results do not depend on unknown dust grain properties.

1.4 Future Prospects in Debris Disks Studies

The discovery of debris disks in 1984, a decade before the detection of extrasolar planets around main-sequence stars, provided the first evidence that a critical step in the process of planet formation (the formation of planetesimals) is taking place around other stars. Since then, our knowledge of debris disks has greatly improved, and this chapter has described how it has shed light on the formation, evolution, and diversity of planetary systems. Debris disk observations with *Herschel* (ongoing), with upcoming observatories like *ALMA* and *JWST* (under development), and with future missions like *SPICA* (proposed), together with new developments in planet-detection techniques, warrant that the field of debris disk studies will keep developing rapidly, enabled not only by the improved sensitivity and spatial and spectral resolution of the observations but also by the interest of the astronomical community. The latter is reflected in the Astro2010 "Decadal Survey"[4] by the National Academies (2010–2010) and the "Cosmic Visions" by ESA (2015–2025), where questions intimately related to debris disks have been identified as priorities for the next decade, namely, how do circumstellar disks form and evolve into planetary systems? What is the diversity of planetary systems? How does it depend on stellar properties? How does the solar system fit in the context of other planetary systems? Is the solar system unique in its formation, characteristics, and/or evolution? What is the composition of primitive planetesimals in the solar system? Is there a radial gradient? What is the frequency of stars with terrestrial planets? Which stars are the best candidates for planet detection? Do they harbor debris disks bright enough to impede planet detection and characterization? To advance in answering some of these question, the Astro2010 "Decadal Survey" made the following specific recommendations related to debris disks: (1) *To carry out debris disk surveys around stars of different ages and spectral types*; the goals are the characterization of the disk properties as a function of the stellar properties, the determination of the necessary conditions for the formation of planetesimals, and the study of the temporal evolution of the systems. (2) *Study of debris disk structure with high spatial resolution observations* in scattered light (optical–near infrared), in thermal emission (mid-infrared–submillimeter), and at different epochs (to allow the detection of proper motions and to exclude features that might be background galaxies); the goals are to identify morphological features that can reveal the presence of planets (allowing to constrain their mass, eccentricity, and period) and can shed light on the dynamical evolution of the system. (3) *Development of theoretical debris disk models* (including planet-disk interactions and collisions), with the goal of interpreting high spatial resolution observations. (4) *Direct detection of planets in protoplanetary and debris disks*, with the goal of studying the planet-disk interaction. (5) *To carry out KBOs surveys* to constrain their size distribution, physical properties (which depend on the size of the body), and their collisional state, with the goal of shedding light on the formation and dynamical evolution of the solar system.

Examples of these studies are the ongoing *Hershel* surveys DUNES and DEBRIS designed to search for debris disks around AFGKM stars at 70, 100, and 160 μm, with follow-up at 350, 450, and 500 μm; these observations are already allowing to characterize a new population of cold

[4]New Worlds, New Horizons in Astronomy and Astrophysics (2010–2020).

disks (Eiroa et al. 2011) and to increase the number of spatially resolved observations (Booth et al. in preparation). *ALMA*'s unprecedented high spatial resolution will be able to advance in the study of debris disk structure and to test the models of planet-disk interactions; its long wavelength observations will allow to better constrain the disks outer radii. Debris disk surveys with *JWST* in the near- to mid-infrared will allow to characterize the warm dust component, setting constrains on the frequency of planetesimal formation in the terrestrial planet region, and identifying stars with low debris dust contamination that may be good targets for terrestrial planet detection. Deep debris disk surveys with *JWST* and *SPICA* (the approval of the latter is pending) will be able to study the debris disk evolution and the dust production rate as a function of stellar age; this will help identify systems undergoing LHB-type of events and to assess whether the dynamical evolution of the solar system was particularly benign (in the sense that it did not affect the orbital stability of the terrestrial planets and it happened early during the solar system evolution). The SAFARI instrument planned for *SPICA* (a proposed telescope similar to *Herschel* but cool down to 5 K allowing a greatly improved sensitivity), covering the wavelength range of 30–210 μm, has debris disks at one of its focus. In fact, three out of its eight proposed core programs are related to debris disk studies, namely, (1) the study of the occurrence and mass of debris disks (carrying out an unbiased survey of all stellar types out to a few hundred pc), (2) the study of the dust mineralogy in debris disks (doing an spectral survey of debris disks and, for nearby disks, spectral imaging), and (3) the study of the composition of the KB (with an unbiased survey of KBOs). Of special interest is that, for the nearby debris disks, it will be possible to trace the variation in the dust mineral content as a function of disk radius that can be compared to the compositional gradient in the solar system. Regarding the solar system, advancements need to be made in the study of the dust-producing planetesimals, e.g., with KBO surveys with *Pan-STARRS*, *LSST*, and *Subaru-HSC*, and in the study of the solar system's dust (which properties in the outer solar system are greatly unknown). For the latter, there are several space missions proposed, including sample return missions to an asteroid and a comet, dust detection experiments, and the study of the dust scattered light using a small telescope on board a spacecraft traveling into the outer solar system. Finally, there are programs to detect planets in stars harboring debris disks using ground-based telescopes (e.g., *Subaru/HiCIAO*, *Gemini/GPI*, and *VLT/SPHERE*) that will allow to study the planet-disk interaction.

These are some of the future research lines in debris disk studies that will help us understand our solar system in the context of the wide diversity of planetary systems: debris disks allow us to "see worlds in grains of sand"[5].

2 Part II: Physical Processes Acting on Dust

Part II reviews the dominant physical processes acting on dust particles and their effect of the particle size and spatial distribution, focussing on radiation and stellar wind forces (Sect. 2.1), gravitational forces in the presence of planets (Sect. 2.2), and collisions (Sect. 2.3). The discussion applies the gas-free environment of the solar system's interplanetary space and extrasolar debris disks. There are many reviews on this topic, e.g., Mukai et al. (2001), Gustafson et al. (2001), Dermott et al. (2001), Wyatt (2008a), and Krivov (2010).

[5] "To see the world in a grain of sand, and to see heaven in a wild flower, hold infinity in the palm of your hands, and eternity in an hou" (William Blake).

2.1 Radiation and Stellar Wind Forces

2.1.1 Radiation Pressure

If the circumstellar dust particle is at rest, the radiation pressure force exerted by the stellar photons on the particle is given by $F_{\rm rad} = \frac{S}{h\nu}\frac{h\nu}{c}Q_{\rm pr}A$, where $\frac{S}{h\nu}$ is the flux of incoming photons, $\frac{h\nu}{c}$ is the momentum per photon, and $Q_{\rm pr}A$ is the particle cross-section for radiation pressure; A is the particle geometric cross-section ($A = \pi s^2$) and $Q_{\rm pr}$ is the dimensionless radiation pressure factor averaged over the stellar spectrum; $Q_{\rm pr}$ is a function of the grain optical properties, size, shape, and chemical composition and a measure of the fractional amount of energy scattered and/or absorbed by the grain. Substituting the energy flux density $S = \frac{L_*}{4\pi r^2}$, one gets that $F_{\rm rad} = \frac{L_* Q_{\rm pr} s^2}{4r^2 c}$, where L_* is the stellar luminosity, s is the particle radius, and r is the heliocentric distance (Burns et al. 1979). Because the radiation pressure force has the same dependency on r as the gravitational force, $F_{\rm grav} = \frac{GM_* m}{r^2}$, it is useful to define the dimensionless parameter β, given by

$$\beta = \frac{F_{\rm rad}}{F_{\rm grav}} = \left(\frac{3L_*}{16\pi G M_* c}\right)\left(\frac{Q_{\rm pr}}{\rho s}\right), \tag{9.1}$$

where $Q_{\rm pr}$ is the radiation pressure factor averaged over the stellar spectrum.

For interplanetary dust particles in the solar system, $\beta = 5.7 \cdot 10^{-5}\left(\frac{Q_{\rm pr}}{\rho s}\right)$, where ρ is the grain density in $\mathrm{g\,cm^{-3}}$ and s is the grain radius in cm (Burns et al. 1979). ❯ *Figure 9-17* and ❯ *Table 9-1* show β as a function of the particle radius for a range of representative dust compositions and assuming spherical grains.

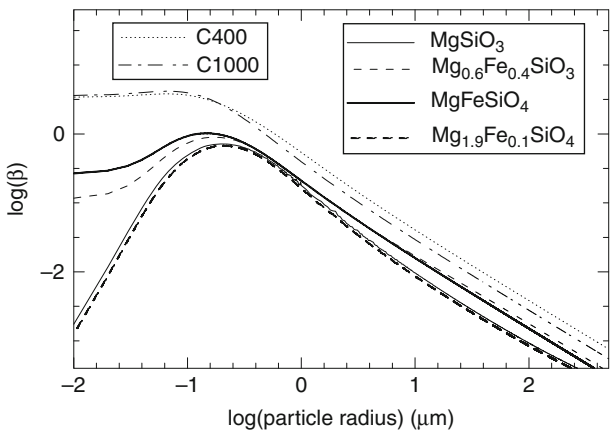


Fig. 9-17

Correspondence between $\beta = \frac{F_{\rm rad}}{F_{\rm grav}}$ and particle size when the central star is solar type, under the assumption of spherical grains and for the following chemical compositions (based on spectroscopic observations of debris disks and evolved stars): $MgSiO_3$ and $Mg_{0.6}Fe_{0.4}SiO_3$ (Fe-poor and Fe-rich pyroxene), $MgFeSiO_4$ and $Mg_{1.9}Fe_{0.1}SiO_4$ (amorphous and crystalline olivine), and C400 and C1000 (graphite-poor and graphite-rich carbon) (From Moro-Martín et al. 2005)

Table 9-1
Correspondence between β and radius (μm) for dust particles orbiting a solar-type star

β	$MgSiO_3$	$Mg_{0.6}Fe_{0.4}SiO_3$	$MgFeSiO_4$	$Mg_{1.9}Fe_{0.1}SiO_4$	C 400	C 1000
0.4	0.53	0.59	0.58	0.50	1.3	0.99
0.2	0.93	1.0	1.1	0.86	2.3	1.8
0.1	1.5	1.8	1.8	1.3	4.3	3.2
0.05	2.5	3.4	3.4	2.3	8.2	6.1
0.025	4.4	6.7	6.4	4.0	15.9	11.7
0.0125	8.0	13.7	12.4	7.1	31.2	22.7
0.00625	14.8	27.8	24.3	13.3	61.6	44.8
0.00312	28.6	54.8	48.1	25.7	122.7	89.0
0.00156	57.3	113.5	95.5	51.4	244.5	177.2

Grain radii are given in μm. Particles are assume to be spherical, with bulk densities in g cm^{-3} of 2.71 ($MgSiO_3$), 3.1 ($Mg_{0.6}Fe_{0.4}SiO_3$), 3.71 ($MgFeSiO_4$), 3.3 ($Mg_{1.9}Fe_{0.1}SiO_4$ – crystalline olivine), 1.435 (C 400), and 1.988 (C1000) (From Moro-Martín et al. 2005)

If the grains are highly nonspherical or have significant porosity (as might be expected for aggregates of smaller grains that result naturally from grain growth via coagulation – see, e.g., *Fig. 9-4*), their increased surface and decreased bulk density will result on an increased value of β (with respect to that of spherical grains).

2.1.2 Poynting-Robertson Drag

If the circumstellar particle is moving with respect to the central star, it experiences a force given by

$$\mu \frac{d^2\mathbf{r}}{dt^2} = \frac{SAQ_{\rm pr}}{c}\left[\left(1-\frac{\dot{r}}{c}\right)\hat{\mathbf{S}} - \frac{\mathbf{v}}{c}\right] \tag{9.2}$$

(to terms of order v/c), where $\mathbf{r}$ and $\mathbf{v}$ are the particle position and velocity with respect to the central star, $\hat{S}$ is the unit vector in the direction of the incident radiation ($\hat{\mathbf{S}} = \mathbf{r}/r$), and μ is the particle mass. The radial term $\frac{SAQ_{\rm pr}}{c}\left(1-\frac{\dot{r}}{c}\right)\hat{\mathbf{S}}$ is the radiation pressure force with the added factor $(1-\frac{\dot{r}}{c})$ to account for the Doppler effect, while the velocity-dependent term, $\frac{SAQ_{\rm pr}}{c}\frac{\mathbf{v}}{c}$, is known as the Poynting-Robertson (P-R) drag. The latter is a relativistic effect that can be intuitively explained in the following way: in the reference frame of the particle, the stellar radiation appears to come at a small angle forward from the radial direction (due to the aberration of light) that results in a force with a component against the direction of motion; in the reference frame of the star, the radiation appears to come from the radial direction, but the particle reemits more momentum into the forward direction due to the photons blueshifted by the Doppler effect, resulting in a drag force (Burns et al. 1979). Using β in (9.1), the equation of motion becomes

$$\frac{d^2\mathbf{r}}{dt^2} = \frac{-GM_*(1-\beta)}{r^3}\mathbf{r} - \frac{\beta}{c}\frac{GM_*}{r^2}\left[\left(\frac{\dot{r}}{r}\right)\mathbf{r} + \mathbf{v}\right]. \tag{9.3}$$

2.1.3 Stellar Wind Forces

Radiation pressure and P-R drag arise from the transfer of momentum between the photons and the dust particle. Similarly, the dust grain interacts with the stellar wind particles giving rise to a corpuscular pressure force and a corpuscular drag force; these forces depend on the stellar wind properties (mass-loss, relative velocity between the stellar wind and the dust particle, and molecular weight) and on how efficiently the stellar wind particles interact with the dust grain. In the solar system, the solar wind carries a momentum flux that is on average about $2 \cdot 10^{-4}$ times the momentum flux carried by radiation, and therefore the corpuscular pressure force can be neglected. On the contrary, the corpuscular drag force is about 35% of the P-R drag force (Gustafson 1994); its increased significance in this case is due to the slower velocity of the solar wind compared to the speed of light which, in the frame of the particle, increases the aberration angle and therefore the component of the force against the direction of motion; the aberration angle of the stellar wind particles is $\arctan(v/v_{\mathrm{sw}})$ compared to $\arctan(v/c)$ for the stellar photos, where v and v_{sw} are the velocity of the particle and stellar wind, respectively (Burns et al. 1979). Defining the ratio of the solar wind drag to the P-R drag as sw, the equation of motion becomes

$$\frac{d^2\mathbf{r}}{dt^2} = \frac{-GM_*(1-\beta)}{r^3}\mathbf{r} - \frac{(1+sw)\beta}{c}\frac{GM_*}{r^2}\left[\left(\frac{\dot{r}}{r}\right)\mathbf{r} + \mathbf{v}\right]. \tag{9.4}$$

For the solar system, $sw = 0.35$ (Gustafson 1994); for M-type stars, the contribution of the corpuscular drag force could be significantly more important because of the increased mass-loss rate and low stellar luminosities (Plavchan et al. 2005).

2.1.4 Effect of Radiation Forces on the Dust Dynamics

As soon as the dust particle is released from its parent body and begins to be subject to radiation forces, its equation of motion changes from that of the parent body, $\frac{d^2\mathbf{r}}{dt^2} = \frac{-GM_*}{r^3}\mathbf{r}$, to (9.3) (or (9.4) if considering stellar wind forces), resulting in a change of the particle orbital elements (Burns et al. 1979). The degree of change will depend on the β-value of the particle (the ratio of the radiation force to the gravitational force acting on the particle). *Figure 9-17* shows that for very large and very small grains $\beta \to 0$; therefore, particles in this size range are unaffected by radiation forces. Intermediate-sized particles might be blown out from the system if their specific orbital energy becomes positive, i.e.,

$$\frac{E}{m} = \frac{v^2}{2} - \frac{GM_*(1-\beta)}{r} \equiv -\frac{GM_*(1-\beta)}{2a} \geq 0, \tag{9.5}$$

where v is the particle velocity and a its semimajor axis. If the particle is released at perihelion, $r = a(1-e), v^2 = \frac{\mu}{a}\frac{1+e}{1-e}$ and ejection occurs for $\beta \geq \frac{(1-e)}{2}$; if the particles are released at aphelion, ejection occurs for $\beta \geq \frac{(1+e)}{2}$. Radiation pressure blowout is very fast, with a timescale similar to the orbital period, $t_{\mathrm{blow}} = \frac{1}{2}\left(\frac{(r/\mathrm{AU})^3}{M_*/M_\odot}\right)^{1/2}$ years. Because β depends on the particle size, e.g., for the solar system $\beta = 5.7 \cdot 10^{-5}\left(\frac{Q_{\mathrm{pr}}}{\rho s}\right)$, the condition above sets up a lower limit for the size of a particle on a bound orbit. In the solar system, the particles smaller than the blowout size are known as "β-meteoroids" (Zook and Berg 1975); these small grains, of asteroidal or cometary origin, are escaping from the solar system on hyperbolic orbits as the result of radiation pressure;

they have been inferred to exist from the lunar microcrater record and from in situ detections on board spacecraft (Grün et al. 1994).

The orbital energy of dust particles with $\beta \leq \frac{(1\pm e)}{2}$, i.e., with sizes larger than the blowout size, will stay negative after release, and therefore these particles will remain on bound orbits; because the dust particle and its parent body are effectively moving under different gravitational potentials, their orbital elements will differ (this is because the particle "feels" a stellar mass reduced by the factor (1–β) (Burns et al. 1979). The position and velocity of the parent body and the dust particle are the same at release,

$$v = \left[GM\left(\frac{2}{r} - \frac{1}{a}\right)\right]^{1/2} = \left[GM(1-\beta)\left(\frac{2}{r} - \frac{1}{a'}\right)\right]^{1/2}, \tag{9.6}$$

from which one gets that the particle semimajor axis right after release (a') is

$$a' = a\frac{1-\beta}{1-2a\beta/r}, \tag{9.7}$$

and its eccentricity (e') is

$$e' = \left|1 - \frac{(1-2a\beta/r)(1-e^2)}{(1-\beta^2)}\right|^{1/2} \tag{9.8}$$

(where a and e are the parent body semimajor axis and eccentricity). The particle inclination remains the same as that of its parent body because radiation pressure is a radial force.

Radiation and stellar wind forces not only change the orbital elements of the dust particles upon release but also affect their evolution with time. The drag forces make the dust particles lose orbital energy and spiral toward the central star, with

$$\left\langle \frac{da}{dt} \right\rangle_{\rm pr} = -\frac{\beta GM_*}{c}\frac{2+3e^2}{a(1-e^2)^{3/2}} \tag{9.9}$$

(averaged over an orbit); for a particle in a circular orbit, $e = 0$, the timescale for orbital collapse can be found from $\int_a^0 ada = \int_0^{t_{\rm pr}} -\frac{\beta 2GM_*}{c}dt$, resulting in $t_{\rm pr} = \frac{a^2 c}{4GM_*\beta} = \frac{4\pi\rho s a^2 c^2}{3L_* Q_{pr}}$ $\approx 690\left(\frac{\rho}{\rm g/cm^3}\right)\left(\frac{s}{\mu\rm m}\right)\left(\frac{a}{\rm AU}\right)^2\left(\frac{L_\odot}{L_*}\right)\frac{1}{Q_{pr}}$year, where ρ is the particle bulk density, s is the particle size, a is the initial heliocentric distance, L_* is the stellar luminosity, and Q_{pr} is the radiation pressure factor; in the solar system, $t_{\rm pr} \approx 400\left(\frac{a}{\rm AU}\right)^2\frac{1}{\beta}$ year $\approx 2{,}000\left(\frac{s}{\mu\rm m}\right)\left(\frac{a}{\rm AU}\right)^2$ year (Burns et al. 1979). A micron-sized dust particle at 40 AU (at the distance of the KB) will spiral into the Sun in ~3 Myr, a timescale much smaller than the age of the Sun; this means that the dust observed in the solar system (and around other mature stars) cannot be primordial but must be replenished by planetesimals.

Because P-R drag is of order $\frac{v}{c}$ and v is highest at perihelion, the orbit not only shrinks but also circularizes, with

$$\left\langle \frac{de}{dt} \right\rangle_{\rm pr} = -\frac{5\beta GM_*}{2c}\frac{e}{a^2(1-e^2)^{1/2}} \tag{9.10}$$

(averaged over an orbit). The particle inclination does not evolve with time because radiation pressure is a radial force (Burns et al. 1979). An example of the evolution of a dust particle under P-R drag can be seen in Fig. 9-18.

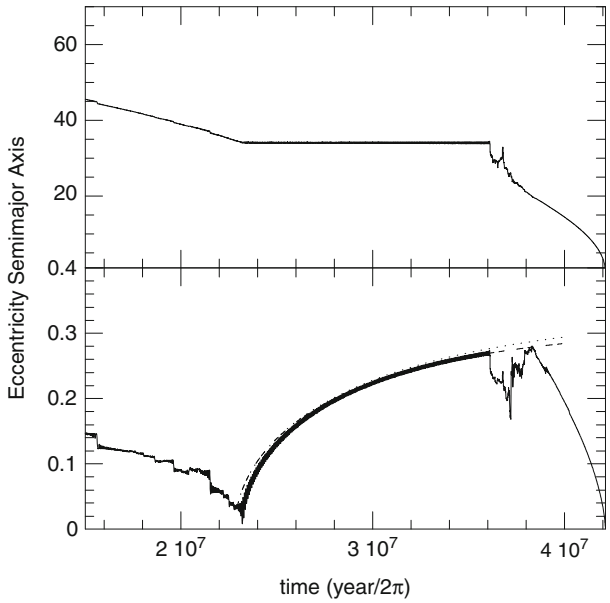


Fig. 9-18

Evolution of semimajor axis and eccentricity for a Kuiper belt dust particle with $\beta = 0.17$ in a planetary system with a solar-type star (with *sw* = 0.35) and a Neptune mass planet in a circular orbit. The *solid line* is the numerical result and the *dotted* and *dashed lines* are two analytical results. At first, the particle drifts inward due to P-R and corpuscular drag; during that time, its semimajor axis and eccentricity decrease. Then the particle is trapped for 14 Myr in the exterior 4:3 MMR with Neptune; during this time, the eccentricity of the particle increases until is sufficiently high to leave the resonance, after which the particle keeps migrating inward (From Moro-Martín and Malhotra 2002)

2.1.5 Effect of Radiation Forces on the Dust Spatial Distribution

In steady state, dust production, and dust loss are balanced, and the amount of mass that crosses a given radius r per unit time is a constant, i.e., $\dot{M} = 2\pi v r \sigma(r) = \text{const.}$, where v is the dust particles velocity and $\sigma(r)$ is the dust disk surface density. For unbound grains being blown out by radiation pressure, $v \sim \text{const.}$, $\sigma(r) \propto \frac{1}{r}$ and $n(r) \propto \frac{1}{r^2}$ (where the latter is the disk number density). For bound grains drifting inward under P-R drag, $v = v_{\rm pr} = \frac{r}{t_{\rm pr}} \propto \frac{1}{r}$, which results in a dust disk with constant surface density, $\sigma(r) \propto \dot{M} = \text{const.}$, and a number density that is inversely proportional to the distance to the central star, $n(r) \propto \frac{1}{r}$.

2.2 Gravitational Forces in the Presence of Planets

If the dust particles are constantly being released by a planetesimal belt, P-R drag would create a dust disk of wide radial extent and uniform surface density. However, if one or more planets are

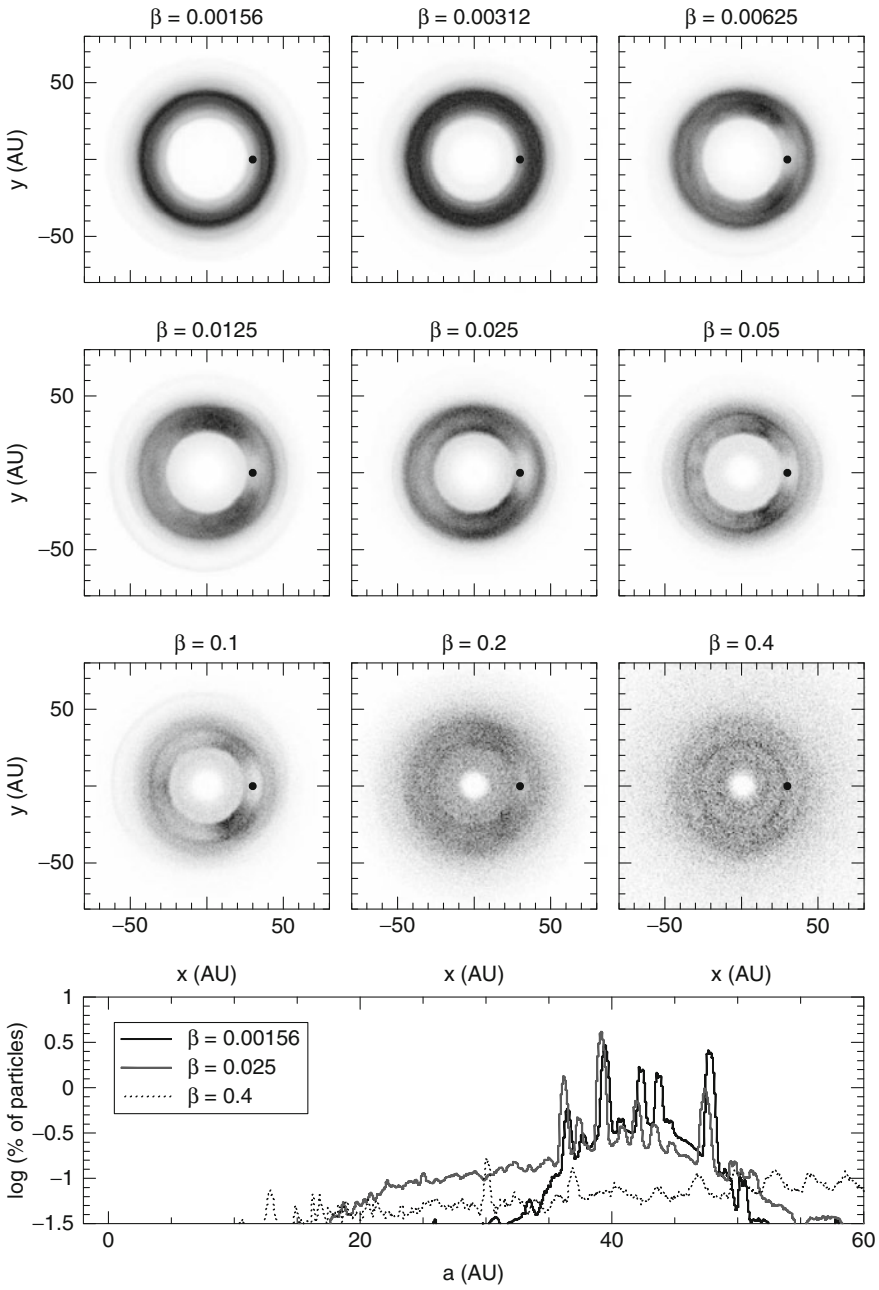


Fig. 9-19

Expected number density distribution of the Kuiper belt dust disk for nine different particle sizes (or β values). Assuming that the grains are composed of spherical astronomical silicates, β values of 0.4, 0.2, 0.1, 0.05, 0.025, 0.0125, 0.00625, 0.00312, and 0.00156 correspond to grain radii of 0.7, 1.3, 2.3, 4.5, 8.8, 17.0, 33.3, 65.9, and 134.7 μm, respectively. The trapping of particles in MMRs with Neptune is responsible for the ring-like structure, the asymmetric clumps along the orbit of

present in the system, on their journey toward the central star, the dust grains will be affected by gravitational perturbations that will modify the spatial distribution of the dust. The following describes the effect of the giant planets on the distribution of Kuiper belt dust in the solar system.

2.2.1 Resonant Perturbations

Figure 9-18 shows the dynamical evolution of a typical dust particle from the Kuiper belt. As the particle drifts inward due to P-R and solar wind drag, its semimajor and eccentricity decrease (following (9.9) and (9.10)). Then the particle might get trapped in a mean motion resonance (MMR) with one of the giant planets – most commonly with Neptune because it is the outermost planet. Mean motion resonances are located where the orbital period of the dust particle is $\frac{p+q}{p}$ times that of the planet, $T_{\rm dust} = \frac{p+q}{p} T_{\rm pl}$, where p and q are integers, $p > 0$, and $p + q \geqslant 1$. While the planet orbital period is $T_{\rm pl} = 2\pi\left(\frac{a_{\rm pl}^3}{GM_*}\right)^{1/2}$, the effect of radiation pressure results in a dust particle orbital period of $T_{\rm dust} = 2\pi\left(\frac{a_{\rm dust}^3}{GM_*(1-\beta)}\right)^{1/2}$ (because the particle "feels" a less massive Sun by a factor of $1-\beta$). Therefore, mean motion resonances take place when $a_{\rm dust\ (p+q):p} = a_{\rm pl}(1-\beta)^{\frac{1}{3}}\left(\frac{p+q}{p}\right)^{\frac{2}{3}}$. To account for corpuscular drag forces, one would substitute β by $\beta(1+sw)$. A particle is trapped when in the reference frame corotating with the planet, the closest approach between the particle and the planet is always at the same point(s) along the particle orbit (one point if $q = 1$, two points if $q = 2$); at these location(s), the particle receives repetitive "kicks" from the perturbing planet that are always in the same direction and can balance the energy loss due to P-R drag, halting the particle migration. While trapped, the particle semimajor axis stays constant, while its eccentricity slowly increases (as it can be seen in *Fig. 9-18*). The amount of time the particle stays trapped in the MMR is highly variable. The particle escapes from the resonance when its eccentricity becomes sufficiently high (~0.3 for the KB dust particle shown in *Fig. 9-18*). After escaping, the particle keeps spiraling inward under P-R and solar wind drag (following again (9.9) and (9.10).

The histogram in *Fig. 9-19* shows the trapping of Kuiper belt dust particles in MMRs. Even though the widths of the resonant regions are finite, these narrow regions of parameter space can become densely populated with dust because they are constantly being fed by the inward migration of dust particles under P-R and stellar wind drag. The features in the histogram are more pronounced for the larger grains (small β) than for the smaller grains (large β) because the larger grains have a higher trapping probability due to their slower migration velocity

Fig. 9-19 (Continued)

Neptune, and the clearing of dust at Neptune's location (indicated with a *black dot*). The disk structure is more prominent for larger particles (smaller β values) because the P-R drift rate is slower and the trapping is more efficient. The disk is more extended in the case of small grains (large β values) because small particles are more strongly affected by radiation pressure. The *histogram* shows the relative occurrence of the different MMRs for different sized grains, where the large majority of the peaks correspond to MMRs with Neptune. The inner depleted region inside ~10 AU is created by gravitational scattering of dust grains with Jupiter and Saturn (From Moro-Martín and Malhotra 2002)

($v_{pr} \propto \frac{1}{t_{pr}} \propto \beta \propto \frac{1}{s}$, where s is the radius of the particle). Generally, more massive planets exert a stronger perturbing force and allow the trapping of dust particles at resonances located further away from the planet. However, as it can be seen in the histogram, Neptune dominates the trapping of Kuiper belt dust because, even though is not the most massive planet, it is the outermost planet and its exterior resonances are not affected by the interior MMRs (located where the orbital period of the planet is $\frac{p+q}{p}$ times that of the particle – cf. Moro-Martín et al. 2008; Wyatt 2008a).

2.2.2 Gravitational Scattering

Figure 9-18) shows that after the Kuiper belt dust particle leaves the resonance, it will continue spiraling inward under the effects of P-R and corpuscular drag. This inward journey will take the particle near the orbit of a giant planet, where neighboring MMRs overlap and make the orbit of the particle chaotic and subject to gravitational ejection. This chaotic region extends from $a_{pl} - \Delta a < a < a_{pl} + \Delta a$, where m_{pl} and a_{pl} are the planet mass and semimajor axis and $\Delta a \simeq \pm 1.5 a_{pl} \left(\frac{m_{pl}}{m_\star}\right)^{2/7}$ (Wisdom 1980). Dynamical models show that most of the Kuiper belt dust grains get ejected from the solar system by gravitational scattering with the giant planets: for particles with $\beta \leq 0.4$, the percentage of particles ejected by Uranus and Neptune is 5–20%, while for Saturn and Jupiter is 25–40% (where the efficiency of ejection refer to each planet); only 10–20% of the Kuiper belt dust particles are able to pass these gravitational barriers and drift into the inner solar system. *Figure 9-20* shows the dependency of the efficiency of gravitational ejection on the planet mass, semimajor axis, and eccentricity: planets with masses of 3–10 M_{Jup} located between 1 and 30 AU in a circular orbit around a solar-type star eject >90% of the dust grains that go past their orbits; a 1 M_{Jup} planet at 30 AU ejects >80% of the grains and about 60% if located at 1 AU, while a 0.3 M_{Jup} planet ejects about 40% if located at 30 AU and <10% if it is at 1 AU. The efficiency of ejection decreases significantly as the planets, eccentricity increases, e.g., for a 1 M_{Jup} planet at 5 AU the efficiency of ejection decreases from 80% for $e = 0$ to 30% for $e = 0.5$. For eccentric planets, the particle ejection preferentially takes place along the major axis of the planet's orbit, with the number of particles ejected in the apoastron direction exceeding that in the periastron direction by a factor of 5 for $e = 0.5$ (because the planet spends more time near apoastron and therefore the probability of encounter with a dust particle is higher near that location).

2.2.3 Secular Perturbations

The gravitational forces on the dust particle exerted by planetary companions are described by a sum of many terms, known as the perturbing function. Secular perturbations are the long-term average of these forces; they result from the terms of the perturbing function that are independent of the mean longitude of the planets and the dust particles. Unlike the resonant perturbations described above, secular perturbations are nonperiodic in nature and act on longer timescales (>0.1 Myr). They can be thought of as the perturbations that would arise if the mass of the perturbing planet were to be spread out along its orbit (like a wire), weighting the mass density to reflect how much time the planet spends in each region. A planet on an eccentric orbit can force an eccentricity on the test particles; if the planet and the test particle are not

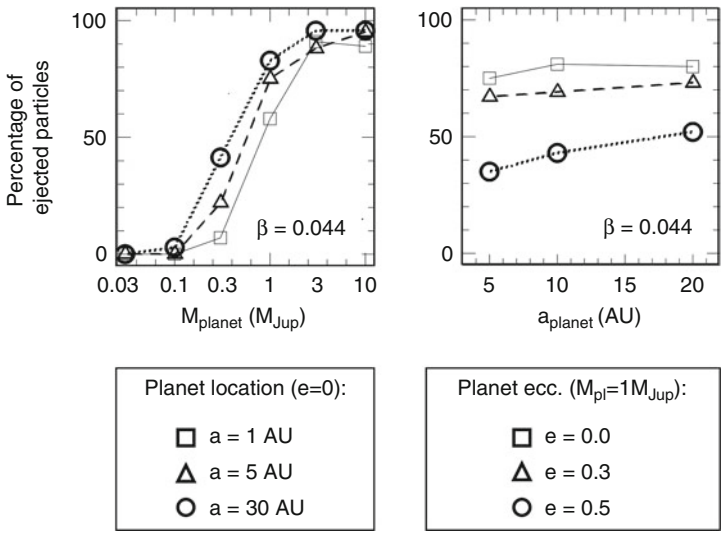


Fig. 9-20

Percentage of dust particles ejected from the system by gravitational scattering with a planet. *Left*: dependency on planet mass (x-axis) and the planet semimajor axis (indicated by the *different symbols*). *Right*: dependency on planet semimajor axis (x-axis) and eccentricity (corresponding to the *different symbols*). The particle size is fixed, corresponding to $\beta = 0.044$. In the case of Kuiper belt dust, ejection efficiencies for particles with $\beta \leq 0.4$ are approximately 5–20% for Uranus or Neptune and 25–40% for Saturn or Jupiter (Moro-Martín and Malhotra 2003, 2005)

co-planar, the secular perturbations will tend to align their orbits; secular perturbations do not affect the semimajor axis of the particles. If there is only one perturbing planet, the strength of the perturbation is independent of its mass, but the smaller the mass, the longer the secular timescale. If there is more than one perturbing planet, particles with precession rates that coincide with the eigenfrequencies of the planetary system (resulting from secular perturbations between the planets) will be strongly affected by secular perturbations (see, e.g., *Fig. 9-21*), resulting in the ejection of these particles from the secular resonant region (cf. Murray and Dermott 1999; Wyatt et al. 1999; Wyatt 2008a).

2.2.4 Effect of Gravitational Forces on the Dust Spatial Distribution

Resonant Perturbations

Figure 9-19 shows that the trapping of Kuiper belt dust in MMRs results in the formation of structure in the dust disk consisting on resonant rings outside the orbit of the perturbing planet, asymmetric clumps and a clearing of dust at the planet position. The rings are created when the dust particles are on nearly circular orbits and because they spend a significant part of their lifetime trapped at certain semimajor axis, corresponding to the most favorable MMRs. For Kuiper belt dust, with a low eccentricity perturbing planet and low eccentricity dust-producing parent bodies, the most favorable MMRs are the first-order resonances 2:1, 3:2, 4:3…The clumps appear when the particles are on eccentric orbits: when trapped in a resonance, the particle

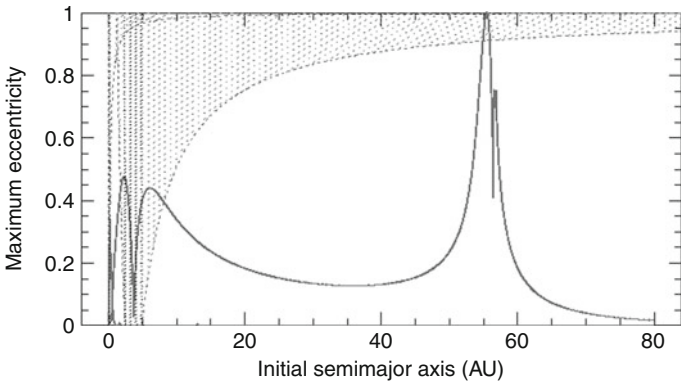


Fig. 9-21

Effect of the secular perturbations created by the two planets in the HD 38529 system. The planetary masses, semimajor axes, and eccentricities are the following: $M_b(\sin i) = 0.8\,M_{Jup}$, $a_b = 0.13$ AU, $e_b = 0.25$, for planet b, and $M_c(\sin i) = 12.2\,M_{Jup}$, $a_c = 3.74$ AU, $e_c = 0.35$, for planet c. The *shaded zones* denote areas that are strongly unstable due to planet-crossing orbits and overlapping first-order mean motion resonances. The y-axis is the maximum eccentricity imposed on the test particles (initially on circular orbits). The secular modes of the two planets excite the eccentricities of the test particles. The effect of these perturbations can be felt at a wide range of distances from the star: secular eccentricity excitation exceeds 0.1 to nearly 60 AU; the sharp peak at 55 AU is due to a resonance with the slow mode. The dust observed in the system is likely produced by planetesimals located in the regions of low eccentricity from ~20 to 50 AU (From Moro-Martín et al. 2007b)

orbits tend to be oriented always in the same direction with respect to the location of the planet, and clumps are created near apocenter, where the particles spend more time; because the clumps are fixed with respect to the planet position, they should follow the orbit of the planet (rotating in the reference frame of the star), and this can be used as an observational test to assess if the proper motion of the clump is consistent with dust particles trapped in an MMR. The clearing of dust near the planet position is created because when trapped in a resonance, the particle avoids being close to the perturbing planet.

The resonant features described above for the Kuiper belt dust disk (*Fig. 9-19*) have yet to be observed because the foreground thermal emission from the dust in the inner solar system (of asteroidal and cometary origin) overwhelms the background emission from the colder Kuiper belt dust. However, resonant features have been observed in the zodiacal cloud itself: a ring of asteroidal dust particles trapped in the 1:1 corotating resonance with the Earth at around 1 AU, with a 10% number density enhancement on the Earth's wake that results from the resonance geometry (Dermott et al. (1994b)). Modeling and observations of the Kuiper belt dust and of the zodiacal cloud indicate that planets with a wide range of masses (down to Neptune and Earth masses) can create high-contrast features via resonant trapping.

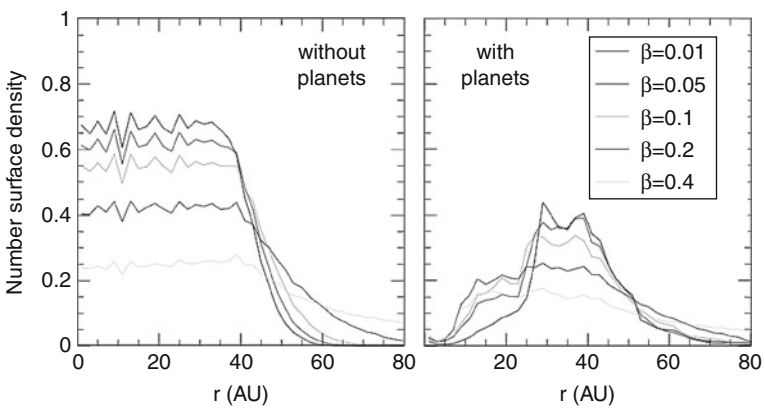
Figure 9-19 shows that the resulting dust disk structure depends on the particle size under consideration. This is because the dynamical evolution of the particle depends on its size: large particles interact weakly with the stellar radiation field, migrate slowly, and can get easily trapped in MMRs; small grains, on the other hand, interact strongly with radiation, their

inward migration is faster, and this gives rise to more extended and more uniform disks (Liou and Zook 1999; Moro-Martín and Malhotra 2002).

Mean motion resonances can also be populated if the perturbing planet migrates outward; in this case, the resonant trapping probability depends on the planet migration rate and the extent of the migration. In the solar system, the early migration of Neptune resulted in the trapping of Kuiper belt objects in MMRs, primarily in the 4:3, 3:2, 5:3, and 2:1 (Malhotra 1993). This means that the dust-producing planetesimals may have an asymmetric distribution (note that the disk models in ❯ *Fig. 9-19* assumed a uniform distribution of the dust-producing bodies, with angular elements uniformly distributed between 0 and 2π). If the parent bodies themselves are trapped in a resonance, the largest dust grains released from them would remain trapped in the same MMR, the smallest grains (with $\beta > 0.5$) would escape the system producing a spiral structure, and the intermediate size grains (with $\beta < 0.5$) would leave the resonance but remain on bound orbits producing an axisymmetric distribution (Wyatt 2008a).

Gravitational Scattering

High-contrast features can also be produced by gravitational scattering: for planetary systems in which the source of the dust is outside the orbit of a massive planet, gravitational ejection can result in the formation of a dust-depleted region inside the planet orbit, where the depletion factor depends on the planet's mass and orbit (❯ *Fig. 9-20*). Models indicate that a dust-depleted region of ~10 AU in radius is expected to be present in the Kuiper belt dust disk due to gravitational scattering by Jupiter and Saturn (❯ *Figs. 9-19* and ❯ *9-22*). There are also indications that large inner cavities are very common in extrasolar debris disks (see ❯ *Fig. 9-11* and the discussion in ❯ Sect. 1.3.3).



❯ **Fig. 9-22**

(*Right*): Expected radial distribution of the number surface density for the Kuiper belt dust disk, taking into account the gravitational perturbations from all the planets in the solar system (excluding Mercury). (*Left*): Same but assuming there are no planets in the system, leading to a uniform surface density. The main features from the comparison of the two panels are the depletion of particles in the inner 10 AU, due to scattering by Jupiter and Neptune, and the enhancement of particles from 30 to 50 AU, due to trapping in MMRs with Neptune (From Moro-Martín and Malhotra 2002)

Secular Perturbations

Secular perturbations can also produce warps, spirals, and brightness asymmetries. A warp is created when the planet and the test particles are on noncoplanar orbits, and because the secular perturbations will tend to align their orbits on a shorter timescale closer to the star. With time, the inflection point of the warp (marking the boundary between perturbed and unperturbed orbits) will move outward at a rate that depends on the planet mass and semimajor axis and is proportional to $M_{pl}\, a_{pl}^2$ (Mouillet et al. 1997), which in principle could be used to constrain the planet mass if the age of the planet were known. With time, the warp will end up disappearing, unless multiple planets on noncoplanar orbits are present in the system. Spirals are created if the planet is in an eccentric orbit; in this case, the secular perturbations can force an eccentricity on the test particles creating two spiral structures that with time propagate away from the planet. This also creates a brightness asymmetry because, after all the test particles have been affected, there is an offset in the dust disk center with respect to the star.

Some of the structure observed in the zodiacal cloud is the result of secular perturbations, namely, the inner edge of the cloud around 2 AU (due to a secular resonance with Saturn that also explains the inner edge of the main asteroid belt), the offset of the cloud center with respect to the Sun, the inclination of the cloud with respect to the ecliptic, and the cloud warp (cf. Murray and Dermott 1999; Wyatt et al. 1999; Wyatt 2008a). And as it was discussed in *Part I*, these features can also be seen in extrasolar dust disks, e.g., the warps in AU Mic and β Pic, the offsets with respect to the central star in ϵ-Eri and Fomalhaut, the brightness asymmetries in Fomalhaut and HD 32297, and the spiral structure in HD 141569, to name a few (see ❯ *Fig. 9-3* and discussion in ❯ Sect. 1.3.3).

Debris Disk Structure Can Unveil the Presence of Planets

Most of the structural features discussed above depend on the mass and orbit of the planet, and as the case of the solar system illustrates, the structure is sensitive to small planets (like the Earth) and to planets located far from the star (like Neptune). As discussed in ❯ Sect. 1.3.3, this opens the possibility of using the study of the dust disk structure as a detection technique of planets of a wide range of masses and semimajor axes. The discovery of the planets around Fomalhaut and β-Pic, that were previously predicted to exist based on the structure of both debris disks, illustrates this idea (see discussion in ❯ Sect. 1.3.3 and ❯ *Fig. 9-3*). Particularly interesting is that this method is complementary to radial velocity and transit surveys (which are limited to planets relatively close to the star) and to direct imaging (which is limited to young and massive planets).

2.3 Collisions

2.3.1 Collisional Lifetimes

The timescale for a collision between two equal-sized grains of radius s is $t_c = \frac{1}{\pi(s+s)^2 F}$, where F is the particle flux. The flux is given by $F = \frac{N}{V}\Delta v$, where N is the total number of particles in the disk, V is the total volume they occupy, and Δv is their velocity dispersion. If the particles in the dust disk have nonzero eccentricities and inclinations (e and i, with i in radians), they will occupy a volume $V = 2\pi a \cdot 2ae \cdot 2ia = 8\pi a^3 ei$, where the factor $2\pi a$ is the disk circumference, $2ae$ is its width (from pericenter, $a(1-e)$, to apocenter, $a(1+e)$), and $2ia$ is its thickness;

the velocity dispersion is $\Delta v = v_K(e^2 + i^2)^{1/2}$, where V_K is the Kepler velocity. This leads to $t_c = \frac{8\pi a^3 ei}{\pi 4s^2 N v_K (e^2+i^2)^{1/2}}$. If there are N grains with a characteristic size s, the optical depth is $\tau = \frac{N\pi s^2}{4\pi a^2 e}$. The collisional timescale in terms of the optical depth and orbital period (for $i = e$) is

$$t_c = \frac{P}{\tau}\frac{i}{4\pi(e^2+i^2)^{1/2}} \sim \frac{1}{9}\frac{P}{\tau} \sim \frac{1}{9\tau}\left(\frac{r}{\mathrm{AU}}\right)^{3/2} \sim \frac{1}{\tau\Omega}, \tag{9.11}$$

where Ω is the angular velocity. The collisional velocity above can also be approximated by

$$t_c \sim 0.1\left(\frac{a}{\mathrm{AU}}\right)^{3/2}\left(\frac{M_\odot}{M_*}\right)^{1/2}\frac{1}{\tau} \sim 1.1\cdot 10^4\left(\frac{a}{\mathrm{AU}}\right)^{3/2}\left(\frac{M_\odot}{M_*}\right)^{1/2}\left(\frac{10^5}{L_{\mathrm{dust}}/L_*}\right), \tag{9.12}$$

where L_{dust}/L_* is the fractional luminosity of the debris disk. All these estimates assume equal-sized grains (the following subsection considers the more realistic case of a distribution of particle sizes). Comparing this collision timescale to the PR timescale,

$$\frac{t_c}{t_{\mathrm{pr}}} \sim \frac{\frac{1}{9\tau}\left(\frac{r}{\mathrm{AU}}\right)^{3/2}}{\frac{400}{\beta}\left(\frac{r}{\mathrm{AU}}\right)^{2}} \sim \frac{\beta}{3{,}600\tau\left(\frac{r}{\mathrm{AU}}\right)^{1/2}}. \tag{9.13}$$

In the Kuiper belt, $r \sim 40$ AU and it is estimated that $\tau \sim 10^{-7}$ so that $\frac{t_c}{t_{\mathrm{pr}}} \sim 440\beta$ (with $t_c \sim 280$ Myr and $t_{\mathrm{pr}} \sim \frac{0.6}{\beta}$ Myr), suggesting that collisional destruction is unimportant and removes only a small fraction of KB dust grain as they drift into the inner solar system (this is referred to as the P-R-dominated regime). Compared to interstellar collisions, mutual collisions in the present-day Kuiper belt are less significant because the relative velocity of Kuiper belt dust grains ($\sim$1.6 km s^{-1}) is significantly smaller than the impact velocity of interstellar grains ($\sim$25 km s^{-1}), making shattering less likely, and because the optical depth of the Kuiper belt is very small so that collisions are infrequent. Due to limitations in the sensitivity of the detectors, the majority of the extrasolar debris disks observed with *Spitzer* have optical depths $\gtrsim$100 times that of the Kuiper belt; unlike in the solar system, in these systems, collisional destruction plays a significant role in the dust dynamics and resulting disk structure (this is referred to as the collision-dominated regime; Wyatt 2005, 2008b). The improved sensitivity of observatories like *ALMA*, *JWST*, and *SPICA* (and to some degree *Herschel*) will enable the study of debris disks that are in the P-R-dominated regime (i.e., KB dust disk analogs).

2.3.2 Effect of Collisions on the Dust Size Distribution

The size distribution most commonly adopted for both interstellar and debris dust is $n(s)\mathrm{d}s \propto s^{-3.5}\mathrm{d}s$ (Dohnanyi 1969; Mathis et al. 1977). It results from a catastrophic quasi-steady state collisional cascade in which the dust is derived from the grinding down of larger bodies, assuming that (1) the strength of the particle does not depend on the target size (where the strength is measured as the energy per unit volume needed to break and disperse the target) and (2) the system is in quasi-steady state, i.e., the same amount of mass that enters one size bin as larger particles break down leaves the size bin as the particles continue breaking in smaller pieces. Obviously, the bins at the two extremes would not be in steady state, and the reservoir of large particles would get depleted with time. For $q = -3.5$, the mass is dominated by the large grains, and the cross-section is dominated by the small grains. For comparison, in a size distribution

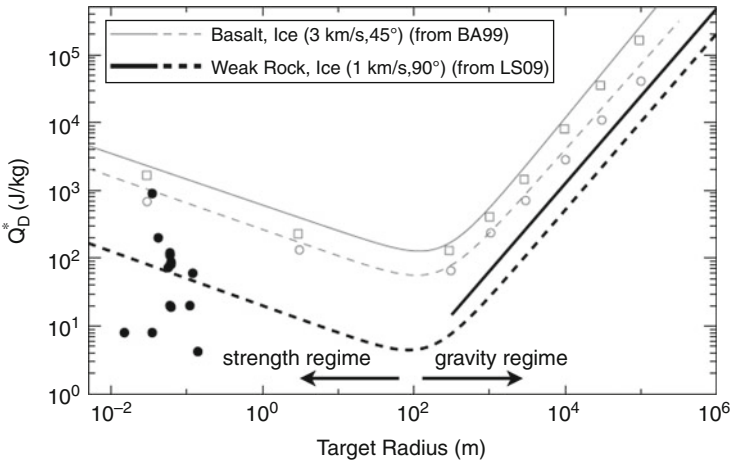


Fig. 9-23

Energy per unit mass necessary for a collision to result in the largest fragment to have half the mass as the original target, with target properties that could be similar to those of KBOs. The *lines* correspond to models (*solid* for rocks and *dashed* for ice) and the symbols to laboratory experiments (*squares* for rocks and *circles* for ice). BA09 is Benz and Asphaug (1999) and LS09 is Leinhardt and Stewart (2009) (Figure from Leinhardt and Stewart (2009))

with a power-law index $q = -3$, each size bin will contain an equal amount of cross-sectional area, while for $q = -4$, each size bin will contain an equal amount of mass.

However, the particle strength is in fact a function of the target size (unlike assumed above). For small objects, this critical specific energy for disruption and dispersal, Q^*_D, is dominated by the material strength, while for large objects, it is dominated by self-gravity, with a turnover of around 100 m (even for relatively small particles, gravity can play an important role because the transfer of momentum is inefficient and more energy than that required to shatter the targets is needed to dispersed the fragments). The turnover results in a sum of two power laws (see *Fig. 9-23*),

$$Q^*_D = Q_s\left(\frac{s}{1\,\mathrm{m}}\right)^{-b_s} + Q_g\left(\frac{s}{1\,\mathrm{km}}\right)^{b_g}, \tag{9.14}$$

with Q_s and Q_g in the range 10^5–10^7 erg g^{-1}, $b_s \sim 0$–0.5, and $b_g \sim 1$–2 (s is for the scattering regime and g for the gravitational regime – Benz and Asphaug 1999; Leinhardt and Stewart 2009). Krivov et al. (2005) estimate that two colliders of mass m_t and m_p are disrupted if their relative velocity is larger than the critical value of $\left(\frac{2(m_t+m_p)^2}{m_t m_p} Q^*_D\right)^{1/2}$, which for equal-sized particles would be $(8Q^*_D)^{1/2}$. Given that $Q^*_D \sim 10^8$ erg g^{-1}, dust grains would get destroyed if their relative velocity exceeds several hundreds m s^{-1}, typical of dust particles with $e \sim 1$ at 10s of AU from the central star (Krivov 2010).

The particle-in-the-box scenario described above that gives rise to the $q = -3.5$ power-law index can be less restrictive by taking into account a size-dependent particle strength, the removal of the smallest particles by radiation pressure, and the effect of collisions with grains

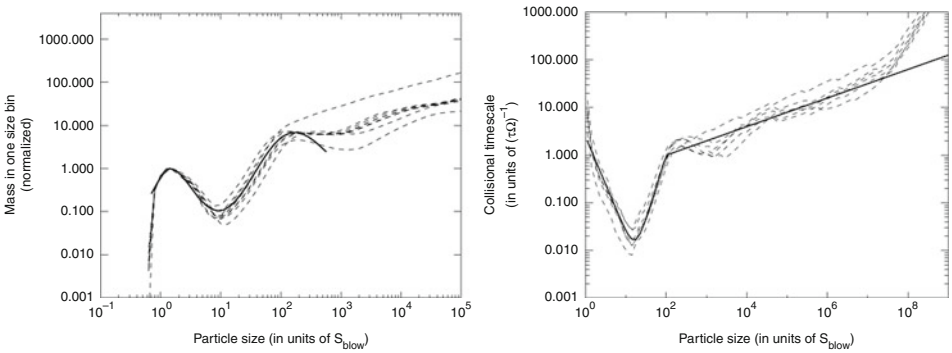


Fig. 9-24

(*Left*): Particle size distribution integrated over the entire disk. The different *dashed lines* correspond to models with different initial disk masses, surface density profiles, and dynamical excitation. The y-axis is normalized to the peak at ~1.5s_{blow}. (*Right*): Collisional lifetime at r = 55 AU. In both panels, the *solid lines* correspond to the empirical approximations described in the text (From Thébault and Augereau (2007))

coming from the inner regions on highly eccentric orbits. Numerical models that take these factors into account find that the size distribution changes from a strict single power law of index −3.5 to a wavy distribution; the resulting size distribution depends on the distance from the star, with the waves being more pronounced in the outer regions because the particles there are more affected by collisions with high-velocity grains (just below the blowout size) coming from the inner disk. The following results from numerical simulations by Thébault and Augereau (2007) illustrate the effect of collisions. Integrated over the entire disk, these models find an overdensity of particles with sizes ~ $2s_{\rm blow}$, and a depletion of particles with sizes in the range (10–50) $s_{\rm blow}$, where $s_{\rm blow}$ is the blowout size (see *Fig. 9-24*). For sizes < $100s_{\rm blow}$, these results are found to be weakly independent of the initial disk mass, the initial surface density profile, and the dynamical excitation, while they depend significantly on the particle strength. An empirical estimate for $s < 100s_{\rm blow}$ based on these numerical simulations gives $\mathrm{d}N \propto G(s)s^{-3.59}\mathrm{d}s$, for $\frac{2}{3}s_{\rm blow} < s \lesssim 100s_{\rm blow}$, with $\log_{10}(G(s)) = \frac{2}{3}\left[\cos\left(2\pi\left[\left|\frac{1}{2}\log_{10}\left(\frac{s}{1.5s_{\rm blow}}\right)\right|\right]^{0.85}\right) - 1\right]$. For $s > 100s_{\rm blow}$, a rough extrapolation is $\mathrm{d}N \propto s^{-3.7}\mathrm{d}s$ (Thébault and Augereau 2007).

Regarding the changes in the collisional lifetime, the expression in (9.11), $t_c \sim \frac{1}{9}\frac{P}{\tau} \sim \frac{1}{\tau\Omega}$, assumed that all impactors are equal-sized and all the collisions are destructive, and ignored both the effect of collisions with grains coming from the inner regions (< r), and the dynamics of the smallest grains affected by radiation pressure. *Figure 9-24* shows that when taking these factors into account, the collisional lifetime depends strongly on the particle size (with a wavy pattern) and in some cases (for particles <100 μm) can be two orders of magnitude smaller than $t_c \sim \frac{1}{9}\frac{P}{\tau} \sim \frac{1}{\tau\Omega}$; the main features are a sharp increase near the blowout size (s_{blow}), a sharp minimum at ~ $10s_{\rm blow}$, a sharp increase between ~ $10s_{\rm blow}$ and ~ $100s_{\rm blow}$, and a slow increase for larger sizes. An empirical estimate of the collisional timescale derived from numerical simulations gives $t_c(s,r) = \frac{1}{\tau\Omega}\left[\left(\frac{s}{s_1}\right)^{-2} + \left(\frac{s}{s_2}\right)^{2.7}\right]$, for $s < s_2$, and $t_c(s,r) = \frac{1}{\tau\Omega}\left(\frac{s}{s_2}\right)^{0.3}$, for $s > s_2$, where

s is the particle size, $s_1 = 1.2 s_{\rm blow}$, $s_2 = 100 s_{\rm blow}$, τ is the geometrical vertical optical depth, r is the distance to the star, and Ω is the angular velocity at r (Thébault and Augereau 2007).

The size distribution of the collisional debris is expected to evolve with time because the collisional timescale depends on the particle size. Small dust-size particles have shorter lifetimes, and they reach collisional equilibrium faster, i.e., their size distribution changes quickly from the primordial one to that of reprocessed material, while larger particles will retain their primordial size distribution for a longer period, creating a "knee" in the size distribution. The critical size at which the transition from a primordial to a collisional equilibrium population takes place increases with time. The largest bodies may not achieve collisional equilibrium within the age of the system. This will result in a combination of power laws for the size distribution, with waves in the smaller end triggered by the dust loss processes.

The size distributions of the small body population in the solar system (asteroids and KBOs) are partly the result of these processes. The top panel of *Fig. 9-25* (from Bottke et al. 2005) shows the results of the collisional evolution models of the asteroid belt: as time goes by, the initial size distribution changes from a power law (solid straight line in the left panel) to the observed wavy distribution (thick dotted line); the observed distribution has two main features, one at $D \sim 120$ km, leftover from the planetesimal accretion process, and another one at $D \sim 200$ m, marking the transition between the regime where particles are held by strength forces and the regime where particles are held by self-gravity. The bottom panel of *Fig. 9-25* (from Bernstein et al. 2004) shows the size distribution of the KBOs; even though this size distribution is more difficult to constrain due to the greater distance (in particular at the small size end), observations show a strong break to a shallower size distribution at $D < 100$ km, the boundary where the particles become more susceptible to collisional destruction. The size distributions of the asteroid and the Kuiper belts indicate that collisions have played an important role in the evolution of the solar system.

2.3.3 Effect of Collisions on the Dust Spatial Distribution

When the optical depth of the disk, τ, is large enough so that collisions dominate the dynamics, i.e., when $t_c \sim \frac{1}{9\tau}\left(\frac{r}{\rm AU}\right)^{3/2} < t_{pr} \sim \frac{r^2 c}{4GM_*\beta}$, the particles will not be able to migrate far from their parent bodies under P-R drag. This is not the case in the solar system, but it is the case in many extrasolar debris disks observed to date (Wyatt 2005). Under these conditions, small particles will not be able to spiral in and sweep a wide range of semimajor axes that may allow them to get trapped in MMRs or get ejected by gravitational scattering; on the contrary, these grains will get quickly ground down into smaller particles that radiation pressure sets on eccentric or hyperbolic orbits. On the other hand, large bodies with longer collisional lifetimes will closely trace the spatial distribution of their parent bodies (which could show structure due to the effect of gravitational perturbations).

Models in Thébault and Augereau (2007) show the effect of collisions on a planetesimal belt with an initial radial distribution $\propto r^{-1.5}$, characteristic of the minimum mass solar nebula. They found that the radial distribution of submillimeter size particles (which dominate the disk optical depth) gets significantly flatter because of two reasons: the smallest grains just over the blowout size are set on eccentric orbits and spend most of their time outside the orbit of their parent planetesimals, and the erosion of the larger 0.05–1 mm grains (the source of the smallest particles) proceeds faster in the inner regions.

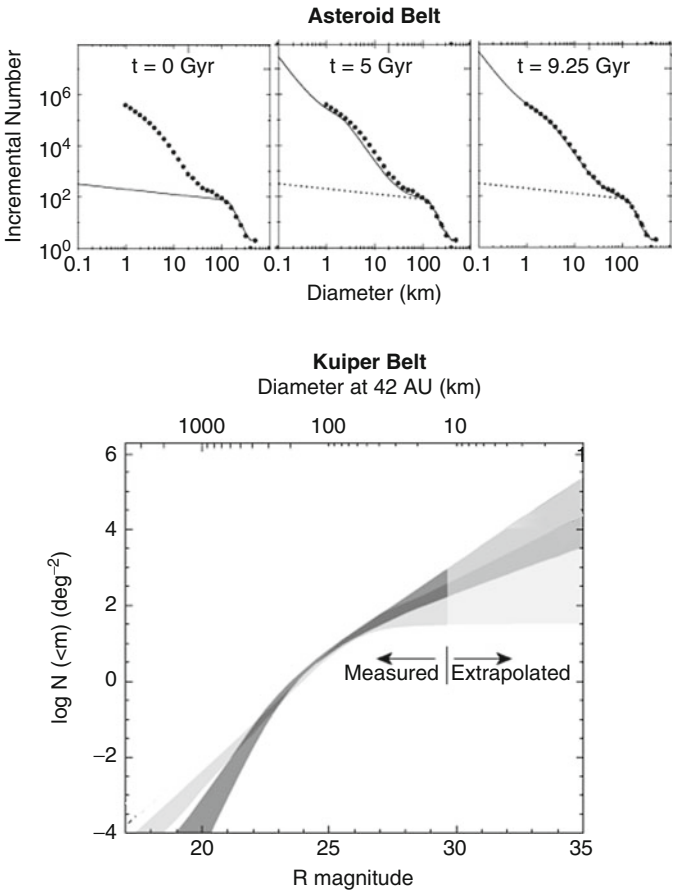


Fig. 9-25

(*Top*): observations (*thick dots*) and collisional evolution models (*thin solid line*) of the asteroid size distribution from Bottke et al. (2005). Note that the best model is for *t* = 9.5 Gyr; *t* is a pseudo-time that measures the degree of collisional evolution; the fact that *t* is greater than the age of the solar system indicates that the asteroid belt was significantly more massive in the past and that dynamical clearing must have played an important role. (*Bottom*): cumulative surface density of the KBOs in the classical disk (*dark grey*) and in the scattered disk (*light grey*) with 95% confidence upper and lower bounds (From Bernstein et al. 2004)

2.3.4 Effect of Collisions on the Dust Disk Evolution

The steady state scenario described in Sect. 2.3.2 cannot be sustained for an indefinite period of time because the reservoir of large particles that feeds the collisional cascade gets depleted, resulting in a decay of the amount of dust in the system. In the simplest scenario, the dust is derived from the grinding down of planetesimals, the planetesimals are destroyed after one collision, and the number of collisions is proportional to the square of the number of planetesimals, N; in this case, $\frac{dN}{dt} \propto N^2$ and $N \propto \frac{1}{t}$. In a collisional cascade, the dust production rate,

$R_{\rm prod}$, would be proportional to the loss rate of planetesimals times a proportionality constant, $R_{\rm prod} \propto \frac{dN}{dt}\left(\frac{s_{\rm dust}}{s_{\rm parent}}\right)^{-3.5}$, where $s_{\rm dust}$ and $s_{\rm parent}$ are the sizes of the dust and the parent bodies, respectively. In this scenario, $\frac{s_{\rm dust}}{s_{\rm parent}}$ is independent of time and one gets $R_{\rm prod} \propto \frac{dN}{dt} \propto N^2 \propto \frac{1}{t^2}$. One can solve for the amount of dust in the disk in steady state by equating the dust production rate to the dust loss rate, $R_{\rm loss}$. Depending on the number density of dust, n, there are two different solutions: (1) If the number density of dust particles is low, the disk is in the P-R drag-dominated regime where the dust loss rate is determined by P-R drag ($t_c > t_{\rm pr}$); in this case, the dust loss rate is proportional to the number density of particles, $R_{\rm loss} \propto n$, and from $R_{\rm prod} = R_{\rm loss}$, one gets $n \propto \frac{1}{t^2}$. (2) If the number density of dust is high, the disk is in the collision-dominated regime, where the main dust removal process is grain–grain collisions ($t_c < t_{\rm pr}$); in this case, the dust loss rate is given by $R_{\rm loss} \propto n^2$, and from $R_{\rm prod} = R_{\rm loss}$, one gets $n \propto \frac{1}{t}$, i.e., both the dust mass and the number of parent bodies (and therefore the total disk mass) decay as $\frac{1}{t}$, with a characteristic timescale that is inversely proportional to the initial disk mass (Dominik and Decin 2003).

The solar system is in the P-R drag-dominated regime, while, due to limited sensitivity of the observations, the majority of the extrasolar debris disks known to date are in the collision-dominated regime; this explains why the intensity of the dust emission of the entire sample of debris disk systems (for similar-type stars, and including stars with a wide range of ages) roughly follows a t^{-1} decay (see *Fig. 9-9*). More specifically, the decay of the dust mass, the fractional luminosity $L_{\rm dust}/L_{\rm star}$, and the thermal excess is better approached by t^{ξ} with $\xi = -0.3$ to -0.4 (Löhne et al. 2008); it deviates from t^{-1} because this corresponds to the assumption that all the bodies have achieved collisional equilibrium; however, the largest planetesimals of some of the debris disks observed (with ages ranging from 10 Myr to 10 Gyr) would not have had enough time to achieve collisional equilibrium within the age of the system. The index depends on the particle strength as a function of particle size and the properties of the largest planetesimals (primordial size distribution, eccentricities, and inclinations – the latter two determining their rate of collisions).

2.4 Other Physical Processes

2.4.1 Dust Sublimation

Silicate dust grains sublimate at ~1,500 K, while for icy grains 1–100 μm in size the destruction temperature is ~120 K (the timescale of sublimation increases dramatically above 100 K). The following is an estimate of the corresponding sublimation distance. If the grain is larger than the peak wavelengths of both the absorbed and emitted radiation, the grain is in the blackbody regime where it emits and absorbs efficiently and its temperature is given by

$$T_{\rm dust} = 278\left(\frac{L_*}{L_\odot}\right)^{1/4}\left(\frac{r}{\rm AU}\right)^{-1/2} \ {\rm K}, \tag{9.15}$$

where r is the heliocentric distance; if the grain is larger than the peak wavelength of the absorbed radiation but smaller than the peak wavelength of the emitted radiation, the grain is in the intermediate size regime and

$$T_{\rm dust} = 468\left(\frac{L_*}{L_\odot}\right)^{1/5}\left(\frac{r}{\rm AU}\right)^{-2/5}\left(\xi s_{\mu\rm m}\right)^{-1/5} \ {\rm K}, \tag{9.16}$$

where $s_{\mu m}$ is the grain radius and $\xi \sim 2\pi$, $\frac{1}{2\pi}$ and 1 for strongly, weakly, and moderately absorbing material, respectively (Backman and Paresce 1993). From Wien's law, the peak wavelength is given by $\lambda_{peak} = \frac{3669}{T}$ μm. The incident solar radiation ($T_\odot = 5{,}778$ K) peaks at 0.6 μm, while the emitted radiation would peak at 2.5 μm if $T_{dust} = 1{,}500$ K or 30 μm if $T_{dust} = 120$ K. In the former case, silicate grains larger than ~2.5 μm would be in the blackbody regime in which case the sublimation temperature of 1,500 K would be achieved at a distance of $r \sim 0.03$ AU. In the latter case, icy grains smaller than 30 μm would be in the intermediate size regime achieving the destruction temperature of 120 K at 4 AU (for 1 μm grains) and 2 AU (for 30 μm grains).

2.4.2 Lorentz Force

Dust grains are expected to be electrically charged due to the accretion of stellar wind ions and electrons, ionization due to impacts with stellar wind particles, and the ejection of electrons due to UV radiation (the latter process dominating the charging of solar system dust). Once charged, the dust grains will be subject to the Lorentz force, $F_L = qvB$, where q is the electric charge, v is the velocity of the grain with respect to the field, and B is the magnetic field flux density.

In the case of the solar system, the grains inside the heliosphere (<150 AU) are subject to the interplanetary magnetic field. This field has a dipole component that changes polarity every 11 years with the 22 year solar cycle; near the ecliptic, these sign reversals take place more rapidly because of the presence of the heliospheric current sheet (the extension of the Sun's magnetic equator into interplanetary space, separating regions of opposite polarity). At solar minimum, the current sheet extends from approximately −25° to 25° from the solar equator; particles within this latitude range will cross the current sheet at least twice every solar rotation (~27 days) or four or even six times if the current sheet is wrapped because of higher order terms in the magnetic field; particles at higher ecliptic latitudes will cross the current sheet at least twice as they orbit the Sun. These sign reversals will cause a random walk in the semimajor axis of the particles; for particles smaller than a few microns, this random walk will dominate over the P-R drift on timescales from a few orbital periods in the inner solar system to a few tens of orbital periods in the outer solar system. In addition to the dipole, the interplanetary magnetic field has a dominant component that is perpendicular to the radial solar wind vector, with a magnitude $\sim \frac{3 \cdot 10^{-5}}{r(\mathrm{AU})}$ Gauss (for heliocentric distances r exceeding a few AU); the Lorentz force in this case will tend to scatter the smallest dust particles out of the ecliptic plane by perturbing the particle inclinations while keeping the energy of the orbit unchanged.

Because, as described above, the circumstellar magnetic field can have a complex structure and time behavior, it is difficult to include the effect of the Lorentz force in the study of the dynamics and spatial distribution of dust particles in extrasolar debris disks for which the magnetic field properties are unknown.

2.4.3 Sputtering

Dust grains erode with time due to the impact of energetic stellar wind particles (a process known as sputtering). The study of exposed lunar rocks allows to estimate the rate of erosion, resulting in values that differ by two orders of magnitude, where a rate of ~0.2 Å year^{-1} at 1 AU would be on the high-end; this rate scales with distance as r^{-2}. Dynamical models indicate that

a dust particle with Kuiper belt origin has a typical lifetime of ~ 10^7 year, most of which is spent at $a > 20$ AU. Assuming the particle spends 10^7 year at 20 AU from the Sun, the fraction of mass loss due to erosion (at the highest measured rate) would be ~50% for a 3 μm particle, scaling as s^{-1}, where s is the particle radius; if the erosion rate is 100 times smaller (which is within the present uncertainties), the mass loss would be negligible. The erosion rate is uncertain because sputtering can cause chemical alteration on the dust grain surface (via the implantation of stellar wind ions) that can create molecular bondings between layers of dissimilar materials making them erosion resistant. These chemical alterations may also change the optical properties of the grain (e.g., by producing a blackened highly carbonized and refractory surface layer from organic and volatile grain mantles), which can affect the particle's response to radiation forces and therefore its lifetime.

2.5 Open Questions

There are many open questions in debris disk modeling, some are related to the model input parameters, while others are related to the modeling procedures. Regarding the input parameters, the lack of spectral features in most of debris disk spectra makes it difficult to constrain the dust properties – like particle size distribution, composition, shape, and porosity – which determine the particle emission properties and dynamics (because they affect how the particle interacts with radiation forces). The properties of the parent planetesimal are also unknown, with the particle strength playing a key role in the dust-producing process via collisions. The origin of the debris dust itself is also uncertain: a component may originate from steady-state erosion, as the models generally assume, but now it has became clear that stochastic collisions are required to explain some of the debris disk observations; however, the origin and ubiquity of stochastic collisions remain unknown; in addition, in some cases, the debris dust could also be due to cometary activity rather than planetesimal collisions. Another obvious caveat is that, in most systems, little is known about the presence of planets, in particular long-period planets and planets with low masses (and even when planets are identified, their masses are generally not well determined); this makes it difficult to constrain the dynamical state of the swarm of dust-producing planetesimal and the effect of the planets on the dust disk structure.

But even if the input parameters were known, there are open questions regarding the modeling procedures and the physics involved. Regarding the physics, collisions are the most difficult to account for because they take place across 12 orders of magnitude and involve a wide range of relative velocities and incoming angles, from head-on to grazing collisions, with their outcome ranging from particle growth, to cratering, to complete disruption.

Regarding the modeling, while the N-body approach (like that in ❯ *Figs. 9-18* and ❯ *9-19*) follows the trajectory of individual particles and can take into account the effect of planetary perturbations, radiation forces, gas drag, and the interstellar medium, the CPU power is limited and such models cannot treat a number of particles sufficiently large to cover a wide range of particle sizes; therefore, the N-body approach cannot model the particle size distribution. On the contrary, statistical methods (like those in ❯ *Figs. 9-15* and ❯ *9-24*), where the particles are replace with packages with given distributions, allow to study the outcome of collisions and the size distributions, but, due to the averaging over angular orbital elements, they lose accuracy in the dynamical modeling and cannot study in detail the spatial distribution of the dust.

References

A'Hearn, M. F. 2008, Deep impact and the origin and evolution of cometary nuclei. Space Sci Rev, 138, 237

Altobelli, N., Dikarev, V., Kempf, S., Srama, R., Helfert, S., Moragas-Klostermeyer, G., Roy, M., & Grün, E. 2007, Cassini/cosmic dust analyzer in situ dust measurements between Jupiter and Saturn. J. Geophys. Res. (Space Phys.), 112, 7105

Aumann, H. H., Beichman, C. A., Gillett, F. C., de Jong, T., & Houck, J. R., et al. 1984, Discovery of a shell around Alpha Lyrae. AJ, 278, L23

Backman, D. E., & Paresce, F. 1993, Main-sequence stars with circumstellar solid material – The VEGA phenomenon, in Protostars and Planets III, ed. E. H. Levy, & J. I. Lunine (Tucson: University of Arizona Press), 1253

Backman, D. E., Dasgupta, A., & Stencel, R. E. 1995, Model of a Kuiper belt small grain population and resulting far-infrared emission. ApJL, 450, L35

Beichman, C. A., Bryden, G., Gautier, T. N., Stapelfeldt, K. R., & Werner, M., et al. 2005, An excess due to small grains around the nearby K0 V star HD 69830: asteroid or cometary debris? ApJ, 626, 1061

Beichman C. A., Tanner, A., Bryden, G., Stapelfeldt, K. R., & Werner, M. W., et al. 2006, IRS spectra of solar-type stars: a search for Asteroid belt analogs. ApJ, 639, 1166

Benz, W., & Asphaug, E. 1999, Catastrophic disruptions revisited. Icarus, 142, 5

Bernstein, G. M., Trilling, D. E., Allen, R. L., Brown, M. E., Holman, M., & Malhotra, R. 2004, The size distribution of trans-neptunian bodies. AJ, 128, 1364

Booth, M., Wyatt, M. C., Morbidelli, A., Moro-Martín, A., & Levison, H. F. 2009, The history of the solar system's debris disc: observable properties of the Kuiper belt. MNRAS, 399, 385

Bottke, W. F., Durda, D. D., Nesvorný, D., Jedicke, R., Morbidelli, A., Vokrouhlický, D., & Levison, H. 2005, The fossilized size distribution of the main asteroid belt. Icarus, 175, 111

Brownlee, D., Tsou, P., Aléon, J., Alexander, C., & Araki, T., et al. 2006, Comet 81P/wild 2 under a microscope. Science, 314, 1711

Bryden, G., Beichman, C. A., Trilling, D. E., Rieke, G. H., & Holmes, E. K., et al. 2006, Frequency of debris disks around solar-type stars: first results from a Spitzer MIPS survey. ApJ, 636, 1098

Bryden, G., Beichman, C. A., Carpenter, J. M., Rieke, G. H., Stapelfeldt, K. R., et al. 2009, Planets and debris disks: results from a Spitzer/MIPS search for infrared excess. ApJ, 705, 1226

Burns, J. A., Lamy, P. L., & Soter, S. 1979, Radiation forces on small particles in the solar system. Icarus, 40, 1

Carpenter, J. M., Bouwman, J., Mamajek, E. E., Meyer, M. R., & Hillenbrand, L. A., et al. 2009, Formation and evolution of planetary systems: properties of debris dust around solar-type stars. ApJS, 181, 197

Chapman, C. R., Cohen, B. A., & Grinspoon, D. H. 2007, What are the real constraints on the existence and magnitude of the late heavy bombardment? Icarus, 189, 233

Chen, C. H., Sargent, B. A., Bohac, C., Kim, K. H., & Leibensperger, E., et al. 2006, Spitzer IRS spectroscopy of IRAS-discovered debris disks. ApJS, 166, 351

Chiang, E., Kite, E., Kalas, P., Graham, J. R., & Clampin, M. 2009, Fomalhaut's debris disk and planet: constraining the mass of fomalhaut b from disk morphology. ApJ, 693, 734

Chyba, C. F. 1990, Impact delivery and erosion of planetary oceans in the early inner solar system. Nature, 343, 129

Clampin, M., Krist, J. E., Ardila, D. R., Golimowski, D. A., & Hartig, G. F., et al. 2003, Hubble space telescope ACS coronagraphic imaging of the circumstellar disk around HD 141569A. AJ, 126, 385

Debes, J. H., Weinberger, A. J., & Schneider, G. 2008, Complex organic materials in the circumstellar disk of HR 4796A. ApJL, 673, L191

Decin, G., Dominik, C., Waters, L. B. F. M., & Waelkens, C. 2003, Age dependence of the vega phenomenon: observations. ApJ, 598, 636

Dermott, S. F., Durda, D.D., Gustafson, B. A. S., Jayaraman, S., Liou, J. C., & Xu, Y. L. 1994a, Zodiacal dust bands, in Asteroids, Comets, Meteors 1993, ed. A. Milani, et al. (Dordrecht/Boston: Kluwer), 127–142

Dermott, S. F., Jayaraman, S., Xu, Y. L., Gustafson, B. Å. S., & Liou, J. C. 1994b, A circumsolar ring of asteroidal dust in resonant lock with the Earth. Nature, 369, 719

Dermott, S. F., Grogan, K., Durda, D. D., Jayaraman, S., & Kehoe, T. J. J., et al. 2001, Orbital evolution of interplanetary dust, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig (Berlin/New York: Springer), 295

Dermott, S. F., Kehoe, T. J. J., Durda, D. D., Grogan, K., & Nesvorný, D. 2002, Recent rubble-pile origin of asteroidal solar system dust bands and asteroidal interplanetary dust particles, in Asteroids, Comets, and Meteors: ACM 2002, Vol. 500, ed. B. Warmbein (Noordwijk: ESA Publications Division), 319

Dohnanyi, J. S. 1969, Collisional model of asteroids and their debris. J. Geophys. Res., 74, 2431

Dominik C., & Decin G. 2003, Age dependence of the vega phenomenon: theory. APJ, 598, 626

Eiroa, C., Marshall, J. P., & Mora, A., et al. 2011, Herschel discovery of a new class of cold, faint debris discs. A&A, 536, L4

Farley, K. A. 1995, Cenozoic variations in the flux of interplanetary dust recorded by 3He in deep sea sediments. Nature, 376, 153

Fischer D. A., & Valenti J. 2005, The planet-metallicity correlation. ApJ, 622, 1102

Frisch, P. C., et al. 1999, Dust in the local interstellar wind. ApJ, 525, 492

Gautier, T. N., III, Rieke, G. H., Stansberry, J., Bryden, G. C., & Stapelfeldt, K. R., et al. 2007, Far-infrared properties of M Dwarfs. ApJ, 667, 527

Gomes, R., Levison, H. F., Tsiganis, K., & Morbidelli, A. 2005, Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Nature, 435, 466

Gurnett, D. A., Ansher, J. A., Kurth, W. S. & Granroth, L. 1997, Micron-sized dust particles detected in the outer solar system by the Voyager 1 and 2 plasma wave instruments. Geophys. Res. Lett., 24, 3125

Greaves J. S., Holland, W. S., Moriarty-Schieven G., Jenness, T., & Dent, W. R. F., et al. 1998, A dust ring around ϵ-eridani: analog to the young solar system. ApJ, 506, L133

Greaves, J. S., Holland, W. S., Jayawardhana, R., Wyatt, M. C., & Dent, W. R. F. 2004, A search for debris discs around stars with giant planets. MNRAS, 348, 1097

Greaves, J. S., et al. 2005, Structure in the ϵ eridani debris disk. ApJ, 619, L187

Greaves, J. S., Fischer, D. A. & Wyatt, M. C. 2006, Metallicity, debris discs and planets. MNRAS, 366, 283

Grün, E., Gustafson, B., Mann, I., Baguhl, M., Morfill, G. E., Staubach, P., Taylor, A., & Zook, H. A. 1994, Interstellar dust in the heliosphere. A&A, 286, 915

Grün, E., Baguhl, M., Svedhem, H., & Zook, H. A. 2001, In situ measurements of cosmic dust, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig (Berlin/New York: Springer), 295

Gustafson, B. A. S. 1994, Physics of zodiacal dust. Ann. Rev. Earth Planet. Sci., 22, 553

Gustafson, B. A. S., Greenbert, J. M., Kolokolova, L., u, Y., & Stognienko, R. 2001, Interactions with electromagnetic radiation: theory and laboratory simulations, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig (Berlin/New York: Springer), 57

Habing, H. J., Dominik, C., & Jourdain de Muizon, M., et al. 2001, Incidence and survival of remnant disks around main-sequence stars. A&A, 365, 545

Heap, S. R., Lindler, D. J., Lanz, T. M., Cornett, R. H., Hubeny, I., Maran, S. P., & Woodgate, B. 2000, space telescope imaging spectrograph coronagraphic observations of β pictoris. ApJ, 539, 435

Hillenbrand, L. A., Carpenter, J. M., Kim, J. S., Meyer, M. R., & Backman, D. E., et al. 2008, The complete census of 70 μm-bright debris disks within "the formation and evolution of planetary systems" Spitzer legacy survey of sun-Like stars. ApJ, 677, 630

Holland, W. S., Greaves, J. S., Zuckerman, B., Webb, R. A., & McCarthy, C. et al. 1998, Submillimetre images of dusty debris around nearby stars. Nature, 392, 788

Holland, W. S., Greaves, J. S., Dent, W. R. F., Wyatt, M. C., & Zucker-man, B. et al. 2003, Submillimeter observations of an asymmetric dust disk around Fomalhaut. ApJ, 582, 1141

Houck, J. R., Roellig, T. L., van Cleve, J., Forrest, W. J., & Herter, T. et al. 2004, The infrared spectrograph (IRS) on the Spitzer space telescope. ApJS, 154, 18

Humes, D. 1980, Results of Pioneer 10 and 11 meteoroid experiments – interplanetary and near-Saturn. J. Geophys. Res., 85(A/II), 5841

Jessberger, E. K., Stephan, T., Rost, D., Arndt, P., & Maetz, M., et al. 2001, Properties of interplanetary dust: information from collected samples, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig (Berlin/New York: Springer) , p. 253

Jewitt, D., & Luu, J. 1993, Discovery of the candidate Kuiper belt object 1992 QB1. Nature, 362, 730

Jewitt, D. C., & Luu, J. X. 2000, Physical Nature of the Kuiper Belt. Protostars and Planets IV, 1201

Jewitt, D., Weaver, H., Agarwal, J., Mutchler, M., & Drahus, M. 2010, A recent disruption of the main-belt asteroid P/2010A2. Nature, 467, 817

Jura, M. 2006, Carbon deficiency in externally polluted white dwarfs: evidence for accretion of asteroids. ApJ, 653, 613

Jura, M., Farihi, J., Zuckerman, B., & Becklin, E. E. 2007, Infrared emission from the dusty disk orbiting GD 362, an externally polluted white dwarf. AJ, 133, 1927

Kalas, P., Graham, J. R., & Clampin, M. 2005, A planetary system as the origin of structure in Fomalhaut's dust belt. Nature, 435, 1067

Kalas, P., Graham, J. R., Clampin, M. C., & Fitzgerald, M. P. 2006, First scattered light images of debris disks around HD 53143 and HD 139664. APJ, 637, 57

Kalas, P., et al. 2008, Optical images of an exosolar planet 25 light-years from Earth. Science, 322, 1345

Kenyon, S. J., & Bromley, B. C. 2005, Prospects for detection of catastrophic collisions in debris disks. AJ, 130, 269

Kóspál, Á., Ardila, D. R., Moór, A., & Ábrahám, P. 2009, On the relationship between debris disks and planets. ApJ, 700, L73

Krist J. E., Ardila D. R., Golimowski D. A., Clampin M. & Ford H. C. 2005, Hubble space telescope advanced camera for surveys coronagraphic imaging of the AU microscopii debris disk. AJ, 129, 1008

Krivov, A. V. 2010, Debris disks: seeing dust, thinking of planetesimals and planets. Res. Astron. Astrophys., 10, 383

Krivov, A. V., Sremčević, M., & Spahn, F. 2005, Evolution of a Keplerian disk of colliding and fragmenting particles: a kinetic model with application to the Edgeworth Kuiper belt. Icarus, 174, 105

Lagrange, A.-M., et al. 2010, A giant planet imaged in the disk of the young star β pictoris. Science, 329, 57

Landgraf, M., Liou, J.-C., Zook, H. A., & Grün, E. 2002, Origins of solar system dust beyond Jupiter. AJ, 123, 2857

Leinhardt, Z. M., & Stewart, S. T. 2009, Full numerical simulations of catastrophic small body collisions. Icarus, 199, 542

Levasseur-Regourd, A. C., Mann, I., Dumont, R. & Hanner, M. 2001, Optical and thermal properties of interplanetary dust, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig, (Berlin/New York: Springer), 57

Li, J., Jewitt, D., Clover, J. M., & Jackson, B. V. 2010, Outburst of comet 17P/Holmes observed With the solar mass ejection imager. arXiv:1012. 1570

Liou, J.-C., & Zook, H. A. 1999, Signatures of the giant planets imprinted on the Edgeworth-Kuiper belt dust disk. AJ, 118, 580

Lisse, C. M., Beichman, C. A., Bryden, G., & Wyatt, M. C. 2007, On the nature of the dust in the debris disk around HD 69830. ApJ, 658, 584

Löhne, T., Krivov, A. V. & Rodmann, J. 2008, Long-term collisional evolution of debris disks. APJ, 673, 1123

Lovis, C., Mayor, M., Pepe, F., Alibert, Y., & Benz, W., et al. 2006, An extrasolar planetary system with three Neptune-mass planets. Nature, 441, 305

Malhotra, R. 1993, The origin of Pluto's peculiar orbit. Nature, 365, 819

Marcy, G., Butler, R. P., Fischer, D., Vogt, S., & Wright, J. T., et al. 2005, Observed properties of exoplanets: masses, orbits, and metallicities. Prog. Theor. Phys. Suppl., 158, 24

Mathis, J. S., Rumpl, W., & Nordsieck, K. H. 1977, The size distribution of interstellar grains. APJ, 217, 425

Maurette, M., Immel, G., Hammer, C., Harvey, R., Kurat, G. & Taylor, S. 1994, Collection and curation of IDPs from the Greenland and Antarctic ice Sheets, in Analysis of Interplanetary Dust, ed. M. E. Zolensky, T. L. Wilson, F. J. M. Rietmeijer, & G. J. Flynn (New York: American Institute of Physics), 277–289

Maurette, M., Engrand, C. & Kurat, G. 1996, Collection and microanalysis of antarctic micrometeorites, in ASP Conf. Ser. 104, Physics, Chemistry and Dynamics of Interplanetary Dust, ed. B. A. S. Gustafson, & M. S. Hanner (San Francisco, CA: ASP), 265–273

Mayor, M., Marmier, M., Lovis, C., Udry, S. Sgransan, D., Pepe, F., Benz, W., Bertaux, J. -L., Bouchy, F., Dumusque, X., Lo Curto, G., Mordasini, C., Queloz, D., & Santos, N. C. 2011, The HARPS search for southern extra-solar planets XXXIV. Occurrence, mass distribution and orbital properties of super-Earths and Neptune-mass planets. arXiv:1109.2497 (A&A in press)

McDonnell, T., McBride, N., Green, S. F., & Ratcliff, P. R., et al. 2001, Near Earth environment, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig (Berlin/New York: Springer), 163

McKeegan, K. D., Aléon, J. B., Bradley, J., Brownlee, D., & Busemann, H. A., et al. 2006, Isotopic compositions of cometary matter returned by stardust. Science, 314, 1724

Meyer, M. R., Carpenter, J. M., Mamajek, E. E., Hillenbrand, L. A., & Hollenbach, D., et al. 2008, Evolution of mid-infrared excess around

sun-like stars: constraints on models of terrestrial planet formation. ApJL, 673, L181

Morales, F. Y., Werner, M. W., Bryden, G., Plavchan, P., & Stapelfeldt, K., et al. 2009, Spitzer Mid-IR spectra of dust debris around A and late B type stars: asteroid belt analogs and power-law dust distributions. ApJ, 699, 1067

Moro-Martín, A., & Malhotra, R. 2002, A study of the dynamics of dust from the Kuiper belt: spatial distribution and spectral energy distribution. AJ, 124, 2305

Moro-Martín, A., & Malhotra, R. 2003, Dynamical models of Kuiper belt dust in the inner and outer Solar system. AJ, 125, 2255

Moro-Martín, A., & Malhotra, R. 2005, Dust outflows and inner gaps generated by massive planets in debris disks. ApJ, 633, 1150

Moro-Martín, A., Wolf, S., & Malhotra, R. 2005, Signatures of planets in spatially unresolved debris disks. ApJ, 621, 1079

Moro-Martín, A., Carpenter, J. M., Meyer, M. R., Hillenbrand, L. A., & Malhotra, R., et al. 2007a, Are debris disks and massive planets correlated? ApJ, 658, 1312

Moro-Martín, A., et al. 2007b, The dust, planetesimals, and planets of HD 38529. ApJ, 668, 1165

Moro-Martín, A., Wyatt, M. C., Malhotra, R., & Trilling, D. E. 2008, Extra-solar Kuiper belt dust disks, in The Solar System Beyond Neptune, ed. A. Barucci, H. Boehnhardt, D. Cruikshank, & A. Morbidelli (Tucson: University of Arizona Press), 465–482 (arXiv:astro-ph/0703383)

Moro-Martín, A., Malhotra, R., Bryden, G., Rieke, G. H., Su, K. Y. L., Beichman, C. A., & Lawler, S. M. 2010, Locating planetesimal belts in the multiple-planet systems HD 128311, HD 202206, HD 82943, and HR 8799. ApJ, 717, 1123

Mouillet, D., Larwood, J. D., Papaloizou, J. C. B., & Lagrange, A. M. 1997, A planet on an inclined orbit as an explanation of the warp in the Beta Pictoris disc. MNRAS, 292, 896

Mukai, T., Blum, J., Nakamura, A. M., Johnson R. E., & Havnes, O. 2001, Physical processes on interplanetary dust, in Springer A&A Libr., Interplanetary Dust, ed. E. Grün, B. A. S. Gustafson, S. F. Dermott, & H. Fechtig (Berlin/New York: Springer), 445

Müller, S., Löhne, T., & Krivov, A. V. 2010, The debris disk of vega: a steady-state collisional cascade, naturally. ApJ, 708, 1728

Murray, C. D., & Dermott, S. F. 1999, Solar System Dynamics (Cambridge: Cambridge University Press)

Nesvorný, D., Jenniskens, P., Levison, H. F., Bottke, W. F., Vokrouhlický, D., & Gounelle, M. 2010, Cometary origin of the zodiacal cloud and carbonaceous micrometeorites. Implications for hot debris disks. ApJ, 713, 816

Plavchan, P., Jura, M., & Lipscy, S. J. 2005, Where are the M dwarf disks older than 10 million years? ApJ, 631, 1161

Quillen, A. C. 2006, Predictions for a planet just inside Fomalhaut's eccentric ring. MNRAS, 372, L14

Raymond, S. N., Armitage, P. J., Moro-Martín, A., Booth, M., Wyatt, M. C., Armstrong, J. C., Mandell, A. M., Selsis, F., & West, A. A. 2011, Debris disks as signposts of terrestrial planet formation. A &A, 530, A62

Raymond, S. N., Armitage, P. J., Moro-Martín, A., Booth, M., Wyatt, M. C., Armstrong, J. C., Mandell, A. M., Selsis, F., & West, A. A. 2012, Debris disks as signposts of terrestrial planet formation. II Dependence of exoplanet architectures on giant planet and disk properties. arXiv:1201.3622 (A&A in press)

Reach, W. T., Franz, B. A., Weiland, J. L., Hauser, M. G., & Kelsall, T. N., et al. 1995, Observational confirmation of a circumsolar dust ring by the COBE satellite. Nature, 374, 521

Reach, W. T., Morris, P., Boulanger, F., & Okumura, K. 2003, The mid-infrared spectrum of the zodiacal and exozodiacal light. Icarus, 164, 384

Rieke, G. H., Young, E. T., Engelbracht, C. W., Kelly, D. M., & Low, F. J., et al. 2004, The multiband imaging photometer for Spitzer (MIPS). ApJS, 154, 25

Sandford, S. A., Aléon, J., Alexander, C., Araki, T., & Bajt, S., et al. 2006, Organics captured from comet 81P/Wild 2 by the stardust spacecraft. Science, 314, 1720

Schneider, G., Silverstone, M. D., & Hines, D. C. 2005, Discovery of a nearly edge-on disk around HD 32297. ApJ, 629, L117

Siegler, N., Muzerolle, J., Young, E. T., Rieke, G. H., Mamajek, E. E., Trilling, D. E., Gorlova, N., & Su, K. Y. L. 2007, Spitzer 24 μm observations of open cluster IC 2391 and debris disk evolution of FGK stars. ApJ, 654, 580

Stapelfeldt, K. R., Holmes, E. K., Chen, C., Rieke, G. H., & Su, K. Y. L., et al. 2004, First look at the Fomalhaut debris disk with the Spitzer space telescope. ApJS, 154, 458

Stark, C. C., & Kuchner, M. J. 2009, A new algorithm for self-consistent three-dimensional modeling of collisions in dusty debris disks. ApJ, 707, 543

Stern, S. A. 1996, Signatures of collisions in the Kuiper disk. A&A, 310, 999

Strom, R. G., Malhotra, R., Ito, T., Yoshida, F., & Kring, D. A. 2005, The origin of planetary

impactors in the inner solar system. Science, 309, 1847

Su, K. Y. L., Rieke, G. H., Misselt, K. A., Stansberry, J. A., & Moro-Martín, A., et al. 2005, The vega debris disk: a surprise from Spitzer. ApJ, 628, 487

Su, K. Y. L., Rieke, G. H., Stansberry, J. A., Bryden, G., & Stapelfeldt, K. R., et al. 2006, Debris disk evolution around A stars. ApJ, 653, 675

Sykes, M. V., & Greenberg, R. 1986, The formation and origin of the IRAS zodiacal dust bands as a consequence of single collisions between asteroids. Icarus, 65, 51

Thébault, P., & Augereau, J.-C. 2007, Collisional processes and size distribution in spatially extended debris discs. A&A, 472, 169

Trilling, D. E., et al. 2008, Debris disks around sun-like stars. ApJ, 674, 1086

Vitense, C., Krivov, A. V., Kobayashi, H., & Löhne, T. 2012, An improved model of the Edgeworth-Kuiper debris disk. A&A, 540, A30

Weingartner, J. C., & Draine, B. T. 2001, Dust grain-size distributions and extinction in the milky way, large magellanic cloud, and small magellanic cloud. ApJ, 548, 296

Wisdom, J. 1980, The resonance overlap criterion and the onset of stochastic behavior in the restricted three-body problem. AJ, 85, 1122

Wyatt M. C. 2005, The insignificance of P-R drag in detectable extrasolar planetesimal belts. A&A, 433, 1007

Wyatt, M. C. 2008a, Dynamics of small bodies in planetary systems. in Lect. Notes Phys. 758, Small Bodies in Planetary Systems, ed. I. Mann, A. Nakamura, & T. Mukai (Berlin/London: Springer) (arXiv:astro-ph/0807.1272)

Wyatt, M. C. 2008b, Evolution of debris disks. ARA&A, 46, 339

Wyatt, M. C., Dermott, S. F., Telesco, C. M., Fisher, R. S., Grogan, K., Holmes, E. K., & Piña, R. K. 1999, How observations of circumstellar disk asymmetries can reveal hidden planets: pericenter glow and its application to the HR 4796 disk. ApJ, 527, 918

Wyatt, M. C., Smith, R., Greaves, J. S., Beichman, C. A., Bryden, G., & Lisse, C. M. 2007, Transience of hot dust around sun-like Stars. ApJ, 658, 569

Yamamoto, S., & Mukai, T. 1998, Dust production by impacts of interstellar dust on Edgeworth-Kuiper belt objects. A&A, 329, 785

Zolensky, M. E., Zega, T. J., Yano, H., Wirick, S., & Westphal, A., et al. 2006a, Mineralogy and petrology of comet 81P/Wild 2 nucleus samples. Science, 314, 1735

Zolensky, M., Bland, P., Brown, P. & Halliday, I. 2006b, Flux of extraterrestrial materials, in Meteorites and the Early Solar System II, ed. D. S. Lauretta, & H. Y. McSween Jr. (Tucson: University of Arizona Press), 869

Zook, H. A., & Berg, O. E. 1975, A source for hyperbolic cosmic dust particles. P&SS, 23, 183

10 Exoplanet Detection Methods

Jason T. Wright [1] · *B. Scott Gaudi* [2]
[1]Department of Astronomy and Astrophysics, Penn State University, University Park, PA, USA
[2]Department of Astronomy, The Ohio State University, Columbus, OH, USA

1 Basic Principles of Planet Detection . . . 491
1.1 Spectroscopic Binaries and Orbital Elements . . . 491
1.2 Radial Velocities . . . 494
1.3 Astrometry . . . 495
1.4 Imaging . . . 496
1.5 Transits . . . 498
1.6 Gravitational Microlensing . . . 500
1.7 Timing . . . 503

2 The Magnitude of the Problem . . . 504
2.1 Radial Velocities . . . 504
2.2 Astrometry . . . 506
2.3 Imaging . . . 507
2.4 Transits . . . 509
2.5 Microlensing . . . 511
2.6 Timing . . . 513

3 Comparisons of the Methods . . . 513
3.1 Sensitivities of the Methods . . . 515
3.2 Habitable Planets . . . 518

4 Early Milestones in the Detection of Exoplanets . . . 522
4.1 Van de Kamp and Barnard's Star . . . 522
4.2 PSR 1257+12 and the Pulsar Planets . . . 522
4.3 Early Radial Velocity Work . . . 523
4.3.1 Campbell and Walker's Survey and γ Cep A*b* . . . 523
4.3.2 Latham's Survey and HD 114762 *b* . . . 524
4.3.3 Marcy and Butler's Iodine Survey . . . 525
4.3.4 Hatzes and Cochran's Survey and β Gem *b* . . . 526
4.3.5 Mayor and Queloz and 51 Pegasi *b* . . . 527
4.4 The First Planetary Transit: HD 209458*b* . . . 527

T.D. Oswalt, L. French, P. Kalas (eds.), *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems*, DOI 10.1007/978-94-007-5606-9_10, © Springer Science+Business Media Dordrecht 2013

4.5 Microlensing 528
4.5.1 Microlensing History 528
4.5.2 First Planet Detections with Microlensing 529

5 State of the Art 529
5.1 Astrometry 529
5.2 Imaging 531
5.2.1 2M1207*b* 531
5.2.2 Fomalhaut *b* 532
5.2.3 β Pictoris *b* 533
5.2.4 The HR 8799 Planetary System 533
5.2.5 SPHERE, GPI, and Project 1640 534
5.3 Rocky and Habitable Worlds 534
5.3.1 HARPS, Keck/HIRES, and the Planet Finding Spectrograph 534
5.3.2 Space-Based Transit Surveys 535
5.3.3 Second-Generation Microlensing Surveys 536

6 Conclusions 537

Acknowledgments 537

References 538

Abstract: This chapter reviews various methods of detecting planetary companions to stars from an observational perspective, focusing on radial velocities, astrometry, direct imaging, transits, and gravitational microlensing. For each method, this chapter first derives or summarizes the basic observable phenomena that are used to infer the existence of planetary companions as well as the physical properties of the planets and host stars that can be derived from the measurement of these signals. This chapter then outlines the general experimental requirements to robustly detect the signals using each method, by comparing their magnitude to the typical sources of measurement uncertainty. This chapter goes on to compare the various methods to each other by outlining the regions of planet and host star parameter space where each method is most sensitive, stressing the complementarity of the ensemble of the methods at our disposal. Finally, there is a brief review of the history of the young exoplanet field, from the first detections to current state-of-the-art surveys for rocky worlds.

1 Basic Principles of Planet Detection

This chapter begins by reviewing the basic phenomena that are used to detect planetary companions to stars using various methods, namely, radial velocities, astrometry, transits, timing, and gravitational microlensing. It derives the generic observables for each method from the physical parameters of the planet/star system. These then determine the physical parameters that can be inferred from the planet/star system for the general case.

Notation with subscripts $*$ and p refers to quantities for the star and planet, respectively. Therefore, a star has mass M_*, radius R_*, mean density ρ_*, surface gravity g_*, and effective temperature T_* and is orbited by a planet of mass M_p, radius R_p, density ρ_p, temperature T_p, and surface gravity g_p. The orbit has a semimajor axis a, period P, and eccentricity e.

1.1 Spectroscopic Binaries and Orbital Elements

Exoplanet detection is essentially the extreme limit of binary star characterization, and so it is unsurprising that the terminology and formalism of planetary orbits derives from that of binaries.

Conservation of momentum requires that as a planet orbits a distant star, the star executes a smaller, opposite orbit about their common center of mass. The size (and velocity) of this orbit is smaller than that of the planet by a factor of the ratio of their masses. The component of this motion along the line of sight to the Earth can, in principle, be detected as a variable radial velocity. The mass of the exoplanet can be calculated from the magnitude of the radial velocity (RV) or astrometric variations and from the mass of the star, determined from stellar models and spectroscopy or astrometry.

Two mutually orbiting bodies revolve in ellipses about a common center of mass, the origin of our coordinate system. Orbital angles in the plane of the bodies' mutual orbit are measured with respect to the line of nodes, formed by the intersection of the orbital plane with the plane of the sky (i.e., the plane perpendicular to a line connecting the observer to the system's center of mass). The position of this line on the sky has angle Ω, representing the position angle (measured east of north) of the ascending (receding) node, where the star (and planet) cross the plane of the sky moving away from Earth. ❯ *Figure 10-1* illustrates the other orbital elements in

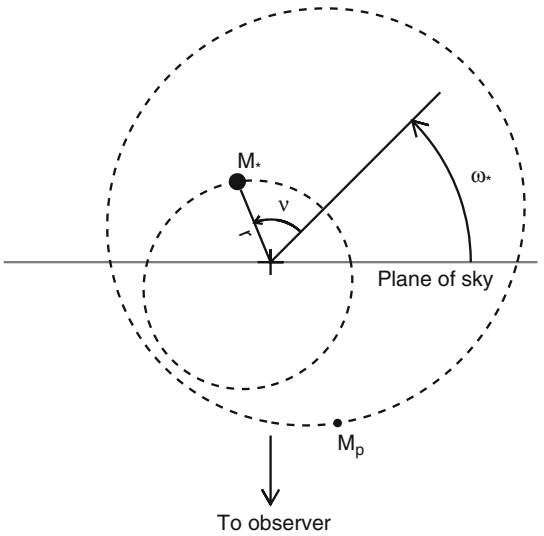


Fig. 10-1

Elements describing orbital motion in a binary with respect to the center of mass (cross). The argument of periastron ω is measured from the ascending (receding) node, and the true anomaly ν is measured with respect to the periastron. Both angles increase along the direction of the star's motion in the plane of the orbit. The longitude of the periastron of the star ω_* is indicated. At a given time in the orbit, true anomalies of planet and star are equal, whereas the longitude of periastron of the planet is related to that of the star by $\omega_p = \omega_* + \pi$. In Doppler planet detection, the orbital elements of the star are conventionally reported, from which the orbital elements of the planet can be inferred

the problem. As indicated, the orientation of the each orbital ellipse with respect to the plane of the sky is specified by the longitude of periastron, ω, which is the angle between the periastron[1] and the ascending node along the orbit in the direction of the motion of the body. Since the orbit of the star is a reflection about the origin of the orbit of the exoplanet, the orbital parameters of the planet are identical to that of the star except that the longitudes of periastron differ by π: $\omega_p = \omega_* + \pi$.

The physical size of the ellipse, given by the semimajor axis, a, is set by Newton's modification of Kepler's third law of planetary motion:

$$P^2 = \frac{4\pi^2}{G(M_* + M_p)} a^3 \tag{10.1}$$

The semimajor axis is $a = a_* + a_p$, where a_* and a_p are the semimajor axes of the two bodies' orbits with respect to the center of mass, given by

$$a_* = \frac{M_p}{M_* + M_p} a; \qquad a_p = \frac{M_*}{M_* + M_p} a. \tag{10.2}$$

[1] Periastron marks the point where the two bodies have their closest approach.

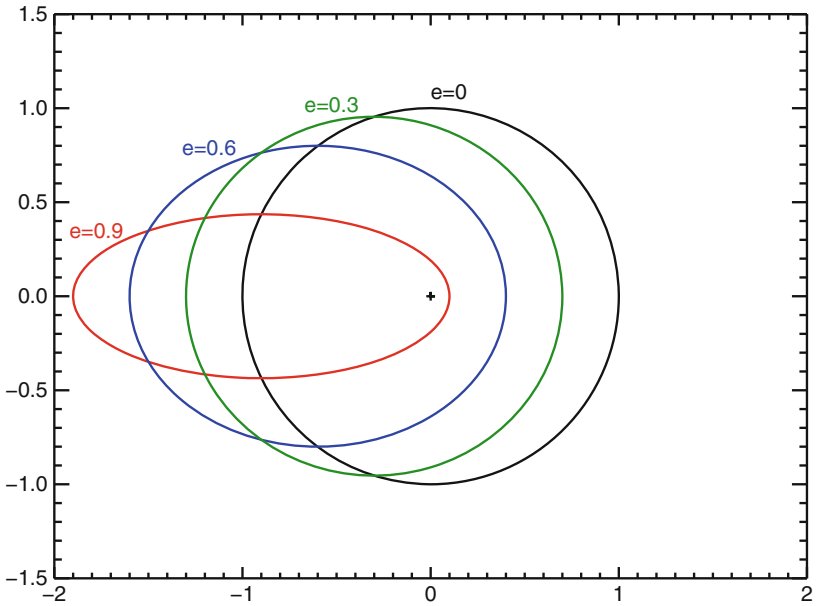


Fig. 10-2
Shape of various eccentric orbits in the orbital plane. A handful of exoplanets with eccentricities above 0.9 have been detected

The position of either body in its orbit about the origin can be expressed in polar coordinates (r, v), where v is the true anomaly, the angle between the location of the object and the periastron. The separation between the star and planet is given by

$$r(1 + e\cos v) = a(1 - e^2), \tag{10.3}$$

where e is called the eccentricity of the orbit and has the domain $[0,1)$ for bound orbits. The observed eccentricities of exoplanets are quite varied: eccentricities above 0.9 have been seen in a few cases and eccentricities above 0.3 are common, at least for Jovian exoplanets (Wright et al. 2011). *Figure 10-2* illustrates the physical shape of such orbits.

Practical computation of a body's position in its orbit with time is usually performed through the intermediate variable E, called the eccentric anomaly. E is related to the time since periastron passage T_0 through the mean anomaly, M:

$$M = \frac{2\pi(t - T_0)}{P} = E - e\sin E. \tag{10.4}$$

and allows the computation of v through the relation

$$\tan\frac{v}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2}. \tag{10.5}$$

The eccentric anomaly is also useful because it is simply related to r:

$$E = \arccos\frac{1 - r/a}{e}. \tag{10.6}$$

1.2 Radial Velocities

The radial reflex motion of a star in response to an orbiting planet can be measured through precise Doppler measurement, and this motion reveals the period, distance, and shape of the orbit and provides information about the orbiting planet's mass. (The treatment of RV and astrometric measurement below follows Wright and Howard (2009)).

Six parameters determine the functional form of the periodic radial velocity variations and thus the observables in a spectroscopic orbit of the star: P, K, e, ω_*, T_0, and γ (it is convention in the Doppler-detection literature to refer to ω without its $*$ subscript, but it is standard to report the star's argument of periastron, not the planet's).

$$V_r = K\left[\cos(\nu + \omega_*) + e\cos\omega_*\right] + \gamma \tag{10.7}$$

with ν related to P, e, and T_0 through E. The semiamplitude of the signal in units of velocity is K (the peak-to-trough RV variation is $2K$). The bulk velocity of the center of mass of the system is given by γ.

For circular orbits $e = 0$, there is no periastron approach, and so T_0 and ω_* are formally undefined; in such cases, a nominal value of ω_* (such as 0 or $\pi/2$) sets T_0 (alternatively, one can specify the value of one of the angles at a given epoch).

In short, the variables P, T_0, and K, respectively, set the period, phase, and amplitude of an RV curve, while the variables ω_* and e determine the shape of the radial velocity signature of an orbiting companion, as shown in ❯ *Fig. 10-3*. Characterization of the orbits of single unseen companions, such as exoplanets, is ultimately an exercise in fitting observed radial velocities to the family of curves in ❯ *Fig. 10-3* to determine the six orbital parameters.

Two additional orbital parameters complete the description of a planet's orbit: the inclination of the orbit, i, which determines the angle between the orbital plane and the plane of the sky, such that $i = 0$ corresponds to a face-on, counterclockwise orbit and Ω, the longitude of the ascending node. These parameters cannot be determined with radial velocity observations alone and can only be measured through astrometry, where the angular displacement of the star on the sky is directly measured.

The effect of the inclination of the orbit is to reduce the radial component of the velocity of the star by $\sin i$. The fundamental observable of a spectroscopic binary which constrains the physical properties of system is thus

$$\frac{PK^3(1-e^2)^{\frac{3}{2}}}{2\pi G} = \frac{M_p^3 \sin^3 i}{(M_p + M_*)^2}, \tag{10.8}$$

where G is Newton's gravitational constant. The right-hand side of this equation is known as the *mass function* of the system. For exoplanets where M_* can be estimated from stellar models, the minimum value for M_p (i.e., its value for $\sin i = 1$ or an edge-on orbit) is called the "minimum mass" of the planet and is usually labeled "$M_p \sin i$" for succinctness (since when $M_p \ll M_*$, its small correction to the denominator is negligible, though not ignored). The true mass of the detected exoplanet is thus higher by a factor of $1/\sin i$, which has a typical (median) value of 1.15 for randomly oriented orbits (all other things being equal).

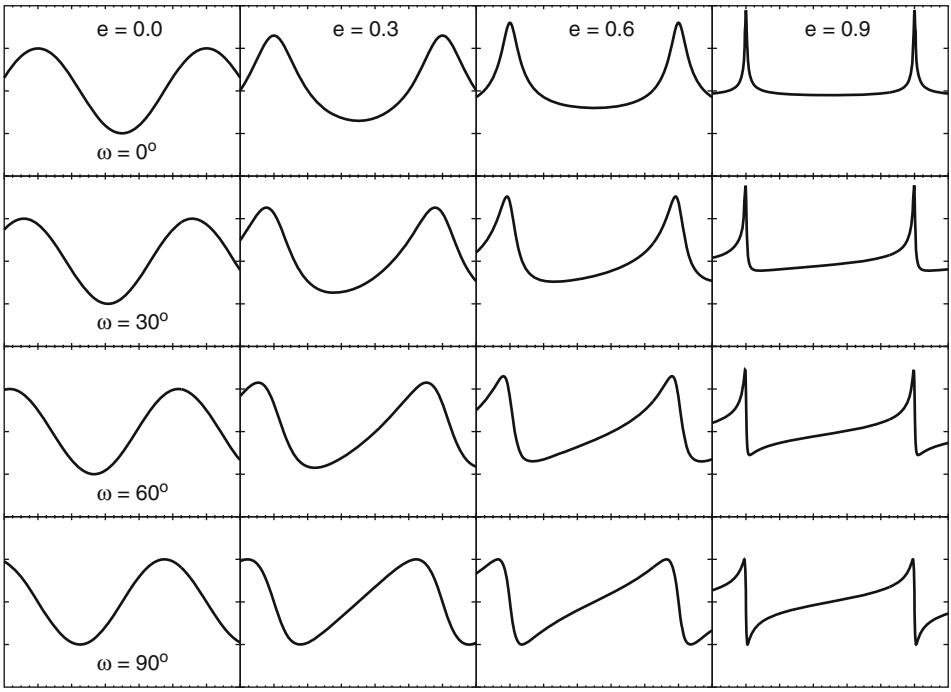


Fig. 10-3

The effects of *e* and ω_* on radial velocity curves. These curves have been scaled to unit *K* and common *P* and T_0. *Each column* shows curves of constant *e* and *each row* curves of constant ω_* as indicated. Other quadrants of ω_* yield reflections of these curves

1.3 Astrometry

Plane-of-sky variations in a star's position provide both redundant and complementary information to radial velocities, yielding the true inclination and orientation of a planetary orbit. Astrometry of the orbits of well-separated binary stars of similar magnitude is a matter of careful instrument calibration to precisely measure the separation and position angle between the stars. For exoplanet detection, the problem is to detect the motions around a star about an unseen companion with respect to a set of (presumably) stable background stars.

For an orbit with semimajor axis a_* of a star at distance d from the Earth, producing an astrometric signal of semiamplitude $\theta_* = a_*/d$, astrometric orbits can be described in terms of the Thiele-Innes constants

$$A = \theta_*(\cos\Omega\cos\omega_* - \sin\Omega\sin\omega_*\cos i) \tag{10.9}$$

$$B = \theta_*(\sin\Omega\cos\omega_* + \cos\Omega\sin\omega_*\cos i) \tag{10.10}$$

$$F = \theta_*(-\cos\Omega\sin\omega_* - \sin\Omega\cos\omega_*\cos i) \tag{10.11}$$

$$G = \theta_*(-\sin\Omega\sin\omega_* + \cos\Omega\cos\omega_*\cos i) \tag{10.12}$$

$$C = \theta_*\sin\omega_*\sin i \tag{10.13}$$

$$H = \theta_*\cos\omega_*\sin i \tag{10.14}$$

which can be quickly computed using rotation matrices:

$$\begin{bmatrix} A & B & C \\ F & G & H \\ \theta_* \sin i \sin \Omega & -\theta_* \sin i \cos \Omega & \theta_* \cos i \end{bmatrix} = \theta_* R_z(\omega_*) R_x(i) R_z(\Omega), \tag{10.15}$$

where R is the 3-D rotation matrix, given in the case of the z-axis by

$$R_z(\Omega) = \begin{bmatrix} \cos\Omega & \sin\Omega & 0 \\ -\sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{10.16}$$

The Thiele-Innes constants are related back to Keplerian orbital elements with the relations:

$$\tan(\omega_* + \Omega) = \frac{B - F}{A + G} \tag{10.17}$$

$$\tan(\omega_* - \Omega) = \frac{-(B + F)}{A - G} \tag{10.18}$$

$$\tan^2\left(\frac{i}{2}\right) = \frac{(A - G)\cos(\omega_* + \Omega)}{(A + G)\cos(\omega_* - \Omega)} \tag{10.19}$$

$$\begin{aligned} \theta_* = \quad & (A\cos\omega_* - F\sin\omega_*)\cos\Omega - \\ & (A\sin\omega_* + F\cos\omega_*)\sin\Omega \sec i \end{aligned} \tag{10.20}$$

$$\theta_*^2 = A^2 + B^2 + C^2 = F^2 + G^2 + H^2. \tag{10.21}$$

The quantities ω_* and Ω have a $\pm\pi$ ambiguity that is resolved with radial velocity measurements, without which convention dictates that we choose $\Omega < \pi$.

The C and H constants are related to the radial component of the motion. These constants can be combined with the elliptical rectangular coordinates, defined as

$$X = \cos E - e \tag{10.22}$$

$$Y = \sqrt{1 - e^2}\sin E \tag{10.23}$$

to describe the astrometric displacements of a star in the north ($\Delta\delta$) and east ($\Delta\alpha\cos\delta$) directions,

$$\begin{aligned} \Delta\delta &= AX + FY \\ \Delta\alpha\cos\delta &= BX + GY, \end{aligned} \tag{10.24}$$

and the magnitude of the astrometric offset from the apparent center of mass is (for small offsets) $\Delta\theta_* \equiv [\Delta\delta^2 + (\Delta\alpha\cos\delta)^2]^{1/2}$. In practice, astrometric motions are small perturbations on the much larger parallactic and proper motions.

1.4 Imaging

The direct detection of planets is the most conceptually straightforward method of detection: essentially, one seeks simply to directly detect photons from the exoplanet, resolved from those of the parent star. Although the emission of exoplanets is indeed quite faint, it is generally the problem of detecting this emission in the proximity of the much brighter stellar source that presents the most severe practical obstacle to direct detection. The disentangling of stellar and

planetary photons is an imperfect process that is easiest at wider separations. The efficiency of this disentangling ultimately determines the detection thresholds of an instrument. Therefore, the most important parameters of the exoplanet for determining the difficulty of direct detection are the planet/star flux ratio f_p and the angular separation between the planet and star. Typically, contrast limits worsen at smaller angular separations.

The angular separation of the planet and star on the sky is given by

$$\Delta\theta = r_\perp/d, \tag{10.25}$$

where $r_\perp$ is the projected separation of the planet from the star and d is the distance to the system. By definition, if d is in parsecs and $r_\perp$ in AU, then θ is in arcseconds. In general, $\Delta\theta = (1+M_*/M_p)\Delta\theta_* = (1+M_*/M_p)\sqrt{(BX+GY)^2+(AX+FY)^2}$. For circular orbits, this reduces to $r_\perp = a(\cos^2\beta + \sin^2\beta\cos^2 i)^{1/2}$, where $\beta = \nu + \omega_p$ is the angle between the position of the planet along its orbit relative to the ascending node. Planets typically orbit stars at distances from hundredths to hundreds of AU. For a hypothetical giant planet orbiting 5 AU from a nearby star sitting at 50 pc, this corresponds to a maximal angular separation of 100 mas.

The emission from an exoplanet can generally be separated into two sources: stellar emission reflected by the planet surface and/or atmosphere and thermal emission from the planet. Thermal emission can be due to either "intrinsic" thermal emission (e.g., the fossil heat of formation) or thermal emission from reprocessed stellar luminosity. Exoplanets may also produce some nonthermal emission, which we will not consider here.

The reflected light will have a spectrum that is broadly similar to that of the star, with additional features arising from the planetary surface and/or atmosphere. Therefore, for solar-type stars, this reflected emission generally peaks at optical wavelengths. The monochromatic planet/star flux ratio for reflected light can generally be written (e.g., Seager 2010)

$$f_{\mathrm{ref},\lambda} = A_{g,\lambda}\left(\frac{R_p}{a}\right)^2 \Phi_{\mathrm{ref},\lambda}(\alpha), \tag{10.26}$$

where $A_{g,\lambda}$ is the monochromatic geometric albedo and $\Phi_{\mathrm{ref},\lambda}$ is the reflected light phase curve, which depends on the planetary phase angle α (the star-planet-observer angle) and the wavelength λ. The geometric albedo is defined as the ratio of the flux emitted from the planet at $\alpha = 0$ relative to that of a perfectly and isotropically scattering uniform disk of equal solid angle. For a circular orbit, $\cos\alpha = \sin\beta\sin i$.

Assuming that the thermal emission from the planet has a roughly blackbody spectrum, the flux ratio is

$$f_{\mathrm{therm},\lambda} = \left(\frac{R_p}{R_*}\right)^2 \frac{B_\lambda(T_p)}{B_\lambda(T_*)}\Phi_{\mathrm{therm},\lambda}(\alpha) \rightarrow \left(\frac{R_p}{R_*}\right)^2 \frac{T_p}{T_*}\Phi_{\mathrm{therm},\lambda}(\alpha), \tag{10.27}$$

where $\Phi_{\mathrm{therm},\lambda}$ is the monochromatic thermal phase curve. For observations in the Raleigh-Jeans tail of the blackbody, $\lambda \gg hc/(k_b T)$, and thus $B_\lambda(T) \propto T$, yielding the limit shown in (➔ 10.27). If the planet is in thermal equilibrium with the stellar radiation, then $T_p = T_{\mathrm{eq}}$ and

$$\frac{T_{\mathrm{eq}}}{T_*} = \left(\frac{R_*}{a}\right)^{1/2}[f(1-A_B)]^{1/4}, \tag{10.28}$$

where A_B is the Bond albedo, the fraction of the total energy incident on the planet that is not absorbed, and f accounts the fraction of the entire planet surface over which the absorbed energy is reemitted, i.e., $f = 1/4$ if the thermal energy is emitted over the entire 4π of the planet.

Of course, planets may be self-luminous as well, particularly if they are young and have retained significant residual heat from formation.

The form for $\Phi_{\mathrm{ref},\lambda}$ depends on the scattering properties of the planetary atmosphere. For the case of a Lambert sphere that scatters all incident radiation equally in all directions,

$$\Phi_{\mathrm{Lambert},\lambda} = \frac{1}{\pi}\left[\sin\alpha + (\pi - \alpha)\cos\alpha\right]. \tag{10.29}$$

Also for a Lambert sphere, $A_B = 1$ and $A_g = 2/3$. The form for $\Phi_{\mathrm{therm},\lambda}$ depends on the surface brightness distribution of the planet, which in turn depends on the amount of heat redistribution. For the case of a tidally locked planet in which the absorbed radiation is promptly and locally reemitted, the phase curve has the same form as for a Lambert sphere (Seager 2010).

Resolved emission of an planet/star system is essentially equivalent to a visual binary. Once the overall scale of the system has been set, measurements of the position of the planet relative to the star at a sufficient number of epochs yield all of the orbital elements of the system, up to the twofold degeneracy in orientation with respect to the sky discussed previously. The scale of the system can be set either by an estimate of the distance to the system or by an external estimate of the primary mass M_* (under the assumption $M_p \ll M_*$). For reflected light measurements, only the product of the geometric albedo and planet cross section can be determined; estimating the planet radius independently generally requires an assumption about the albedo. For thermal light measurements, the temperature T_p can (in principle) be estimated from the flux at multiple wavelengths, and then, the surface brightness can be estimated from T_p, and thus the radius can be inferred from the planetary flux. The planet mass cannot be determined from the planet flux or its relative orbit and must be inferred indirectly through coupled atmosphere/evolutionary models. In some favorable cases, mutual gravitational perturbations in multiplanet systems may allow the determination of the planet masses directly.

Of course, the real power of direct imaging lies in the ability to acquire spectra of the planets once they are discovered and thus characterize the constituents of the planetary atmosphere. This provides one of the only feasible routes to assessing the habitability of terrestrial planets in the Habitable Zones (Kasting et al. 1993) of the parent stars and likely the *only* feasible route to do so for Earthlike planets orbiting solar-type stars.

1.5 Transits

The presence of a planetary companion to a star gives rise to a multitude of physical phenomena that manifest themselves via temporal variations of the flux of the system relative to that of an otherwise identical isolated star. Typically, the largest of these occurs if a fortunate alignment allows a planet to transit (pass in front of) its host star from our perspective. In this case, the star will exhibit brief, periodic dimmings which signal the presence of the planet. Transits offer a intriguingly simple way to detect planets.

The condition for a transit is roughly that the projected separation between the planet and host star at the time of inferior conjunction of the planet is less than the sum of the radii of planet and star, i.e., $r(t_c)\cos i \leq R_* + R_p$, where $r(t_c)$ is the separation of the planet from the host star at conjunction, and R_* and R_p are the radii of the star and planet, respectively. Given the definition of ω_*, $r(t_c) = a(1 - e^2)/(1 + e\sin\omega_*)$, and so transits occur when the

impact parameter of the planet's orbit with respect to the star in units of the host radius,

$$b \equiv \frac{a \cos i}{R} \frac{1 - e^2}{1 + e \sin \omega_*}, \tag{10.30}$$

is less than the sum of the (normalized) radii, $b \leq 1 + k$, where $k \equiv R_p/R_*$. Integrating over i assuming isotropic orbits and thus a uniform distribution of cos i yields the *transit probability*

$$P_{\rm tr} \equiv \left(\frac{R_* + R_p}{a} \right) \frac{1 + e \sin \omega_*}{1 - e^2}. \tag{10.31}$$

For a circular orbit and assuming $k \ll 1$, this reduces to the simple expression $P_{\rm tr} = R_*/a$. Note that in these expressions, we have used the longitude of the periastron of the orbit of star rather than the (perhaps more intuitive) value for the planet, because the former is generally adopted for fits to the stellar reflex radial velocity data.

When the planet transits in front of its parent star, the flux of the star will decrease by an amount that is proportional to the ratio of the areas of the planet and star. For the purposes of exposition, in the following, we assume a circular orbit, uniform host surface brightness, and $R_p \ll R_* \ll a$ and $M_p \ll M_*$. In the general case of a limb-darkened star, eccentric orbit, and arbitrary scales for R_p, R_*, and a, the expressions for the shape of the transit are considerably more complicated, as are the arguments for the kinds of information that can be extracted from transit and RV signals (see Winn 2010 and references therein). However, the basic structure of the problem is the same under our approximations, and what follows serves to illustrate the essential concepts.

Under these assumptions, the planet follows a rectilinear trajectory across the face of the star with an impact parameter b, and the transit signature will have an approximately trapezoidal shape, which can be characterized by the duration T, ingress/egress time τ, and fractional flux depth δ. The depth of the transit relative to the out-of-transit flux is

$$\delta = k^2. \tag{10.32}$$

The duration of the transit can be quantified by its full width at half maximum, which is roughly the time interval T between the two points where the center of the planet appears to touch the edges of the star. This is approximately

$$T \simeq T_{\rm eq}(1 - b^2)^{1/2}, \tag{10.33}$$

where it is useful to define the equatorial crossing time (i.e., the transit duration for $b = 0$),

$$T_{\rm eq} \equiv \frac{R_* P}{\pi a} = f_{\rm tr} P \simeq \left(\frac{3P}{\pi^2 G \rho_*} \right)^{1/3}. \tag{10.34}$$

Here, ρ_* is the mean density of the host star and $f_{\rm tr} \equiv P/\pi a = P_{\rm tr}/\pi$ is the transit *duty cycle*, or the fraction of planet orbit in transit. The last equality, which assumes $M_p \ll M_*$, also implies that, to an order of magnitude, the equatorial transit duration is the cube root of the product of the orbital period and the stellar dynamical or free-fall time ($t_{\rm dyn} \sim (G\rho_*)^{-1/2}$) squared.

The ingress/egress time (these are equal for a circular orbit) τ is the time between when the edge of the planet just appears to touch the star for the first and second time (ingress or the time between first and second "contact") or third and fourth time (egress, the time between third and fourth "contact") and is given by

$$\tau \sim T_{\rm eq} \delta^{1/2} (1 - b^2)^{-1/2}. \tag{10.35}$$

One of the most useful aspects of transiting planets is that, when combined with radial velocity data, they allow one to infer the masses and radii of the star and planet up to a one-parameter degeneracy, as follows. Measuring T, t, and δ from a single transit allows one to infer b, $T_{\rm eq}$, and k:

$$b^2 = 1 - \delta^{1/2}\frac{T}{\tau}, \qquad T_{\rm eq}^2 = \frac{T\tau}{\delta^{1/2}}, \qquad k = \delta^{1/2}. \tag{10.36}$$

The impact parameter is related to the orbital inclination i via $b = a\cos i/R_*$, but a and R_* cannot be determined from light curves alone. With the detection of multiple transits, one can further infer the period P and thus the stellar density ρ_* via (❯ 10.34). As reviewed in ❯ Sect. 1.2, the reflex radial velocity orbit of the star allows one to infer K and P which can be combined to determine the mass function, $\sim (M_* \sin i)^3/M_*^{2/3}$, but a determination of the planet mass requires both a measurement of i and M_*. Thus, one additional parameter is needed to break the degeneracy and set the overall scale of the system. This can be accomplished by imposing external constraints on the properties of the primary, either through parameters measured from high-resolution spectroscopy or parallax, or invoking theoretical relations between the mass and radius of the star through isochrones or both. For illustration, if we assume the primary mass is precisely known, then we can infer R_* through ρ_*, and a through P, and thus determine i from the impact parameter measurement. Finally, we can measure R_p from k and the planet mass from the mass function, i, and M_*.

1.6 Gravitational Microlensing

The gravitational microlensing method detects planets via the direct gravitational perturbation of a background source of light by a foreground planet (see Gaudi 2010, 2012 for thorough reviews). When a foreground compact object (either a star or stellar remnant) happens to pass very close to our line of sight to a more distant star, the light from the background star will be split into two images. These images are typically unresolved, but they are magnified by an amount that depends on the angular separation between the lens and source. Since this separation is a function of time, the background source exhibits a smooth, symmetric time-variable magnification: a microlensing event. If the foreground planet happens to have a planetary companion and the planetary companion happens to have a projected separation from the primary lens near the paths of the two primary images, the gravity of the planet will further perturb the light, resulting in a short-lived perturbation from the primary microlensing event, revealing the planetary companion. Free-floating planets and planets widely separated from their parent star can also be detected as isolated, short timescale microlensing events.

Consider a planet/star system acting as a lens located at a distance d and source located at a distance d_s. Light from the source is deflected, split into multiple images, and magnified by the gravity of the foreground lens. The fundamental equation that is used to derive the observable properties of a gravitational microlensing event is the *lens equation*, which relates the angular separation $\boldsymbol{\beta}$ between the lens and source in the absence of lensing to the angular positions $\boldsymbol{\theta}$ of the images of the source created due to lensing. For a general lens system, these are vector quantities, but for a single lens, the lens, source, and image positions are all colinear, so we can drop the vector notation. The lens equation for an isolated point lens is (Einstein 1936)

$$\beta = \theta - \frac{4GM_*}{c^2 d_{\rm rel}\theta}, \tag{10.37}$$

where $d_{\rm rel}^{-1} \equiv d^{-1} - d_s^{-1}$. If the lens and source are perfectly aligned ($\beta = 0$), the source is imaged into a ring of radius equal to

$$\theta_{\rm E} \equiv \left(\frac{4GM_*}{d_{\rm rel}c^2}\right)^{1/2} \simeq 713\ \mu{\rm as} \left(\frac{M}{0.5M_\odot}\right)^{1/2} \left(\frac{d_{\rm rel}}{8\ {\rm kpc}}\right)^{-1/2}. \tag{10.38}$$

The Einstein ring radius is the fundamental scale of gravitational microlensing and depends on the distances to the lens and source, and the mass of the lens. At the distance of the lens, the linear Einstein ring radius is

$$r_{\rm E} \equiv \theta_{\rm E} d \simeq 2.85\ {\rm AU} \left(\frac{M_*}{0.5M_\odot}\right)^{1/2} \left(\frac{d_s}{8\ {\rm kpc}}\right)^{1/2} \left[\frac{x(1-x)}{0.25}\right]^{1/2}, \tag{10.39}$$

where $x \equiv d/d_s$.

Normalizing by $\theta_{\rm E}$, the lens (10.37) simplifies to

$$u = y - y^{-1}, \tag{10.40}$$

where $u \equiv \beta/\theta_{\rm E}$ and $y \equiv \theta/\theta_{\rm E}$. If $u \neq 1$, this has two solutions, $y_\pm = \pm\frac{1}{2}(\sqrt{u^2+4} \pm u)$, and thus, in general, an isolated point lens create two images. One of these images is always separated by more than $\theta_{\rm E}$ from the lens ($y_+ \geq 1$) and the other is always separated by less than $\theta_E(|y_-| \leq 1)$. The separation between the two images is ~$2\theta_{\rm E}$, and thus they are typically unresolved. Because the images are distorted relative to the source, they are also (de-)magnified. The total magnification for the sum of the two unresolved images is

$$A(u) = \frac{u^2+2}{u\sqrt{u^2+4}}. \tag{10.41}$$

The magnification increases for decreasing u (better source-lens alignment) and formally diverges as $u \to 0$ for a point source.

The source, lens, and observer are all in relative motion, and thus the angular separation between the source and lens is a function of time: a microlensing event. If we approximate the relative proper motion $\mu_{\rm rel}$ of the lens and source as constant, then we can parameterize the trajectory of the source relative to the lens as

$$u(t) = \left[\left(\frac{t-t_0}{t_{\rm E}}\right)^2 + u_0^2\right]^{1/2}, \tag{10.42}$$

where u_0 is the dimensionless angular separation at the time of closest approach to the lens (the impact parameter), t_0 is the time when $u = u_0$ (also the time of maximum magnification for a point lens), and $t_{\rm E}$ is the Einstein ring crossing time,

$$t_{\rm E} \equiv \frac{\theta_{\rm E}}{\mu_{\rm rel}}. \tag{10.43}$$

Figure 10-4 shows the source positions, image positions, and magnification of an example single-lens microlensing event with $u_0 = 0.2$. In general, the magnification as a function of time for a single-lens event has a smooth, symmetric form that is described by three parameters $(t_{\rm E}, t_0, u_0)$. Events with lower u_0 lead to more distorted images and higher magnification near peak. For $u_0 \ll 1$, the peak magnification is $A_{\rm max} \propto u_0^{-1}$. Events with $A_{\rm max} \gtrsim 100$ are typically referred to as "high-magnification events."

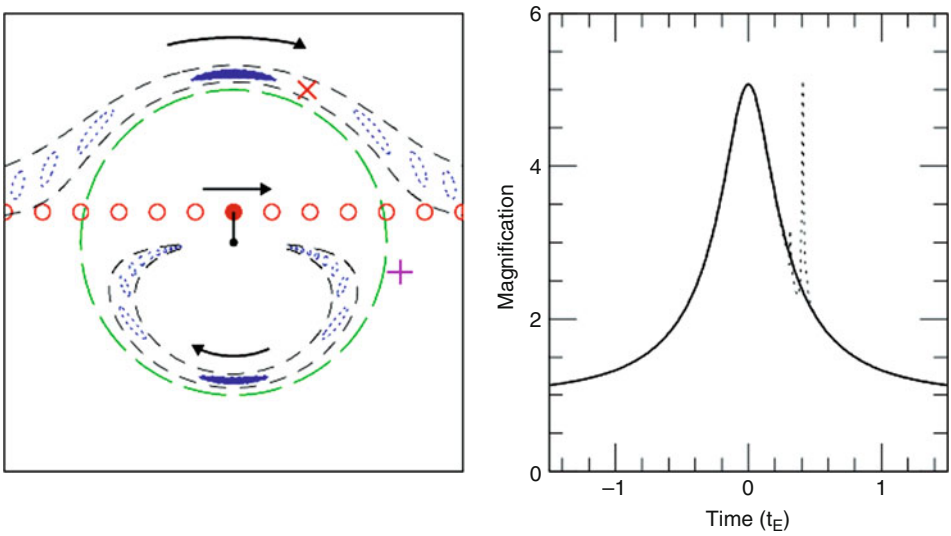


Fig. 10-4

The *left panel* shows the images (*dotted ovals*) for several different positions of the source (*solid circles*) for a microlensing event with an impact parameter of 0.2 Einstein ring radii. The primary lens is indicated as a small *black dot*, and the primary lens Einstein ring is indicated a *green long-dashed circle*. If the primary lens happen to have a planet near the path of one of the images (roughly within the *short-dashed lines*), then the planet will perturb the light from the source, creating a deviation to the single-lens light curve. *Right panel*: The magnification as a function of time is shown for the case of a single lens (*solid*) and accompanying planet (*dotted*) located at the position of the X in the *left panel*. If the planet was located at the + instead, then there would be no detectable perturbation, and the resulting light curve would be essentially identical to the *solid curve* (Adapted from Gaudi (2010) in Exoplanets, edited by Sara Seager, copyright 2012, The Arizona Board of Regents. Reprinted by permission of the University of Arizona Press)

Planetary companions to the lens star can be detected in a microlensing event if they happen to have a projected separation in the paths of one or both of the images created by the primary lens. As the image sweeps by the planet, the gravity of the planet will further perturb the light from the source associated with the image, creating a short-lived deviation from the single-lens form (Mao and Paczynski 1991; Gould and Loeb 1992).

Unfortunately, there are no simple analytic expressions relating the observable features of planetary perturbations to the underlying physical parameters of the planet and host star. Adding another body to the lens system increases the complexity of the lensing behavior significantly and in particular inverting the lens equation for a binary lens to obtain the image positions for a given source position cannot be done analytically. Furthermore, the binary gravitational lens has a rich and complex phenomenology, which we will not attempt to explore in this brief review. We refer the reader to more comprehensive summaries by Bennett (2008) and Gaudi (2012). Here, we will simply provide a qualitative discussion of planet detection with microlensing.

Three additional parameters are required to uniquely specify the light curve due to a binary lens (of which star/planet lenses are a subset). The planet/star mass ratio $q = M_p/M_*$ and instantaneous projected separation $s = r_\perp/r_\mathrm{E}$ between the planet and star in units of r_E at the time of the event together specify the magnification structure of the lens, i.e., the magnification as a function of the (vector) source position $\boldsymbol{u} \equiv \boldsymbol{\beta}/\theta_\mathrm{E}$. Finally, the parameter α (not to be confused with the phase angle) describes the orientation of the source trajectory relative to the projected planet/star axis. Thus a total of six parameters $(t_\mathrm{E}, t_0, u_0, q, s, \alpha)$ describe the magnification as a function of time for a binary lens and are thus generically observable.

Single-lens microlensing events yield only one parameter that depends on the physical properties of the lens star, namely, the time scale t_E. The time scale provides only a weak constraint on the lens mass, because it depends not only on the mass but also on the lens and source distances, and the relative lens-source proper motion, all of which are relatively broadly distributed for a typical microlensing survey. In addition, the lens stars are typically quite faint and are blended with other stars (including the lensed source). Thus, little is generally known about the host star properties. Planetary microlensing events generally yield two parameters that are related to the planet properties, q and s. While q is of interest in its own right, s is generally not, because it depends on the phase, orientation, and eccentricity of the orbit, as well as on the Einstein ring radius, all of which are a priori unknown. Therefore, s is only weakly correlated with the semimajor axis of the orbit and provides essentially no constraint on the other orbital elements.

Although this "baseline" situation sounds quite dire, in fact, it has been shown that with additional effort, it is possible to obtain substantially more information about the host star, planet, and its orbit for the majority of detected systems using a combination of subtle, higher-order effects that are detectable in precise microlensing light curves, and follow-up high-resolution imaging in order to isolate the light from the lens (Bennett et al. 2007).

1.7 Timing

A star or stellar remnant that exhibits regular, periodic photometric variability, such as pulsars, pulsating white dwarfs, eclipsing binary stars, pulsating hot subdwarfs, or even stars with transiting planets, can show evidence of a planetary companion through timing variations in those periodic phenomena. There are three principal sources of such variations: the Doppler shift, light-travel time, and gravitational perturbations.

The first of these sources is exactly analogous to the radial velocity method but measures changes in frequencies of some property other than photons. If the period of the pulsations or eclipses can be measured to sufficient precision, then the interpretation of those variations is identical to that in the radial velocity method.

The light-travel time effect comes about when the reflex orbit of a star about the center of mass of the star-planet system is sufficiently large that the additional light-travel time across this orbit is detectable as a timing variation. This is not a truly distinct phenomenon from the Doppler shift timing method, since it is essentially the accumulated effects of the Doppler-shifted period that produce the timing anomaly. Depending on the period of the intrinsic variation and the physical size of the star's reflex orbit, either effect, or both, may be detectable.

The above methods have been most successfully employed with pulsars (through the pulse arrival times) and eclipsing binary systems (through the timing of eclipse ingress and egress) and were responsible for the detection of the first exoplanets (Wolszczan and Frail 1992).

Finally, in the case of an eclipsing system, such as an eclipsing binary or a transiting planet, additional bodies in the system will perturb the orbits of the eclipsing bodies. These perturbations can be especially large if the perturbing body is near a mean-motion resonance with the other bodies. When applied to systems of transiting planets, this method is called *transit timing variations* (TTVs, Agol et al. 2005; Steffen and Agol 2005; Holman and Murray 2005) and has been most successfully employed by *Kepler* (e.g., Ford et al. 2012).

2 The Magnitude of the Problem

By almost any physical measure, planets are small in comparison to their parent stars, and the observable phenomena that are used to directly or indirectly detect them are likewise small. In this section, we attempt (where possible) to provide order-of-magnitude estimates of the precisions of the relevant observations that are required to detect planets using various methods. We then use these estimates, along with additional requirements imposed by the specifics of the detection method (i.e., the detection efficiencies) to provide a broad outline of the practical requirements that must be met for planet surveys to successfully detect planets with a given set of properties.

In general, specifying the criteria needed to detect a planet requires a detailed analysis of the signal and data properties as well as a quantitative definition of the meaning of a detection. However, for many of the detection methods, a rough estimate can be obtained by decomposing the primary observable signal into two conceptually different contributions: an overall scale and detailed signal waveform. The overall scale, which depends on the physical parameters of the system, encodes the order of magnitude of the signal and largely dictates its detectability. The waveform itself depends on more subtle details of the system (i.e., the precise shape of the planet orbit), but typically takes on values of order unity and thus has a relatively small effect on the detectability of the signal. Therefore, in most cases, these two contributions can be fairly cleanly separated. In this approximation, the detectability of a planet with a given set of properties therefore primarily depends primarily on the overall signal scale, and the data quality and quantity, i.e., the typical observational uncertainties and the total number of observations. With this in mind, given a signal amplitude A, number of observations N, and typical measurement uncertainty σ, the detectability will depend primarily on the total signal-to-noise ratio S/N, which scales as

$$\mathrm{S/N} \simeq g\sqrt{N}\frac{A}{\sigma}, \tag{10.44}$$

where g is a factor of order unity that depends on the details of the signal.

2.1 Radial Velocities

The differential radial velocity signal of a planet has the form $\Delta V_r = KF(t; e, \omega_*, T_0, P)$, where $F(t)$ encodes the detailed shape of the RV signal. Assuming uniform and dense sampling of the RV curve over a time span that is long compared to P and assuming a total of N observations each with measurement uncertainty σ_{RV}, the total signal-to-noise ratio is

$$(\mathrm{S/N})_{\mathrm{RV}} \simeq g(e, \omega_*)\sqrt{N}\frac{K}{\sigma_{\mathrm{RV}}}. \tag{10.45}$$

For a circular orbit, $g = 2^{-1/2}$ and is generally a weak function of e for $e \lesssim 0.6$. For larger eccentricities, g declines gradually, but more importantly, the stochastic effects of finite sampling become significant for typical values of N (e.g., O'Toole et al. 2009; Cumming 2004). For planets with periods larger than the duration of the observations, the detectability depends additionally on the period and phase of the planet, and generally decreases dramatically with increasing period, typically as $(\mathrm{S/N})_{\mathrm{RV}} \propto P^{-1}$ (e.g., Eisner and Kulkarni 2001; Cumming 2004).

Thus, a robust detection of a planet via RV typically requires achieving radial velocity precisions of $\sigma_{\mathrm{RV}} \ll KN^{1/2}$. For $M_p \ll M_*$, the semiamplitude K is

$$K = \left(\frac{P}{2\pi G}\right)^{-1/3} \frac{M_p \sin i}{M_*^{2/3}} (1 - e^2)^{-1/2} \tag{10.46}$$

Thus, to detect a true Jupiter analogue (i.e., a Jupiter-mass planet in a 11.8 year, circular orbit around a solar-mass star), for which $K \simeq (12.5\ \mathrm{m/s}) \sin i$, requires a few dozen observations with precisions of a few m/s. An RV precision of 3 m/s corresponds to a Doppler shift of $K/c \simeq 10^{-8}$. The motion induced by an Earth analogue is smaller by a factor of $318/(11.8)^{\frac{1}{3}} \sim 100$ so requires an additional two orders of magnitude in precision.

Typically, the centroid of stellar spectral lines at fixed equivalent width can be measured with a precision of $\propto \sigma_V^{3/2}/N_{\mathrm{eff}}^{1/2}$, where σ_V is the effective velocity width of the spectral line and N_{eff} is the effective number of photons in the line (i.e., the equivalent width of the line times the photon rate per unit wavelength). Maximizing the precision requires that the lines are well resolved, and thus that the instrumental velocity resolution is less than the intrinsic velocity width of the star. For reference, the typical width of a spectral feature in a slowly rotating star is of order a few kilometers per second ($\sim 10^{-5}$), and thus resolving powers of ($R = \Delta\lambda/\lambda \sim 10^5$) are needed, comparable to the resolving power of a typical high-resolution astronomical echelle spectrograph. The velocity precision per line is generally insufficient to detect planetary companions, and thus averaging over many lines is required. The statistical signal-to-noise ratio requirements are quite stringent, and thus bright stars and/or large apertures are generally needed.

Because the velocity precisions needed to detect planetary companions are well below the intrinsic widths of the spectral lines and even below the velocity precisions that can be obtained for individual lines, getting close to the photon limit requires excellent control of systematics. One of the most severe requirements is that the wavelength calibration must be both more precise than the desired velocity precisions and stable over many times the orbital period of the planet. For a Jupiter analogue, this wavelength calibration must be at a level of better than 10^{-3} of a resolution element and stable over the course of decades. Since the Earth's motion about the Sun imparts a periodic Doppler shift of order 30 km/s ($v/c = 10^{-4}$), this accuracy and precision must be maintained even as the spectral lines move annually by 10^4 times the measurement precision.

There are at least two proven[2] paths to surmounting this challenge: though precise instrumental calibration with an absorption cell (the iodine technique) and through instrumental ultra-stability (as exemplified by HARPS), both of which are briefly described in Sects. 4.3.1–4.3.5.

[2]Another technique, externally dispersed interferometry (EDI, Erskine and Ge 2000), has shown promise as a third path to precise velocimetry. It employs an interferometer in front of a spectrograph at modest resolution, generating a known, unresolved, sinusoidal transmission function, somewhat analogous to a gas cell's absorption properties. The phase of the beating of the stellar spectrum against this pattern is a measure of radial velocity.

2.2 Astrometry

The magnitude of the differential astrometric offset of a star at a distance d due to a planetary companion has the general form $\Delta\theta_* = \theta_* F(t; e, \omega_*, i, T_0, P)$, where the semi-amplitude of the astrometric offset for a circular, face-on orbit is

$$\theta_* \equiv \frac{a}{d}\frac{M_p}{M_*}, \tag{10.47}$$

and we have assumed $M_p \ll M_*$. Again, assuming uniform and dense sampling of the astrometric curve over a time span that is long compared to P and assuming a total of N observations each with measurement uncertainty $\sigma_{\rm AST}$, the total signal-to-noise ratio is

$$({\rm S/N})_{\rm AST} \simeq g(e, \omega_*, i)\sqrt{N}\frac{\theta_*}{\sigma_{\rm AST}}. \tag{10.48}$$

We note for simplicity we have assumed that each observation yields a given uncertainty $\sigma_{\rm AST}$ on the magnitude of the vector position of the star relative to some reference frame; in reality, each of these measurements may require a separate measurement for each of the two orthogonal directions. For $e = 0$, $g(i) = [0.5(1 + \cos^2 i)]^{1/2}$. For more general cases, the behavior of g is qualitatively similar to that for radial velocity signals. For $e \neq 0$, g depends additionally on ω_* and e but is a relatively weak function of e for $e \lesssim 0.6$. However, the effects of finite sampling start to become more important as e increases, particularly for low N. When P is greater than the span of observations, the detectability also depends on T_0 and P, generally decreasing rapidly with increasing period, also typically as P^{-1}.

The magnitude of the astrometric signal of a Jupiter analogue orbiting a nearby solar-type star at a distance of $D \sim 20$ pc is $\theta_* \simeq 0.25$ mas, whereas for an Earth analogue, the astrometric wobble is over 1,500 times smaller or around 0.15 μas. Thus, astrometric precisions of order a mas or a μas are needed to detect gas giants or terrestrial planets, respectively. Since astrometry is most sensitive to planets orbiting the nearest stars, which have typical proper motions of $\sim 10^3$ mas/year and annual parallactic motion of $\sim 10^2$ mas, the target stars typically move by more 10^3 times the required measurement precision over the course of a year and secularly at 10^5 times the measurement precision per decade.

The photon limit of an astrometric measurement of a star depends on the signal-to-noise ratio and width of the point spread function (PSF), and scales as $\sigma_{\rm AST} \sim {\rm FWHM}/\sqrt{N}$, where N is the total number of photons in the measurement. As mentioned previously, diffraction-limited PSFs, ${\rm FWHM} \sim \lambda/D$, where D is either the aperture of the telescope or the baseline of the interferometer. Baselines of $\lesssim$100 m therefore yield single-measurement precisions of $\lesssim$4 mas. Therefore, the astrometric detection of planets generally relies on the ability to achieve both nearly photon-limited performance when measuring the centroid of individual images and the ability to average many individual measurements to improve the final precision. As is the case with RV, excellent control of systematics is therefore required. There are a number of ways to achieve this, depending on the nature of the observing setup (direct imaging, interferometry, etc.).

Interferometric methods in particular allow precisions below 1 mas from the ground around bright stars with good, nearby reference stars, putting astrometric exoplanet detection within reach. Much better control of systematics is in principle possible from space, and thus, space-based interferometers should be able to achieve precisions of 1–10 μas, making them a potential route for the detection of nearby true Earth analogues (Unwin et al. 2008).

2.3 Imaging

The flux ratio of a planet (or planet/star contrast) at a given wavelength λ and epoch can be expressed as $f_\lambda = f_{0,\lambda}\Phi(\alpha)$, where $\Phi(\alpha)$ describes the phase curve, whose form depends on the properties of the planet atmosphere, and is a function of the phase angle α, which in turn depends on the measurement epoch and orbital elements e, ω_p, i, T_0, P. The phase curve typically takes on values $\lesssim 1$, and thus the magnitude of the reflected light signal is characterized by $f_{0,\lambda}$. This factor depends on the nature of the planetary emission, but for reflected light and thermal emission it takes the form (see (❯ 10.26) and (❯ 10.27))

$$f_{0,\lambda} = A_{g,\lambda}\left(\frac{R_p}{a}\right)^2 \quad \text{(reflected)}, \qquad f_{0,\lambda} \simeq \left(\frac{R_p}{R_*}\right)^2 \frac{T_p}{T_*} \quad \text{(thermal)}, \tag{10.49}$$

where the latter equality assumes observations on the Rayleigh-Jeans tail, which yields the largest flux ratio. Further, for thermal emission arising from reprocessed starlight,

$$f_{0,\lambda} \simeq \left(\frac{R_p}{R_*}\right)^2 \left(\frac{R_*}{a}\right)^{1/2} [f(1-A_B)]^{1/4}. \quad \text{(thermal, equilibrium)} \tag{10.50}$$

The signal-to-noise ratio with which a planet can be directly detected in N measurements is (Kasdin et al. 2003; Brown 2005; Agol 2007)

$$(\mathrm{S/N})_{\mathrm{dir}} \simeq g\sqrt{N}\frac{f_{0,\lambda}}{\sigma_{\mathrm{eff}}}. \tag{10.51}$$

Here, $g = \left[N^{-1}\sum_k \Phi(\alpha_k)^2\right]^{1/2}$ is the root mean square of the phase function values at the times k of the observations, and σ_{eff} is the average effective per-measurement photon noise uncertainty normalized to the total stellar flux. In the usual background-limited case, the primary contributions to the uncertainty are residual light from the stellar point spread function, and local and exozodiacal light. In the case where the scattered light from the star is dominant, $\sigma_{\mathrm{eff}} \sim \sqrt{C/N_*}$ (Kasdin et al. 2003), where C is the contrast ratio between the intensity of the scattered light from the star in the wings of the point spread function relative to the peak and N_* is the total number of photons collected from the star in the measurement.

In contrast to many radial velocity and astrometric surveys, direct imaging surveys are generally designed with the requirement that such target signal-to-noise ratio *per measurement* is $\gtrsim 1$ (Kasdin et al.), and thus $N \sim 1$. Achieving a sufficient S/N per measurement then typically translates into a requirement that $C \lesssim f_{0,\lambda}$, i.e., the flux from the planet within a given aperture is larger than the local background from the stellar PSF in the same aperture.

Young (<1 Gyr old), self-luminous planets can still be quite warm (1,000–2,000 K), even at arbitrarily large separations from their parent stars, making them in some sense the easiest targets for direct imaging surveys. For these temperatures and roughly Jupiter radii, the planet/star flux ratios are $f_{0,\lambda} \sim 10^{-4}$–$10^{-6}$ at near-IR wavelengths or $\Delta m \sim 10$–15 mag. Purely in terms of overall brightness, young exoplanets are rather easily detectable with large telescopes at infrared wavelengths; the primary difficulty therefore lies in suppressing the residual starlight at the position of the planet in order to achieve the contrast ratios C needed to distinguish the planetary light from the star's.

Since the albedos of exoplanets at distances of $\gtrsim 0.1$ AU are typically expected to be of order unity, the flux ratio of a planet in reflected starlight is $f_{0,\lambda} \sim (R_p/a)^2$. For a Jupiter analogue, this is $\sim 10^{-8}$ or about 20 mag, whereas for an Earth analogue, this is $\sim 10^{-9}$ or about 23 mag. The bolometric thermal flux ratio of an exoplanet in equilibrium with the starlight will be of the

same order of magnitude as the reflected light flux ratio; however, the monochromatic thermal flux ratio may be substantially larger, since the planet is cooler, and so, its thermal emission will peak in the Rayleigh-Jeans tail of the stellar blackbody emission. For a Jupiter analogue at ~10μm, $f_{0,\lambda} \sim 10^{-8}$, whereas for a Earth analogue, it is $f_{0,\lambda} \sim 10^{-7}$.

Achieving a given contrast ratio is generally more difficult for small angular separations from the host star and becomes generally impossible closer than some minimum *inner working angle*, $\theta_{\rm IWA}$. Thus, the angular separation $\Delta\theta = r_\perp/d$ is another important parameter that determines the detectability of planets by direct imaging. The probability distribution of the projected separation $r_\perp$ given a value of a for random orbital phases and viewing geometries is generally sharply peaked at a. Thus, the typical angular separation of a planet with $a = 5.2$ AU orbiting a star at 20 pc is $\sim a/d = 250$ mas, whereas it is 50 mas for $a = 1$ AU. The inner working angle typically scales as (and is generally similar to) the diffraction limit of telescope

$$\theta_{\rm diff} \sim \lambda/D, \tag{10.52}$$

where D is the diameter of the telescope (or the most widely-separated components of an interferometer). This corresponds to 60 mas at 2 μm on an 8 m telescope. Thus, surveys for Earth analogues are generally only feasible for the nearest stars.

The detectability of a planet by direct imaging depends on a complicated interplay between many variables, including the semimajor axis and size of the planet, age and distance to the star, and wavelength and capabilities of the imaging system. For example, for reflected light surveys, the orbital separation effects the detectability of the planet through the opposing effects of contrast and angular separation. As another example, while younger planets tend to be more luminous, younger stars are also less common and so typically more distant. Additional factors may also contribute to these interplays, such as the brightness of the exozodiacal light as a function of semimajor axis and variations in the planetary atmospheric properties (e.g., albedo and absorption bands) as a function of semimajor axis, age, and surface gravity. Direct imaging surveys therefore need to be designed carefully in order to maximize the discovery space and so chance of success. Combined with the technical challenges associated with achieving the contrast ratios and inner working angles needed for planet detection briefly described below, it is clear that direct imaging is a generally expensive and challenging detection method. Nevertheless, the potential payoff is enormous, particularly when considering the goal of directly imaging Earth analogues.

The technical aspects of imaging exoplanets comprise surmounting three challenges: corralling starlight into a nearly diffraction-limited PSF (and away from the planet image), mechanically blocking the starlight before it can diffract into the planet image, and subtracting the remaining starlight at the position of the planet image on the detector to reveal the planet image beneath. These three challenges are most forcefully attacked using adaptive optics, coronagraphy, and various forms differential imaging, respectively.

Adaptive optics (AO) refers to controlling the wavefronts of the incoming starlight and planet light, which ideally consist of parallel planes propagating toward the telescope. The atmosphere and telescope optics both introduce aberrations to this wavefront which result in a PSF that differs significantly from that which a theoretically perfect optical system would produce (for an unocculted circular aperture, this would be an airy function). For most ground-based telescopes, the primary source of wavefront aberrations is the atmosphere. Adaptive optics use movable or deformable mirrors which can be rapidly actuated in response to measured atmospheric aberrations, usually at tens to thousands of Hertz. These systems dramatically reduce the effects of atmospheric blurring, and the best of them can collect most of a star's light into

the shape dictated by optical diffraction. This heightens the peak of faint sources and reduces noise from the star outside the diffraction limit.

The technique of blocking the light of a bright source to reveal faint surrounding features is called coronagraphy. A coronagraph uses a series of masks in an optical system to block, reorganize, or alter the phase of incoming light such that "on-axis" light from the star is almost entirely blocked or caused to destructively interfere, while "off-axis" light (for instance, from a nearby planet) is relatively unaffected. Because important aspects of this technique happen in the pupil plane, stellar photons can be distinguished from planetary photons and rejected before they arrive on the same pixels on the detector. The effect is to reduce the contamination from stellar photons at the detector position of the planet, enhancing its detectability outside of the diffraction limit. There has been a proliferation of coronagraph designs in recent years, but they share the common feature of reducing or controlling the nature of the diffraction of the light into the planet image.

Adaptive optics systems and coronagraphs are not perfect. Their limitations, the aberrations introduced by the telescope, and diffraction spikes and rings from the aperture can result in significant amounts of starlight outside of the diffraction limit. The most insidious of these effects is the semi-static patterns of "speckles" from residual wavefront errors. Differential imaging is the process of precisely determining the PSF of the starlight and attempting to subtract it, leaving only the planet light to be detected. In principle, differential imaging is limited by the quality of the model PSF and the unavoidable photon noise in the residuals to that model. The reference image being subtracted can be determined from a reference star (RDI) or from the data themselves through angular modulation (ADI), spectral analysis (SDI), polarization analysis (PDI), or other some other method or combination of methods.

A conceptual cousin of coronagraphy is interferometry, which allows widely separated apertures to combine incoming light to form interference fringes whose amplitudes and phases are sensitive to the presence of faint, off-axis companions. Such work is common at radio wavelengths and in the infrared, can be especially profitable just inside of the traditional diffraction limit of the telescope. Two such techniques are aperture masking interferometry, where a single telescope pupil is divided into small sub-apertures and the light is combined at the focal plane, and nulling interferometry where light from two telescopes is combined such that the starlight undergoes destructive interference, while the planet light, incoming at a slightly different angle, interferes constructively.

2.4 Transits

The fractional change in the flux of a star when it is transited by a planetary companion has the form $\Delta F_*/F_* = -\delta\, F(t; R_p, M_*, R_*, a, i, e, \omega_p)$, where $\delta = k^2$ is the square of the planet/star radius ratio and the function $F(t)$ describes the detailed shape of the transit curve, and also generally depends on the surface brightness profile of the star. In the case of circular orbits, no limb darkening, and $R_p \ll R_*$, the form for $F(t)$ can be approximated by a box car with a depth of unity and duration of $T = T_{\rm eq}(1 - b^2)^{1/2}$, where $T_{\rm eq}$ and b are the equatorial crossing time and impact parameter, respectively, as defined in Sect. 1.5. The fraction of time the planet is in transit (the transit duty cycle) is then $f_{\rm tr} = T/P$.

Transit surveys generally operate by obtaining many observations of the target stars over a given time span. Of course, in order to be detectable, a planet must be favorably aligned such that it transits. The transit probability is roughly $P_{\rm tr} \sim R_*/a$. Then, assuming uniform sampling

over a time span that is long compared to P, the signal-to-noise ratio of the transit when folded about the correct planet period is

$$(\mathrm{S/N})_{\mathrm{tr}} \simeq \sqrt{N f_{\mathrm{tr}}} \frac{\delta}{\sigma_{\mathrm{ph}}}, \tag{10.53}$$

where N is the total number of observations and σ_{ph} is the fractional photometric uncertainty. Therefore, the probability of detecting a given planet via transits can be roughly quantified by three characteristics of the planetary system that depend primarily on R_p, R_*, and a,

$$\delta = \left(\frac{R_p}{R_*}\right)^2, \qquad P_{\mathrm{tr}} \sim \frac{R_*}{a}, \qquad f_{\mathrm{tr}} \sim \frac{P_{\mathrm{tr}}}{\pi}. \tag{10.54}$$

For a typical hot Jupiter with $R_p \simeq R_J$ and $P \simeq 3$ days orbiting a solar-type star, the transit probability is $P_{\mathrm{tr}} \sim 10\%$, the transit depth is $\delta \sim 1\%$, and the duty cycle is $f_{\mathrm{tr}} \sim 3\%$. These parameters place Hot Jupiters well within the capabilities of ground-based surveys, although the requirements are not trivial. First, since Hot Jupiters are only found around ~0.5% of solar-type stars (Gould et al. 2006), many thousands of stars must be surveyed to guarantee a transiting Hot Jupiter. Obtaining relative photometry at precisions of less than a few millimagnitudes for thousands of stars simultaneously from the ground is generally difficult, and thus ground-based transit surveys operate close to the limit where $\delta/\sigma_{\mathrm{ph}} \sim 1$. Therefore, hundreds of epochs during transit are needed for robust detections, corresponding to many thousands of total measurements. Aliasing effects arising from the diurnal constraints make achieving the required number of points in transit more challenging. All of these requirements are most easily met with relatively small aperture but very wide field-of-view telescopes (e.g., Pepper et al. 2003; Bakos et al. 2004; Pollacco et al. 2006; McCullough et al. 2005).

In fact, finding transiting planets in wide-field surveys has proven even more difficult than simply meeting these (already difficult) requirements. First, wide-field transit surveys must contend with a huge fraction of false positives in the form of grazing eclipsing binaries (EBs), eclipsing binaries blended with brighter stars, and more exotic variables. Furthermore, even those signals that are consistent with a Jupiter-sized transiting object can, in principle, be much more massive companions, since the radius of compact objects is essentially constant from the mass of Saturn through ~$0.1 M_\odot$ (e.g., Burrows et al. 1997). Thus radial velocity follow-up is needed to eliminate these false positives. Finally, high-precision (few m/s) radial velocity follow-up is needed to precisely measure the planet mass. The most successful searches achieve reliably high photometric accuracy over large fields and employ multiple sites with good longitudinal coverage, sophisticated and automated transit identification algorithms, and thorough follow-up campaigns that use multiband photometry, multi-band astrometry (to rule out close, chance alignments of EBs and foreground stars) , and radial velocity work. Further characterization of a transiting planet is most successfully done using photometry and high signal-to-noise ratio spectroscopy with larger ground and/or space telescopes.

The requirements for the detection of Earth analogues orbiting solar-type stars are especially challenging. In this case, the fractional transit depth is ~10^{-4}, the transit probability is ~0.5%, and the duty cycle is ~0.1% (i.e., the planet transits for $T \lesssim 13$ h once a year). The detection of transiting Earth analogues requires essentially continuous observations of hundreds of stars, and precisions of better than ~0.1 mmag for periods of several years. These requirements cannot be met from the ground and require space-based photometric monitoring. Indeed, the *Kepler* mission was designed to detect such planets, achieving the required photometric precision to detect Earth-sized planets on tens of thousands of stars (Borucki et al. 2010, see also Sect. 5.3.2).

Transiting planets can also be found via photometric follow-up of known radial velocity companions; indeed, the first transiting planet was discovered in this way (Henry et al. 2000; Charbonneau et al. 2000). Here, the challenges are somewhat different than the "traditional" method of discovering transiting planets through their photometric signature. First, the probability that a given radial velocity companion will also transit its parent star is low, $\lesssim$10%. Second, radial velocity searches have traditionally been limited to relatively bright stars, and so the total number of stars have targeted for precision RV searches is ~10^3, making the total yield of transiting planets from this sample also low. Furthermore, achieving photometry at the level of precision needed to detect the transit signature may be challenging for very bright stars, primarily because of the lack of suitable comparison stars. Finally, the uncertainties in the predicted times of inferior conjunction from the radial velocity fits can be quite large, from several hours to several days. Nevertheless, seven transiting systems have been discovered among the sample of companions first discovered via radial velocity, and there are ongoing projects that aim to increase this sample by first refining the radial velocity emphemerides of promising systems, and then performing photometric follow-up (Kane et al. 2009).

2.5 Microlensing

Unlike the detection methods discussed above, the signals caused by planetary companions in microlensing events cannot be described analytically except in a few specific limits that are not generally applicable. Nevertheless, we can provide some qualitative guidelines and approximate scaling behaviors that will elucidate the general requirements for successful surveys for planets with microlensing. We stress that, because of the large diversity in the properties of the systems that give rise to gravitational microlensing events, the expressions provided should be treated as very rough estimates only.

A somewhat unusual attribute of the microlensing method is that the magnitude of a microlensing perturbation does not depend on the properties of the planet in the general case. Rather, the magnitude depends primarily on the angular separation of the planet from the image(s) it is perturbing. However, the duration of the planetary perturbation does depend on the planet properties, in particular, the mass ratio q. Very approximately, the duration of the planetary deviation is $\Delta t_p \sim q^{1/2} t_{\rm E}$, where $t_{\rm E}$ is the primary event time scale. The primary event light curves must be sampled on a time scale significantly smaller than Δt_p in order to detect and characterize the planetary perturbation. Furthermore, the detection probability also depends on the planet-mass ratio, such that $P_{\rm det} \sim 20\%(q/0.001)^{\sim 5/8}$ (Horne et al. 2009). This detection probability is averaged over a uniform distribution of impact parameters and is appropriate for planets with projected separations that are within a factor of ~2.6 of the Einstein ring radius, $r_\perp \sim [0.6-1.6] d\theta_{\rm E}$, sometimes called the "lensing zone." Planets with separations much smaller or much larger than this range have substantially lower probability of detection. As discussed in the context of direct imaging, the distribution of projected separation $r_\perp$ for random viewing geometries and orbital phase is sharply peaked at $r_\perp \sim a$.

In addition, there is a minimum mass that can be detected in microlensing surveys that is set by the finite size of the source stars. When the angular size of the planet perturbation region is substantially smaller than the angular size of the source, the planet perturbs only a small fraction of the source and the magnitude of the resulting deviation is strongly suppressed. A rough limit on the mass ratio can be established when the angular size of the source θ_* is a factor of ~3 times larger than the angular Einstein ring radius of the planet $\theta_p = q^{1/2}\theta_{\rm E}$, corresponding to roughly

an order-of-magnitude suppression of the planet signal. Thus, $q_{\rm min} \sim 0.1\rho_*^2$, where $\rho_* \equiv \theta_*/\theta_{\rm E}$ (Gould and Gaucherel 1997).

Thus, the parameters that determine the detectability of planets with microlensing are

$$\Delta t_p \sim \left(\frac{M_p}{M_*}\right)^{1/2} t_{\rm E}, \quad P_{\rm det} \sim 20\% \left(\frac{M_p/M_*}{0.001}\right)^{5/8}, \quad a \sim [0.6-1.6] d\theta_{\rm E}, \quad q_{\rm min} \sim 0.1\rho_*^2, \tag{10.55}$$

where the parameters $t_{\rm E}$ and ρ_* additionally depend on the mass and distance to the host star via the angular Einstein ring radius (see (❯ 10.38)). The distributions of M_*, $t_{\rm E}$, d, and $\theta_{\rm E}$ for microlensing events toward the Galactic bulge are quite broad, but we can take typical values of $M_* \simeq 0.5\,M_\odot$, $t_{\rm E} \simeq 25$ days, $d \sim 4$ kpc, and $\theta_{\rm E} \sim 0.7$ mas. Thus, microlensing planet surveys are most sensitive to planets with semimajor axes of $a \simeq [1-5]\,{\rm AU}(M/0.5\,M_\odot)^{1/2}$. For a Jupiter-mass planet, the typical planet perturbation duration is $\Delta t_p \sim 1$ day and the typical detection probability is ~30% in the lensing zone. For an Earth-mass planet, $\Delta t_p \sim 1.5$ h, whereas the detection probability in the lensing zone is ~1%.

The typical dimensionless source size for a clump giant star (~13 $R_\odot$) in the Galactic bulge is $\rho_* \sim 0.01$, whereas for a turnoff star (~$R_\odot$), it is ~0.001. Thus, the minimum mass ratio that can be detected by monitoring clump giant sources is $q_{\rm min} \sim 10^{-5}$, corresponding to ~1.7× mass of the Earth for a typical primary lens of 0.5 $M_\odot$. For main-sequence stars in the bulge, $q_{\rm min} \sim 10^{-7}$ corresponding to just over the mass of the Moon! Thus, detecting planets with mass of the Earth or less requires monitoring main-sequence stars. The difficulty lies in the fact, in the crowded fields toward the Galactic bulge, most main-sequence stars are severely blended with other unrelated background stars in typical ground-based seeing conditions, dramatically increasing the photometric noise. Therefore, detecting planets with mass substantially less than that of the Earth generally requires a space-based survey (Bennett and Rhie 2002; Bennett 2008).

A final difficulty within microlensing surveys is the low overall event rate of gravitational microlensing events. Toward the Galactic bulge, the rate of microlensing events is roughly $\Gamma \sim 10^{-5}$ per star per year (e.g., Kiraga and Paczynski 1994). Thus, in order to detect ~10^3 events per year (the current number of microlensing events that are detected per year toward the Galactic bulge by the Optical Gravitational Lensing Experiment (OGLE) collaboration[3]), of order 100 million source stars must be monitored. There are 3 million stars per square degree down to an I magnitude of 19 in Baade's window (Holtzman et al. 1998), where $I \sim 19$ is roughly the peak of the distribution of baseline magnitudes for microlensing events. Thus, several tens of square degrees of the bulge must be monitored.

The unpredictability of microlensing events requires monitoring the potential source stars with a cadence that is substantially smaller than the timescale of interest. For the primary microlensing events, which have a typical $t_{\rm E} \sim 25$ days, this means roughly daily observations. Detecting the planetary perturbations on these events requires much higher cadences of a few hours or less. Furthermore, since the total durations of the planetary perturbations are of order a day or less, networks of longitudinally distributed telescopes must be employed in order to avoid missing part or all of the perturbations. Given these requirements, traditional microlensing planet surveys have used a two-tier approach, where collaborations with dedicated access to telescopes with a relatively wide fields of view of ~0.5–2 square degrees monitor the tens of square degrees needed to detect the primary microlensing events, but with cadences that are generally insufficient to detect planetary perturbations on these events (Udalski 2003;

[3] See http://ogle.astrouw.edu.pl/ogle4/ews/ews.html.

Sako et al. 2008). These survey collaborations alert the microlensing events real time before the peak magnification, thus allowing "follow-up" collaborations with access to narrow-angle telescopes on several continents to monitor only a subset of the stars that display ongoing microlensing events with the cadences needed to detect planetary perturbations (Albrow et al. 1998; Tsapras et al. 2009; Dominik et al. 2010; Gould et al. 2010). Future surveys will operate on a very different principle, as described in Sect. 5.3.3.

2.6 Timing

The magnitude of the signal in other planet detection techniques varies. Timing for millisecond pulsars like PSR 1257+12 can in principle detect extremely low-mass objects (significantly $<1\ M_{\oplus}$) given a sufficient amount of data, limited primarily by pulsar timing noise.

Other timing techniques, such as eclipsing binary times or pulsating hot subdwarfs, rely on timing variations being correctly interpreted as a light-travel time effect of a star or stellar system orbiting a common center of mass with an unseen companion. A summary of such detections is listed in Schuh (2010); most of these detections imply minimum companion masses of several times that of Jupiter. The sensitivities of these methods is difficult to determine, however, since they depend on the magnitude of all non-orbital origins in timing variations, which have not been well quantified. Most of the current detections are of fewer than two full apparent orbits (periods are 3–16 years), and so, the strict periodicity that is characteristic of Keplerian signals cannot yet be confirmed. Further, quasi-cyclical timing variations may be generated by poorly understood internal mechanisms such as the "Applegate effect" (Applegate 1992). Following these apparent planetary systems for multiple orbits will help illuminate the true sensitivities of these methods.

Transit timing variations (TTVs) provide an extremely sensitive method of detecting new planets or characterizing known planets in a transiting system. The sensitivity is a complex function of the orbital parameters of the planets involved but is optimized when the planets are in mean-motion resonances (Veras et al. 2011). *Kepler* is sufficiently precise in its timing to measure variations of order minutes in the ingress and egress times of transiting planets, which in principle allows it to reach mass precisions of order $1\ M_{\oplus}$ over several years of observation.

In known multi-transit systems, these variations can be used to infer the masses of the planets involved (e.g., Lissauer et al. 2011), and in apparently single systems, they can be used to detect non-transiting planets (e.g., Ballard et al. 2011). Ground-based planet transit timing will generally be limited to precisions of a tens of minutes and so have correspondingly weaker sensitivities.

3 Comparisons of the Methods

In the previous sections, we reviewed the primary exoplanet detection methods in some detail, outlining the principles of each method, including the primary physical observables and practical challenges associated with achieving robust detections. In this section, we place these discussions in the context of the larger goal of constraining exoplanet demographics by outlining the regions of planet and host star parameter space where each method is most sensitive. When then compare the methods with one another in this context, highlighting the strong complementarity

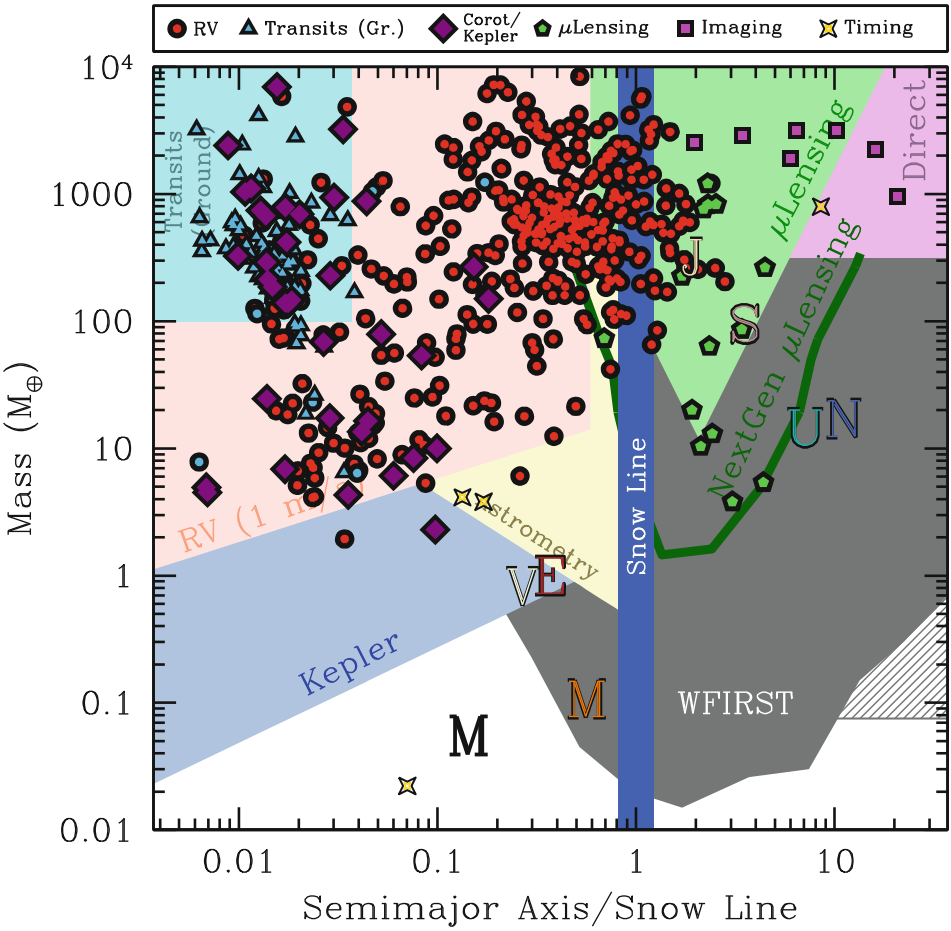


Fig. 10-5

The *points* show the masses versus semimajor axis in units of the snow line distance for the exoplanets that have been discovered by various methods as of 12/2011. See the Extrasolar Planets Encyclopedia (http://exoplanet.eu/) and the Exoplanet Data Explorer (http://exoplanets.org/). Here, we have taken the snow line distance to be $a_{sl} = 2.7\ \mathrm{AU}(M_*/M_\odot)$. Radial velocity detections (here, what is actually plotted is $M_p \sin i$) are indicated by *red circles* (*blue* for those also known to be transiting), transit detections are indicated by *blue triangles* if detected from the ground and as *purple diamonds* if detected from space, microlensing detections are indicated by *green pentagons*, direct detections are indicated by *magenta squares*, and detections from pulsar timing are indicated by *yellow stars*. The *letters* indicate the locations of the Solar System planets. The *shaded regions* show rough estimates of the sensitivity of various surveys using various methods, demonstrating their complementarity (Adapted from Gaudi 2012)

of the methods. We also briefly discuss and compare how the intrinsic sensitivity of each method to planets in the Habitable Zones of their parent stars scales with host star mass.

The sensitivity of the various detection methods as a function of planet mass and separation is illustrated in *Fig. 10-5*. We show the masses and semimajor axes of the exoplanets discovered by radial velocities, direct imaging, timing, transits, and microlensing as of 12/2011.

In addition, we show estimates of the sensitivity of various surveys using radial velocities, direct imaging, transits, microlensing, and astrometry. In the following subsection, we explain the scaling of these survey sensitivities with planet parameters and explain our specific choice for their normalization. Host star mass is a third parameter that can strongly influence the sensitivity of these methods but is suppressed in this figure. Therefore, in order to provide a somewhat fairer comparison across the broad range of host star masses represented in this figure, we normalize the semimajor axis by an estimate of the snow line distance (e.g., Kennedy and Kenyon 2008),

$$a_{\rm sl} = 2.7\ {\rm AU}\frac{M_*}{M_\odot}. \tag{10.56}$$

The snow line is the location in the protoplanetary disk where the temperature is below the sublimation temperature of water. In the currently favored paradigm of planet formation, the snow line distance plays an important role, as the larger surface density of solids beyond the snow line facilitates giant planet formation, whereas primarily rocky planets are expected to form interior to the snow line (e.g., Lissauer 1987; Pollack et al. 1996; Ida and Lin 2004; Mordasini et al. 2009).

Both the locations of the known planets and the regions of sensitivity in *Fig. 10-5* serve to illustrate the complementarity of the various detection methods. In particular, radial velocity and transit surveys are most sensitive to relatively short-period exoplanets with separations interior to the snow line. State-of-the-art surveys using these methods are sensitive to planets as small as the Earth. In contrast, direct imaging and microlensing surveys are most sensitive to planets beyond the snow line. Current direct imaging surveys are sensitive to very widely separated, massive planets. Current and near-term microlensing surveys are sensitive to planets with masses greater than that of the Earth, whereas a space-based microlensing survey would be sensitive to planets with mass greater than that of Mars and separations greater than a few AU, including free-floating planets. The combined sensitivity of current and proposed surveys using the methods discussed in the chapter encompasses an extremely broad volume of parameter space, including planets with masses greater than that of the Earth for arbitrary separations (including free-floating planets), host stars with masses from the bottom of main sequence to several times the mass of the Sun, and distances from the solar neighborhood to the Galactic center. Thus, the full complement of these methods can potentially provide a nearly complete picture of the demographics of exoplanets.

3.1 Sensitivities of the Methods

- **Radial Velocities**. The radial velocity signal and signal-to-noise ratio scale as

$$({\rm S/N})_{\rm RV} \propto M_p P^{-1/3} M_*^{-2/3} \propto M_p a^{-1/2} M_*^{-1/2}. \tag{10.57}$$

  Thus, at fixed host star mass, RV surveys are generally more sensitive to more massive planets and have a weak preference for shorter-period planets, at least for periods shorter than the span of the measurements. At fixed S/N, the minimum detectable mass scales as $M_{p,\rm min} \propto P^{1/3} M_*^{2/3} \propto a^{1/2} M_*^{1/2}$. This mass limit is shown in *Fig. 10-5* assuming $M_* = M_\odot$ and a minimum detectable RV signal of $K \sim 1$ m/s (i.e., corresponding to a $({\rm S/N})_{\rm RV} \sim 10$ detection with 100 observations each with single measurement precisions of ~1 m/s). In principle, sufficiently massive planets can be detected even when they have periods longer than the duration of observations; however, it will generally not be possible

to uniquely measure $M_p \sin i$ and P in these cases, and thus the usefulness of such "detections" is significantly compromised. Therefore, we simply assume an upper limit on the period of $P = 2{,}000$ days.

At fixed planet properties, the RV signal increases with decreasing host star masses as $M_*^{-2/3}$. However, there are many additional factors that enter into the overall detectability as a function of mass, through the radial velocity uncertainty $\sigma_{\rm RV}$. At fixed distance to the host star, the velocity uncertainty due to photon noise increases with decreasing mass for main-sequence stars, both because of the strong mass-bolometric luminosity relation for main-sequence stars and because of the difficulties of continuum normalization in very cool stars, where most RV surveys are carried out. The intrinsic velocity information in stellar spectra also varies as a function of spectral type. In particular, hot ($T_{\rm eff} \gtrsim 6{,}500$ K) stars have few spectral lines and typically rotate much more rapidly than cooler stars with convective envelopes. Finally, the astrophysical radial velocity noise (i.e., "jitter" due to spots) also depends on spectral type and stellar age (Wright 2005). When taken together, these factors tend to favor late G or early K quiet main-sequence stars as the most promising for detecting low-mass planets (e.g., Pepe et al. 2011).

- **Astrometry**. The astrometric signal and signal-to-noise ratio scale as

$$({\rm S/N})_{\rm AST} \propto M_p P^{2/3} M_*^{-2/3} d^{-1} \propto M_p a M_*^{-1} d^{-1}, \tag{10.58}$$

and so, at fixed stellar mass, astrometric surveys are more sensitive to massive, long-period planets. At fixed S/N, the minimum detectable mass scales as $M_{p,\rm min} \propto P^{-2/3} M_*^{2/3} d \propto a^{-1} M_* d$. This limit is shown in ❯ *Fig. 10-5* assuming $M_* = M_\odot$ and $d = 10$ pc, and a minimum detectable astrometric signal $\theta_* \simeq 0.35$ μas (i.e., a $({\rm S/N})_{\rm AST} \sim 5$ detection with 200 2-D observations each with $\sigma_{\rm AST} \sim 1$ μas precision). As with radial velocity observations, although it is possible to detect the astrometric signal of planets with period larger than the duration of observations, it is generally not possible to independently measure the mass and orbital parameters with observations that do not cover a complete orbit. This is particularly problematic for astrometric observations, because in this regime, the signal of the proper motion of the star is partially degenerate with the astrometric signal of the planetary companion. Astrometric surveys are therefore expected to have the most sensitivity to planets with periods similar to the survey duration, and increasing the survey duration has a strong effect on the survey yield.

At fixed planet properties and host star distance, the astrometric signal increases with decreasing host star mass as $M_*^{-2/3}$. Stellar spots are generally not expected to be an important source of astrometric noise (Makarov et al. 2009), and thus, the only additional dependence of the sensitivity of astrometric surveys on stellar mass enters through the effects of the typical distance and photon noise uncertainty of the host stars. Specifically, low-mass stars are more common and thus have a smaller typical distance but are less luminous and thus yield poorer astrometric precision.

- **Imaging** For direct detection in reflected and equilibrium thermal emission light, the planet/star flux ratio and thus the signal-to-noise ratio scale as

$$({\rm S/N})_{\rm dir} \propto R_p^2 a^{-2} \qquad ({\rm reflected}), \tag{10.59}$$

$$({\rm S/N})_{\rm dir} \propto R_p^2 T_p R_*^{-2} T_*^{-1} \qquad ({\rm thermal}) \tag{10.60}$$

$$({\rm S/N})_{\rm dir} \propto R_p^2 R_*^{-3/2} a^{-1/2} \qquad ({\rm thermal, equilibrium}), \tag{10.61}$$

where the last two forms again assume observations in the Rayleigh-Jeans tail. The other primary requirement for direct imaging is that the angular separation between the planet and star is larger than inner working angle, and thus $a \gtrsim \theta_{\rm IWA} d^{-1}$. Thus, at fixed primary properties and distance, larger and hotter planets are more readily detectable. As discussed in Sect. 2.3, the dependence of detectability on semimajor axis is not trivial: planets with larger orbits generally have smaller flux ratios; however, they must have an angular separation greater than the inner working angle to be detectable. Furthermore, additional effects that are likely to depend on the planet semimajor axis may affect the detectability, such as the variation of the planetary albedo with separation.

At fixed planet properties, the signal-to-noise ratio with which a planet can be detected in reflected light is largely independent of the host star properties. For detection in thermal emission, larger and/or hotter host stars generally give rise to smaller flux ratios. The other effect of host star mass enters through the dependence on distance: less massive host stars are more numerous and so have a smaller average distance, whereas more massive host stars are more luminous and thus give rise to smaller photon noise uncertainties at fixed distance. Finally, the age of the host star plays an important role in the detectability of planets, particularly with current surveys: young, self-luminous planets have flux ratios that are both larger than would be expected for planets whose emission is dominated by reflected light or equilibrium thermal emission and are independent of their semimajor axis. Thus, relatively luminous planets with separations well beyond the inner working angle can be found around young stars.

Current ground-based imaging surveys are most sensitive to young, massive ($M_p \gtrsim M_{\rm JUP}$), self-luminous planets with semimajor axes of $\gtrsim$10 AU around the nearest stars. Thus, these surveys are sensitive to planets in a regime of parameter space that is not currently being probed by other methods. We illustrate the current region of sensitivity of imaging surveys in *Fig. 10-5*, assuming planets with $M_p \gtrsim M_{\rm JUP}$ and $a \gtrsim 10$ AU can be detected. Future surveys (some of which will be initiated very soon) will be sensitive to a much broader region of planet parameter space.

- **Transits**. Assuming uniformly sampled observations over a time span T that is long compared to the planet period, the signal-to-noise ratio with which a transiting planet can be detected scales as

$$({\rm S/N})_{\rm tr} \propto R_p^2 P^{-1/3} M_*^{-5/3} \propto R_p^2 a^{-1/2} M_*^{-3/2}, \tag{10.62}$$

where we have assumed $R_* \propto M_*$ as appropriate for stars on the main sequence with $M \lesssim M_\odot$. In addition, the transit probability scales as $P_{\rm tr} \propto P^{-2/3} M_*^{2/3} \propto a^{-1} M_*$, and the requirement to detect at least n transits sets a strict lower limit on the period $P \leq T/n$. Thus, for fixed host star properties, the sensitivity of transit surveys is strongly weighted toward short-period, large-radius planets. At fixed S/N, the minimum detectable planet radius scales as $R_{p,\rm min} \propto P^{1/6} M_*^{5/6} \propto a^{1/4} M_*^{3/4}$. For planets with a constant density, this translates into a minimum mass of $M_{p,\rm min} \propto R_{p,\rm min}^3 \propto P^{1/2} M_*^{5/2} \propto a^{3/4} M_*^{9/4}$. This limit is shown in *Fig. 10-5*, assuming a minimum $({\rm S/N})_{\rm tr} = 8$ and a midlatitude transit and photon noise-limited precision for a $M_* = M_\odot$, $V = 12$ star from *Kepler* (Gilliland et al. 2011). The relatively strong dependence of the signal-to-noise ratio on planet radius, combined with the decreasing transit probability with increasing planet period, generally implies that the yield of a transit survey is a relatively weak function of the total duration T.

Main-sequence stars are clearly the best targets for transit searches. At fixed planet radius and period, low-mass main-sequence stars yield larger transit signals. However, for photon

noise-limited uncertainties, the signal-to-noise ratio also depends on the stellar luminosity in the wavelength of the observation and the distance to the star. Low-mass stars are more common and are therefore closer on average. These net results of these various countervailing effects on the scaling of the sensitivity of transit surveys with host star mass depend on the survey parameters, particularly the wavelength, target field, and magnitude limit (Pepper and Gaudi 2005). All-sky surveys in the visual at fairly bright magnitudes tend to be dominated by F and G stars (Pepper et al. 2003), but surveys for fainter stars, or surveys in near-IR wavelengths, tend to be more heavily weighted to toward low-mass stars.

- **Microlensing**. As discussed previously, microlensing surveys are less amenable to analytic characterization of the sensitivity and scaling with planet and host star parameters. Nevertheless, we have the general qualitative result that the sensitivity peaks for semimajor axes that are similar to the Einstein ring of the host star lens, $a_{\rm opt} \sim 2.85\ {\rm AU}(M_*/0.5M_\odot)^{1/2}$. Both the detection probability at fixed semimajor axis and range of sensitivity around this peak increases for larger mass ratio. The range of planet mass and semimajor axis to which a given microlensing survey is most sensitive depends on the details of the survey design, in particular the number and peak magnification of the primary events that are monitored, and the precision and cadence of the observations. In ❯ *Fig. 10-5*, we show representative estimates for the sensitivity of a current ground-based microlensing survey based on the analysis of Gould et al. (2010), a next-generation ground-based survey as described in ❯ Sect. 5.3.3, and a space-based survey similar to WFIRST as described Green et al. (2011). Generally, current microlensing surveys are most sensitive to planets with $M_p \gtrsim 10\ M_\oplus$, with separations just beyond the snow line spanning a factor of a few in semimajor axis. Next-generation ground-based surveys will lower the sensitivity in mass by roughly an order of magnitude to $M_p \gtrsim M_\oplus$ and broaden the range of semimajor axis by a factor of ~2. A space-based survey would be sensitive to planets with mass greater than the mass of Mars with separations greater than a few AU, including analogues to all of the Solar System planets except Mercury.

  Because microlensing does not rely on the detection of photons from the host star, the sensitivity of a microlensing survey to the host star parameters enters primarily through the microlensing event rate as a function of host star mass. The microlensing event rate is given by the integral of the number density of lenses along the line of sight to the bulge, weighted by the cross section for lensing, which depends on the host star Einstein ring radius ($\propto M_*^{1/2}$) and its projected transverse velocity. ❯ *Figure 10-6* shows a theoretical estimate of the microlensing event rate toward the Galactic bulge, as a function of the lens mass. The sensitivity of microlensing planet surveys is roughly proportional to this event rate. Thus microlensing surveys are sensitive to main-sequence hosts with mass from below the hydrogen-burning limit up to the turn off of the stellar population (~$M_\odot$) as well as remnant hosts. The sensitivity to brown dwarf and main-sequence hosts is roughly constant per log M_*.

3.2 Habitable Planets

Of particular interest is the detection of terrestrial planets in the Habitable Zones of their parent stars. In ❯ Sect. 2, we discussed the magnitude of the signal expected for an Earthlike planet separated by 1 AU from a solar-type star for various detection methods. These signals are generally quite small. However, we can and should also consider the signals and detectability of Earthlike

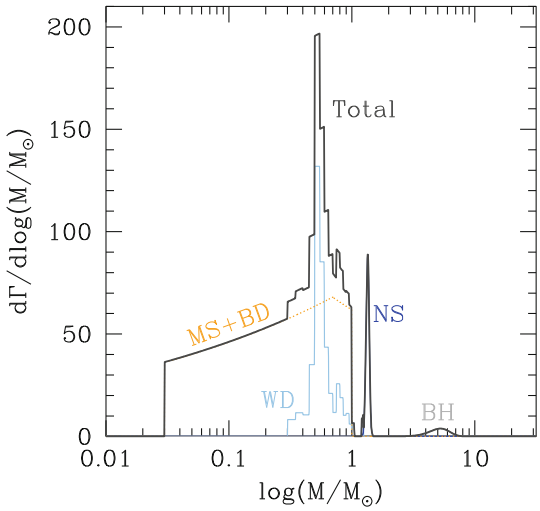


Fig. 10-6

The *solid black curve* shows the theoretical estimate of total event rate of microlensing events toward the Galactic bulge from Gould (2000), as a function of the mass of the lens. This event rate is decomposed into its contribution from main-sequence and brown dwarf lenses, white dwarfs, neutron stars, and black holes, as indicated. The sensitivity of microlensing planet surveys is roughly proportional to this event rate (Adapted from Gould (2000), reproduced by permission of the AAS)

planets orbiting in the Habitable Zones of main-sequence stars with mass significantly different than the Sun. Because the masses, radii, and luminosities of stars on the main sequence can span an order of magnitude or more, both the location of the Habitable Zone and the magnitude of the signal expected from an Earthlike habitable planet can vary substantially.

In this subsection, we derive the scaling of the detectability of Earthlike habitable planets with host star mass. To this end, we adopt a power-law form for the stellar mass-bolometric luminosity relation,

$$\frac{L_*}{L_\odot} \sim \left(\frac{M_*}{M_\odot}\right)^4 . \tag{10.63}$$

Our adopted exponent of 4 roughly corresponds to that found from a weighted fit to the data for benchmark binary systems analyzed in Torres et al. (2010). This form provides a reasonable approximation of this data set over the full span of masses of ~0.2–50 $M_\odot$. We note, however, that there are significant deviations from this form within the range, particularly for stars below the fully convective boundary, which are generally more luminous that predicted by (10.63). Nevertheless, we will adopt this simple form for the purposes of illustration. With this adopted mass-luminosity relation, we can define the location of the Habitable Zone as

$$a_{\rm HZ} = {\rm AU}\left(\frac{L_*}{L_\odot}\right)^{1/2} \sim {\rm AU}\left(\frac{M_*}{M_\odot}\right)^2 , \qquad P_{\rm HZ} \sim {\rm year}\left(\frac{M_*}{M_\odot}\right)^{5/2} . \tag{10.64}$$

Therefore, the Habitable Zone location ranges from a semimajor axis of ~0.01 AU and period of ~1 day for a star at the bottom of the main sequence to ~4 AU and ~6 years for a $M_* \sim 2\,M_\odot$ star. We adopt a mass-radius relation of the form

$$\frac{R_*}{R_\odot} \sim \frac{M_*}{M_\odot}, \tag{10.65}$$

which describes the data in Torres et al. (2010) reasonably well for non-evolved stars with $M_* \lesssim 2M_\odot$.

With these assumptions, we can now use the results from Sect. 3.1 to derive the scaling of the signal of a habitable planet with host star mass for the various methods we have discussed.

- **Radial Velocities**. For radial velocity surveys, the radial velocity and signal-to-noise ratio for planets in the Habitable Zone scale as

$$(\mathrm{S/N})_{\mathrm{RV}} \propto M_p M_*^{-3/2} \qquad \text{(habitable)}. \tag{10.66}$$

  Therefore, all else being equal, Habitable Zone planets are significantly easier to detect around lower-mass stars. In particular, for stars with $M \lesssim 0.2\ M_\odot$, the radial velocity amplitude for an Earth-mass planet in the Habitable Zone is expected to be $\gtrsim$1 m/s, which is within the reach of current instrumentation (Bean et al. 2010b).
- **Astrometry**. The astrometric signal and signal-to-noise ratio for habitable planets scale as

$$(\mathrm{S/N})_{\mathrm{AST}} \propto M_p M_* d^{-1} \qquad \text{(habitable)}, \tag{10.67}$$

  and thus at fixed distance and planet-mass, astrometric surveys are more sensitive to habitable planets orbiting higher-mass stars, provided that the period of the planets is less than the duration of the survey. In addition, higher-mass stars are more luminous and thus have smaller photon noise uncertainties. On the other hand, higher-mass stars are also less common and thus are typically more distant. The net result of these factors is that A and F stars are the most promising targets for astrometric searches for planets in the Habitable Zones of nearby stars (Gould et al. 2003a).
- **Imaging**. For direct detection of habitable planets in thermal equilibrium with their host stars, the planet/star flux ratio and signal-to-noise ratio scale as

$$(\mathrm{S/N})_{\mathrm{dir}} \propto R_p^2 M_*^{-4} \qquad \text{(reflected, habitable)}, \tag{10.68}$$

$$(\mathrm{S/N})_{\mathrm{dir}} \propto R_p^2 M_*^{-5/2} \qquad \text{(thermal, habitable)}, \tag{10.69}$$

  strongly favoring low-mass stars. Note that, by definition, the amount of stellar irradiation for a planet in the Habitable Zone is independent of the mass of host star, and thus for fixed planet properties, the thermal or reflected flux of the planet is also independent of the mass of the host star. The dependence on stellar mass in the above scaling relations therefore arises simply from the change in the flux of the star. However, the second requirement for direct detection is that the angular separation of the planet from its parent star must be larger than the inner working angle of system. At fixed mass, this translates into a maximum distance that a Habitable Zone planet can be detected,

$$d_{\mathrm{max}} = 10\ \mathrm{pc} \left(\frac{\theta_{\mathrm{IWA}}}{100\ \mathrm{mas}}\right)^{-1} \left(\frac{M_*}{M_\odot}\right)^2 \qquad \text{(habitable)}. \tag{10.70}$$

  The number of available targets is $\propto n(M_*) d_{\mathrm{max}}^3$, where $n(M_*)$ is the volume density of stars of a given mass, i.e., the mass function. Since the exponent of the mass function in the local solar neighborhood is generally $\gtrsim -2$, this requirement strongly favors high-mass stars. The optimal mass will depend on the precise details of the survey and the nature of the

noise sources (see, e.g., Agol 2007), but these arguments demonstrate that we can generically expect the sensitivity of imaging surveys for habitable planets to be fairly strongly peaked at intermediate masses.

- **Transits**. The signal-to-noise ratio, transit probability, and period of a transiting habitable planet scale as

$$(\mathrm{S/N})_{\mathrm{tr}} \propto R_p^2 M_*^{-5/2}, \qquad P_{\mathrm{tr}} \propto M_*^{-1}, \qquad P \propto M_*^{5/2}, \qquad (\text{habitable}) \tag{10.71}$$

all of which favor or strongly favor low-mass stars. Furthermore, as discussed above, the radial velocity signals of Habitable Zone planets around low-mass stars are also larger and within reach. Finally, low-mass stars are more common. These various considerations have led to the suggestion that transit surveys of low-mass stars may be the most promising route to detecting habitable Earthlike planets (Gould et al. 2003b; Nutzman and Charbonneau 2008; Blake et al. 2008). Indeed, several such surveys are underway or are being planned (e.g., Charbonneau et al. 2009), with the ultimate goal of finding a Earthlike system whose atmosphere can be characterized with, e.g., the James Webb Space Telescope (Deming et al. 2009).

- **Microlensing**. The system parameters which determine the detectability of a given planetary system with gravitational microlensing are the mass ratio q and projected separation s in units of r_{E}. For a habitable Earthlike planet, these are

$$q \sim 5 \times 10^{-5} \left(\frac{M_p}{M_\oplus}\right)\left(\frac{M_*}{0.5 M_\odot}\right), \tag{10.72}$$

$$s_{\mathrm{HZ}} \equiv \frac{a_{\mathrm{HZ},\perp}}{r_{\mathrm{E}}} \sim 0.1 \left(\frac{M}{0.5 M_\odot}\right)^{3/2} \left(\frac{d_s}{8\ \mathrm{kpc}}\right)^{-1/2} \left[\frac{x(1-x)}{0.25}\right]^{-1/2}, \quad (\text{habitable}) \tag{10.73}$$

where in the expression for s_{HZ}, we have assumed a median projection factor such that $a_{\mathrm{HZ},\perp} = 0.866 a_{\mathrm{HZ}}$. Therefore, for typical microlensing host stars, the Habitable Zone distance is much smaller than the Einstein ring. While mass ratios of $q \sim 10^{-5}$ are readily detectable for planets with separations near the Einstein ring ($s \sim 1$), they are much more difficult to detect for planets with $s \ll 1$. This is because these such planets can only be detected when they perturb the inner image created by the primary host star, and then only when this image is close to the primary and thus highly demagnified (see *Fig. 10-4*). These perturbations are therefore generally quite small. Furthermore, perturbations of the inner image are more strongly suppressed for large source stars (Gould and Gaucherel 1997; Bennett and Rhie 1996).

From (10.73), we see that microlensing favors the detection of habitable planets around higher-mass stars (Di Stefano and Night 2008), and stars that are close to the source or close to the Earth (i.e., such that $x(1-x)$ is small). While current and next-generation ground-based microlensing surveys have essentially no sensitivity to habitable Earthlike planets, specialized surveys for nearby microlensing events, or space-based surveys which boast much larger event rates and detection efficiencies, could potentially detect such systems (Di Stefano and Night 2008; Park et al. 2006; Bennett et al. 2010b; Green et al. 2011).

4 Early Milestones in the Detection of Exoplanets

4.1 Van de Kamp and Barnard's Star

The pre-1995 literature is scattered with several (presumably) spurious claims of detections of planets around nearby stars. Perhaps the best known early claim is that of van de Kamp, who conducted an astrometric campaign to detect "dark" companions to nearby stars (van de Kamp 1986). Van de Kamp's lower limits were impressive, typically ruling out Jupiter-mass objects in periods of years to a couple decades, and he reported several stars as having barely detectable companions of apparently substellar mass. Most intriguing was his report of first one, then later two Jupiter-mass companions to Barnard's star (GJ 699), the second closest stellar system to Earth.

Van de Kamp made astrometric measurements from the positions of the apparent centroids of stellar images on photographic plates and targets such as Barnard's star suffered from having a constantly changing set of astrometric references over his multi-decade survey due to its record-high proper motion (over $10''$/year). Subsequent astrometric and radial velocity work have ruled out his claims (Choi et al. 2012, and references therein).

4.2 PSR 1257+12 and the Pulsar Planets

Pulsars are exquisite clocks, typically producing pulses with periods of order ~1–10^{-3} s. Once these periods are corrected for well-measured linear drifts with time and occasional "glitches" (sudden shifts in period), their precision can rival and even surpass the best atomic clocks on Earth. Successful analysis of pulse arrival times requires carefully solving for the distance and space motion of the pulsar and the removal of the effects of the motion of the observatory.

In 1991, two teams announced having contemporaneously observed unexplained residuals to their timing models indicative of the first known exoplanets: very small, terrestrial planet mass companions orbiting their pulsars. Matthew Bailes and Andrew Lyne (Bailes et al. 1991) reported a timing variation with a period of 6 months apparently due to a 10 $M_\oplus$ companion orbiting the pulsar PSR 1829–10. Meanwhile, Cornell astronomer Alexander Wolszczan had observed a similar sort of signal from a millisecond pulsar, PSR 1257+12, and had recruited Dale Frail of the National Radio Astronomical Observatory to help confirm its position on the sky to perfect the position model. By November 1991, Wolszczan and Frail (1992) had submitted a manuscript on their discovery, and both teams planned to describe their work at the January 1992 meeting of the American Astronomical Society.

At the meeting, Lyne announced that just days earlier, he had discovered an error in his timing model. A tiny positional error combined with an insufficiently precise description of the Earth's orbit had led to the small, periodic, 6 month signal in their residuals that they had mistaken for a planet. With the correct timing model, there was no evidence of a planetary perturber on the pulsar. Lyne's public and frank admission was acknowledged as a laudable demonstration of scientific integrity by a standing ovation at the meeting and an editorial in *Nature* (Lyne and Bailes 1992; Nature 1992; Wolszczan 2012).

The world would not be long with its first exoplanets, however. The very next speaker at the AAS meeting was Wolszczan, whose timing model had correctly accounted for all important effects. Wolszczan described the first planets known outside the Solar System: a pair of bodies with ~4 $M_\oplus$ orbiting the millisecond pulsar PSR 1257+12 with periods of 66 and 98 days.

This system would continue to impress, revealing a third low-mass planet to Wolszczan's team, as well (Wolszczan et al. 2000).

These planets' formation mechanism is still not understood, and to date, no similar system of low-mass planets is known. Signals of higher-mass planets orbiting pulsars would continue to be found, however: in particular, Bailes would go on to discover an apparently high-density 1.4 M_{Jup} object orbiting pulsar PSR J1719-1438 (Bailes et al. 2011).

4.3 Early Radial Velocity Work

4.3.1 Campbell and Walker's Survey and γ Cep A*b*

In the 1981, Bruce Campbell of the University of Victoria, Gordon Walker of the University of British Columbia, and their team began an ambitious survey for substellar objects orbiting 20 of the brightest nearby stars. They sought to exploit the recently available technologies of high-resolution (R > 40,000) echelle spectrographs and digital detectors in the form of Reticons and CCDs to achieve the best possible Doppler precision.

To establish an accurate and robust wavelength solution, they employed an absorption cell of hydrogen fluoride (HF) which imposed a "picket fence" of regular, strong, narrow, and widely spaced (~15 Å) absorption lines from the $3 \rightarrow 0R$ branch transitions near 8,700 Å, far from any telluric features (Campbell et al. 1988). Campbell and Walker had sought a chemical cell that would provide such lines in the optical that would not contaminate the stellar lines and with few extraneous absorption features from isotopic impurities or other effects. Campbell and Walker found no other chemicals that fit their needs and so used HF despite the serious difficulties of working with that chemical.[4] To achieve the necessary optical depth and wavelength stability of HF lines, Campbell and Walker stabilized the temperature and pressure of their 1-m-long cell at 373 K and connected it to a vessel containing liquid HF in an ice bath, yielding a pressure of roughly 0.5 atm.

The well-known and well-spaced HF lines granted Cambpell and Walker unprecedented optical Doppler precision, independent of mechanical and thermal changes in the optics of their spectrograph. This method was, at the time, greatly superior to emission line calibration because the reference lines were measured simultaneously to the stellar exposure and through the same optical path as the stellar photons. Campbell and Walker achieved 10–15 m/s radial velocity precision on most of their sample of 20 stars. This precision is, in retrospect, more than sufficient to detect close-in giant exoplanets, but their sample of stars was simply too small, and close-in planets are too rare for them to have discovered a strong exoplanetary signal with any significant likelihood.

Campbell et al. (1988) were able to confirm that objects with $m \sin i$ = 10–80 M_{Jup} are rare and noted several interesting signals near the limits of their detectability. They noted an especially intriguing apparent signal for γ Cep A with a 25 m/s amplitude and 2.7-year period, superimposed on a long-term acceleration from the star's binary companion. The implication of this periodicity was that γ Cep A was orbited by a ~1.7 M_{Jup} mass planet at a few AU.

[4]Campbell and Walker understatedly reported the "obnoxious" nature of HF and that "standard safety precautions of chemical laboratories are appropriate" for this rather dangerous chemical which reacts with glass, forms hydrofluoric acid on contact with water, can painlessly penetrate human tissues, and causes burns that can necessitate amputation (OSHA Occupational Safety and Health Guidelines for HF, US Department of Labor, 2012).

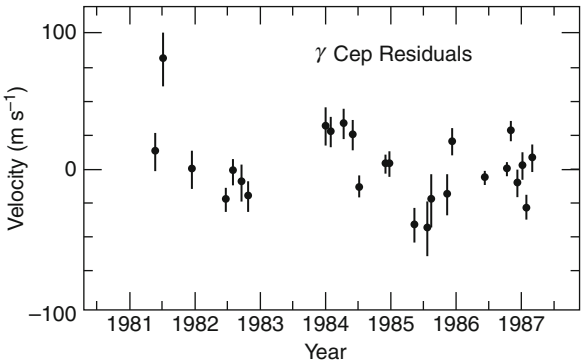


Fig. 10-7
Figure from Campbell et al. (1988) illustrating the first tentative detection of a real exoplanet from the pioneering radial velocity survey (Reproduced by permission of the AAS)

Four years later, Walker et al. (1992) reported on their monitoring of the Ca II 8,662 Å line and determined that γ Cep had a weak 2.52-year activity period, uncomfortably close to the purported planetary signal, and cautiously noted that the RV signal was likely due to stellar activity.

Eleven years later, after nearly 100 bona fide exoplanets had been discovered by teams around the world, Hatzes et al. (2003) announced RV monitoring at McDonald Observatory had confirmed Campbell, Walker, and Young's original detection: the variations were indeed due to a $m \sin i = 1.7\,M_{Jup}$, $P = 2.5$-year planet, and there was no longer any evidence of a 2.5-year activity period. In retrospect, Campbell and Walker's planet search had been a sort of success, after all. It had also inspired subsequent teams to continue their efforts (◉ *Figs. 10-7* and ◉ *10-8*).

4.3.2 Latham's Survey and HD 114762 *b*

In 1988, David Latham of the Harvard-Smithsonian Center for Astrophysics and his team described a new result from their precise radial velocity instrument, the CfA Digital Speedometer at Oak Ridge Observatory. In 1984, Latham and Israeli astronomer Tsevi Mazeh had conducted a short survey of about three dozen early M dwarfs for short-period, massive planets using this instrument but had found nothing (Latham 2012). For this new survey, Latham et al. sought to further stabilize their spectrograph to achieve ~100 m/s precision to measure accurate binary orbits and improve the IAU system of radial velocity standards. Latham et al. achieved this stabilization by removing the Cassegrain instrument from the telescope, stabilizing its temperature, and feeding it with a 100 μ optical fiber. This provided a constant gravity vector, thermal and mechanical stability, and a (relatively) uniformly illuminated entrance slit, robust against guiding errors.

On its first night of operation, Latham et al. (1989) observed several RV standard stars including HD 114762. They noted a large (390 m/s) radial velocity discrepancy from the known value. Curious, they compared their previous measurements made at lower precision and found a highly significant signal at near 84 d with 530 m/s semiamplitude, corresponding to a 13 M_{Jup} companion.

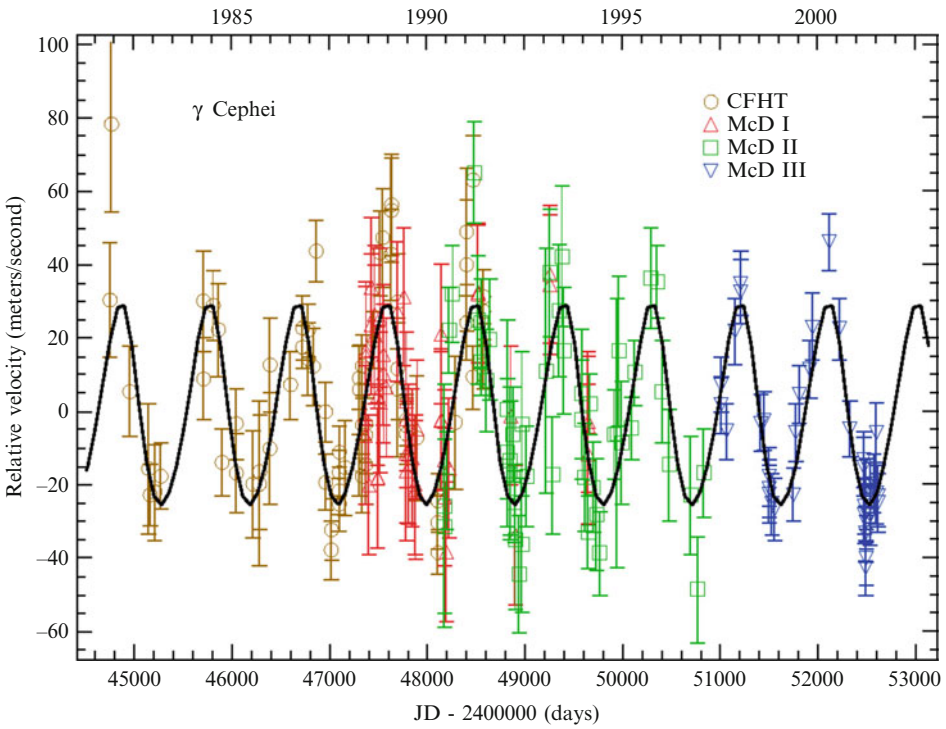


Fig. 10-8

Figure 3 of Hatzes et al. (2003) showing confirmation data for the planet γ Cep *b*, including the original Campbell et al. data (Reproduced by permission of the AAS)

Subsequent observations at high precision confirmed the reality of the signal. Latham et al. (1989) had discovered, serendipitously, on the first night of observation, and in a sample of only seven objects, what today would be considered by many to be an exoplanet. At the time, no objects with a mass anything like 13 M_{Jup} were known, and Latham et al. cautiously referred to their object as "a probable brown dwarf"; however, they noted that the companion "might even be a giant planet,"[5] a point that was picked up by the media but criticized by many of their colleagues (Latham 2012). Regardless of its taxonomic class, it was the first firm detection of a substellar object beyond the Solar System and today is often included in catalogs of exoplanets (e.g., Wright et al. 2011).

4.3.3 Marcy and Butler's Iodine Survey

In 1992, Geoffrey Marcy and Paul Butler of San Francisco State University announced their survey of 70 nearby stars using an iodine (I_2) cell for wavelength calibration. Previous Doppler work by Marcy and Benitz (1989) had used ThAr emission lamps to search for brown dwarfs

[5]Later, the IAU would provisionally set the deuterium-burning limit for star-like objects, 13 M_{Jup}, as the border between planets and brown dwarfs found in orbit around stars. This is useful for purposes of nomenclature but bears little on issues of formation and composition, and today the distinction is not always honored.

orbiting nearby M dwarfs with radial velocity precision of ~250 m/s; Marcy and Butler sought to improve this precision by 1–2 orders of magnitude with their iodine cell and the CCD at the high-resolution Hamilton spectrograph, designed by Vogt (1987).

Unlike Campbell and Walker, Marcy and Butler sought an absorption cell that would provide absorption features *throughout* a broad region of wavelength space. Following Libbrecht (1988) (who had employed an iodine cell to study solar sunspots and had extended their studies to p-mode measurements of stars), Marcy and Butler settled on iodine as the ideal absorption gas. Their rationale was that this provided the wavelength reference at every point in the spectrum, not just every 15 Å, and that the potentially problematic blending of stellar features with iodine features could be modeled given a sufficiently accurate template spectrum of the star and the iodine cell (the latter obtained with a Fourier transform spectrograph).

Marcy and Butler further sought to model the variable instrumental profile of the spectrograph as a function of wavelength to account for the not-insubstantial thermal and mechanical variations in the spectrograph and for the nonuniform illumination of the entrance slit. Marcy and Butler's sealed cell was ~10 cm long, held at a constant 323 K, and 0.001 atm and had the further advantage that it was relatively easy to construct and its contents were benign (indeed, medicinal!). Marcy and Butler (1992) reported that they had achieved 25 m/s precision at the beginning of their survey and foresaw significant improvement through more sophisticated instrumental profile modeling. Indeed, by 1996 Marcy and Butler would demonstrate 3 m/s precision (Butler et al. 1996), and they and their collaborators would go on to be responsible for over half of the exoplanets discovered over the next 15 years.

4.3.4 Hatzes and Cochran's Survey and β Gem *b*

Hatzes and Cochran (1993) reported their results from precise Doppler monitoring of three bright K giants as part of a broader planet detection effort. Their primary technique was to use telluric (atmospheric) O_2 bands as an absorption wavelength reference, which had been reported by Griffin (1973) to be sufficiently stable to allow 10 m/s precision. They found that typical long-term stabilities were more like 20 m/s. They had also begun employing an iodine cell (Cochran and Hatzes 1993) and at this point had obtained a small number of iodine observations.

Hatzes and Cochran found that all three K giants in their sample, Arcturus (α Boo), Aldebaran (α Tau), and Pollux (β Gem), displayed large, periodic radial velocity variations with semiamplitudes of 50–200 m/s. Comparison with prior radial velocities obtained by Cambpell's group (Walker et al. 1989) revealed that the variations were coherent over 10 years. While both α Boo and α Tau showed significant day-to-day RV variations indicative of radial pulsation modes and correlated variations in the 10,830 Å He I line, β Gem seemed to have a clean signal, consistent with a 554 d planet with a minimum mass of 3 M_{Jup}.

Observations by the Canadian team (Larson et al. 1993) showed that the Ca II 8,662 Å line showed periodic variation at the same frequency as the RV variations (which they had measured independently). This coincidence cast strong doubt on the planetary interpretation of the β Gem RV variations, especially in light of the much larger and more clearly activity-related variations in α Boo and α Tau.

Hatzes et al. (2006) combined literature data with subsequent iodine observations from McDonald Observatory and Tautenburg Observatory to show that the RV variations continued coherently into 2006 and that the Ca II H and K lines showed no coincident variation. They concluded that the variations in β Gem were likely due to a minimum mass 3 M_{Jup} companion with period 590 d.

4.3.5 Mayor and Queloz and 51 Pegasi *b*

The first unambiguous detection of a planet-mass object orbiting a normal star was by Mayor and Queloz (1995) of Geneva Observatory. Mayor and Queloz used the ELODIE spectrograph, which achieved 13 m/s precision through outstanding mechanical stability. ELODIE (the successor to CORAVEL) was a fiber-fed spectrograph within a stable, temperature-controlled environment (Queloz et al. 1998). Wavelength calibration was achieved through use of a simultaneous observation of a thorium-argon (ThAr) emission lamp. Mayor and Queloz used cross-correlation with a binary mask to determine the velocity of the stellar spectrum with respect to the known wavelengths from the emission line spectrum. The mechanical stability of the instrument ensured that the offset between the stellar and comparison lamp spectra was fixed and stable, and the scrambling inherent to the fiber ensured that the position of the stellar spectrum did not suffer significantly from variations in illumination or guiding on the fiber tip.

51 Peg *b* has a semiamplitude of only 59 m/s and period of only 4.2 d, implying a minimum mass of 0.5 M_{Jup}. This was a shocking development – planet formation theory had not predicted the existence of such close-in planets,[6] and indeed the tiny mass implied by the detection was smaller than any known binary companion by more than an order of magnitude.[7] Immediately, Marcy and Butler (1995) confirmed the detection, as did Hatzes et al. (1997) soon thereafter. Debate ensued about the nature of the variations and whether they could be due to nonradial pulsation modes (Gray 1997), but the detection of planetary transits would put these concerns to rest: the field of exoplanetary science had begun in earnest. The Geneva team would expand and find great success over the next decades, eventually pushing their precision below 1 m/s with the HARPS spectrograph.

4.4 The First Planetary Transit: HD 209458*b*

The presence of close-in planets provided an opportunity to detect exoplanets directly through transits. The probability that a planet will transit its hosts star is inversely proportional to its orbital distance, and since 51 Peg *b* and similar "Hot Jupiters" orbited at ~10 stellar radii from their host stars, their transit probability was around 10%. Photometrists began to monitor these planets' hosts stars for such events, expecting to find one once the number of known systems approached 10. Concerns over nonradial pulsations also contributed to the desire to monitor stars for photometric evidence of such effects.

Two teams detected the $m \sin i = 0.7\ M_{Jup}$, $P = 3.5$ d planet orbiting HD 209458 independently (Mazeh et al. 2000; Henry et al. 2000) and collaborated with photometrists to conduct

[6] But see the remarkably prescient article by Struve (1952), which all but foresaw this detection, how it would be made, and the subsequent detection of planetary transits.

[7] The third firm detection of a substellar object, the imaging of brown dwarf GJ 299 B, was announced at the same conference as 51 Peg *b*!

the now-standard photometric follow-up prior to publication. Two teams succeeded contemporaneously: their announcements of the detection of the transits of HD 209458 appeared in the literature simultaneously, having been submitted to the Astrophysical Journal within 1 day of each other (Henry et al. 2000; Charbonneau et al. 2000) exemplifying the intense competition to produce exoplanetary "firsts." This measurement of the orbital inclination and radius (and thus the true mass and density) of the planet dispelled any remaining doubt as to the origins of most of the similar RV variations of stars and provided the necessary impetus for large-scale efforts to detect more planets with radial velocities, transits, and, soon, microlensing and direct imaging.

4.5 Microlensing

4.5.1 Microlensing History

While the idea of gravitational microlensing by individual stars was considered sporadically over the past century (Einstein 1936; Eddington 1920; Chwolson 1924; Lodge 1919; Liebes 1964; Refsdal 1964), it was the seminal paper by Paczynski (1986) that gave birth to the modern microlensing field. In this paper, Paczyński argued that it would be feasible to monitor several million stars toward the Magellanic clouds on timescales of a few hours to a few years, in order to search for gravitational microlensing events due to foreground massive compact objects that could make up a substantial fraction of the mass of the dark matter halo of the Milky Way. Within a few years, several collaborations were initiated to survey regions in the Large Magellenic Clouds and Galactic bulge to search for microlensing events (Alcock et al. 1993; Aubourg et al. 1993; Udalski et al. 1993). The first detections followed shortly thereafter and to date, microlensing events of the order 10^4 have been detected, with the majority seen along the line of the sight to the bulge.

Although the original motivation for microlensing surveys was the search for dark matter, it was soon realized that it would be possible to search for planetary companions to the stars and remnants that provided a guaranteed signal for these experiments. Mao and Paczynski (1991) first pointed out that binary lenses whose components were separated by roughly their Einstein ring radius would give to sharp, distinctive light curve features associated with the presence of caustic curves in such systems. Caustics are the set of source positions where extra image pairs are created or destroyed with the source crosses the caustic, resulting in large changes in the total magnification. They also noted that the probability of a source crossing these caustics for a binary lens remained substantial down to mass ratios of $q \sim 10^{-3}$ therefore suggesting that planetary companions could also be detected in this way. Gould and Loeb (1992) consider this idea in detail, refining the estimates for the detection probabilities for different planet-mass ratios and discussing the practical requirements for carrying out an exoplanet survey with microlensing. In particular, they advocated a two-tier strategy, whereby survey collaborations use a single dedicated telescope equipped with a wide-field camera monitor large areas of the sky to identify and alert stellar microlensing events before peak, and follow-up collaborations with access to several longitudinally distributed, narrow-angle telescopes follow particularly promising events at much higher cadence to search for the brief planetary deviations. The first microlensing planet surveys began in 1995, with the first real-time alerts from

the survey collaborations (e.g., Udalski et al. 1994; Alcock et al. 1996) and subsequent monitoring of these alerts by several follow-up collaborations (Alcock et al. 1996; Albrow et al. 1998; Rhie et al. 2000).

Ongoing surveys over the next 6 years (1995–2001) failed to detect any planetary microlensing events. The primary reason for this is that the total number of events alerted by the survey collaborations was relatively low ($\lesssim$100), and therefore there were typically few events ongoing at any given time that were both suitable for follow-up and very sensitive to planets. In particular, there were only a handful of high-magnification events per year, which had been recognized to be intrinsically very sensitive to planetary companions (Griest and Safizadeh 1998). Nevertheless, this phase was important for the field, as the real-world struggles involved with carrying out and analyzing the results from these surveys, including the ensemble of nondetections (Gaudi et al. 2002; Snodgrass et al. 2004), naturally led to the development and maturation of both the theory and practice of the method.

4.5.2 First Planet Detections with Microlensing

The first detections of planets with microlensing were enabled primarily by a series of upgrades by several survey collaborations to their observational setups. In 2001, the OGLE collaboration initiated their third phase with an upgrade to a new camera with a 16 times larger field of view. With this larger field of view, they were able to monitor a larger area of the Galactic bulge with higher cadence and as a result began alerting ~500 microlensing events per year. This higher event rate, combined with improved cooperation between the survey collaborations, led to the first planet discovery in 2003 by the Microlensing Observations in Astrophysics (MOA) and OGLE collaborations (Bond et al. 2004). Shortly thereafter, the MOA collaboration upgraded to a 1.8 m telescope with a 2 deg^2 camera (Sako et al. 2008). By 2007, the MOA and OGLE collaborations were sending alerts for nearly 1,000 microlensing events per year, thus enabling a substantial increase in the rate of planet detections.

The light curve data and best-fit model for the first microlensing planet discovery are shown in ❯ *Fig. 10-9*. This is a "cold Jupiter": the planet has a mass $M_p \sim 3\ \mathrm{M_{Jup}}$ and orbits a star with $M_* \simeq 0.6\ M_\odot$ at a separation of $a \sim 4$ AU or ~2.5 times the snow line distance (Bennett et al. 2006).

To date, 14 microlensing planet detections have been published. The masses and semimajor axes of these planets are shown in ❯ *Fig. 10-5*; they span nearly three decades in mass from a few times the mass of the Earth to several times the mass of Jupiter and are spread over a factor of ~5 in separation, centered at a few times the snow line distance. Notable among these detections are the first discovery of a "cold Super-Earth" (Beaulieu et al. 2006) and the first discovery of Jupiter/Saturn analogue (Gaudi et al. 2008; Bennett et al. 2010a).

5 State of the Art

5.1 Astrometry

Astrometric precision has improved considerably since van de Kamp's work, and the first verifiable astrometric discovery of an exoplanet appears imminent. Pravdo and Shaklan (2009)

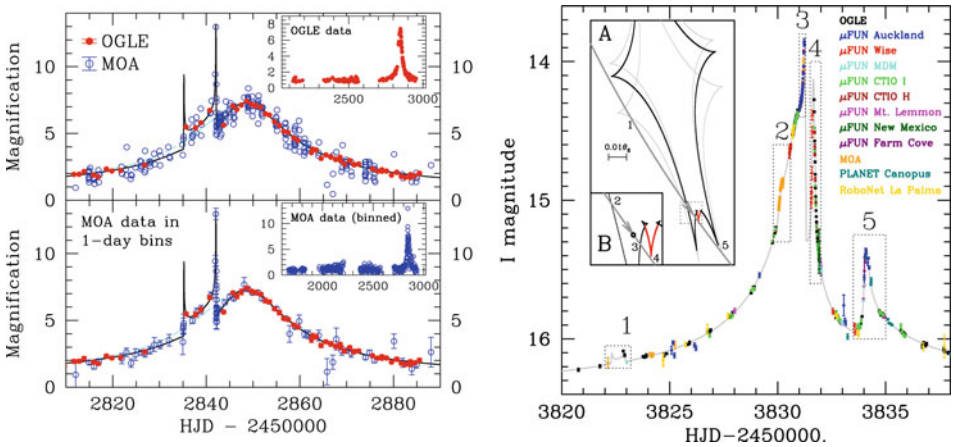


Fig. 10-9

(*Left*) The first discovery of an exoplanet with microlensing in the OGLE 2003-BLG-235/MOA 2003-BLG-53 event (Bond et al. 2004). The *red* and *blue points* show the data from the OGLE and MOA collaborations, respectively. The *top large panel* shows the native data, whereas the MOA data have been binned into 1 day bins in the *bottom panel*. The *black* and *cyan curves* show the best-fit planetary and single-lens model, respectively. The planetary companion is revealed through the brief deviation from the smooth symmetric curve arising from the host star, including the well-covered sharp spike near HJD-2450000~2842 caused by the source crossing a caustic created by the planetary companion. The small insets show the full, multiyear data spans for the OGLE and MOA data (From Bond et al. (2004), reproduced by permission of the AAS). (*Right*) The OGLE-2005-BLG-109 microlensing event, arising from a star with a Jupiter/Saturn analogue two-planet system. *Panel A* shows the source trajectory through the caustic created by the two planets (*dark-gray curve*). The five light curve features are caused by the source crossing or approaching the caustic, and the locations of these features indicated with *numbers*. The majority of the caustic (in *black*) is due to the Saturn analogue planet, which explains 4 of the 5 features. The portion of the caustic arising from the Jupiter analogue planet is shown in *red*. This additional caustic is required to explain the fourth feature in the light curve. The *light-gray curves* show the caustic at the time of features 1 and 5. *Panel B* shows the detail of the source trajectory and caustic near the times of the second, third, and fourth features (From Gaudi et al. 2008)

announced that their long-term astrometric monitoring of the ultracool dwarf star vB 10 had revealed a ~6 M_{Jup} companion in a 9 month orbit; however, subsequent follow-up with radial velocities determined that the signal was spurious (Bean et al. 2010a; Anglada-Escudé et al. 2010; Lazorenko et al. 2011).

More promisingly, Muterspaugh et al. (2010) used an optical interferometer to carefully measure the astrometric motions of binary stars and combined these measurements with radial velocities of the systems to search for low-mass companions to stars in tight binaries. The project concluded with six planetary candidates, including two "high confidence" members that could prove to be the first astrometrically detected exoplanets. If real, these planets will put strong constraints on planet formation theories in binary systems.

The astrometric detection of planets discovered by other means has produced substantially more results, primarily because the approximate astrometric signals are known from prior radial velocity work, and so searches are more efficient. Most fruitful has been work on nearby stars employing the *Hubble Space Telescope* Fine Guidance Sensor, which is capable of precise astrometry of bright stars. This has revealed some high-mass planet candidates from radial velocity surveys to be binary stars in face-on orbits and has revealed the mutual inclinations of planets in multiplanet systems (Bean et al. 2007; Bean and Seifahrt 2009; McArthur et al. 2010).

5.2 Imaging

5.2.1 2M1207*b*

Chauvin et al. (2010) describe their survey of young, nearby stars for low mass, possibly planetary companions using the ESO/VLT 8 m telescope equipped with the NACO AO system and infrared camera. They began their survey in 2002 during commissioning and targeted, among other things, the lowest mass members of known, nearby, young stellar associations. This allowed them to maximize the separation and contrast of companions and push coronagraphy into the planetary-mass regime.

After some promising detections of higher-mass objects, the survey bore fruit when Chauvin et al. (2004) detected a very low-mass companion to the M8 TW Hydra association (TWA) brown dwarf 2MASSW J1207334-393254 (called 2M1207; *Fig. 10-10*) at a separation of only

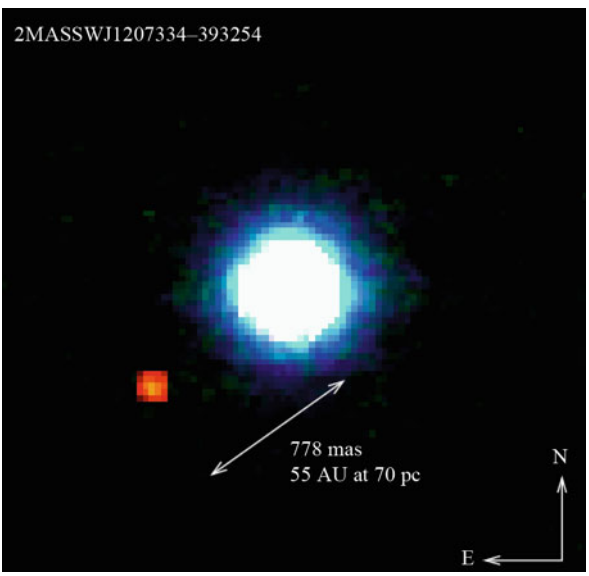


Fig. 10-10
NACO image of 2M1207*b*. The primary is a M8 dwarf at ~70 pc; the secondary is much cooler late L dwarf with probable mass ~5 M_{Jup} (From Chauvin et al. (Fig. 11 2004, credit G. Chauvin and ESO))

0.″8 (~55 AU). Membership in the TWA yielded an age for the companion, and a distance was estimated from the colors and brightness of 2M1207. Comparison with models of the thermal evolution of young objects (Burrows et al. 1997; Chabrier et al. 2000; Baraffe et al. 2002) allowed Chauvin et al. (2004) to estimate the companions mass to be ~5 M_{Jup}, assuming that it was indeed a bound, coeval object and not a background contaminant.

Chauvin et al. (2005) followed up their prior observations to confirm common proper motion using the same instrument. They demonstrated that the two objects share proper motion and parallactic motion with this star, all but proving that they form a bound pair.

The nature of 2M1270*b* was, and still is, unclear. Its status as a "planetary-mass" object seems secure, but its wide separation (55 AU) and ~0.2 mass ratio with its primary made the pair perhaps more analogous to a scaled-down binary star system than a planet-star system. Chauvin et al. (2005) noted that a protoplanetary disk origin for the *b* component seemed unlikely. Nonetheless, they had acquired the first image of a planet, by mass if not by formation mechanism, and this presaged the many more successes to come from high contrast imaging. The ESO/VLT group would go on to detect the planet-mass object AB Pic *b* and many faint stellar companions to nearby stars, including many planet hosts (Chauvin et al. 2010).

5.2.2 Fomalhaut *b*

Kalas et al. (2005) used the *Hubble Space Telescope (HST)* Advanced Camera for Surveys (ACS) to image the dust belt orbiting Fomalhaut (α Piscis Austrini), a ~400 Myr-old A4 dwarf at 7.7 pc. Debris disks or belts are a common feature of many young main-sequence stars, typically with structure consistent with dynamical interactions from unseen planets (see chapter by Moro-Martin). The coronagraphic mode of ACS allowed the *HST* team to make a spectacular and detailed image of the disk in reflected optical light, revealing that its dominant feature is a highly inclined, off-center belt with an apparently sharp inner edge at 133 AU radius. The 15 AU geometric offset of the belt from the star and the sharp inner edge could be explained by the presence of a planetary companion with nonzero eccentricity "sculpting" the inner edge of the belt and maintaining the geometric offset via secular perturbation (Wyatt et al. 1999).

Follow-up observations to determine the structure of the disk using a variety of PSF subtraction techniques allowed Kalas et al. (2008) to confirm a persistent source with optical brightness ~25 mag and located ~13″ away from the primary star, just inside the belt, and consistent with the "sculpting" hypothesis. Comparison of multiple epochs allowed Kalas's team to measure the orbital motion of the object astrometrically and confirm that it is a proper motion companion of Fomalhaut.

Puzzlingly, this very faint companion did not appear in their infrared imaging at the Keck 10 m and Gemini 8 m telescopes, which is inconsistent with the optical emission being thermal in origin according to current models. Kalas et al. were able to use the dust disk itself to constrain the object's mass dynamically to be <3 M_{Jup} (Chiang et al. 2009) and thus the lowest mass directly imaged planet candidate to date. Kalas et al. (2008) proposed several possible explanations for Fomalhaut b's unusual optical brightness, such as reflection from a circumplanetary dust ring that could be as large as 35 planetary radii, though still significantly smaller than Saturn's Phoebe ring. They also proposed an alternative model where reflected light is due to a transient dust cloud produced by a rare destructive collision between two analogues of Solar System Kuiper Belt objects.

As with 2M1207*b*, ascertaining the nature of Fomalhaut *b* will require additional observations and theoretical input. It orbits an intermediate mass star with semimajor axis ~120 AU and is still undetected in the infrared. Whatever its nature, its presence in a disk of material strongly implicates a disk origin for the object and provides hope for more secure similar detections of similarly bright planetary objects in the future.

5.2.3 β Pictoris *b*

β Pic is an A star with a prominent, edge-on debris disk that was the first to be imaged in optical scattered light (Smith and Terrile 1984). One unexpected result was that the disk had several asymmetries in structure (Kalas and Jewitt 1995), including a vertical warp in the disk midplane at <100 AU projected radius from the star (Burrows et al. 1995). These and other phenomenon observed toward β Pic indirectly suggested the existence of a planetary system, and β Pic *b* was finally directly imaged using VLT/NACO (Lagrange et al. 2009). As with Fomalhaut, the debris disk structure could be used to constrain the planet mass through dynamical theory (Mouillet et al. 1997), as an alternative to mass estimates based on planet luminosity models. Its measured L' brightness of 11th mag corresponds to a ~8 M_{Jup} planet at age ~10 Myr.

β Pic b is currently unique among the directly imaged exoplanets for having the smallest semimajor axis, which, at ~8 AU, corresponds to the approximate ice line of the system. Unfortunately, the projected separation is also very small, 0.″4, and follow-up spectroscopic study has yet to be obtained. Systems such as Beta Pic are therefore ideal targets for the next generation of extreme adaptive optics instrumentation discussed below. These future results will provide important tests of various planet formation and luminosity evolution models, which, in the 10 Myr-age regime, offer significantly different predictions for the physical properties of Jupiter-mass planets.

5.2.4 The HR 8799 Planetary System

A US/Canadian team led by Marois used AO coronagraphy on the Keck and Gemini telescopes in angular differential imaging mode (ADI, Marois et al. 2006), which exploits the rotation of the field of view on an altitude/azimuth telescope with time to distinguish (and subtract) PSF features from astrophysical sources. Marois' team observed the nearby (40 pc) A star HR 8799, which was known to have an IR excess and to be relatively young (20–160 Myr), making any orbiting planets likely to be bright in the near infrared. Marois et al. (2008) reported the discovery of three faint companions to HR 8799, with proper motions consistent with HR 8799 and detectable orbital motion. Comparison with Pleiades brown dwarf brightnesses demonstrated that these objects had likely masses below 11 M_{Jup}. Further observations at Keck Observatory allowed Marois et al. (2010) to discover a fourth, *e* component to the system. The projected orbital separations of the four planets range from 14 to 70 AU.

This family of objects, the first imaged multiplanet system, poses special challenges for planet formation theory. Marois et al. (2010) argue that their masses and orbital radii (and the relative scarcity of other systems such as this) are inconsistent with both in situ gravitational instability disk fragmentation and core-accretion scenarios. Like the close-in "Hot Jupiters," these distant planets would seem to implicate migration in a disk as a primary architectural factor in planetary system formation.

5.2.5 SPHERE, GPI, and Project 1640

Several "extreme" adaptive optics systems with coronagraphs are in development or operation and will be capable of detecting young (<1 Gyr) giant planets at a few diffraction widths from the position of a bright star. Such systems employ deformable mirrors with several hundreds to thousands of actuators and typically observe bright stars in the near infrared. They use integral field spectrographs that produce data cubes (i.e., a low resolution spectrum at each angular position) and can exploit field rotation to employ a variety of PSF subtraction and speckle suppression techniques. They thus produce large data volumes, including spectra of their imaged planets and will typically have narrow fully-corrected fields of view ($<1''$).

Project 1640 (Hinkley et al. 2011), already in operation, is a collaboration between the American Museum of Natural History and the University of Cambridge. It employs the PALM 3000 adaptive optics system on the Hale 200 in. telescope at Palomar Observatory. SPHERE (Beuzit et al. 2006) and GPI (Macintosh et al. 2008) are next-generation coronagraphic imagers on the VLT and Gemini South telescopes, respectively. Both will employ thousands of actuators and execute campaigns to discover young Jupiter-mass planets orbiting at several AU from the nearest Sun-like stars. Depending on the adopted planet formation and luminosity evolution model, the detection rate predicted for GPI ranges between 10% and 25%, given a target sample with age <100 Myr within 75 pc (McBride et al. 2011). Therefore, if these instruments observe ~500 stars from this sample, then there will be at least 50 new exoplanets discovered and characterized via direct imaging.

5.3 Rocky and Habitable Worlds

5.3.1 HARPS, Keck/HIRES, and the Planet Finding Spectrograph

The first RV-discovered planets had typical Doppler amplitudes of ~50–500 m/s; the 10 m/s barrier was breached several times between 2000 and 2005, and detections between 2 and 5 m/s were common between 2005 and 2010. The primary instruments making these detections were the HIRES instrument on the Keck I telescope (operated by various teams of Marcy and Butler) with precision as low as 1–3 m/s and the HARPS spectrograph (with its heritage from the Geneva team) which regularly achieves precision below 1 m/s on bright stars. Several next-generation planet finding spectrographs are being built or commissioned at this writing, including the Planet Finding Spectrograph at Magellan, HARPS-North, and ESPRESSO.

There are two primary obstacles to further precision in radial velocity surveys toward the 10 cm/s necessary for the detection of true Earth analogues. The first is an instrumental stability issue: calibrating the wavelength solution of spectrographs to an order of magnitude better precision than previously possible. For emission-lamp calibration, a fundamental limit is the lifetime and stability of the thorium-argon lamps used as wavelength fiducial. A promising solution is the use of laser frequency combs, which provide essentially arbitrary levels of wavelength precision. Such devices will be used in HARPS-North and ESPRESSO. Also of importance is understanding and maximizing the consistency of the illumination of the output fibers used to guide light to the spectrograph.

For absorption-cell instruments, the practical limit is one's ability to model the system, in particular, the absorption cell, the slit illumination function, the instrumental profile, and the star itself. Progress here is primarily made through careful FTS scans of the actual cell used

at the telescope, improved modeling techniques that account for scattered light, and better deconvolution techniques for acquiring stellar templates.

The second obstacle is that of the fundamental stability of the stars themselves. Stars experience p-mode oscillations at the few m/s level that must be either modeled or averaged over. Stellar magnetic surface activity also contributes to radial velocity signatures in many ways, most importantly through rotationally modulated spots and plage, and perhaps through long-term stellar cycles. Modeling and mitigating these effects, perhaps though careful use of contemporaneous photometry, will be necessary to achieve the next large improvement in RV precision for the next generation of planet searches.

5.3.2 Space-Based Transit Surveys

Transit surveys for rocky planets around solar-type stars require extremely precise relative photometry. The fractional depth of the transit of an Earth-sized planet passing in front of a solar radius star is only $\delta \sim 8 \times 10^{-5}$. While relative photometry at a few times this level has been achieved from the ground for individual bright stars with specialized techniques (e.g., Johnson et al. 2009; Southworth et al. 2009; Colón et al. 2010), obtaining $\lesssim 10^{-4}$ relative photometry for the large ensembles of stars required to detect numerous transiting systems is probably out of reach for ground-based surveys, due to unavoidable systematics arising from variations due to the Earth's atmosphere. The stability afforded by space-based surveys, on the other hand, enables relative photometry for large numbers of stars that is limited primarily by photon and astrophysical noise. Furthermore, for space-based surveys, it is possible to obtain continuous photometry for very long periods of time, without diurnal or weather interruptions. This eliminates the aliasing problems that are germane to single-site ground-based transit surveys and enables the detection of long-period transiting systems, which, due to the low duty cycles and long transit durations, are extremely difficult to detect from the ground.

CoRoT is a CNES-led space mission with participation from ESA and other international partners, with the primary goals of studying stars via asteroseismology and detecting transiting planets (Baglin 2003). The 27 cm telescope was launched in December 2006 and is located in a low Earth orbit. CoRoT is equipped with a 3.05° by 2.7° camera that primarily monitors fields in two different areas of the sky, located toward Galactic longitudes of ~40° and ~210° (Auvergne et al. 2009). There are two dwell times for the fields; long fields are typically observed for ~150 days, whereas short fields are observed for ~30 days. To date, more than 130,000 stars in ~20 fields have been monitored with a cadence of 8.5 min (Michel and Baglin 2012). The stellar populations of these fields vary dramatically, but anywhere from 40% to 60% of the targets are expected to be dwarf stars suitable for transit surveys (Cabrera et al. 2009; Erikson et al. 2012). Over the typical $R \sim 12-16$ magnitude range of the targets, CoRoT achieves relative photometry on time scales of ~2 h at the level of $\sim 10^{-4}$ at the bright end, degrading to $\sim 10^{-3}$ at $R \sim 16$ (Aigrain et al. 2009). The precision and cadence is sufficient to detect Jupiter-sized companions over the entire magnitude range, whereas Neptunes and Super-Earths can be detected around the brighter stars (e.g., Cabrera et al. 2009).

To date, CoRoT has announced over 20 detections of transiting planets and brown dwarfs. Notable among these discoveries are the detection of a transiting brown dwarf with a mass of ~60 M_{Jup} (Bouchy et al. 2011), the detection of a Jupiter-sized planet with a relatively long period of ~95 days (Deeg et al. 2010), and the first detection of a transiting Super-Earth, with a radius of ~1.7 $R_\oplus$ and a mass of $3 - 10\ M_\oplus$ (Léger et al. 2009).

Kepler is a NASA mission launched in March of 2009, with the primary goal of measuring the frequency of rocky planets in the Habitable Zones of sun-like stars (Borucki et al. 2010). To accomplish this, the 0.95 m *Kepler* telescope situated in an Earth-trailing orbit is monitoring a 105 $\deg^2$ field of view near the constellation Cygnus continuously over the 7+ year lifetime of the mission. Light curves of ~200,000 stars have been obtained over the course of the mission, with typical sampling cadences of ~30 min, amounting to ~24,000 observations over the first ~500 days of the mission for the subset of stars monitored continuously during this time (Batalha et al. 2012). A subset of stars have higher cadences of ~1 min. The majority of the target stars are solar-type dwarfs with $T_{\rm eff} \sim 5{,}000-6{,}500$ K and $\log g \sim 4.5$, but the full range of target properties span $T_{\rm eff} \simeq 3{,}500-10{,}000$ K and $\log g \simeq 3-5$ (Batalha et al. 2012). The photometric precision and intrinsic stellar variability of this sample is discussed in Gilliland et al. (2011). Relevant to the primary mission goal, the photometric variability for $V \lesssim 12$ stars on transit time scales is $\sim 3 \times 10^{-5}$, quite an impressive figure, but ~50% larger than originally anticipated, primarily due to the fact that typical target stars turn out to be a factor of ~2 times more variable than the Sun, probably due to their relative youth. Although this increased noise reduces the expected sensitivity and so yield of the mission for habitable Earthlike planets, this reduced sensitivity was offset with a mission extension (Gilliland et al. 2011).

Although the data set is not yet sufficient to reliably detect true Earth analogues, the exquisite photometric precision and large sample size have already enabled an fantastic array of science. In particular, based on the first ~500 days of data, a total of ~2,300 transiting planet candidates have been identified (Batalha et al. 2012), including ~360 multiplanet systems (Fabrycky et al. 2012). The majority of these candidates are smaller than Neptune. Despite the fact that a small fraction of these signals have been confirmed by either radial velocity or transiting timing methods, the overwhelming majority of these signals are expected to be due to real planets, from a number of lines of evidence (e.g., Morton and Johnson 2011; Lissauer et al. 2012). Over 30 systems have been confirmed from various methods, including a system with six transiting planets (Lissauer et al. 2011), the first discovery of a circumbinary planet (Doyle et al. 2011), and several planets with radius of $\lesssim R_\oplus$ (Muirhead et al. 2012; Fressin et al. 2012).

The masses and semimajor axes of the confirmed *CoRoT* and *Kepler* systems with mass measurements as of Dec. 2011 are shown in ❯ *Fig. 10-5*. Comparing these to the planets discovered by ground-based transit surveys highlights the large expansion of discovery space that is enabled by going to space.

5.3.3 Second-Generation Microlensing Surveys

Microlensing exoplanet searches are currently in the midst of a transition to the second generation of surveys that will enable the routine detection of rocky planets. Although there have been substantial modifications and upgrades to the details of the "two-tier" survey strategy initially suggested by Gould and Loeb (1992), up until very recently, this basic approach has been in use. Second-generation surveys will operate in a very different manner. With the development of very large format CCD cameras with fields of view of a square degree or greater, it becomes possible to monitor tens of square degrees of the Galactic bulge containing roughly 100 million stars with cadences of tens of minutes. These cadences are sufficient to detect both the primary event *and* detect the perturbations from low-mass ($\sim M_\oplus$) planets therefore obviating the need for a follow-up observations and allowing for more uniform data and a more objective detection criteria. In order to detect all the planetary perturbations, including those that last less than

a day, a longitudinally distributed network of 1–2 m class telescopes equipped with such wide FOV cameras is required.

The transition to the next-generation survey model began in 2006 when MOA upgraded to the dedicated MOA-II telescope in New Zealand, which has a diameter of 1.8 m and 2.2 deg^2 FOV (Sako et al. 2008; Hearnshaw et al. 2006). In 2010, OGLE upgraded to the 1.4 deg^2 OGLE-IV camera on their dedicated 1.3 m telescope in Chile (Udalski 2009). The Wise Observatory's 1.0 m telescope in Israel has recently been equipped with a 1 deg^2 camera (Gorbikov et al. 2010). These three groups are collaborating to continuously monitor an ~8 deg^2 region of the bulge with cadences of 15–30 min (Shvartzvald and Maoz 2012), and the first planet detection with this strategy was recently announced (Yee et al. 2012).

The next milestone in the development of second-generation microlensing surveys will be the completion of the Korean Microlensing Telescope Network (KMTNet). KMTNet is an ambitious, fully funded plan by the Korean government to build three identical 1.6 m telescopes with 4 deg^2 FOV cameras. These will be located in South America, South Africa, and Australia. First light for the final telescope is scheduled for late 2014.

These second generation surveys are expected to increase the planet yields by roughly an order of magnitude over current surveys (Bennett 2004; Shvartzvald and Maoz 2012) and enable the detection of Earth-mass planets as well as free-floating planets with masses greater than ~10 $M_\oplus$.

6 Conclusions

In the roughly two decades since the first detections of planets outside the Solar System, the field of exoplanets has grown enormously, developing into one of the forefront research areas in astronomy. The count of confirmed planets is now over 700, with the sample doubling in size every few years at the current rate. New techniques, methods, experiments, instruments, telescopes, and satellites to detect exoplanets are continually being developed and are enabling the detection and characterization of an increasingly broad diversity of planets orbiting a wider and wider range of hosts. These efforts are not only constantly uncovering new and unexpected types of planetary systems but are beginning to allow for the robust statistical characterization of the demographics of large samples of exoplanets spanning a wide range of parameter space. These efforts ultimately serve to allow us achieve the more general goals of placing our Solar System in the context of planetary systems through the Galaxy, understanding the physics of planetary formation and evolution, and determining the frequency of habitable and inhabited worlds.

Acknowledgments

B.S.G. would like to thank Thomas Beatty and Karen Mogren for permission to discuss results in advance of submission, and acknowledges support from NSF CAREER Grant AST-1056524.

J.T.W. acknowledges support by funding from the Center for Exoplanets and Habitable Worlds. The Center for Exoplanets and Habitable Worlds is supported by the Pennsylvania State University, the Eberly College of Science, and the Pennsylvania Space Grant Consortium.

This work makes extensive use of NASA's Astrophysics Data System, the Exoplanet Orbit Database at exoplanets.org, and the Extrasolar Planets Encyclopedia at exoplanet.eu.

References

Agol, E. 2007, MNRAS, 374, 1271

Agol, E., Steffen, J., Sari, R., & Clarkson, W. 2005, MNRAS, 359, 567

Aigrain, S., Pont, F., Fressin, F., et al. 2009, A&A, 506, 425

Albrow, M., Beaulieu, J.-P., Birch, P., et al. 1998, ApJ, 509, 687

Alcock, C., Akerlof, C. W., Allsman, R. A., et al. 1993, Nature, 365, 621

Alcock, C., Allsman, R. A., Alves, D., et al. 1996, ApJ, 463, L67

Anglada-Escudé, G., Shkolnik, E. L., Weinerger, A. J., Thompson, I. B., Osip, D. J., & Debes, J. H. 2010, ApJ, 711, L24

Applegate, J. H. 1992, ApJ, 385, 621–629

Aubourg, E., Bareyre, P., Bréhin, S., et al. 1993, Nature, 365, 623

Auvergne, M., Bodin, P., Boisnard, L., et al. 2009, A&A, 506, 411

Baglin, A. 2003, Adv. Space Res., 31, 345

Bailes, M., Lyne, A. G., & Shemar, S. L. 1991, Nature, 352, 311

Bailes, M., Bates, S. D., Bhalerao, V., et al. 2011, Science, 333, 1717

Bakos, G., Noyes, R. W., Kovács, G., et al. 2004, PASP, 116, 266

Ballard, S., Fabrycky, D., Fressin, F., et al. 2011, ApJ, 743, 200

Baraffe, I., Chabrier, G., Allard, F., & Hauschildt, P. H. 2002, A&A, 382, 563

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2012, arXiv:1202.5852

Bean, J. L., & Seifahrt, A. 2009, A&A, 496, 249

Bean, J. L., McArthur, B. E., Benedict, G. F., et al. 2007, AJ, 134, 749

Bean, J. L., Seifahrt, A., Hartman, H., Nilsson, H., Reiners, A., Dreizler, S., Henry, T. J., & Wiedemann, G. 2010a, ApJ, 711, L19

Bean, J. L., Seifahrt, A., Hartman, H., et al. 2010b, ApJ, 713, 410

Beaulieu, J.-P., Bennett, D. P., Fouqué, P., et al. 2006, Nature, 439, 437

Bennett, D. P. 2004, Extrasol. Planet, 321, 59

Bennett, D. P. 2008, Exoplanets: Detection, Formation, Properties, Habitability, ed. J. Mason (Berlin: Springer), 47–88

Bennett, D. P., & Rhie, S. H. 1996, ApJ, 472, 660

Bennett, D. P., & Rhie, S. H. 2002, ApJ, 574, 985

Bennett, D. P., Anderson, J., Bond, I. A., Udalski, A., & Gould, A. 2006, ApJ, 647, L171

Bennett, D. P., Anderson, J., & Gaudi, B. S. 2007, ApJ, 660, 781

Bennett, D. P., Rhie, S. H., Nikolaev, S., et al. 2010a, ApJ, 713, 837

Bennett, D. P., Anderson, J., Beaulieu, J. P., et al. 2010b, arXiv:1012.4486

Beuzit, J.-L., Feldt, M., Dohlen, K., et al. 2006, Messenger, 125, 29

Blake, C. H., Bloom, J. S., Latham, D. W., et al. 2008, PASP, 120, 860

Bond, I. A., Udalski, A., Jaroszyński, M., et al. 2004, ApJ, 606, L155

Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977

Bouchy, F., Deleuil, M., Guillot, T., et al. 2011, A&A, 525, A68

Brown, R. A. 2005, ApJ, 624, 1010

Burrows, C. J., Krist, J. E., Stapelfeldt, K. R., & The WFPC2 Investigation Definition Team 1995, BAAS, 187, 3205

Burrows, A., Marley, M., Hubbard, W. B., et al. 1997, ApJ, 491, 856

Butler, R. P., Marcy, G. W., Williams, E., McCarthy, C., Dosanjh, P., & Vogt, S. S. 1996, PASP, 108, 500

Cabrera, J., Fridlund, M., Ollivier, M., et al. 2009, A&A, 506, 501

Campbell, B., Walker, G. A. H., & Yang, S. 1988, ApJ, 331, 902

Chabrier, G., Baraffe, I., Allard, F., & Hauschildt, P. 2000, ApJ, 542, 464

Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, ApJ, 529, L45

Charbonneau, D., Berta, Z. K., Irwin, J., et al. 2009, Nature, 462, 891

Chauvin, G., Lagrange, A.-M., Dumas, C., Zuckerman, B., Mouillet, D., Song, I., Beuzit, J.-L., & Lowrance, P. 2004, A&A, 425, L29

Chauvin, G., Lagrange, A.-M., Dumas, C., Zuckerman, B., Mouillet, D., Song, I., Beuzit, J.-L., & Lowrance, P. 2005, A&A, 438, L25

Chauvin, G., et al. 2010, A&A, 509, A52+

Chiang, E., Kite, E., Kalas, P., Graham, J. R., & Clampin, M. 2009, ApJ, 693, 734

Choi, J., et al. 2012, ApJ, arXiv:1208.2273

Chwolson, O. 1924, Astronomische Nachrichten, 221, 329

Cochran, W. D., & Hatzes, A. P. 1993, in ASP Conf. Ser. 36, Planets Around Pulsars, ed. J. A. Phillips, S. E. Thorsett, & S. R. Kulkarni (San Francisco, CA: ASP), 267–273

Colón, K. D., Ford, E. B., Lee, B., Mahadevan, S., & Blake, C. H. 2010, MNRAS, 408, 1494

Cumming, A. 2004, MNRAS, 354, 1165

Deeg, H. J., Moutou, C., Erikson, A., et al. 2010, Nature, 464, 384

Deming, D., Seager, S., Winn, J., et al. 2009, PASP, 121, 952
Di Stefano, R., & Night, C. 2008, arXiv:0801.1510
Dominik, M., Jørgensen, U. G., Rattenbury, N. J., et al. 2010, Astronomische Nachrichten, 331, 671
Doyle, L. R., Carter, J. A., Fabrycky, D. C., et al. 2011, Science, 333, 1602
Eddington, A. S. 1920, Cambridge Science Classics (Cambridge: Cambridge University Press), 134
Einstein, A. 1936, Science, 84, 506
Eisner, J. A., & Kulkarni, S. R. 2001, ApJ, 550, 871
Erikson, A., Santerne, A., Renner, S., et al. 2012, A&A, 539, A14
Erskine, D. J., & Ge, J. 2000, in ASP Conf. Ser. 195, Imaging the Universe in Three Dimensions, ed. W. van Breugel & J. Bland-Hawthorn (San Francisco, CA: ASP), 501–+
Fabrycky, D. C., Lissauer, J. J., Ragozzine, D., et al. 2012, arXiv:1202.6328
Ford, E. B., Ragozzine, D., Rowe, J. F., et al. 2012, arXiv:1201.1892
Fressin, F., Torres, G., Rowe, J. F., et al. 2012, Nature, 482, 195
Gaudi, B. S. 2010, in Exoplanets, ed. S. Seager (Tucson: University of Arizona Press), 79
Gaudi, B. S. 2012, ARAA, in press
Gaudi, B. S., Albrow, M. D., An, J., et al. 2002, ApJ, 566, 463
Gaudi, B. S., Bennett, D. P., Udalski, A., et al. 2008, Science, 319, 927
Gilliland, R. L., Chaplin, W. J., Dunham, E. W., et al. 2011, ApJS, 197, 6
Gorbikov, E., Brosch, N., & Afonso, C. 2010, Ap&SS, 326, 203
Gould, A. 2000, ApJ, 535, 928
Gould, A., & Gaucherel, C. 1997, ApJ, 477, 580
Gould, A., Loeb, A. 1992, ApJ, 396, 104
Gould, A., Ford, E. B., & Fischer, D. A. 2003a, ApJ, 591, L155
Gould, A., Pepper, J., & DePoy, D. L. 2003b, ApJ, 594, 533
Gould, A., Dorsher, S., Gaudi, B. S., & Udalski, A. 2006, Acta Astron., 56, 1
Gould, A., Dong, S., Gaudi, B. S., et al. 2010, ApJ, 720, 1073
Gray, D. F. 1997, Nature, 385, 795
Green, J., Schechter, P., Baltay, C., et al. 2011, arXiv: 1108.1374
Griest, K., & Safizadeh, N. 1998, ApJ, 500, 37
Griffin, R. 1973, MNRAS, 162, 243
Hatzes, A. P., & Cochran, W. D. 1993, ApJ, 413, 339
Hatzes, A. P., Cochran, W. D., & Johns-Krull, C. M. 1997, ApJ, 478, 374
Hatzes, A. P., Cochran, W. D., Endl, M., McArthur, B., Paulson, D. B., Walker, G. A. H., Campbell, B., & Yang, S. 2003, ApJ, 599, 1383
Hatzes, A. P., Cochran, W. D., Endl, M., et al. 2006, A&A, 457, 335
Hearnshaw, J. B., Abe, F., Bond, I. A., et al. 2006, The 9th Asian-Pacific Regional IAU Meeting, (Indonesia, Bali: Nusa Dua), 272
Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2000, ApJ, 529, L41
Hinkley, S., Oppenheimer, B. R., Zimmerman, N., et al. 2011, PASP, 123, 74
Holman, M. J., & Murray, N. W. 2005, Science, 307, 1288
Holtzman, J. A., Watson, A. M., Baum, W. A., et al. 1998, AJ, 115, 1946
Horne, K., Snodgrass, C., & Tsapras, Y. 2009, MNRAS, 396, 2087
Ida, S., & Lin, D. N. C. 2004, ApJ, 604, 388
Johnson, J. A., Winn, J. N., Cabrera, N. E., & Carter, J. A. 2009, ApJ, 692, L100
Kalas, P., & Jewitt, D. 1995, AJ, 110, 794
Kalas, P., Graham, J. R., & Clampin, M. 2005, Nature, 435, 1067
Kalas, P., et al. 2008, Science, 322, 1345
Kane, S. R., Mahadevan, S., von Braun, K., Laughlin, G., & Ciardi, D. R. 2009, PASP, 121, 1386
Kasdin, N. J., Vanderbei, R. J., Spergel, D. N., & Littman, M. G. 2003, ApJ, 582, 1147
Kasting, J. F., Whitmire, D. P., & Reynolds, R. T. 1993, Icarus, 101, 108–128
Kennedy, G. M., & Kenyon, S. J. 2008, ApJ, 673, 502
Kiraga, M., & Paczynski, B. 1994, ApJ, 430, L101
Lagrange, A.-M., Gratadour, D., Chauvin, G., et al. 2009, A&A, 493, L21
Larson, A. M., Irwin, A. W., Yang, S. L. S., Goodenough, C., Walker, G. A. H., Walker, A. R., & Bohlender, D. A. 1993, PASP, 105, 825
Latham, D. W. 2012, New A Rev., 56, 16
Latham, D. W., Stefanik, R. P., Mazeh, T., Mayor, M., & Burki, G. 1989, Nature, 339, 38
Lazorenko, P. F., et al. 2011, A&A, 527, A25+
Léger, A., Rouan, D., Schneider, J., et al. 2009, A&A, 506, 287
Libbrecht, K. G. 1988, ApJ, 330, L51
Liebes, S. 1964, Phys. Rev., 133, 835
Lissauer, J. J. 1987, Icarus, 69, 249
Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al. 2011, Nature, 470, 53
Lissauer, J. J., Marcy, G. W., Rowe, J. F., et al. 2012, arXiv:1201.5424
Lodge, O. J. 1919, Nature, 104, 354
Lyne, A. G., & Bailes, M. 1992, Nature, 355, 213
Macintosh, B. A., Graham, J. R., Palmer, D. W., et al. 2008, The Gemini Planet Imager: from science to design to construction, Proc. SPIE 7015, 701518 Proc. SPIE, 7015

Makarov, V. V., Beichman, C. A., Catanzarite, J. H., et al. 2009, ApJ, 707, L73
Mao, S., Paczynski, B. 1991, ApJ, 374, L37
Marcy, G. W., & Benitz, K. J. 1989, ApJ, 344, 441
Marcy, G. W., & Butler, R. P. 1992, PASP, 104, 270
Marcy, G. W., & Butler, R. P. 1995, in Bulletin of the American Astronomical Society, 27, Bulletin of the American Astronomical Society, (Washington, DC: AAS), 1379–+
Marois, C., Lafrenière, D., Doyon, R., Macintosh, B., & Nadeau, D. 2006, ApJ, 641, 556
Marois, C., Macintosh, B., Barman, T., Zuckerman, B., Song, I., Patience, J., Lafrenière, D., & Doyon, R. 2008, Science, 322, 1348
Marois, C., Zuckerman, B., Konopacky, Q. M., Macintosh, B., & Barman, T. 2010, Nature, 468, 1080
Mayor, M., & Queloz, D. 1995, Nature, 378, 355
Mazeh, T., et al. 2000, ApJ, 532, L55
McArthur, B. E., Benedict, G. F., Barnes, R., et al. 2010, ApJ, 715, 1203
McBride, J., Graham, J. R., Macintosh, B., Beckwith, S. V. W. B., Marois, C., Poyneer, L. A., & Wiktorowicz, S. J. 2011, PASP, 123, 692
McCullough, P. R., Stys, J. E., Valenti, J. A., et al. 2005, PASP, 117, 783
Michel, E., & Baglin, A. 2012, Proceedings of the Second CoRoT Symposium, arXiv:1202.1422
Mordasini, C., Alibert, Y., & Benz, W. 2009, A&A, 501, 1139
Morton, T. D., & Johnson, J. A. 2011, ApJ, 738, 170
Mouillet, D., Larwood, J. D., Papaloizou, J. C. B., & Lagrange, A. M. 1997, MNRAS, 292, 896
Muirhead, P. S., Johnson, J. A., Apps, K., et al. 2012, ApJ, 747, 144
Muterspaugh, M. W., et al. 2010, AJ, 140, 1657
Nature 1992, Nature, 355, 187
Nutzman, P., & Charbonneau, D. 2008, PASP, 120, 317
O'Toole, S. J., Tinney, C. G., Jones, H. R. A., et al. 2009, MNRAS, 392, 641
Paczynski, B. 1986, ApJ, 304, 1
Park, B.-G., Jeon, Y.-B., Lee, C.-U., & Han, C. 2006, ApJ, 643, 1233
Pepe, F., Lovis, C., Ségransan, D., et al. 2011, A&A, 534, A58
Pepper, J., & Gaudi, B. S. 2005, ApJ, 631, 581
Pepper, J., Gould, A., & Depoy, D. L. 2003, Acta Astron., 53, 213
Pollacco, D. L., Skillen, I., Collier Cameron, A., et al. 2006, PASP, 118, 1407
Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, Icarus, 124, 62
Pravdo, S. H., & Shaklan, S. B. 2009, ApJ, 700, 623
Queloz, D., Mayor, M., Sivan, J. P., Kohler, D., Perrier, C., Mariotti, J. M., & Beuzit, J. L. 1998, in ASP Conf. Ser. 134, Brown Dwarfs and Extrasolar Planets, ed. R. Rebolo, E. L. Martin, & M. R. Zapatero Osorio, 324–+
Refsdal, S. 1964, MNRAS, 128, 295
Rhie, S. H., Bennett, D. P., Becker, A. C., et al. 2000, ApJ, 533, 378
Sako, T., Sekiguchi, T., Sasaki, M., et al. 2008, Exp. Astron., 22, 51
Schuh, S. 2010, Astronomische Nachrichten, 331, 489
Seager, S. 2010, Exoplanet Atmospheres: Physical Processes (Princeton: Princeton University Press)
Shvartzvald, Y., & Maoz, D. 2012, MNRAS, 419, 3631
Smith, B. A., & Terrile, R. J. 1984, Science, 226, 1421
Snodgrass, C., Horne, K., & Tsapras, Y. 2004, MNRAS, 351, 967
Southworth, J., Hinse, T. C., Jørgensen, U. G., et al. 2009, MNRAS, 396, 1023
Steffen, J. H., & Agol, E. 2005, MNRAS, 364, L96
Struve, O. 1952, Observatory, 72, 199
Torres, G., Andersen, J., & Giménez, A. 2010, A&A Rev., 18, 67
Tsapras, Y., Street, R., Horne, K., et al. 2009, Astronomische Nachrichten, 330, 4
Udalski, A. 2003, Acta Astron., 53, 291
Udalski, A. 2009, Var. Universe, 403, 110
Udalski, A., Szymanski, M., Kaluzny, J., et al. 1993, Acta Astron., 43, 289
Udalski, A., Szymanski, M., Kaluzny, J., et al. 1994, Acta Astron., 44, 227
Unwin, S. C., Shao, M., Tanner, A. M., et al. 2008, PASP, 120, 38
van de Kamp, P. 1986, Space Sci. Rev., 43, 211
Veras, D., Ford, E. B., & Payne, M. J. 2011, ApJ, 727, 74
Vogt, S. S. 1987, PASP, 99, 1214
Walker, G. A. H., Yang, S., Campbell, B., & Irwin, A. W. 1989, ApJ, 343, L21
Walker, G. A. H., Bohlender, D. A., Walker, A. R., et al. 1992, ApJ, 396, L91
Winn, J. N. 2010, Exoplanets, 55
Wolszczan, A. 2012, New A Rev., 56, 2
Wolszczan, A., & Frail, D. A. 1992, Nature, 355, 145
Wolszczan, A., Hoffman, I. M., Konacki, M., Anderson, S. B., & Xilouris, K. M. 2000, ApJ, 540, L41
Wright, J. T. 2005, PASP, 117, 657
Wright, J. T., & Howard, A. W. 2009, ApJS, 182, 205
Wright, J. T., et al. 2011, PASP, 123, 412
Wyatt, M. C., et al., 1999, ApJ, 527, 918
Yee, J. C., Shvartzvald, Y., Gal-Yam, A., et al. 2012, arXiv:1201.1002

Index

A

Accrete (accretion), 313–315, 322, 323, 328, 337, 350, 363, 367, 368
Adiabatic hypothesis, 207
Adrastea, 315, 318, 352, 366
Aegaeon, 360
Aeolian geology, 126, 155
Aerodynamic migration, 22
Aerodynamic particle concentration, 28–30
α-disk model, 13–14
Amalthea, 318, 351, 366
Ammonia, 227, 229–232, 240, 245, 247
Angular momentum, 9–15, 22, 32, 35, 53
Anomalous viscosity, 13
Anthe, 356, 361
Asteroids, 66, 70, 88, 94–97, 99–101, 378–421
  definition, 378
  spectral classes, 414
Astrometry, 491, 494–496, 506, 510, 515, 516, 520, 529–531
Aurora, 266, 273, 281–285, 288–290, 294, 298, 303

B

Ballistic transport, 328, 337
Bending wave (spiral bending wave), 329, 330, 332, 340
Binary and multiple systems, 388–389
Bondi radius, 46, 54
Bond ringlet, 347, 348
Bow shock, 252, 253

C

Cassini, 225, 226
Cassini divisions, 319, 320, 328, 330, 333, 335, 337, 349
Catastrophic disruption, 39
Ceres, 378–380, 382, 394, 400–402, 404, 406–408, 419
Charged particles, 317, 325, 326, 350, 353, 361, 366
Charming ringlet, 349
Chondrites, 382, 391, 394, 400, 402, 403, 407–412, 417–421
Chondrules, 5–7, 29–31
Climate change, 135, 190
Clouds, 227–247
Coagulation models, 43, 57
Coefficient of restitution, 341, 364–365
Collisional cascade, 38, 39, 41, 42, 45, 46, 57
Collisional families, 386–388
Collisional growth, 22, 25–26, 31
Colombo Gap, 320, 349
Comet Shoemaker-Levy, 9, 245, 361
Cooling time, 14, 55, 56
Cordelia, 321
Core accretion, 224, 225, 228
  instability, 47–51
Corotation resonances, 357, 359, 361
Corrugations, 332, 335, 361
Cosmic abundances, 197, 198, 210, 219
Craters, 380, 391–394, 397–400, 418

D

Daphnis, 334, 335
Debris disks, 65, 90–93, 101
  composition, 458–461
  evolution, 447–449, 462
  extrasolar, 433–462, 473, 475, 478, 480, 481
  fractional luminosity, 446, 475
  frequency, 444–447
  solar system, 433, 435–444, 446, 454, 457, 459
  spatial structure, 436
  spectral energy distribution (SED), 450, 451, 454
  Spitzer surveys, 446, 447, 449
  white dwarfs, 445, 446, 460
Density wave (spiral density wave), 329, 330, 332, 333, 337, 348, 362
Disk instabilities, 19–20, 57
Disks, 3–58
Drag forces, 21–25, 41
Dust
  circumstellar, 433, 434, 460, 463
  collisions, 441–443, 455, 474–480, 482
  composition, 458
  dynamics, 437, 450, 454, 465–467
  interplanetary, 433, 437, 438, 443, 444, 463
  lifetime, 433, 434
  production rate, 439–444, 447, 449, 462, 479, 480
  size distribution, 475–478
  solar system
    asteroidal, 438, 439, 443, 444, 472
    beta-meteoroids, 465
    cometary, 439–441, 472
    Kuiper belt, 434, 435, 440–442, 467–473, 475, 478, 482
    micrometeorites, 438
  spatial distribution, 433, 440, 467, 469, 471–474, 478–479, 481, 482

sputtering, 481–482
sublimation, 438, 439, 480–481
thermal emission, 433, 437, 440, 442, 472
zodiacal, 433, 435–438, 440, 441, 444, 446
Dwarf planets, definition, 379, 380
Dynamical friction, 37–39, 41, 43, 44, 84, 85, 93, 94
Dynamical optical depth, 316, 341
Dynamo, 252–254, 257–260, 289, 290, 292, 295, 303

E

Earth, 112–137, 139, 140, 144, 149, 150, 153, 155, 158, 159, 163, 173, 174, 178, 179, 181, 182, 189, 190
Eccentricity gradient, 348
Edges, 311, 323, 333–335, 337, 348–350, 352, 359, 362, 363, 365, 366
Einstein ring radius, 501, 503, 511, 512, 518, 528
Enceladus, 319, 355, 356, 359, 360
Encke Gap, 335, 342, 343, 349, 356, 359
Encke ringlets, 349, 350
Energy balance, 226, 236, 239, 245
Epicyclic orbital elements, 313
Equilibrium temperature, 497, 507, 516, 520
Equinox, 332, 335, 342–344, 347, 350, 354, 361, 362
Eros, 383, 391, 392, 394, 396–400, 402, 403, 417, 418
Escape velocity, 33, 34, 37, 38, 41
Exoplanets, 3, 8–9, 32, 33, 46, 47, 52, 54, 55, 57, 248, 489–537
Direct imaging, 498, 506–508, 511, 514, 515, 517, 528, 534
Iodine cell, 525, 526
PSR 1257+12, 513, 522–523
Radial velocities, 491, 494–496, 499, 500, 503–507, 510, 511, 514–516, 520–528, 530, 531, 534–536
Transits, 491, 498–500, 504, 509–511, 513–515, 517, 518, 521, 527–528, 535–536

F

51 Pegasi *b*, 527
Fluvial geology, 122, 155
Flux rope, 295, 296
Force
gravitational, 433, 462, 463, 465, 467–474
Lorentz, 481
Poynting-Robertson (P-R) drag, 437, 464–467, 469, 470, 475, 480, 481
radiation pressure, 463–467
stellar wind, 433, 462–467
Formation
giant planets, 65–68, 73, 101
moon, 87, 95
terrestrial planets, 87, 93, 94, 97, 98, 101

G

Galatea, 323, 356–359, 361
arc, 356
Galileo, 226, 229, 230, 246
probe, 230, 233, 234, 247
Gas giant, 224, 225, 228, 235, 236, 246–248
Giant propellers, 343
Gravitational collapse of solids, 26–28
Gravitational focusing, 35–39, 42, 51, 54
factor, 66
Gravitational instability, 224, 225
Gravitational microlensing, 491, 500–503, 511, 512, 521, 528
Gravitational perturbations
resonant, 469–473
scattering, 455, 456, 469–471, 473, 478
secular, 453–455, 470–472, 474
Gravitational unstable gas disk, 55
Great Dark Spot, 244
Great Red Spot (GRS), 239, 240
GRS. *See* Great Red Spot (GRS)

H

Habitable Zone, 498, 514, 518–521, 536
Hadley circulation, 238
Hayabusa, 383, 391, 395, 402, 417
Haze, 228, 241
HD209458*b*, 527–528
Heat sources, 210, 212
Hill sphere, 34, 35, 51, 54, 314, 327, 366
Hubble Space Telescope, 236, 237, 240, 242, 243, 435–436, 532
Huygens ringlet, 347–349
Hybrid code, 43
Hydrostatic equilibrium, 197, 200, 203, 204, 206, 207, 209

I

Iapetus, 320, 327
Ice, 113, 118, 122, 124–128, 141, 154–159, 161, 164, 174, 180, 184, 185
giant, 224, 225, 238, 239, 248
Impact craters, 113, 116, 131–136, 140–142, 153, 155–157, 161, 164, 174, 176–179, 183–185, 189
Impacting, 327
Impactors (impact), 311, 318, 320, 324, 344, 347, 352–354, 361, 362, 366
Infrared excess, 91, 93, 442–452
Infrared Interferometer Spectrometer and Radiometer (IRIS), 230
Inner working angle, 508, 517, 520
Insolation, 235, 236, 238, 239, 243, 244, 246

IRIS. *See* Infrared Interferometer Spectrometer and Radiometer (IRIS)
Irradiated disks, 4, 7, 17, 46, 56
Isotopic timescales, 5–6

J

James Webb Space Telescope (JWST), 247, 248
Janus Epimetheus, 335, 351, 362
Jupiter, 4, 5, 29, 31, 34, 37, 51, 52, 224–230
Jupiter's atmosphere, 71
JWST. *See* James Webb Space Telescope (JWST)

K

Keeler Gap, 334, 335
Kelvin-Helmoltz contraction, 50
Kepler, 492, 504, 510, 513, 517, 536
Keplerian shear, 333, 347, 356, 362
Kepler's third law, 333, 334, 339, 357
Kuiper belt, 5–7, 30–33, 57

L

Lambert sphere (phase function), 498
Laplace planes, 313, 327, 361
Laplace ringlet, 347, 348
Late Heavy Bombardment (LHB), 64, 86–90, 101, 116, 117, 134, 136, 150, 155, 160, 174, 177, 183
Lens equation, 500, 502
LHB. *See* Late Heavy Bombardment (LHB)
Life on mars, 171
Lindblad, 357
Lindblad resonances (LR), 329–330, 333, 335, 344, 346, 348, 357, 361
LR. *See* Lindblad resonances (LR)
Luminosity, 205, 206, 209, 211–213

M

Mab, 322, 355, 356
Magnetic fields, 113, 118, 119, 137, 151, 174, 177, 181, 189, 196, 218, 219
Magnetopause, 252, 261, 263, 267, 270, 271, 273–277, 281, 284–286, 288, 293
Magnetorotational instability, 13
Magnetosheath, 252
Magnetosphere, 252–303
Magnetotail, 252, 262, 267, 269, 274–278, 281, 283–286, 288, 290, 291, 293, 295, 296, 302, 303
Mars, 112–115, 117, 120, 134, 149–172, 189
Martian meteorites, 153, 154, 172
Mass function, 494, 500, 520
Mass wasting, 112, 122, 131, 143, 155, 164, 171, 189
Maxwell ringlet, 347, 348
Mercury, 112–115, 117, 120, 134, 150, 173–178, 184, 189, 190
Meridional wind, 235
Metallic hydrogen, 196, 207, 213, 218
Metallicity, 4, 8, 9, 18, 28, 30, 32, 36, 53, 65, 71
Meteorites, 5, 6, 21, 30, 31, 378–421
Meter-sized barrier, 21–26, 28, 31, 57
Methone, 351, 356, 360, 361
Metis, 315, 318, 352, 359, 366
Micrometeoroid, 337, 347, 352, 366
  impacts, 318, 346, 347
Migration, 328, 344, 346
  planetesimal driven, 81–85, 98, 99
  type-I, 67, 68, 75, 81
  type-II, 68–72
Milankovitch cycles, 125, 126, 159
Mimas, 320, 327, 332, 333, 335, 339, 348, 360, 361
Minimum-mass solar nebula (MMSN), 4, 7, 13, 17, 21, 23, 27, 28, 37, 42, 43, 53
MMSN. *See* Minimum-mass solar nebula (MMSN)
Moment of inertia, 215, 217
Moon, 112–117, 131, 134, 150, 151, 153, 155, 175, 178–190
Moonlet wakes, 333–335

N

N-body code, 43, 57
Near-Earth object (NEOs), 379, 380, 383, 385–392, 395–398, 402, 413, 417, 418, 421
Near-Infrared Mapping Spectrometer (NIMS), 229
NEAR Shoemaker, 382, 383, 392, 394, 396, 397, 399, 402, 418
Nebular hypothesis, 4
NEOs. *See* Near-Earth object (NEOs)
Neptune, 224–230, 232, 234–239, 243–244, 247, 248
Nice model, 64, 84, 90–92
NIMS. *See* Near-Infrared Mapping Spectrometer (NIMS)
Normal optical depth, 316, 325

O

Oblateness, 224, 226
Occultations, 230, 245, 317, 332, 333, 336, 338, 347, 348, 350, 356, 357, 365
Oligarchic growth, 66, 68, 93
Optical depth, 316–317, 325, 326, 328, 329, 332, 333, 336–339, 341, 343, 349, 351, 354, 366
Orbital elements, 312–313, 329
Osculating orbital elements, 313
Oval BA, 232, 239–241
Overstability, 336

P

Pallene, 351, 360
Pan, 329, 344, 346, 349
Pandora, 348, 350, 359
Phase (phase angles), 317–319, 325, 330, 331, 350, 354
Phoebe, 320, 354
Planet
  eccentricity, 89
  formation, 3–9, 21, 22, 32, 43–47, 51–55, 57
  instability, 86–90, 98–100
  scattering, 64, 66, 72–78
Planetary atmospheres, 137
Planetary cores, 67, 68, 76, 78, 93
Planetary embryos, 68, 81, 93, 94, 101
Planetary interiors, 113, 115, 116
Planetary models, 197, 201
Planetesimals, 3, 4, 6, 10, 20–57
Planetesimal size distribution, 84
Plasma, 252–254, 261, 263–295, 297–301, 303
Plasmasphere, 264, 271, 272, 281, 293
Plasmoid, 277–278, 283–285, 290, 293
Plateau, 349
Plate tectonics, 116, 119, 149, 169
Polytropes, 203, 209
Portia group, 321–323
Poynting-Robertson drag, 8, 39, 42, 46, 318, 320, 327, 352
Precess (precession), 313, 318, 330, 348, 361
Prometheus, 335, 350, 351, 359
Propellers, 339–346, 363, 365, 368
  belts, 341–343
  moons, 343, 346, 368
Protoplanetary atmospheres, 46, 48
Protoplanets, 6, 10, 32–52, 54
Protostellar disk, 3, 7, 8, 14, 15, 17, 51, 57

R

Radial drift, 21–29, 57
Radiation belts, 266–271, 293
Radius-mass relationship, 205
Reaccretion, 322
Reconnection, 267, 274, 276–278, 283, 284, 288–290, 293, 294, 303
Regolith, 388, 391, 392, 394–401, 408, 417, 418, 420
Resonance
  argument, 329
  capture, 70, 80
  locking, 83
Rhea, 325–327
Rings
  Adams rings, 323, 349–351, 356–359, 361
  α ring, 347
  Arago rings, 349, 353
  A ring, 311, 319, 321, 328, 335–337, 340, 341, 343, 347, 363
  β ring, 347, 348
  B ring, 311, 316, 319, 320, 328, 335–337, 339, 343, 344, 346, 348, 363, 366, 368
  C rings, 311, 320, 328, 333, 336, 337, 347–349, 352, 353, 361, 362
  δ ring, 347, 348, 355
  D ring, 320, 351–353, 359, 361
  ε ring, 321, 322, 347, 348, 350, 354, 368
  E ring, 355, 359
  η ring, 347, 354
  F ring, 348, 350, 353
  Galle ring, 351, 352
  G rings, 319, 353, 355, 359
  γ ring, 347, 348
  Gossamer rings, 317, 318, 351, 353
  Halo ring, 319, 351, 352
  λ ring, 349, 350, 354
  Lassell rings, 323, 351, 353
  Le Verrier ring, 349, 351, 358
  Main ring, 311, 314, 315, 317–322, 328, 329, 333, 348–352, 356, 359, 361, 366
  μ ring, 322, 351, 355, 356
  ν ring, 322, 351, 355
  Ramp, 337
  ζ ring, 351, 352, 355
Roche critical density, 313–315, 321, 323
Roche division, 320, 351, 353, 361
Roche lobe, 313–315, 364
Roche radius (Roche limit), 313–315, 321, 328, 367
Rubble pile, 40
Runaway growth, 66

S

Saturn, 224–232, 234, 235, 241–243, 246–248
Self-gravity wakes (SGWs), 316, 317, 332, 335–341, 343, 344, 346, 362, 363, 368
  clump, 344
Shadows, 318, 332, 335, 342–344, 350, 353
Shepherding, 348, 359
Snow line, 66, 68, 75, 93, 97, 514, 515, 518, 529
Solar nebula, 3–5
Solar system
  abundances, 198
  debris disk, 433–444, 446, 454, 457, 459
  dust production rate, 439–444
  late heavy bombardment (LHB), 443, 449, 456, 459, 462
Solar wind, 252–256, 261–267, 270–277, 281, 283–286, 288–291, 293, 295–298, 302, 303
Space weathering, 396, 397, 418
Spiral bending waves, 329, 330, 332

Spiral density waves, 329, 330, 332, 333, 337, 348, 362
Spiral waves, 329–332, 337, 362, 368
Spokes, 328, 346–347
Strange ringlet, 347, 349
Streaming instability, 21, 28–32, 57
Surface densities, 316, 320–323, 330, 335, 337, 338, 341, 344, 346, 363

T

Tectonism, 112, 140, 169, 189
Terrestrial planets, 112–190
Tethys, 327
Thermochemical equilibrium, 227, 229, 245
Tisserand parameter, 378, 385
Titan, 320, 348, 349
  ringlet, 347–349
Toomre critical wavelength, 337, 344
Toomre Q criterion, 55, 337
Transition disks, 8
Transit probability, 499, 509, 510, 517, 521, 527
Transparency, 316, 317, 339
Triton, 261, 264, 315
Tropopause, 228

U

Uranus, 224–230, 232, 234–239, 243–244, 247, 248

V

Venus, 112–115, 117, 120, 137–150, 153, 155, 169, 189
Vertical resonance (VR), 329, 330
Vesta, 379–382, 387, 391, 394, 397, 402, 404, 407, 409, 411, 412, 414, 415, 418, 421
Virial theorem, 207–209, 212
Viscosity (viscosities), 10–14, 16–18, 20, 51, 54, 57, 330, 332, 336, 363
Viscous disk model, 11
Viscous overstability, 336
Viscous spreading, 348, 367
Viscous stirring, 37, 38, 45
Volcanism, 112, 119, 131, 135, 140, 143, 144, 148, 152, 166, 168, 169, 175, 177, 187–190
Volume filling factor, 317
Voyager 1, 226, 241
Voyager 2, 225, 226, 235–238, 241, 243, 244
VR. *See* Vertical resonance (VR)

W

Water, 4–7, 66, 88, 95, 98
Wavelet (wavelet transform), 330, 331

Y

Yarkovsky force, 385, 389, 409

Z

Zodiacal cloud, 93
Zonal wind, 234–236, 238, 242, 244, 246