Cosmic Catastrophes

Exploding Stars, Black Holes,
and Mapping the Universe
Second Edition

J. CRAIG WHEELER

The University of Texas at Austin

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521857147

First published in print format 2007

ISBN-13 978-0-511-26911-0 eBook (EBL)
ISBN-10 0-511-26911-0 eBook (EBL)

ISBN-13 978-0-521-85714-7 hardback
ISBN-10 0-521-85714-7 hardback

To my sons,

Diek W., the scientist,

and J. Robinson, the artist.

Contents

Preface

The core of this book concerns supernovae, my principal research interest, but the broader theme is the connection of these cosmic catastrophes with the sweep of intellectual ferment in astrophysics. The story leads from the birth, evolution, and death of stars to the notion of complete collapse in a black hole, to wormhole time machines, the possible birth of new universes, and the prospect of a conceptual revolution in our views of space and time in a ten-dimensional string theory. It is all one glorious, interconnected Universe, both physically and intellectually. Or maybe there are more than one.

In terms of astrophysical connections, the book reaches back to the origins of stars and how they evolve, treats the mechanisms of supernovae, and then moves forward to the compact progeny of supernovae – neutron stars and black holes. Neutron stars are presented in all the variety we know today – pulsars, millisecond pulsars, binary pulsars, magnetars, and X-ray sources both steady and transient. The concrete manifestation of black holes in observational astronomy, especially in binary stellar systems, is described. Topics that have come to light as the book was being written, soft gamma-ray repeaters and the revolution in cosmic gamma-ray bursts, are presented. The scientific background is given in order to understand what kind of supernovae are used to produce the radical notion of the acceleration of the Universe, and how and why. Similar background aids in making the connection between flaring gamma-ray sources and compact objects.

A parallel theme is not the objects themselves, but the intellectual framework that underlies our study and the limits to which it

currently extrapolates. This involves discussions of the physics of the twentieth century, the quantum theory and Einstein's gravity, how they collide, and the prospects for reconciliation. In the process, the concept of gravity as curved space is shown to lead to radical notions, such as time machines and baby bubble universes. The promise of string theory to give a unifying view and to open new conceptual windows is illustrated.

Because I have used and intend to use this book for classes, I have, for completeness, written about topics that have been presented before: the basics of stellar evolution, the discovery and interpretation of pulsars, the nature of space and time in the vicinity of black holes, and the more recent topics, such as wormholes and the promise of string theory. I have presented this material in my own style and hope that there is some benefit to seeing it again. In addition, I have tried to present this material in a broad context that gives it a different perspective to that of previous treatments.

There are other topics that I have stressed here because they are of crucial importance and because they tend to get overlooked. One of these is binary-star evolution. When I began to teach this material, there was scarcely any mention of binary stars in introductory astronomy texts, save perhaps for a mention of eclipses and visual and spectroscopic binaries. Current texts are much better, but this topic is so fundamental that I am compelled to present it in some detail. Supernova researchers believe many supernovae depend incidentally or critically upon their being in binary systems. Much of what we know about neutron stars follows from their being in binaries. The only way we know about stellar-mass black holes is by discovering them in binary systems. Many books on black holes concentrate on the supermassive variety in galactic nuclei and scarcely mention those in binary systems, never mind the amazing array of phenomenology associated with them and the reasons for it. I have thus devoted a chapter to discussing the systematics of Roche lobes, mass transfer, and common envelopes, the language of this field that is often passed over in books of this kind.

A closely related topic is that of accretion disks. The study of disks has become an industry unto itself, but these objects are rarely presented with the background of how they work and why they are so important to the topics of this book, from the evolution of Type Ia supernovae to binary neutron stars to binary black holes to the cosmic

gamma-ray bursts. Accretion disks have a life of their own, with instabilities that cause them to flare and attract the attention of astronomers. With the exception of venerable old Cygnus X-1 and a few others, all the host of new black-hole candidate discoveries are due to flaring systems. The most plausible mechanism for the flaring is associated with the disk. Accretion disks also merit a separate chapter.

I have also included topics that, although the subject of many articles in popular science literature, have not, to my knowledge, been incorporated in a book where the relevant background can be laid out in advance and the story told as an integral part of modern astrophysics. There are three examples of that, all of which have ''exploded'' in the past year. One is the proof that the soft gamma-ray repeaters involve exceedingly strongly magnetized neutron stars – magnetars in the language of my colleague Robert Duncan. Another story is the amazing array of developments that have followed since the discovery of the first optical counterparts of the cosmic gamma-ray bursts, not the least of which, to someone of my bent, is the association of one with a supernova. In each of these cases, to understand the story behind the headlines fully, one needs to know the relation of the topic to stellar evolution, the ideas behind the birth of neutron stars and black holes, the significance of supernovae that show a paucity of hydrogen and helium, and the nature of binary star evolution. Last, but certainly not least, is the use of supernovae to measure distances on cosmological scales. The tentative result, that the Universe is accelerating, was recently proclaimed the scientific breakthrough of the year 1998 by *Science Magazine*. Here I have the opportunity to tell the story in terms of the history of the topic as well as the astrophysical background involving binary-star evolution, specific supernova mechanisms, and the elements of cosmology.

The seeds of this book were planted in 1975. My colleague, R. Edward Nather, invented a course at the University of Texas called Astronomy Bizarre. The purpose of this course was to tell the story of the Universe from the big bang onward, rather than from the Solar System outward as is traditional for introductory astronomy courses, and to introduce some of the exotica of astronomy for which one has little time in the standard introductory course for nonscience majors. Nather taught the first version of this course just after I arrived at the University of Texas. The prerequisite of a standard introductory astronomy course was omitted from the catalog. More than 300

students registered, and a second section had to be opened. I was assigned that section and have been teaching some version of the course for the last 25 years. This book represents some of the material I have developed for the course.

Nather and I planned to write a book based on his original Astronomy Bizarre syllabus. We wrote a draft, but the project foundered for various reasons. The material that ended up in this book is very different from that first draft, but the early introduction of the notion of conserved quantities is a vestige of that work, and I thank Ed for that idea.

Astronomy Bizarre was such a successful course that it evolved to encompass several versions. Over the years, I inherited the course that concentrated on stars. To keep my teaching fresh, I have regularly changed the content of the course. Sometimes I concentrate on supernovae and closely related topics. Other times, I have taught the whole course just on black holes and related ideas. I have taught it sometimes to a small class required to do substantial writing. To stay current, I have added new material as new developments have come along, a never-ending process in astrophysics.

As I have taught the course, I have had to wrestle with how to portray the complex and fascinating ideas of astrophysics to classes of bright, interested, but nontechnically trained students. This book also represents a compilation of the ideas I like to try to explain to popular audiences and the techniques I have developed to accomplish this. One of the ideas with which I am most pleased is blowing up a balloon and turning it inside out to portray the embedding diagram of the curved space around a black hole. I have also tinkered with the vocabulary. In many cases, I adopt the jargon of astronomy and endeavor to define and explain it. In other cases, I have invented new phrases. I did not think that the term ''degeneracy'' carried much import for a popular audience, even after an attempt to explain it. I have thus referred to a ''quantum pressure'' rather than ''degeneracy pressure,'' feeling that this term gets the basic point across that this pressure is different in a fundamental way from that exerted by a gas of hot plasma. I trust that these attempts to make the material accessible to nonscience-major students have some value for audiences beyond the lecture hall.

In addition to the various themes of the book I outlined earlier, I have emphasized several physical themes that tie together various

topics of the course. I stress the difference between stars supported by thermal pressure and those supported by the quantum pressure, why one results in regulated nuclear burning and one leads to stellar explosions. These lessons are used throughout stellar evolution, from star formation to hydrogen burning to red-giant formation to the formation of iron cores and the contrasting examples of classical novae and Type Ia supernovae. The nature of the weak interaction and the intimate connection to neutrinos is introduced early and used to relate the topics of the solar-neutrino problem, massive core collapse, and the radioactive decay that powers the light curves of supernovae devoid of extended envelopes of matter at the time of explosion.

Over the years, many friends and colleagues have helped me to understand the material I have tried to synthesize in this book. Any errors of fact or interpretation are mine, not theirs. I am indebted to Ed Fenimore for clarifying the early history of gamma-ray bursts. Special thanks go to Stirling Colgate for his contributions to the research depicted here and for his intensity and wide-ranging imagination that have stimulated me both scientifically and otherwise.

I am grateful to all my students over the years as I have developed and altered the course. Their feedback has allowed me to better understand what works and what does not. In the spring of 1998, I made this feedback more concrete by offering extra credit to students in my Astronomy Bizarre class who would make comments on clarity and errors in the draft of the book I was using for class. Many of them made very valuable suggestions that I have incorporated. Among these people were Ramesh Dhanaraj, Angela Entzminger, Laura Tamayo Gamborino, John Going, Jonathan Hurley, John Kendall, Sara Keyes, Rubi Melchor, Siddarth Ranganathan, Natalie Sidarous, Benjamin Tong, and Victor Yiu.

I am also grateful to Adam Black of Cambridge University Press for his enthusiasm for this book and especially to Timothy Jones whose magic with computer illustration has brought many ideas to life.

PREFACE TO THE SECOND EDITION

I was very distracted with supernova 1987A and chairing my department when Kip Thorne and Igor Novikov wrought the revolution in thinking about wormholes and time machines that is

now the topic of Chapter 13 in this revised edition. I was rather chagrined that I had been so myopic as to miss this development. As it happened, another intellectual revolution occurred in the late 1990s that I also missed out on, partly because I was laboring to finish the first edition of this book. That was the startling understanding by Lisa Randall and Raman Sundrum that there might exist large extra dimensions that nevertheless leave gravity acting essentially as an agent of three-dimensional space. I am not, nor will ever be, an expert in this, but this sort of intellectual development is just the type of thing that I like to try to capture and describe to the students in my class. The topic belonged in the book, but I missed out. In this edition I have tried to capture some of the spirit of this development and the reasoning behind it.

While little else can compete with this dramatic breakthrough, astronomy, astrophysics, and cosmology rush on. There were plenty of other developments over the last few years that required modification of my lecture notes and the first edition of the book. In addition, I have attempted to correct all the errors that ''alert readers'' brought to my attention in the first edition. Any remaining are my responsibility.

The change that draws most deeply on my personal research is the growing understanding that supernovae are aspherical. Core-collapse supernovae are especially so, but the thermonuclear explosions of Type Ia supernovae are also showing significant and fascinating irregularities. The first edition contained glimmers of the asymmetries in core collapse, but the current edition contains a whole section in Chapter 6 on the observational and theoretical developments pertaining to that deepening understanding. The opening discussion in Chapter 6 of observations of supernovae has also been modified appropriately to elucidate the apparent correlation of compact objects and asymmetric, jet-like, extended remnants, a point not yet made in the formal research literature. The section on Type Ia supernovae has also been lightly updated to reflect this aspect and other developments. Chapter 7 on supernova 1987A has also been updated to emphasize the ongoing collision of the ejecta with the inner ring and the evidence for the asymmetry of the ejected matter. I added an arrow to the photograph showing the location of the star that blew up as SN 1987A. This allowed me to replace the associated impossibly obscure figure caption that attempted to describe the location of the

small black dot in words (backwards giraffe heads entered here), that *no one* understood, with the simple expedient of a graphical aid.

Chapter 8 on neutron stars has been updated to reflect the dramatic observations of recent giant flares from soft gamma-ray repeaters, otherwise known as magnetars. I have left Chapter 9 on black-hole theory virtually unchanged, with the exception of adding a much-needed schematic figure of the insides of a rotating black hole. For Chapter 10 on observing black holes, I added some discussion of supermassive black holes that was needed for context, even though this book is mostly stellar in theme. The remarkable discovery that the mass of these black holes is directly connected in some way to the mass and structure of the much more massive galactic bulges that house them was too important to pass up. That also set the context for a new and important section on the possible existence of intermediate-mass black holes.

To make the rest of the book work and give me room to talk about the Randall/Sundrum revolution, I had to do some wholesale re-structuring of the remainder of the book. I split off the discussion of gamma-ray bursts to be the sole topic of a new Chapter 11. That gave me space to describe the onrush of developments in that field. One was the proof in 2003 that long gamma-ray bursts are intimately related to supernovae. Another was the establishment that gamma-ray bursts emit their intense energy in tightly collimated beams, a notion that was just being developed as the first edition went to press. I also dawdled getting the second edition revised long enough to be able to describe the most recent revelation in this game: that the short gamma-ray bursts are also explosions in very distant galaxies, but with properties that distinguish them from their observationally more common long cousins.

The material in Chapter 12 is mostly that from the first edition on the discovery with supernovae of the remarkable acceleration of the Universe, but now set out in its own chapter. That gave me room to expand on the conceptual background of this topic: what we knew, or thought we knew, about the age, shape and fate of the Universe. I have also included a discussion of dark matter. This topic does not relate to the theme of stars very directly, but it is so important in modern cosmology, and its quantity was also elucidated by the supernova cosmology and related work, that this was a required addition. Discussing dark matter is also necessary to compare and

contrast it with dark energy. While writing this section and pondering the tiny fraction of the Universe that is composed of stuff like us, I had the minor epiphany that, while the dark energy and dark matter dominate the energy density of the Universe, unlike baryonic matter, they cannot write books. There is some solace in that. I have also expanded somewhat the discussion of dark energy and our revised notions of the shape and fate of the Universe.

I have not made any substantial changes to the material on wormholes and time machines, but have separated that out in its own Chapter 13.

This brings me to the real reason a second edition was needed, and that is to capture some of the dramatic nature of our expanding view of space and time. I have made the discussion of string theory and associated topics a separate Chapter 14. Most of the material from the first edition is there, but re-organized somewhat. In the discussion of hyperspace, I have added some of the history of the ''fourth dimension'' and its role in the world of art. For this, I thank my colleague and friend, art historian Linda Henderson. I understand branes a bit more now, though not deeply, and have expanded that discussion. There is a new section on brane worlds, the reasons why physicists argued that if there were higher dimensions they must be curled up, and the intellectual (and paper writing!) revolution that Randall and Sundrum unleashed with their insights that higher dimensions need not be curled up. Lastly, in a feat of reckless overextension of my understanding of the topic, but again in the spirit that it is just too intellectually fun to pass on, I have added a section on the ideas concerning holographic universes.

I am modestly content with the current content of the book, but I also know full well that a year from now I will decry the lack of some new, amazing development. Astrophysics is like that.

1

Setting the stage: star formation and hydrogen burning in single stars

1.1 INTRODUCTION

We look up on a dark night and wonder at the stars in their brilliant isolation. The stars are not, however, truly isolated. They are one remarkable phase in a web of interconnections that unite them with the Universe and with us as human beings. These connections range from physics on the tiniest microscopic scale to the grandest reaches in the Universe. Stars can live for times that span the age of the Universe, but they can also undergo dramatic changes on human timescales. They are born from great clouds of gas and return matter to those clouds, seeding new stars. They produce the heavy elements necessary to make not only planets but also life as we know it. The elements forged in stars compose humans who wonder at the nature of it all. Our origin and fate are bound to that of the stars. To study and understand the stars in all their manifestations, from our life-giving Sun to black holes, is to deepen our understanding of the role of humans in the unfolding drama of nature.

This book will focus on the exotica of stars, their catastrophic deaths, and their transfigurations into bizarre objects like white dwarfs, neutron stars, and black holes. This will lead us from the stellar mundane to the frontiers of physics. We will see how stars work, how astronomers have come to understand them, how new knowledge of them is sought, how they are used to explore the Universe, and how they lead us to contemplate some of the grandest questions ever posed.

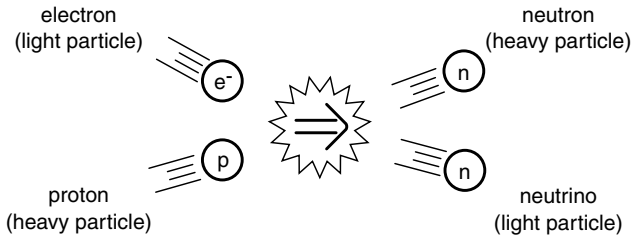
We will begin by laying out some of the fundamental principles by which stars and, indeed, the Universe function.

1.2 BACKGROUND

1.2.1 *The basic forces of Nature*

The nature of stars is governed by the push and pull of various forces. The traditional list of the basic forces of Nature is as follows:

- *Electromagnetic force* – long-range force that affects particles of positive (+) and negative (–) electrical charge, as shown in Figure 1.1 (top). *Protons* (p) are examples of positive charges, and *electrons* (e^-), negative charges.
- *Strong or nuclear force* – short-range force that affects heavy (high-mass) particles such as protons (p) and *neutrons* (n). The strong force binds protons and neutrons together in the atomic nucleus, as shown in Figure 1.1 (middle). The strong force turns repulsive at very small distances between the particles.
- *Weak force* – short-range force that affects interactions between light (low-mass) particles such as electrons (e^-) and *neutrinos* (ν). The weak force converts one light particle into another and one heavy particle into another; for instance, as shown in Figure 1.1 (bottom).



- *Gravity* – long-range force that affects all matter and is only attractive.

The particle known as the neutrino is a special one with no electrical charge. It interacts only by means of the weak force (and gravity), that is to say, scarcely at all. Its properties and its role in nature will be explained in more detail below and in later chapters.

The results of theoretical work in the 1960s by Steven Weinberg, Abdus Salaam, and Sheldon Glashow, followed by experimental verification in the 1970s and 1980s by a large team led by Carlo Rubbia and Simon van der Meer, showed that the electromagnetic and weak forces are actually manifestations of the same basic force, which has

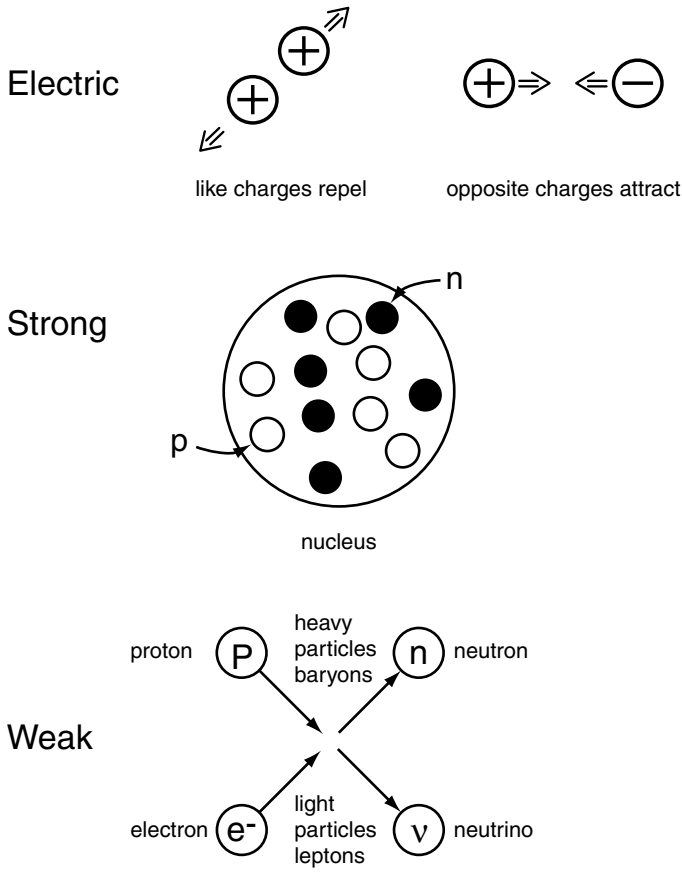


Figure 1.1 The action of the basic forces: (top) opposite electrical forces attract, and like charges repel; (middle) the attractive nature of the strong force holds protons and neutrons together in atomic nuclei despite the charge repulsion among the protons; (bottom) the weak force causes protons to convert into neutrons and electrons into neutrinos and vice versa.

come to be called the *electroweak force*. This unification is analogous to the recognition, based on the work of Thompson and Maxwell in the nineteenth century, that electrical effects and magnetic effects are actually intimately interwoven in what we now call the electromagnetic force. Nobel Prizes are only the celebrated tip of the ferment that leads to scientific progress; however, their winners deserve their credit, and the prizes are signposts of major progress. Weinberg, Salaam, and Glashow won the 1979 Nobel Prize in Physics for their work; Rubbia and van der Meer, for theirs in 1984.

Current research is aimed at the goal of showing that the strong force is also related to the electroweak force, and that both are manifestations of some yet more fundamental force. Definite progress has already been made toward this goal of constructing a *grand unified theory*. Another dream is to show how gravity may also be understood as intrinsically related to the other forces. The story of gravity is a complex one at the heart of modern physics, and even its role in the pantheon of forces requires some interpretation. Newton interpreted gravity as a force, but, as will be elaborated in Chapter 9, Einstein's theory leads to the interpretation that gravity is a property of curved space and time, that there is no ''force of gravity'' in the sense that Newton conceived it. Recent dramatic progress has been made toward a unified picture of gravity and the other forces by envisaging particles as one-dimensional strings, rather than as points, as we will see in Chapter 12. In this evolving theory, gravity is again interpreted as a force, but one Newton would scarcely recognize. In practice, we will often refer to these forces in their four traditional categories, as given earlier, with emphasis where appropriate on the interpretation of gravity as a property of curved space.

1.2.2 *Conservation laws*

To a physicist, conservation does not mean careful use to ensure future supplies, but that some quantity is constant and does not change during an interaction. Physicists have learned to make powerful use of principles of conservation, which are stated in roughly the following manner: ''I don't care what goes on in detail; when all is said and done, quantity X is going to be the same.'' Conservation laws do not help to untangle the details of a given physical process; rather, they help to avoid complex details. Conservation laws are of great help exactly when the details are complicated because one can proceed with confidence that certain basic quantities are known and unchanging, despite the details. How this works will be more clear when we see how these conservation laws are used in various ways. They are employed to help understand why stars get hotter when energy is radiated away, the nature of nuclear reactions that power the stars, why stars become red giants and white dwarfs, the very existence and role of the elusive neutrino, how stars circle one another in binary orbits, why disks of matter form around black holes, and why some supernovae shine by radioactive decay. For now we will describe some of the conservation laws most frequently used in the astrophysics of stars.

One of the most fundamental conservation laws is the *conservation of energy*. Energy can be converted from one form to another so understanding energy conservation can sometimes be tricky, but, for all physical interactions, energy is conserved. The energy can be converted from energy of directed motion to random thermal energy and from, or to, gravitational energy. Even mass can be converted to energy and energy to mass according to Einstein's most famous formula, $E = mc^2$. Despite all these potential conversions in form, the energy of a physical system is conserved. When you drop a piece of chalk, it shatters with a small crash, as illustrated in Figure 1.2 (top). The potential gravitational energy goes first into the kinetic energy of falling, then into the energy of breaking electrical bonds among the particles of chalk, and even into

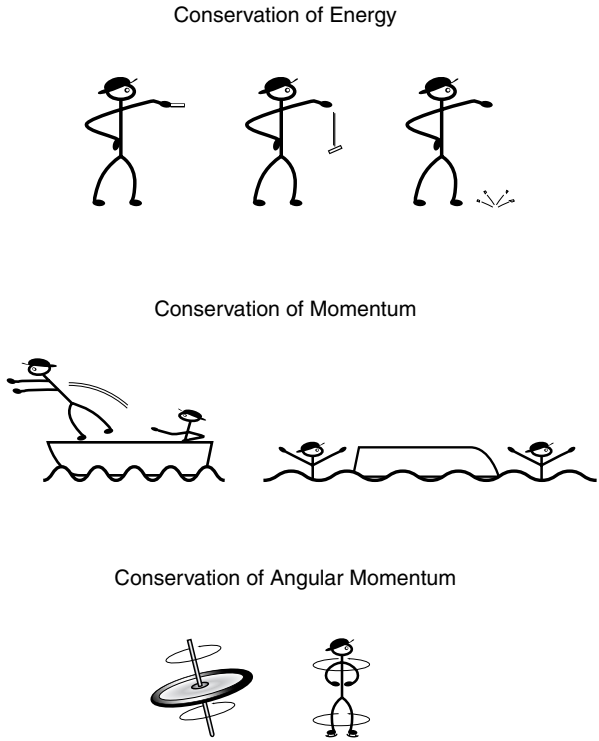


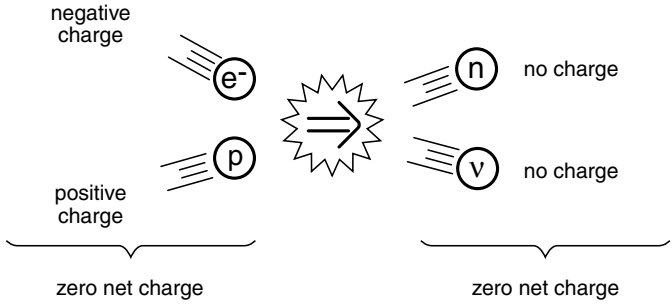
Figure 1.2 The principles of conservation: (top) dropping and shattering a piece of chalk is a complicated process, but the energy of breaking, motion, heat, and noise is exactly that gained by falling; (middle) a person leaping from a boat will send the boat and his companion rapidly in the opposite direction, illustrating conservation of momentum; (bottom) a skater drawing in his arms will spin faster, conserving angular momentum.

the energy of the sound waves of the noise that is made. Despite the complicated details, the total energy of everything is conserved.

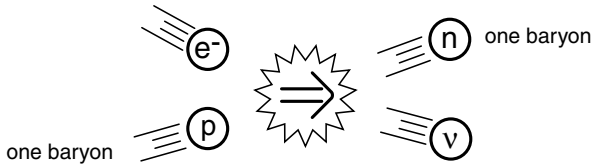
Momentum is a measure of the tendency of an object to move in a straight line. The measure of the momentum is not which team scored the last touchdown or goal, a common usage of the phrase in a sports context, but the product of the mass of an object with its velocity. The *mass* is a measure of the total amount of stuff in an object. The *velocity* is the speed in a given direction. Momentum characterized as mass times velocity is also conserved. A mass moving with a certain speed in a certain direction will continue to do so unless acted upon by a force. A given mass may be sped up or slowed down by the action of a force, but the agent supplying the force must suffer an equal and opposite reaction so as to conserve the momentum as a whole. Try jumping suddenly out of a boat (Figure 1.2, middle) and ask your companions if they appreciate the overwhelming verity of the principle of conservation of momentum. If you leap out one side, the boat must react by moving in the opposite direction with the same momentum as your leap. The boat will inevitably tip and leave everyone in the drink.

Angular momentum is a property related to ordinary momentum, but it measures the tendency of an object of a given mass to continue to spin at a certain rate. The measure of the angular momentum is the mass times the velocity of spin times the size of the object. A popular demonstration of conservation of angular momentum is an ice skater. When a spinning skater draws his arms in closer to his body, his ''size'' gets smaller. Because his mass does not change, his rate of spin must increase to ensure that his total angular momentum will be constant. In detail, this is a complex process involving the contraction and torsion of muscles and ligaments. You do not have to understand the details of how muscles and ligaments work, however, to see that the skater must end up in a dizzying spin when he pulls his arms in, and that he will slow again by simply extending his arms (Figure 1.2, bottom).

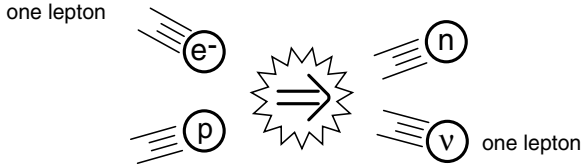
Other conservation laws are important to physics but are not reflected so easily in everyday life. An especially powerful example is that of *conservation of charge*. Electrical charge, the total number of positively and negatively charged particles, is conserved. Physical processes can cancel charges, a positive charge against a negative one, but the net positive or negative charge cannot change in a physical process. Neither positive nor negative charges can simply appear or disappear. In a reaction involving a bunch of particles, the total charge at the end of the reaction must be the same as at the beginning of the reaction. Here is an example:



Elementary particles have other properties, akin to electrical charge, that are conserved. The heavy particles like protons and neutrons that constitute atomic nuclei are called *baryons* (from the Greek ''bary'' meaning heavy). In a nuclear reaction, the number of baryons is conserved. The baryons may be changed from one kind to another, protons to neutrons for instance, but the number of baryons does not change. If there were four baryons at the start, there will be four at the end. The same example applies to baryons:



There are other elementary particles that do not belong to the baryon family. The ones in which we will be especially interested are the low-mass particles known as *leptons*. Electrons and neutrinos are members of this class. As for baryons, nuclear reactions conserve the total number of leptons, even though individual particles may be created or destroyed. Common reactions will involve both baryons and leptons, and both classes of particles are separately conserved. That is true in our sample reaction:



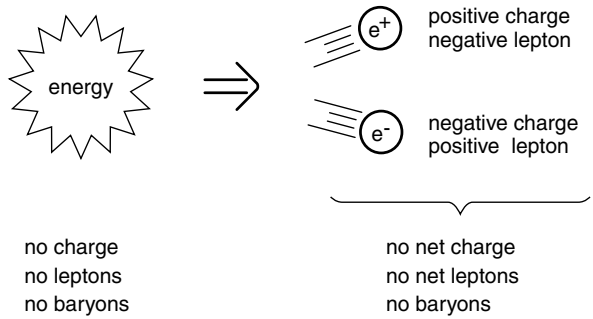
These last two conservation laws, of baryon number and lepton number, are highly accurate. These laws were once thought inviolate.

Recent theoretical developments have suggested that this is not strictly true. One of the suggestions arising from the work of constructing a grand unified theory of the strong and electroweak forces is that baryons may not be completely conserved. The big bang itself may depend on the breakdown of these conservation laws. On timescales vastly longer than the age of the Universe, baryons, including all the protons and neutrons that make up the normal matter of stars, may decay into photons and light particles. For all ''normal'' physics, and hence for all practical purposes, baryons and leptons are conserved, and we will use these conservation laws to understand some of the reactions that are crucial to understand the nature of stars.

An important offshoot of the ideas of conservation of energy, charge, baryon number, and lepton number is the existence of matter and antimatter. For all ordinary particles – electrons, neutrinos, protons, and neutrons – there are antiparticles – antielectrons, antineutrinos, antiprotons, and antineutrons. These are not fantasy propositions; they are made routinely in what are loosely called ''atom smashers,'' and more formally, particle accelerators, and they rain down continually on the Earth in the form of cosmic rays. The connection to the conservation of charge is that antiparticles always have the opposite charge of the ''normal'' particle. The antielectron, also called a *positron*, has a positive electrical charge. An antiproton has a negative charge. Because neutrinos and neutrons have no electrical charge, neither do their antiparticles; but they have other complementary properties. For instance, to make sense of the way physics works, it is necessary to consider an antielectron to count as a ''negative'' lepton and an antiproton to count as a ''negative'' baryon. In that sense, assigning the property of ''leptonness'' or ''baryonness'' to a particle is like assigning an electrical charge; it can be positive or negative and is opposite for particles and their antiparticles.

A remarkable property of particles and antiparticles is that they can be produced from pure energy and can annihilate to produce pure energy. Carl David Anderson won the Nobel Prize in Physics in 1934 for the discovery of positrons. Positrons were first created in a laboratory by applying a very strong electric field, the energy source, to an empty chamber, a vacuum. When the electric field reached a critical value, out popped electrons and positrons. You can see the connection with conservation of energy, charge, and leptons here. The energy of the electric field must be strong enough to provide the energy equivalent of the mass of an electron and a positron, twice the mass of a single electron. Because the original vacuum, even with

the imposed electrical field, had no net electrical charge, the final product, the electrons and positrons, also must have no net electrical charge. For every negatively charged electron that is created in this way, there must be a particle with the opposite electrical charge, an antielectron, a positron. Likewise, the original apparatus had no ''leptons,'' just the electrical field and vacuum. When an electron and positron appear, the electron must count as plus one lepton, and the positron as minus one lepton, so that the net number of leptons is still zero, in analogy with the way one keeps track of electrical charge. Here is a schematic reaction:



This experiment can also be run backward. If an electron and positron collide, they annihilate to produce pure energy – photons of electromagnetic energy – with no net electrical charge and no net number of leptons. The same is true of any particle and antiparticle. When they collide, they annihilate and produce pure energy; all the mass disappears. This is a very dramatic example of conservation of energy and of Einstein's formula, $E = mc^2$; pure energy can be converted into matter, and matter can be converted into pure energy. In the process, the total number of electrical charges, the total number of leptons, and the total number of baryons does not change. The total of each is always zero.

You might wonder, if antiprotons annihilate protons on contact and hence are antimatter, do they antigravitate? If I make an antiproton in a particle accelerator, will it tend to float upward? The answer is no. Energy is directly related to mass by the formula $E = mc^2$. One implication of this relation is that because mass falls in a gravitational field, energy also falls in a gravitational field. Because particles and antiparticles annihilate to form a finite, positive amount of energy that will fall in a gravitational field, so the individual particles

and antiparticles must fall. An antigravitating particle might annihilate with a gravitating particle to produce no energy, but we do not know of any such particles. Current physics does give some hints of the existence of antigravity which we will discuss in Chapter 12.

1.2.3 *The energy of stellar contraction*

We can now apply these various conservation laws to stars. We will start with the principle of conservation of energy. The result is a little surprising at first glance, but crucial to understanding the way in which stars evolve.

Let us first consider the nature of a star. A star is a hot ball of gas in *dynamic equilibrium*. This means that a pressure of some kind pushes outward and balances the gravity that pulls inward. The Sun does not have the same size day after day because there are no forces on it that might alter its size; rather there are great forces both inward and outward at every point in the Sun. The structure of the Sun has adapted so that the forces just balance. The equilibrium is such that the pressure force keeps gravity from collapsing the star, and gravity keeps the pressure from exploding the star. We will see in Chapter 6 that this condition of delicate balance can be interrupted and either collapse or explosion can result, depending on the circumstances. The mass and size of a star determine the gravity and hence the pressure and heat needed to arrange the balance of forces.

The Sun and most stars we see scattered in the night sky are supported by the pressure of a hot gas. The pressure, in turn, is directly related to the thermal energy in the star. At the same time, the star is held together by gravity. As the star radiates energy into space, it loses a net amount of energy. What happens to the temperature in the star? The answer is dictated by the principle of conservation of energy.

If the star were like a brick, the answer would be simple. As energy is radiated away, a brick just cools off. Gravity plays a crucial role in the makeup of a star, however. If the star were to cool, the pressure would tend to drop, and then gravity would squeeze the star, compressing and heating it. A star responds to a loss of radiant energy in just this paradoxical way. As the star loses energy, it contracts under the compression of gravity and actually heats up! This process, illustrated in Figure 1.3, is completely in accord with the conservation of energy. One must remember only that the squeezing by gravity is an important energy source that cannot be ignored when counting

up all the energy, just as the energy of falling breaks the chalk in Figure 1.2.

If nuclear reactions happen by accident to momentarily put more energy into the star than it radiates, the star gains energy. What happens to its temperature in this case? If you were bitten in the first case, you should be wise by now. As shown in Figure 1.3, if the temperature were to go up, the pressure would rise and push outward against gravity. The expansion would cause the star to cool. That is just what a star does; if you put in an excess of energy, it expands and gets cooler.

This apparently contradictory behavior of a star to heat up when it loses energy and cool off when it gains energy is a direct application of the law of conservation of energy. This behavior is crucial for the evolution of stars as various nuclear fuels flare up and burn out.

1.2.4 *Quantum theory*

Things work differently in the microscopic world of atoms and elementary particles than would seem to be ''normal'' from our everyday experience. On the scale of very small things, behavior is described by *quantum theory*. On this scale, changes do not occur smoothly, but in jumps. The behavior of matter on the quantum level does, however, have important implications for big things like stars.

In our ordinary macroscopic world, the old argument about the impenetrability of matter is approximately true; you cannot put your fist through a concrete wall. Your fist and the wall cannot occupy the same volume. The notion of impenetrability is very different in the microscopic world of the quantum theory. According to the quantum theory, elementary particles are not hard little balls, but also have wave-like qualities to them. Particles can, in principle, occupy exactly the same position in the same way that two ripples on a pond can occupy the same position momentarily as they pass through one another. Another aspect of the wave-like nature of particles is that their position cannot be specified. Think of the task of specifying where an ocean wave is: is it where the surface starts to curl upward, where the froth breaks on the crest, or in the wake? According to the *uncertainty principle* of the quantum theory, the positions of particles cannot be specified exactly. More precisely, there is complementary uncertainty between the position and the momentum of a particle. If the location of a particle is limited in some way, for instance, by being confined in an atom, the momentum and the energy become very

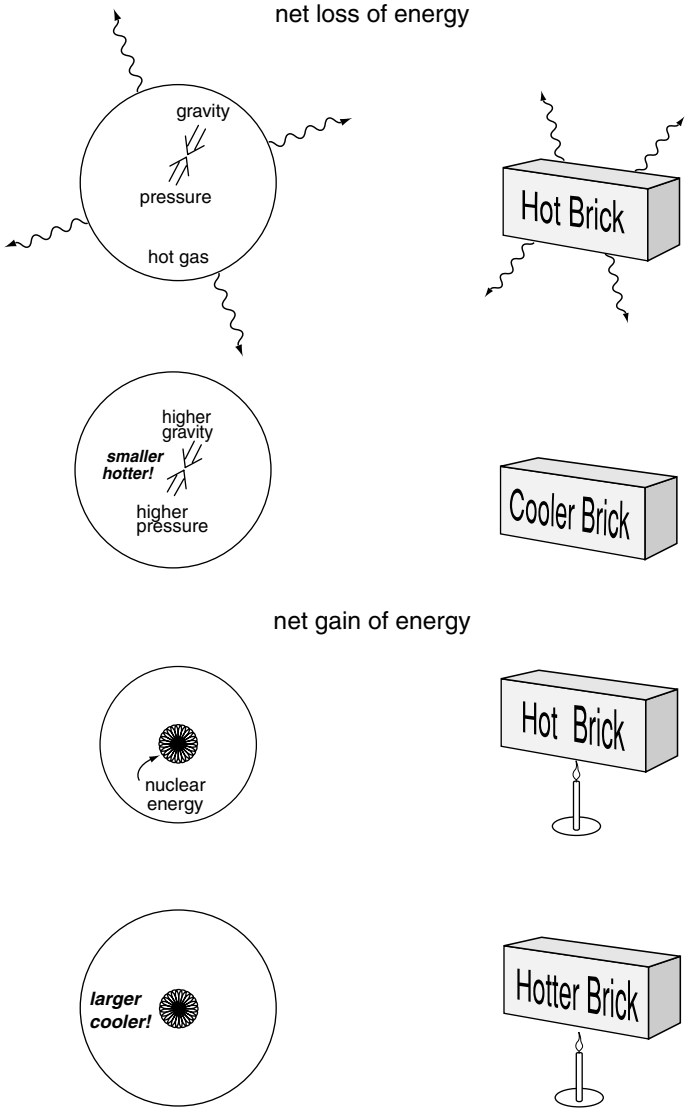


Figure 1.3 Stars supported by the pressure of a hot gas behave differently than a solid object like a brick. A brick will cool off as it radiates energy and heat up if a source of energy is added. Because of the action of gravity, a star held up by the thermal pressure of a hot gas will heat up when it loses a net amount of energy by radiation and cool off if it gains a net amount of energy from nuclear reactions.

uncertain. If the momentum is made more certain, you do not know where the particle is. According to the quantum theory, the position of a particle is the place where it might be and the volume it occupies is a measure of the uncertainty of its position. Rather than hard spheres, particles are more like little fuzzy balls or collections of waves. This property of uncertain position, momentum, and energy allows more than one of them to occupy the same volume in the right circumstances.

There are particles in the quantum world, however, that in special situations possess a property of absolute impenetrability. Among the particles that possess this property are familiar ones – electrons, protons, neutrons, and neutrinos. Particles of this class cannot occupy the same little smeared-out uncertain region of space if they have the same momentum, or, rather loosely, the same energy. This property is known formally in the quantum theory as the *exclusion principle*. Curiously, these particles can occupy the exact same volume as long as they have different momentum or energy. Two electrons, for instance, cannot occupy the same place if they have the same momentum, but they can if they have different momentum, as shown in Figure 1.4 (top). A common particle that does not obey the exclusion principle is the photon; two photons of the same energy can occupy the same volume at the same instant.

The uncertainty and exclusion principles determine the structure of atoms. The electrons exist in a smeared volume surrounding the atomic nucleus. The size of this volume is in accord with the uncertainty principle and the fact that electrons are wave-like and their positions cannot be specified precisely. The electrons are confined into a restricted volume by the positive attraction of the protons in the nucleus. The electrons can all occupy nearly the same volume because some have higher energy, thus satisfying the constraint of the exclusion principle.

These quantum properties of particles come into play in a very important way as stars evolve. Normally the particles in a star are spread out in space and in energy, as shown in Figure 1.4 (bottom left). In this situation, the gas exerts a *thermal pressure* as the particles randomly collide and bounce off one another and generally tend to move apart. This thermal pressure associated with a hot gas supports the Sun and stars like it.

As the stars burn out their nuclear fuels, they contract and become very dense. The electrons in the stars are squeezed tightly together. The electrons get compacted into a state where the volume

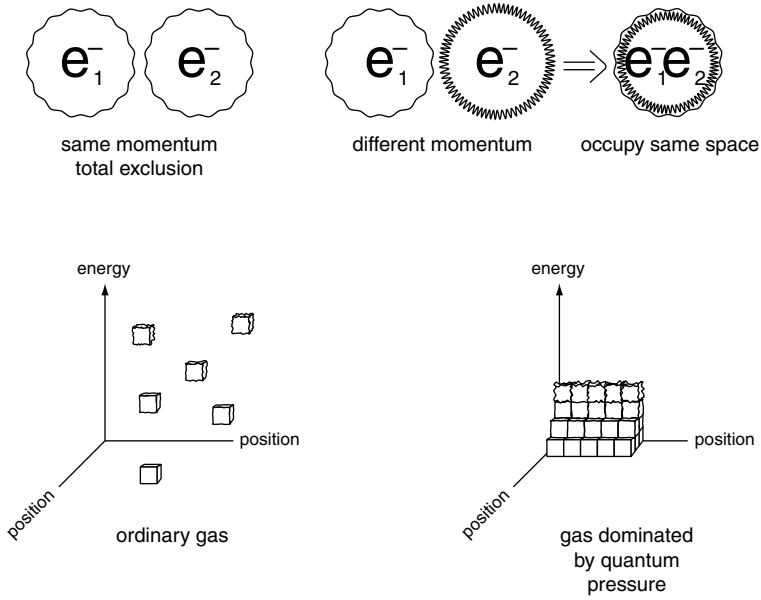


Figure 1.4 Aspects of the quantum behavior of particles: (top left) two electrons with the same momentum are absolutely excluded from being in the same place, and from occupying the same volume; (top right) if one electron has a different momentum and hence energy, its ''waves'' are in a different state and this allows the two electrons to occupy exactly the same volume; (bottom left) a normal gas of hot particles has the particles spread out in position and energy so that quantum effects are not important and the resulting thermal pressure depends on the temperature as well as the density of the particles; (bottom right) if particles are packed tightly enough by having a very high density, then particles with the same energy occupy volumes dictated by the uncertainty principle but, according to the exclusion principle, cannot occupy the same volume unless they have different energies. The energy acquired by the particles depends only on the density and not the temperature, but it can provide a quantum pressure that can support a star.

of quantum uncertainty occupied by each electron is bumping up against that of its neighbor. Electrons of the same energy would then absolutely resist any more compaction. That state of the star would be the maximum compression allowed according to the exclusion principle if no two electrons could occupy the same volume. Many electrons can, however, occupy the same volume if some of the electrons have extra energy. Extra energy does arise in this circumstance as a

result of the compaction by gravity and the action of the uncertainty principle. As the space that the electrons occupy becomes more confined, their positions become more ''certain.'' To satisfy the uncertainty principle, the energy (strictly speaking the momentum) must become more uncertain. As the uncertainty in the electron energy becomes higher, the effective average energy of the electron increases. Thus the compaction squeezes the electrons together, the exclusion principle prevents two electrons with the same energy from occupying the same volume, and the restricted volume gives the electrons more energy in accord with the uncertainty principle. With more energy, some electrons can now occupy the same volume, as illustrated in Figure 1.4 (bottom right). The fact that electrons can gain energy and hence overlap in the same volume allows greater compaction of the star.

The net effect is that the squeezing of the electrons gives them an energy that derives purely from quantum effects. The ''quantum energy'' that results from stellar compaction depends only on the density and is completely independent of the temperature. This quantum energy can exceed the normal thermal energy due to random motion of gas particles by great amounts. The electrons that acquire this quantum energy can also exert a *quantum pressure*. This quantum pressure can provide the pressure to hold the star up even when the thermal pressure is insufficient.

The fact that the quantum pressure is independent of the temperature has major implications for the thermal behavior of compact stars for which this pressure dominates. When a star is supported by the quantum pressure, it does not contract upon losing energy by radiating into space. The reason is that as the temperature drops, the quantum pressure is unaffected and remains constant. A star supported by quantum pressure behaves like ''normal'' matter; when it radiates away energy, it cools off. In this sense, such a star is more like a brick that just cools off when it radiates its heat, as illustrated in Figure 1.3.

Stars supported by the quantum pressure of electrons are known as *white dwarfs*. They will be discussed in more detail in Chapter 5. Only so much mass can be supported by the quantum pressure of electrons. This limiting mass is called the *Chandrasekhar mass* after the Indian physicist, Subramanyan Chandrasekhar, who first worked out the concept, shortly after the birth of the quantum theory. Chandrasekhar did this work as a very young man and was finally awarded the Nobel Prize for it in 1983. Chandrasekhar and his

work have been honored once again by naming a major NASA orbiting observatory, the *Advanced X-ray Astronomy Facility*, the *Chandra Observatory*. The maximum mass a white dwarf can have for an ordinary composition is 1.4 solar masses, not much more massive than the Sun. If mass were to be piled onto a white dwarf so that its mass exceeded that limit, the white dwarf would collapse, or perhaps explode if it were composed of the right stuff. That notion will be explored in Chapter 6.

1.3 EVOLUTION

The mass of a star sets its fate. The structure and evolution of a star of typical composition follow from the mass with which it is born. The mass determines the pressure required to hold the star up. The condition that the pressure balances gravity determines the temperature and the temperature determines the rate of nuclear burning and hence the lifetime of the star. For much of a star's life, the pressure to support it comes from the thermal pressure of a hot gas. This means that when a star loses a net amount of energy it heats up and when it gains a net amount of energy it cools off, as described in Section 1.2.3. This fundamental property controls the development of the star.

1.3.1 *Birth*

Stars first come into existence as protostars. Protostars are thought to form by some sort of intrinsic instability in the cold molecular gas that pervades the interstellar medium. Sufficiently massive clumps of this matter have an inward gravity that exceeds the pressure they can exert, so they contract and become ever more dense and hot until nuclear reactions start and the clump becomes a star. Alternatively, there are processes involving energetic shock waves that may cause the matter floating through space to clump together. The shocks may come from the passage of the interstellar gas through the spiral arm of a galaxy, from the explosion of a supernova, or from the flaring birth of another nearby star.

When a protostar forms, it is not yet hot enough to burn nuclear fuel. To burn nuclear fuel, the protostar must get hotter. The wonderful property of stars, even as protostars, is that if they must become hotter to yield nuclear input, they will automatically do so. That is the nature of the star machine, a machine controlled by conservation of energy under the influence of gravity. For the protostar,

this works because the protostar is warmer than the cold space around it. Under this circumstance, the protostar will radiate energy into space. Because a protostar has no energy input from nuclear burning, it loses a net amount of energy into space. This is exactly the circumstance in which a star will heat up! As shown in Figure 1.5, the protostar will continue to lose energy and heat until it becomes hot enough to ignite its nuclear fuel. At this point, the protostar becomes a real star, shining with its own nuclear fire.

1.3.2 *The main sequence*

If you point at a person in a crowded shopping mall, the probability is that the person is middle aged, neither an infant nor very aged. The stars about us in space have a similar property. If you pick a star in the night sky at random and ask what it is doing, the probability is that it will be in the phase where stars spend most of their active lives. When stars were first categorized, most were empirically found to fall in one category in terms of the basic observable criteria of temperature and luminosity. This category is called the *main sequence*. We now understand the physical meaning of the main sequence. Stars are composed mostly of hydrogen, and the main sequence represents the phase of the thermonuclear burning of that hydrogen. Hydrogen burns for a long time compared to other elements. For this reason, stars spend most of their active lifetimes not as protostars or highly evolved stars but as hydrogen-burning stars, just as humans spend most of their lifetime as adults, not as infants or octogenarians.

The Sun is in the main sequence hydrogen-burning phase. It is about halfway through its allotted span of 10 billion years. Stars more massive than the Sun burn hydrogen for a shorter time. This may seem strange because massive stars contain more hydrogen fuel to burn. The reason is that massive stars require a greater pressure to support them and hence have a higher temperature. This causes them to burn their extra fuel at a far more prodigious rate than the Sun and so spend their extra fuel in a very short time. Likewise, stars with less mass than the Sun have lower pressure and temperature. They burn their smaller ration of fuel very slowly and live on it far longer than even the Sun will. Stars with less than about 80 percent of the mass of the Sun that were born when the Galaxy first formed have scarcely begun to evolve; the Universe is not old enough yet.

A given star burns its hydrogen at a very steady rate. This is because the star acts to regulate its burning to a very precise level,

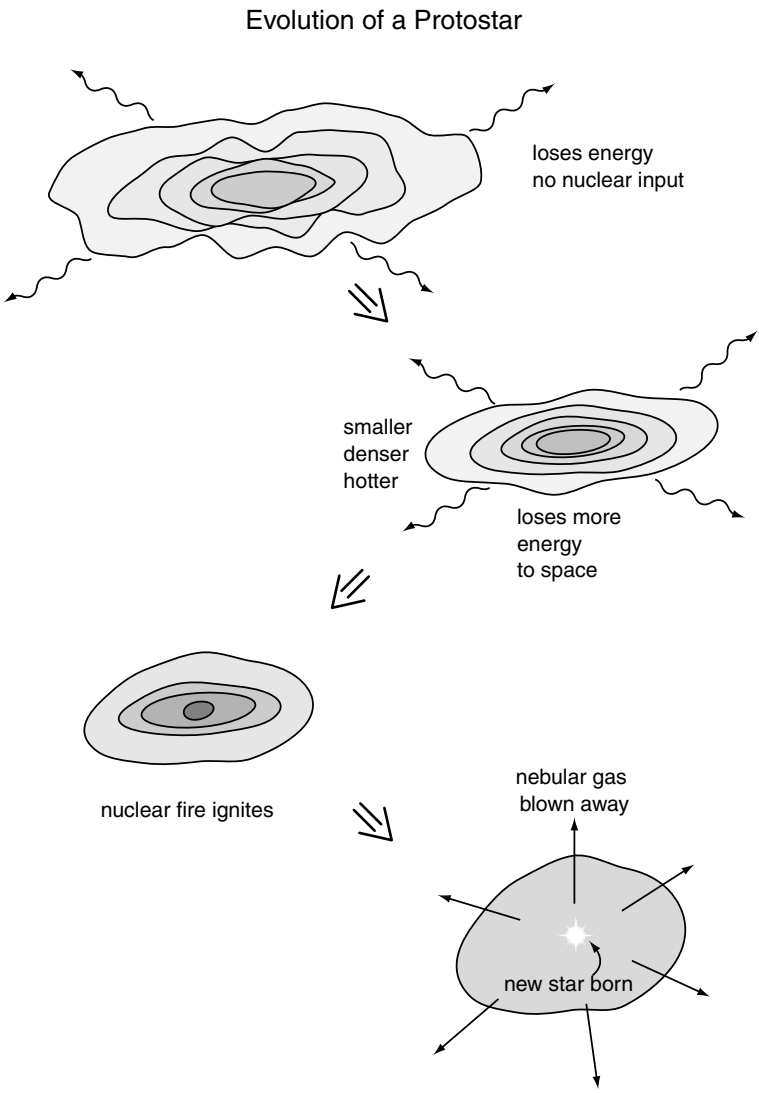


Figure 1.5 A protostar forms from a swirling cloud of cold interstellar gas. Because it radiates into space, but has no nuclear input, the protostar will contract under the pull of gravity and become smaller, denser, and hotter. This process will continue until the center of the star becomes hot enough to light the nuclear fire. The excess gas is blown away and the star emerges from its cocoon to shine with its own nuclear energy.

using the same principle of energy conservation that ignites the fuel in the first place. If the nuclear furnace belches slightly and puts forth a little more heat than can be carried off by the radiation from the star, the excess heat increases the pressure and causes the star to expand. The excess energy is spent in making the star expand. More energy goes into the expansion than was produced in the nuclear belch, and the star actually ends up slightly cooler as explained in Section 1.2.3 (Figure 1.3). The nuclear reaction rates are sensitive to the temperature, and so the nuclear burning slows as the expansion occurs and the temperature drops. The net effect is a highly efficient process of negative feedback. If the star temporarily produces an iota too much heat, the nuclear fires are automatically damped a bit by the expansion to restore equilibrium. The opposite is also true. If the nuclear burning should fail to keep up with the energy radiated away for an instant, the heat would be insufficient, the pressure would drop, the star would contract, and the temperature would rise. The result is that the nuclear burning would be increased to the equilibrium value. A star burning hydrogen on the main sequence thus works in a manner similar to the thermostat and furnace in a house. If the temperature drops, the furnace kicks in and restores the lost energy. If the house gets too hot, the furnace turns off temporarily until the desired temperature is restored.

The process of hydrogen burning on the main sequence is one of thermonuclear fusion. Nuclei of hydrogen atoms, protons, are fused together to make the nucleus of the heavier element helium, which consists of two protons and two neutrons. Burning hydrogen to helium depends primarily on the nuclear force. The role of the nuclear force is to bind the four particles in the helium nucleus. The energy left over from combining the particles is available as heat. This process is not different in principle from ordinary burning, where chemical forces bind the combined products together and liberate the energy of combining the molecules as heat. Chemical forces are based on the electrical force. The reason that nuclear burning is so much more powerful than chemical burning is because the nuclear force is so much stronger than the electrical force. The energy released in the fusion of hydrogen into helium is an appreciable fraction, about 1 percent, of the maximum amount of energy that could be released if all the mass of hydrogen were turned into pure energy in accordance with $E = mc^2$. That very high efficiency of energy release is why thermonuclear bombs are such a fearful weapon and why the promise

of controlled thermonuclear fusion is so enticing as an ultimate energy source.

Look more closely at the process of turning hydrogen into helium. There are many ways in which this can be done in practice, but they all have a common link. The process of thermonuclear fusion consists of combining four protons to make helium. Of necessity, some step in this process requires that two of the protons be converted into two neutrons. Protons are converted into neutrons (and vice versa) by the influence of the weak force. To understand how this process works, and to reveal an important practical consequence, we must also invoke the laws of conservation of charge and of baryons and leptons, as introduced in Section 1.2.2.

The conversion of two protons into two neutrons during hydrogen fusion conserves the number of heavy, baryon, particles; there are two to start and two in the end. That process cannot occur alone, however, because charge is not conserved; the charge on the protons cannot just disappear. One way to get around this is to produce two positively charged particles to balance the charge on the protons and to give no net change in the electrical charge. These positive particles cannot be baryons of any kind because the number of baryons in the reaction is already balanced. Nature solves this problem by providing leptons in the form of positrons. If two protons are converted into two neutrons and two positrons by the weak force, we have no net charge. Now, however, we are making two new leptons, and to conserve the lepton number, the reaction must spit out two other leptons along with the two neutrons. Recall from Section 1.2.2 that positrons have the opposite charge and the opposite leptonness from electrons. Algebraically, they each count as ''minus one'' lepton in the exit channel. The other leptons coming out of the reaction must carry no charge, because the charge is already properly balanced, but must count as ''plus one'' in terms of leptons in order to offset the positrons. To balance charge, baryons, and leptons all at once in this reaction, nature provides the neutrino!

The fact that the neutrino was needed to conserve all the relevant quantities in certain nuclear reactions was first realized by the Italian physicist, Enrico Fermi. It was Fermi who gave the particle its name, meaning little neutral one. Fermi was awarded the Nobel Prize for this and related work in 1938 as he prepared the world's first nuclear reactor and took seminal steps that would lead to the Manhattan Project in World War II. The neutrino was not directly detected until after the war, in the 1950s, when Fred Reines and colleagues

registered neutrinos coming from a nuclear reactor. Reines was given the Nobel Prize for this discovery in 1995.

Figure 1.6 summarizes the essential processes that occur when hydrogen undergoes thermonuclear fusion to make helium. In that conversion, a neutrino must be made for every neutron that is produced in order to conserve baryons, leptons, and electrical charge simultaneously. For every atom of helium produced, two neutrinos must be generated. That fact represents both an opportunity and a challenge to astronomers and physicists.

1.3.3 *The solar-neutrino problem*

Hydrogen burns and neutrinos are produced in the centers of stars because that is where the temperature is the highest. Because neutrinos interact only by the weak force, normal stellar matter is virtually transparent to them. The neutrinos that are produced in the central hydrogen-burning reactions immediately flow out of the star at nearly the speed of light, as shown in Figure 1.7. They carry off a small amount of energy that would otherwise be available to heat the star, but this energy is not of great import. The importance of the neutrinos to astronomers is that they come directly from the center of the star, carrying information about conditions in the stellar core. Otherwise, astronomers are limited to studying photons of light that come only from the outer surface of the stars. Study of these photons is a powerful tool to deduce the nature of the inner portions of a star, but that is no substitute for being able to directly ''see'' inside. Neutrinos from the Sun provide that opportunity.

The problem with observing the heart of the Sun by means of neutrinos is that the neutrinos will stream through any detector unimpeded, for the same reason that they stream freely out of the star. Detection of the neutrinos depends on amassing a huge detector and then waiting for that rare time when the weak force causes a reaction within the detector. This process is totally impractical for any star but the Sun, because the great distance dilutes the neutrino ''brightness'' from a distant star as rapidly as it does visible photons.

The first successful effort to detect neutrinos from the Sun was the result of a multi-decade effort by Ray Davis and his collaborators (see Figure 1.7). In the first edition of this book, I said ''This work has not yet won a Nobel Prize, but it should.'' I am happy to write in the

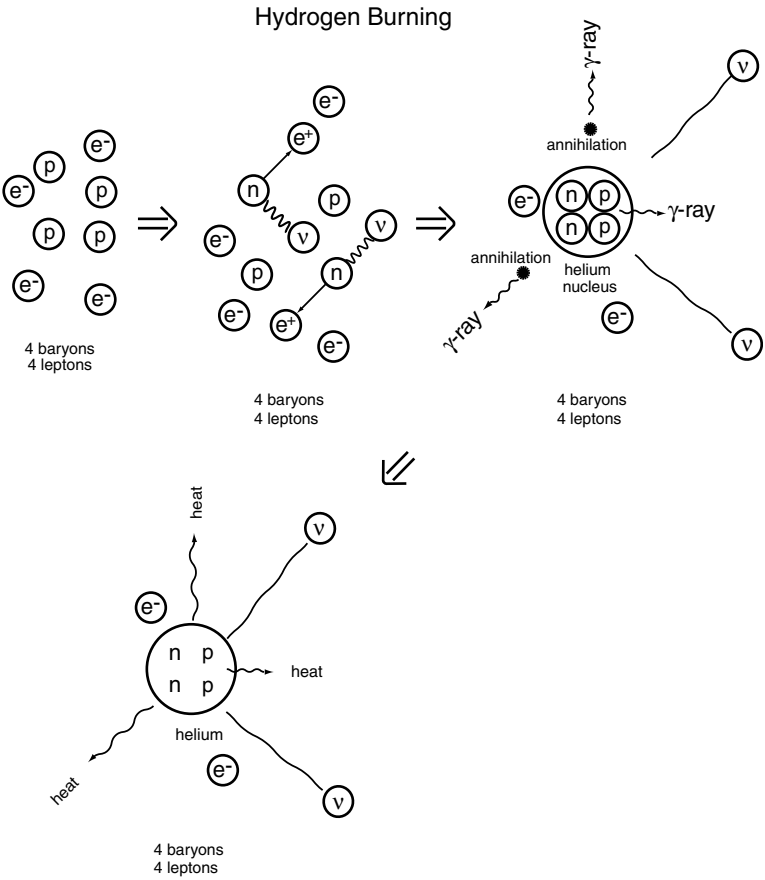


Figure 1.6 The process of hydrogen burning involves the thermonuclear fusion of hydrogen to make helium: (top left) the original hydrogen gas consists of equal numbers of protons and electrons, four baryons and four leptons; (top middle) under the combined action of the strong and weak forces, two of the protons are converted to two neutrons plus two positrons and two neutrinos. The net electrical charge is still zero and because positrons represent antileptons, there are still only four baryons and a net of four leptons; (top right) the strong force binds the two remaining protons and the two newly created neutrons into a nucleus of helium. This process releases a large amount of heat in the form of the radiant energy of gamma rays. The positrons annihilate upon collision with two of the initial electrons and produce a little more gamma-ray energy. The net result is still four baryons – two protons and two neutrons – and four leptons – two of the remaining initial electrons and two newly made neutrinos; (bottom) the final product is a new helium nucleus, heat, and two neutrinos that race out of the star and into space.

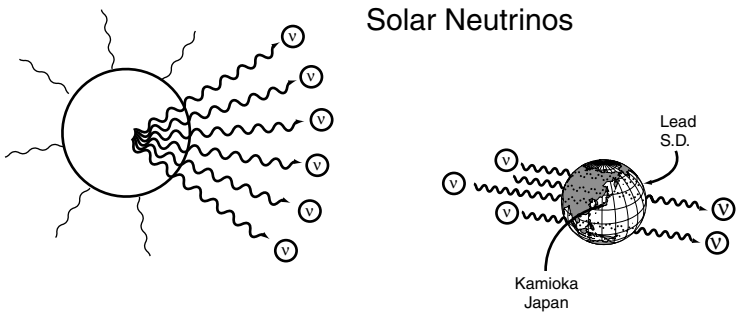


Figure 1.7 Neutrinos produced in the thermonuclear burning of hydrogen to helium in the center of the Sun flood into space. Some of the neutrinos head in the direction of the Earth. Most of the neutrinos that reach the Earth also pass right through it, but a few can be stopped and studied in special detectors. The pioneering solar-neutrino detector was in Lead, South Dakota. The currently most successful detector is in Kamioka, Japan.

second edition that Ray Davis was awarded the 2002 Nobel Prize in Physics for this revolutionary undertaking.

The detector consisted of a hundred thousand gallons of chlorine-rich cleaning fluid. The chlorine undergoes an interaction with a neutrino by means of the weak force. This interaction turns a neutron within a chlorine nucleus into a proton, just the opposite of the reaction that produced the neutrino in the Sun. Changing a neutron in chlorine into a proton converts an atom of chlorine into an atom of radioactive argon. The argon can be collected efficiently because it is a noble gas and does not combine chemically. The tank containing the cleaning fluid was at the bottom of the Homestake gold mine in Lead, South Dakota. The underground operation is necessary to screen out cosmic ray particles that could induce spurious transitions of the chlorine to argon. The mine was vacant until the price of gold soared to astronomical highs several years ago. The Homestake company reactivated it, and for a while the scientists had to work to the sound of dynamite explosions as new veins were developed. More recently, mining stopped again and the mine has flooded. There are attempts to get the whole mine dedicated to underground physics, but it is not clear they will be successful.

At first, the solar-neutrino experiment gave no signal at all above the background ''noise'' of extraneous reactions. This caused a great deal of anguish in the astronomical community because the first opportunity to peer directly inside the Sun gave a result inconsistent

with apparently straightforward theoretical predictions. With patience, a positive signal was detected. A few hundred atoms of argon are collected each month from the hundred thousand gallons of fluid! Detection of some neutrinos is more reassuring than detection of none at all, but a new and serious problem still arose. The most careful analysis of a standard computer model of the Sun predicts several times more neutrinos than are observed.

The discrepancy could lie in several areas. The nuclear reactions could proceed in a different manner than we envisage. The structure of the Sun could be somehow different. Perhaps the composition, particularly the heavy elements, is not spread uniformly through the volume, as assumed. Perhaps the fundamental properties of the neutrinos themselves are different. The gold-mine experiment is looking for the particular type of neutrino produced when protons change to neutrons. There are (at least) two other kinds of neutrinos. If the neutrinos have undergone a Jekyll and Hyde transformation in flight and are one of the other types when they arrive at Earth, they would not induce the desired transformation of chlorine to argon and would go undetected.

Recent developments may have given the key to this mystery. One reassuring result came from an underground neutrino detector constructed in Kamioka, Japan, called Kamiokande (Figure 1.7). This detector is a massive vat of water. Unlike the chlorine experiment, it can see neutrinos in real time and can tell the direction in which the neutrinos are moving and hence the direction from which they came. The neutrinos can trigger the conversion of a neutron to a proton in the oxygen in the water or collide with one of the electrons in the water. In either case, the particle that is hit is given substantial energy and flies rapidly through the water in the direction that the neutrino was traveling. The recoil particles give a flash of blue light known as Cerenkov radiation in the direction in which they are moving. From this flare of light in the detector, the direction of the neutrinos can be tracked. The Kamiokande experiment saw the same kind of neutrinos as the chlorine experiment and at the same low rate, but, to everyone's great relief, the neutrinos were definitely coming from the direction of the Sun! Without that confirmation, there was a small probability that the Homestake detection was some local contamination and not solar neutrinos at all. That would have made the problem even worse.

The second development may have given the real answer. The Homestake and Kamiokande experiments detect only the stream of

the few high-energy, relatively easy to detect neutrinos that come from a rare version of the hydrogen-burning process. That rare process might be affected by subtle changes in the interior of the Sun that would not affect the overall power output. The chlorine and water experiments cannot detect the far more numerous neutrinos that must be produced in the basic reaction by which a proton is turned into a neutron at a rate that is directly proportional to the power that flows in radiation from the surface of the Sun. Another experiment, carefully planned for a decade in collaboration between Ray Davis and Russian physicists, uses the element gallium as a detector. This substance is sensitive to the basic flood of low-energy neutrinos that must be there because the Sun, after all, is shining. The gallium experiment also failed to see the predicted rate of neutrinos! The only remaining conclusion is that something is omitted from our simplest physical picture of the neutrinos.

As mentioned earlier, there are three different types of neutrinos, each with their antineutrinos. That there are three types of neutrinos is related to the fact that there are three types of quarks that make up other particles like protons and neutrons. When neutrinos were first discovered, it was suspected that they had no mass. If that were the case, each type of neutrino would always be the same. The fledgling grand unified theory combining the strong and electroweak forces suggests that neutrinos must have a small mass. In that case, the theory predicts, there are circumstances in which one type of neutrino can be converted to another type. If this happens round and round and back and forth among the three types of neutrinos, then by the time the neutrinos arrive at the Earth there might be roughly equal amounts of all three. In this case, only one-third of the type originally produced in the Sun that the experiments were specifically designed to register would reach the detectors. The fact that about one-third of the expected rate is observed is consistent with this notion.

This interpretation of the solar-neutrino experiments strongly suggests that we not only have at last the solution to the solar-neutrino problem but also have strong evidence for the grand unified theory of elementary particles. This is probably the answer, but it also raises the challenge of building more experiments to test the hypothesis.

A major step in this saga was announced in the summer of 1998 by the teams of scientists working on the new, larger underground experiment in Japan known as Super Kamiokande. This experiment

found evidence that neutrinos do shift from one type to another as they interact with the Earth's atmosphere, and hence that they must have a mass, as expected from theory. The mass is not measured directly, only the difference in the masses, but this is a major breakthrough. On the other hand, to account for all the data from all the experiments, there is some discussion of the need to introduce yet another type of neutrino called a ''sterile'' neutrino that interacts only with neutrinos and with no other particles at all. This seems a step backward. Study of solar neutrinos still has much to teach us. We will return to neutrinos in another context in Chapters 6 and 7.

2

Stellar death: the inexorable grip of gravity

2.1 RED GIANTS

The Sun looks the same to us, unchanging, day after day. A simple observation, however, tells us that it is evolving and must be changing in some manner. That observation is just the warmth on our upturned faces on a sunny day. The radiation that flows from the Sun carries energy out into space. There is nothing from space replacing that energy. The Sun must, therefore, be losing energy overall. Something must be going on within the Sun that is slowly, inevitably altering it. The lesson from Chapter 1 is that the change in the Sun involves its composition. The Sun is irrevocably transmuting some of its hydrogen into helium. That transformation cannot be undone. The alteration of the structure of the Sun is slow, but it is steady. Eventually, the changes will be drastic.

As remarked in Chapter 1, the hydrogen burns only in the center of a star, where the temperatures are highest. That means that the central region is where the hydrogen is consumed and the helium builds up. Even when the hydrogen is fully transformed in the central region, the outer, cooler portions of the star will not have burned. They retain their original composition. This causes the star to become schizoid and to do two things simultaneously: shrink and swell. This development is in strict accord with the principle of conservation of energy, but the application of this principle is more complex than for stars with a homogeneous composition.

When hydrogen is exhausted in the center, the star has a central volume of nearly pure helium (along with the scattering of heavy elements initially present in the star). The remainder of the star is original material, composed mostly of hydrogen. The difference between the inner parts of the star, where the composition has been

altered, and the outer part, where the composition is unchanged, become ever more distinct as the star evolves. To distinguish these two portions of the star, the inner part is called the *core*, and the outer part, the *envelope*.

The helium in the stellar core can become a thermonuclear fuel. Helium burning does not happen spontaneously, however, any more than hydrogen burning did. The nuclear force is strong, but it only acts over very short distances. The particles to be combined must be brought close together. There is, however, a force that inhibits the particles from getting close to one another. This is just the electrical force of the repulsion of like charges. The nuclei of atoms, such as hydrogen, helium, or heavier elements, are composed of positively charged protons and neutrons with no electrical charge. All nuclei thus have a net positive charge. If the electromagnetic force and gravity were the only forces in the Universe, this charge repulsion would prevail, and there would never be any nuclear reactions.

To initiate thermonuclear burning, the charge repulsion among the protons must be overcome. The electrical repulsion is not as strong as the nuclear force, but it acts over greater distances and dominates while the particles are far apart. At close distances, the nuclear force is stronger, and it can grab the particles and bind them tightly together. To bring like-charged particles together so the nuclear force can grab them and liberate energy, some energy must first be expended to fling the particles together despite the resistance of the electrical repulsion. You do not get something for nothing, but the nuclear payoff is worth the investment of some energy to overcome the charge repulsion. This principle is illustrated in Figure 2.1.

In practice, the charge repulsion is overcome by investing the particles with heat energy. This gives them more random energy of motion so they collide more fiercely and come closer within the grasp of the nuclear force during an encounter. To burn a nuclear fuel, you have to heat it first by raising the temperature, just as you need a match and kindling for the wood in a fireplace. A protostar must contract sufficiently to heat the hydrogen to get burning started initially. Helium nuclei have two protons, whereas hydrogen nuclei have only one. The charge repulsion is stronger for helium than for hydrogen, so helium must be heated to higher temperatures than hydrogen before it undergoes thermonuclear reactions.

When the last of the hydrogen burns out in the center of a star, the star must get even hotter to burn helium. It solves this problem in a natural way, using energy conservation. After the hydrogen burns

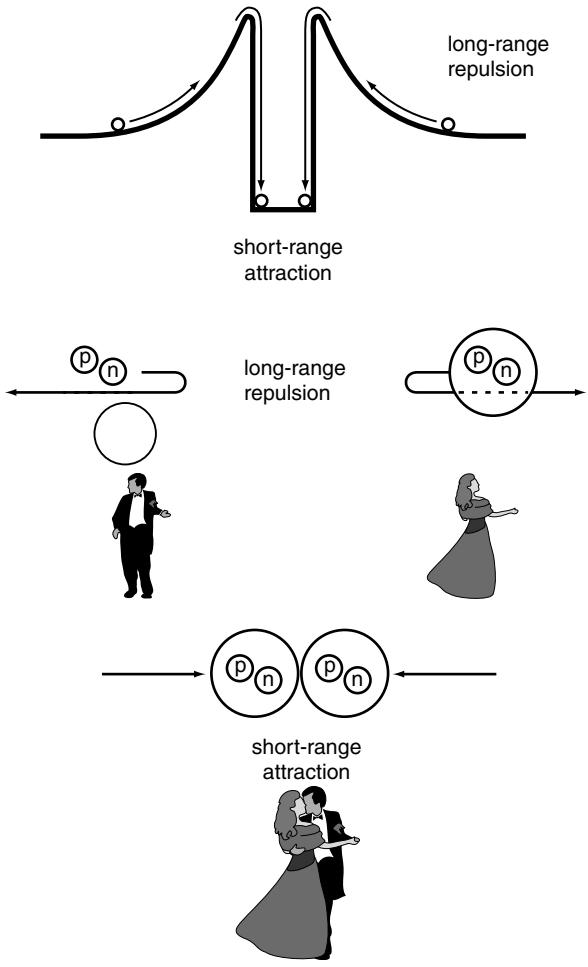


Figure 2.1 Positively charged atomic nuclei repel one another at long distances but are strongly attracted at short distances by the nuclear force. In analogy to a deep hole surrounded by a raised lip (a ''volcano''), some energy must be invested as heat to force the nuclei close to one another or to roll a ball up the hill. After the nuclei are sufficiently close together, the short-range nuclear force can bind the nuclei together to make a new element and liberate energy, just as a ball, having reached the precipice, can plunge down into the crater, yielding more energy than it took to roll it up the hill. In practice, the atomic nuclei need to have some neutrons so that their nuclear attraction can overcome the charge repulsion of the protons.

out in the center, no energy is being produced. Without the input of heat, the pressure cannot support the star. The star thus contracts and derives heat that way until the helium becomes hot enough to burn. The same mechanism that is responsible for igniting and regulating hydrogen burning on the main sequence causes the helium to ignite after the hydrogen is exhausted in the center. When the star has insufficient heat, it naturally contracts until that heat can be provided, whether by hydrogen, helium, or ultimately other sources of nuclear fuel.

Now comes the schizophrenia. The helium core contracts and heats until helium ignites. In its inimitable way, the gravitational contraction liberates more heat energy in the core than the core needs. The excess heat flows out into the overlying envelope of pristine material. The envelope responds in its own natural, but opposite, way. The envelope feels that it is getting an excess input of heat. The excess pressure causes the envelope to expand against gravity and cool to lower temperatures. The star thus does both things at once. The core loses energy, contracts, and heats, and the envelope gains energy, expands, and cools.

The contracting core is more important for the ultimate evolution of the star, but what astronomers actually see in their telescopes is the outside of the envelope. The outside, like the whole envelope, gets cooler and hence more red in color. Inside, the helium burns at a high rate and provides a high brightness for the star. At a given surface temperature, astronomers categorize the brightest stars as giants and the rather dim stars as dwarfs. The stars we are describing have become what astronomers call *red giants*. The size of such stars also becomes very large as the envelope expands, so the star is also a giant in terms of its extent, even though this is not technically what an astronomer means by giant. For instance, a blue supergiant is much brighter, but much smaller than a red giant. In any case, red-giant stars swell from the size of the Sun to extend well beyond the radii of the inner planets of the Solar System. We expect the Sun to undergo this transition in about another 5 billion years, at which time the inner planets should be engulfed and evaporated. The Sun will live about 1 billion years, about 10 percent of its total lifetime, as a red giant and then die.

To be fair, this explanation for the formation of a red giant by exchange of energy from the core to the envelope is a little simplistic. The exchange of energy does happen and is one factor, but experts still argue about the best way to understand why red giants form. The

computer models show that it happens, but the process is a complex, nonlinear interaction of the star with gravity and is not that susceptible to simple, this-is-the-key-factor-type explanations. In a certain sense, the formation of a red giant involves an instability. It is as if you push a book toward the edge of a table. Nothing much happens for quite a while. If you push too far, however, the book will land on the floor. As the core shrinks in a star that has consumed its central hydrogen, there comes a point where the envelope ''falls'' outward, coming to a lower-energy solution that couples the pressure in the core and envelope with gravity.

Stars with appreciable mass pass through several burning stages after they become red giants. They also spend about 10 percent of their total life in this phase, with each stage progressing faster than the one before. After each successive fuel is exhausted in the center, the star finds itself without a source of heat, so the core contracts until the material that was formed by the previous burning phase becomes hot enough to burn. The core must become hot enough to overcome the charge repulsion among the greater number of protons in ever more complex nuclei. In massive stars, hydrogen burns to become helium in the basic way we described in Chapter 1. The details are different than those for the Sun, but the net outcome is the same: four protons must combine to make a helium nucleus with the creation of two neutrinos.

In stars with the mass of the Sun and in more massive stars, helium burns to become carbon and then oxygen. The reason for this is that the simplest interaction one can imagine, combining two helium nuclei, makes a nucleus with four protons and four neutrons. For reasons that have to do with the details of how the nuclear force works, the nuclear attraction of that combination of protons and neutrons is not able to overcome the charge repulsion of the four protons. The combination of four protons and four neutrons is unstable. A nucleus with four protons and four neutrons falls apart and hence cannot be one of the steps in nuclear burning to produce a heavier ''ash'' from a given fuel.

Nature finds a way around this bottleneck by utilizing the more rare process by which three helium nuclei occasionally become close enough to combine under the control of the nuclear force. The result is a nucleus with six protons and six neutrons, the element carbon! This is where all the carbon necessary for life arises. As the helium burns in this way, some of the as yet unconsumed helium can combine with the newly formed carbon to make an element with eight

protons and eight neutrons, the element oxygen, another critical agent for life as we know it.

In the Sun, thermonuclear burning is expected to halt with the production of carbon and oxygen for reasons that will be addressed in Section 2.3. For sufficiently massive stars, the process continues. Ultimately, a complex of heavier elements forms. Prominent among these substances are the elements neon, magnesium, silicon, sulfur, argon, calcium, and titanium. That may seem an odd assortment, from a noble gas to the stuff in your bones to a metal used in submarine hulls, but there is a common factor. Each of those successive elements consists of two more protons and two more neutrons than the one before. Stars produce this chain of elements in especially large abundance because each is essentially made up of the basic building blocks of helium nuclei: three for carbon, five for neon, and ten for calcium. Each successive element contains more protons than the last because each phase of burning is one of fusing lighter nuclei into heavier ones. More protons means more charge repulsion to be overcome by higher temperatures. The star obligingly provides the higher temperature in the core by contracting whenever it finds itself without any nuclear energy input to balance the radiation energy lost to space.

This seductive process by which a star prolongs its life actually just puts it deeper and deeper in the grip of gravity. Gravity will ultimately win the battle.

2.2 STELLAR WINDS

Before delving into the depths of the stellar cores, let us consider some of the important processes in the outer parts of the star by which some stellar matter can escape the grip of gravity.

On the Earth, a wind is the actual motion of matter, air molecules moving en masse from one place to another. In addition to radiation, the Sun emits a wind of particles, mostly hydrogen, that flows out into space in all directions. For the Sun, the cause is not precisely known. It may be due to the turbulent, boiling surface pumping magnetic energy into the outer layers and expelling them. Evidence for the solar wind is in the tails of comets. Comet tails always point away from the Sun, wafted by the stellar breeze, whether the comet is headed toward or away from the Sun. The solar wind is interesting, but the total amount of matter expected to be lost from the Sun during its lifetime on the main sequence is negligible. The nature of a wind from a star is illustrated schematically in Figure 2.2.

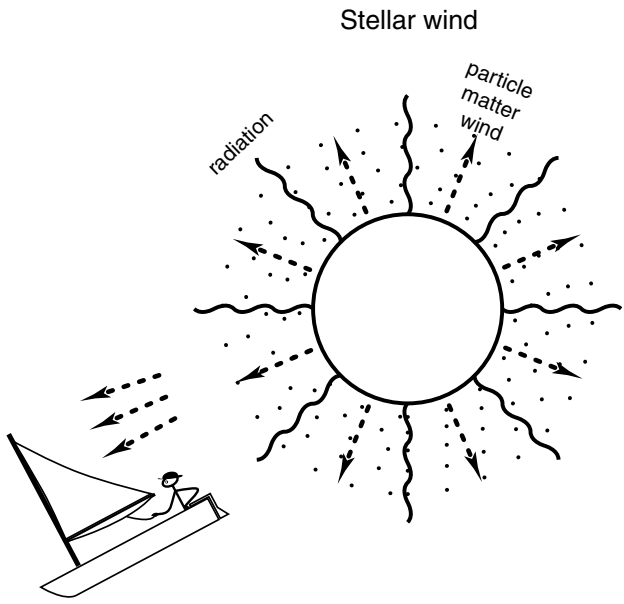


Figure 2.2 In addition to the flow of radiation from the surface of the Sun or other stars, there can also be a flow of matter, a stellar wind.

For more massive stars, the story is different because the loss of mass to a wind can substantially alter the evolution of the star. For massive stars, the mechanism to expel matter is thought to be the pressure of the intense radiation that flows from the star. Although we turn to the Sun for warmth, we do not usually think of the pressure of the sunlight on our faces. It is there, but it is very small. In space, with no competing effects, the pressure exerted by the photons of radiation streaming out from the Sun can be appreciable. There are dreams to have a sail-plane race in space with all the craft powered by the pressure of the solar radiation.

The power emitted in radiation from a star is known as the star's *luminosity*. The luminosity is the amount of radiation energy that flows from a star in a given time. The pressure exerted by the radiation is proportional to the luminosity. As the mass of a star goes up, the luminosity and the pressure exerted by the radiation increase by about the third power. That means that if you consider a star of twice the mass, the luminosity goes up by a factor of eight. For a sufficiently large stellar mass, the large radiation pressure associated with the large luminosity becomes a dominant process. In massive, bright stars, the pressure of the radiation flow is much greater than it is for

the Sun. For massive stars, the radiation pressure in the outer parts of the star can be so great that matter is actually blown off the surface of the star in appreciable quantities. This is thought to be the mechanism behind the large stellar winds from massive stars.

Because of the very strong stellar winds, massive stars can lose a large part of their mass while they slowly burn hydrogen on the main sequence. After a massive star leaves the main sequence, the lifetime gets shorter, but the rate of loss of mass in a wind is much higher. The result is that appreciable mass can be lost in the red-giant phase, even if relatively little has been shed on the main sequence. Large mass loss can affect the evolution of the star. If the wind is strong enough, the entire hydrogen envelope can be expelled, thus exposing the core of helium and heavier elements.

Stars with less than about 30 solar masses can lose enough mass in a wind that they end up with substantially less mass than they had when they were born. This does not affect the qualitative behavior of the star, but it can alter details of the evolution. Stars of this relatively low mass do not have sufficiently strong winds to expose the core. In some cases, however, a binary stellar companion can tug the outer mass off and still produce a bare core with little or no hydrogen blanket. This and other effects of binary companions will be discussed in Chapter 3. Stars with mass between about 30 and 50 solar masses do become red giants but then are thought to undergo such an appreciable loss of mass to a stellar wind that the red-giant envelope is ejected anyway, exposing the core. For stars in excess of about 50 solar masses, there is no observed red-giant phase. The interpretation is that so much mass is lost on the main sequence due to a strong stellar wind that no outer hydrogen envelope is left to expand and become a red giant.

If the entire hydrogen envelope is lost to a wind, the bare core composed of helium and heavier elements should be exposed to view. We observe stars with just these properties. The *Wolf–Rayet stars* have little or no hydrogen on their surfaces and are seen to have strong winds themselves. Wolf–Rayet stars are thus thought to be the result of strong mass loss by winds from massive stars. This means that massive stars may not be red giants when they undergo core collapse but rather Wolf–Rayet stars. Whether Wolf–Rayet stars explode as supernovae or collapse to make black holes or some mix of both is not known.

As already noted, radiation pressure exerted by a star is proportional to its luminosity. There is a critical luminosity above which the outward radiation pressure exceeds the inward pull of gravity. In

this case, the result is not just a wind but rather a complete disruption of the balance of pressure and gravity in the star. This limit to the luminosity is called the *Eddington limit*, after the early British astrophysicist, Sir Arthur Eddington, who first realized the key role radiation could play in stars. The critical Eddington-limit luminosity is proportional to the mass of the gravitating star; it is the gravity of that mass that the radiation pressure must overcome. A star of fifty times the mass of the Sun is so bright that it is near the Eddington limit. That is why it blows such a substantial wind.

In Chapters 5, 8, and 10, we will also talk about circumstances when matter is dropped onto a compact, high-gravity star, like a white dwarf, a neutron star, or a black hole. Radiation pressure can also play a crucial role in these circumstances. If matter falls onto a star of high gravity, a great deal of heat and luminosity are generated. The resulting luminosity can exceed the Eddington limit, and the associated radiation pressure can actually prevent matter from falling onto the star at any higher rate. If the rate were higher, the excess matter would be blown away rather than falling on the star. The rate of infall of mass that just provides the Eddington luminosity is known as the *Eddington mass accretion rate*. In principle, a balance can be achieved in which the radiation pressure allows only enough mass to fall onto a compact star to generate the Eddington-limit luminosity that provides the pressure. A star in such a balance will automatically radiate precisely the Eddington-limit luminosity and the mass infall onto it will be precisely the Eddington mass accretion rate.

2.3 QUANTUM DEREGULATION

Let us now return to what happens in the guts of a star as it evolves. Section 2.1 described thermonuclear burning in conditions where the thermal pressure dominated over the quantum pressure. In this situation, the star can regulate its burning because the star will heat up when it loses energy and cool off if it gains energy. The process of contracting and heating and passing from burning phase to burning phase is halted if the core of the star gets too dense. At high density, the electrons are squeezed together so much that the exclusion and uncertainty principles come into play as described in Chapter 1. In this circumstance, the quantum pressure of the electrons exceeds the thermal pressure of the electrons and atomic nuclei. This happens first for lower-mass stars that are denser than high-mass stars at a given burning phase.

In this compact state governed by the quantum pressure, the star loses the ability to heat and ignite a new, heavier nuclear fuel. Any nuclear fuel that does burn under these conditions is not regulated. The star loses the ability to control its burning and its temperature. The quantum pressure deregulates the temperature; the thermostat of the star is broken.

The reason for this quantum deregulation is that the quantum pressure does not depend on temperature. If the star, supported by the quantum pressure, loses a net amount of energy because the nuclear fires have gone out, the pressure remains unchanged. There is no contraction to provide heat, so the temperature just drops as the heat is lost. A star, or portion of a star supported by the quantum pressure, behaves as you would think normal matter should: when it radiates away heat, it cools off, as illustrated in Figure 1.3. If a nuclear fuel ignites in a star supported by quantum pressure, the burning adds some heat. The pressure does not rise, so there is no expansion to absorb the heat. The temperature simply rises. The nuclear burning is very sensitive to the temperature, however. Thus at the new higher temperature, the burning proceeds even faster, raising the temperature even more. The nuclear rates can become so fast that the energy they produce can blow the star to smithereens. A star supported by the quantum pressure has an unstable, unregulated temperature. The temperature will decline toward absolute zero if there is no nuclear burning. The temperature will rise sharply if there is nuclear burning. The star has a broken thermostat. Even more, it is as if, when your house gets a little hot, you set the rafters on fire. The way in which the quantum deregulation sets stellar rafters aflame is given in Chapter 6.

Most stars reach this state of unregulated temperature and burning after helium has burned out in the core. The core is then composed of a mixture of carbon and oxygen. The core typically has a mass about 60 percent of the mass of the Sun, independent of the total mass of the star. This applies to all stars with mass up to about ten times the mass of the Sun, and that is most of the stars. The remaining mass is in the extended red-giant envelope. While the envelope is as big as the Earth's orbit, the core is very tiny by the time the quantum pressure becomes dominant – a few thousand kilometers in diameter, about the size of the Earth. The resultant density can be a million to a billion grams per cubic centimeter. Ordinary earthly matter, or that in normal stars, is about one gram per cubic centimeter. To get such high densities that the quantum pressure comes into play, a whole building, such as the seventeen-floor physics

building in which I work, would have to be packed into the volume of a sugar cube. Only gigantic gravitational forces can achieve such a compaction.

This small dense core is immediately surrounded by two narrow, very bright shells of matter where helium and hydrogen are burning. These shells are the last remnants of the stages of hydrogen and helium burning in the center of the core through which the star has already passed. The pressure of radiation from these burning shells causes the envelope to pulsate violently and blow matter from the star. The outer envelope is ejected in this process. Astronomers see the outcome of this process as a shell of gas proceeding outward from the star. These expanding, ejected shells are called *planetary nebulae*. They have nothing to do with planets except that they are often sufficiently extended in photographs that, like planets, they do not have a ''star-like'' point image. Planetary nebulae were misnamed by early astronomers, but the name has stuck.

When the envelope is ejected, the core of the star is left behind. Supported by the quantum pressure of its squeezed electrons, the core cools off to become what is known as a *white dwarf*. When a white dwarf forms, it still has a great deal of heat and looks blue-white to an astronomer. The term ''dwarf'' comes from the low luminosity. The white dwarf has such a small surface area that the white dwarf is dim despite its high temperature. White dwarfs are also tiny in size and hence dwarf-like in that sense, even though, again, that is not the meaning astronomers have attached to the word. We will return to white dwarfs in Chapter 5.

2.4 CORE COLLAPSE

Massive stars continue to evolve, forming cores within cores of ever heavier elements until the innermost regions are turned into iron. Iron is a very special element in the Universe. It is almost composed of fourteen helium nuclei but is a little more complex because two of the protons have converted to neutrons, so iron has four more neutrons than protons. By the happenstance of the nature of the strong nuclear force among protons and neutrons, the fifty-six particles of an iron nucleus are more tightly bound together than in any other element (with the possible exception of a couple of exotic elements like rare isotopes of nickel, which cannot easily be formed in nature). Iron happens to be at the bottom of a nuclear ''valley'' toward which all other elements would like to fall, just as rocks roll down a

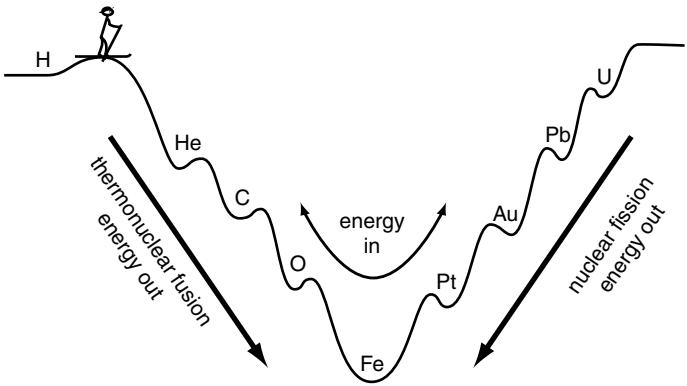


Figure 2.3 The element iron sits at the bottom of the nuclear "valley" defined by the nuclear and electromagnetic forces. Light elements, shown here schematically as hydrogen, helium, carbon, and oxygen, need to be heated to overcome the "bumps" representing charge repulsion, but then they can fuse into heavier elements, end up deeper in the valley, and thereby release a net amount of energy. Heavier elements, shown here schematically as platinum, gold, lead, and uranium, will liberate energy, slipping deeper into the valley, if they fission into lighter elements. Iron can be transmuted only by putting energy into it, either to break it apart into lighter elements or to fuse it into heavier elements. The result is that iron can only absorb energy from a star, never produce energy.

mountainside, as shown in Figure 2.3. The difference is that the force causing the settling toward the "bottom" is the nuclear force, not gravity. All elements lighter than iron would energetically prefer to merge together to form iron. They are prevented from doing so only by the repulsion of the electric charge on the protons. Stars are Nature's way of overcoming the electrical repulsion and rolling the elements down the nuclear hillside to the bottom where iron comfortably sits.

As rocks roll downhill, they turn their gravitational energy into other forms, such as noise, breaking trees, dislodging other rocks, and compacting and heating the soil where they land. This complex process conserves the total energy. When the rock is at the bottom of the valley, it can roll no farther, and no more energy can be obtained from it. A similar process occurs in forming iron. Energy is released as light elements fuse together to form heavier ones closer to iron. Elements heavier than iron are on the other side of the valley from the light

elements, but their protons and neutrons are also less tightly bound than those of iron. These elements approach iron by splitting apart into lighter elements in the process called *nuclear fission*. This process is the one that powers nuclear reactors, but it does not occur naturally in stars to any great extent because the stars are composed of elements lighter than iron.

Energy cannot be obtained from a rock at the bottom of a valley. On the contrary, to move the rock, energy must be invested to lift or roll the rock back up the hillside from which it originally fell. What about a stellar core made of iron? No more nuclear energy can be derived from that core. With no nuclear energy input, the star radiates a net amount of energy into space. The massive stars that develop iron cores are typically hot enough that thermal, not quantum, pressure dominates their structure. Thus when such stars lose energy, gravity squeezes them, and they heat up. Gravity naturally makes energy available to the iron. The response of the iron is to roll up the nuclear hillside. Most of it breaks apart into the lighter nuclei from which it originally formed. Some of the iron will undergo fusion reactions that lead to the heavier particles on the other side of the valley. Both of these processes require energy. Rather than firing up a new nuclear reaction to repel the squeeze of gravity, the iron absorbs heat energy from the star. The hot particles exerted the thermal pressure to support the star. When the particles lose energy to the breakup of iron, the pressure cannot rise. Gravity then compresses the iron core even more, but the iron continues to break apart, absorbing the energy and preventing a rise in pressure to withstand the stronger gravity.

The result is another example of energy conservation, with iron playing the negative role of a sponge rather than a source of energy. With iron absorbing energy, gravity overwhelms the weakened pressure. The formation of an iron core in a massive star signals the end of the thermonuclear life of the star. At that point, the star is doomed. Gravity prepares to deal the death blow. The core collapses in a mighty implosion!

2.5 TRANSFIGURATION

As the iron disintegrates into lighter elements in the collapse, the core plunges to a smaller size, and the density skyrockets. The rising quantum pressure of the electrons is too feeble. The electrons stop fighting the gravity and disappear. They do this by combining with a

convenient proton (a mutual suicide pact determined by the conservation of charge) and forming a neutron. To conserve lepton number, a neutrino must be produced for every electron that disappears, as discussed in Chapter 1. The result is that in the collapse of the iron core, the electrons and protons disappear to be replaced by neutrons and a flood of neutrinos. The result is the formation of an entirely new type of astronomical object, a *neutron star*.

A neutron star is composed almost entirely of neutrons. The mass of a typical neutron star is somewhat more than that of the Sun, and its radius is about 10 kilometers. This is only about the size of a small city like Austin, Texas. The density at the center of a neutron star exceeds that in the nucleus of an atom. In a sense, a neutron star is a gigantic atomic nucleus held together by gravity. A typical density would be about 10^{14} grams per cubic centimeter. To attain such a density, an entire city like Austin would have to be packed into the size of a sugar cube.

The gravity of a neutron star is fantastically large and must be balanced by an equally large pressure. At a large enough density, the quantum pressure of the neutrons can become sufficiently great to overcome the force of gravity and restore the condition of dynamic equilibrium. The quantum pressure of the neutrons is aided by the nuclear force. As described in Section 2.1, the nuclear force has no effect on particles that are a large distance apart; however, when they get quite near, the nuclear force pulls them together. The nuclear force is ''attractive,'' like gravity or opposite charges. An important detail mentioned in Chapter 1 comes into play when nuclear particles are packed very close together. At very small distances between particles, the nuclear force drives baryons apart. The nuclear force becomes ''repulsive,'' like similar charges. This repulsive force on closely packed neutrons helps to hold them apart and contributes to the pressure that supports a neutron star. As for white dwarfs, there is a maximum mass to neutron stars, a maximum mass that can be supported by the combined quantum and nuclear pressure of neutrons. The quantum effects are known precisely, but the nuclear force is not exactly established, so this pressure, and hence the total mass it can support, is still somewhat uncertain. The best guesses based on sophisticated calculations of nuclear matter are that the maximum mass for a neutron star is of order 1.5–2 solar masses.

The process of collapse and renewed support by the quantum pressure of the neutrons and the repulsive nuclear force among very compact neutrons is quite rapid. It requires only about a second in a

star that has lived for millions and millions of years in tranquillity. The details of this process will be explored in Chapters 6 and 7. A summary of what we have learned about neutron stars will be given in Chapter 8.

There is no guarantee that the process of core collapse will result in the formation of a neutron star. A tremendous amount of gravitational energy is released in the collapse, a hundred times more energy than is necessary to blow the outer layers away from the star. One reason that the nature of neutrinos was stressed in Chapter 1 is that they play a dominant role in core collapse. The majority of gravitational energy produced in the creation of a neutron star, more than 99 percent, is given to the neutrinos. The neutrinos escape from the collapsing iron core and the newly formed neutron star and carry most of the energy off into space.

The degree to which the remaining energy available from collapse may be directed outward rather than inward is not clear. If a fraction of the energy is used to blow off the layers of the star surrounding the original iron core, then a neutron star can be left behind. On the other hand, if insufficient energy is directed outward to eject the outer portions of the star, then the outer layers rain inward. A neutron star may form momentarily from the collapsed iron core, but then the rest of the star falls inward. Because we are talking about a process that occurs in massive stars, the mass that falls in will far exceed the maximum mass a neutron star can support. The neutron star will rapidly be crushed out of existence in a process of total, ultimate collapse. The result will be the unique gravitational entity that astrophysicists call a *black hole*. A black hole is an object for which all the mass has been crushed to what is effectively zero volume. All that remains is the gravitational field that becomes overwhelming at distances close to the center of the collapse. We will study the details of these fantastic objects in Chapters 9 and 10.

3

Dancing with stars: binary stellar evolution

3.1 MULTIPLE STARS

Cecelia Payne-Gaposhkin was a pioneer of modern astronomy. She devoted much of her research to the study of multiple star systems and coined a comic adage to describe one of the basic tenets of that work: ''Three out of every two stars are in a binary system.'' By this she meant to illustrate that roughly half the stars in the sky have companion stars in orbit. If you were to look closely at half the stars you would find that there are two stars, where a more casual examination would have revealed only one point of light. Many people know that the nearest star to the Sun is Alpha Centauri. Less well known is that Alpha Centauri has a companion in wide orbit, known as Proxima Centauri. A closer examination shows that Alpha Centauri itself is not a single star but has a closely orbiting companion as well. Of the ''two'' stars closest to the Sun, three are in the same mutually orbiting stellar system.

Stars occur in many combinations. Single stars and pairs are most common, but some systems contain four or five stars in mutual orbit. In this chapter, we will concentrate on the systems with a pair of stars, double stars, or, somewhat more technically, binary stars (but we try to refer to the phenomenon of *duplicity*, not the word ''binarity'' born of mangled jargon that has crept into the literature). Binary stars come in two basic classes: wide and close. Wide binaries are stars in large, long-period orbits. Such systems probably formed by the accidental gravitational capture of two stars born separately. These stars will evolve independently, as two separate single stars. That they are a gravitational pair will not concern us much here. Of greater interest, because of the effect on the evolution of the stars, are the close binaries. These systems probably formed by the fragmentation

of an initial single protostellar clump. Triple and quadruple systems probably formed in the same way. These close pairs are of particular significance because the presence of a nearby companion profoundly alters the course of stellar evolution.

3.2 STELLAR ORBITS

The force of gravity and the principles of conservation of linear and angular momentum govern the orbits of a pair of stars. Recall from Chapter 1 that linear momentum is the product of mass multiplied by velocity, whereas angular momentum, or spin, is the product of the mass, the velocity, and the size of the object under consideration.

Orbits of stars are very nearly ellipses. This is not exactly true if one considers the small effects of the complete theory of gravity as described by Einstein's general theory of relativity, but the assumption that orbits are ellipses is adequate for all our purposes now. We will mostly consider orbits that are the simplest special case of ellipses, namely circles. Two stars orbit one another on elliptical paths around a common *center of mass*. This center of mass can drift through space, but for simplicity we will pretend that there is no net motion of the two stars. Although the two stars share the same sense of the orbit, for instance clockwise, at any given moment, the individual stars move in opposite directions in their mutual orbital dance. They must do so to conserve the linear momentum, to keep the net momentum constant and equal to zero. If they moved in the same direction, the momentum would be first directed in one direction and later in another, in violation of the principle of conservation of momentum. Nature does not allow such behavior. The sizes of the orbits are different if the masses of the two stars are different. Again, to balance momentum, the smaller-mass star must move faster in the opposite direction to offset the momentum of the larger-mass star. The *period*, or the time for the stars to complete an orbit, must be the same for both. When the first star has traveled all around the second, the second cannot have traveled only part way around the first. If the smaller star moves faster but takes the same amount of time to complete an orbit as the more massive star, then the smaller star must cover more distance. The orbit of the smaller star must be larger.

Similar laws govern the orbits of the planets around the Sun. The planets move in relatively large orbits about the center of mass that lies between the planets and the Sun. At the same time, the Sun is not completely stationary but moves in a tiny orbit about the center of

mass. The size of the Sun's orbit is about the same as the physical size of the Sun itself. The Sun moves at about 30 miles per hour, a small but measurable speed. The presence of large planets around nearby stars was recently established with techniques to measure such speeds. The Sun's orbit is fairly complex in detail. Although the Sun mostly responds to Jupiter, the Sun is trying to orbit around the center of mass of nine planets at once.

Using the data on planetary motion carefully garnered by his mentor, the Danish astronomer Tycho Brahe, Johannes Kepler deduced empirically that planets move on ellipses (his first law) and that the period of the orbit is simply related to the size of the orbit (his third law). The angular momentum of the orbital motion of two stars depends on their mass and velocity, just as the linear momentum does. The angular momentum also depends on the size of the orbits. For this reason, the angular momentum helps to determine exactly how big the orbits will be for two stars of given mass and velocity. Kepler's second law of orbital motion comes about because the angular momentum of each star about the center of mass is constant.

With the help of Newton's law of gravity, we now interpret Kepler's third law as saying that the square of the period, P, of an orbit is proportional to the cube of the size, a, of the orbit divided by the total mass of the two orbiting stars. This law and the understanding of it are crucial in astronomy. The relation between period, orbital size, and mass provides the only reasonably direct way to measure the masses of stars. For two stars in a binary system, astronomers can measure the period fairly easily and the separation between the two stars with some difficulty. These two pieces of information and Kepler's third law as codified by Newton determine the total mass of the system. Astronomers must obtain other information to suggest how much of the mass is in each star. One of the reasons why the study of double-star systems is so important is that double stars provide direct information on the masses of stars.

3.3 Roche Lobes: The Cult Symbol

Before reading this section you must assume the posture and repeat the oath of secrecy. Curl your right arm over your head and place the fingers of your right hand on your left shoulder. Then curl your left arm so that the fingers of your left hand also touch your left shoulder. Now whisper loudly, ''I solemnly swear not to reveal what I am about

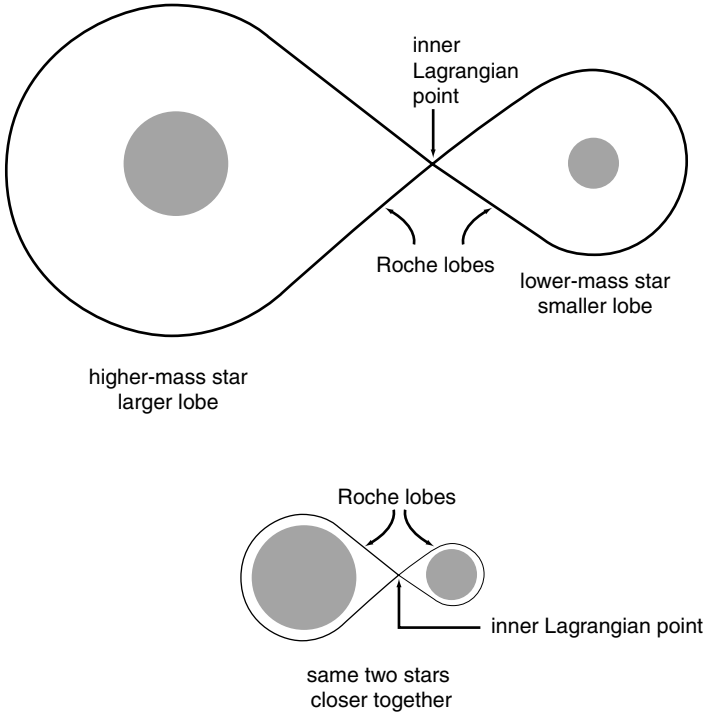


Figure 3.1 The upper diagram shows the Roche lobes, the regions of gravitational domain, around two orbiting stars. The lower diagram shows the same stars in closer orbit. Note that the Roche lobes are always roughly as large as the distance between the stars, but that the star with the larger mass always has the larger gravitational domain and hence the larger lobe. Both of the lobes are smaller if the stars are closer together.

to learn to anyone upon penalty of being ridiculed by my peers.'' As we proceed with this chapter you will find that the significance of the posture is that your brains were about to undergo mass transfer onto your shoulder.

For two stars in a binary system, each reaches out to gravitationally dominate some region beyond its own surface, as shown in Figure 3.1. The more massive star, the star on the left in Figure 3.1, has a larger sphere of influence. If one carefully maps the regions of influence of each star, accounting for the complexities of the fact that each star is moving in orbit, you find that the boundary of the two regions, seen in cross section, resembles a figure eight turned on its side. The two halves of the figure are called *Roche lobes* after the

German scientist who first worked out their mathematical form. The physical importance of these gravitational lobes is so great that no lecture on binary stars can continue without a sketch of the famous figure. For this reason one of our colleagues refers to this sketch as the ''cult symbol'' of the priesthood of the binary-star specialists.

The neck of the figure where the two lobes join is called the first or *inner Lagrangian point*, after the French mathematician Lagrange who also studied these systems. This point represents the position in space where the pull of gravity from the two stars just balances. A slight tip in either direction will send a bit of matter falling toward one star or the other. Beyond the surface of the Roche lobes, matter would belong to neither star but would be comfortable to orbit both of them. On a line extending out through the stars are the second and third Lagrangian points. Beyond these points, centrifugal forces overwhelm the combined pull of gravity of the two stars and tend to throw matter out of the system completely. At right angles to the line between the stars, one finds the fourth and fifth Lagrangian points. These are of little interest to us in the present context, but these Lagrangian points are potentially important in the subject of space colonization, as past members of the L5 Society will know (the fifth Lagrangian point was their cult symbol). The fourth and fifth points are locations at which a third body is locked in a stable position in the gravity of the two main objects. The idea is that this would be a good place to locate an artificial space colony between Earth and the Moon.

3.4 The First Stage of Binary Evolution: The Algol Paradox

One of the first lessons learned in the study of binary star systems is that the presence of a companion alters the course of evolution. Recall one of the most important aspects of the evolution of single stars. More massive stars have more fuel to burn, but they burn the fuel at a profligate rate. As a result, massive stars live a much shorter time than smaller-mass stars that hoard their meager allotment of hydrogen fuel. Given this most important lesson, how are we to understand the demon star Algol?

The star Algol presents a blue-white appearance to the eye. Algol also appears to be brighter and then dimmer every few days. When it is dimmer, it appears to be a little redder. In some early cultures a red, winking light in the sky did not bode well. Thus Algol acquired the name the demon star, ''Algol'' being the Arabic word for demon. We

now understand that Algol is a binary system. The red appearance comes because one of the stars is an evolved red giant. The winking derives from the fact that we happen to be looking almost edge-on to the orbits of the stars and hence witness the eclipses as each star in turn moves in front of the other. The slight reddening occurs because one of the stars is a red giant, and we see more of its light and less from that of the blue-white companion when the red giant is in front. We can go a step farther. Because one star has already evolved and has become a red giant, and the other star is still on the main sequence, we know which is the more massive. The red giant has evolved first so the red giant must be the more massive.

Wrong! From the measured period, some astronomical tricks, and Kepler's ever-handy third law, we can work out the masses and find that the red giant has a mass of about 0.5 solar mass, whereas the main sequence star has 2–3 solar masses. This is the *Algol paradox*. How can the evolved star be the less massive one?

To resolve the paradox, we hold firm to the idea that the red giant must originally have been the more massive in order for it to have evolved first. Our basic lessons are impeccable there. The key to resolving the paradox is that, unlike most single stars, close binary stars do not retain the mass with which they were born. When two stars are close together, as in the Algol system, one star can transfer mass to the other. The star that was the most massive became a red giant and then transferred mass to the other star until the mass ratio reversed completely: the originally more massive star became the less massive, and the originally less massive became the more massive.

3.5 MASS TRANSFER

To see how this process of mass transfer occurs, we must return to the meaning of the cult symbol, the Roche lobes. Even in a binary system, evolution begins on its normal course. Two stars in a close binary system are presumably born out of the same fragment of interstellar gas, and hence born at the same time. These are fraternal, not identical, twins, however. The chances of the stars being of identical mass are virtually nil. One star will be appreciably more massive than the other. The more massive star uses up its supply of hydrogen in the center and begins to evolve first. The core shrinks, the envelope expands, and the star begins to become a red giant. The more massive star has a greater gravitational domain and hence the larger Roche lobe. However, the size of the lobe is still finite – roughly the same

size as the distance between the stars, as you can see from Figure 3.1. As long as the stars are closer together than the eventual size the red giant would normally attain, the presence of the companion star interrupts the normal evolution. This interruption of the evolution is the basic criterion for whether a given binary system is categorized as a close binary system.

The story must change when the more massive star expands to the point where that star fills its Roche lobe. The internal forces of core contraction continue to cause the envelope to expand. As the outer parts of the star pass beyond the Roche lobe, however, they are beyond the gravitational influence of the star from which they came. When that happens, the matter that has moved out beyond the star's Roche lobe no longer belongs to that star. Some of the mass will take up a swirling orbit around both stars, but a great deal will find itself forced through the neck at the inner Lagrangian point joining the Roche lobes of the two stars. Matter that passes through the inner Lagrangian point now finds itself within the gravitational region of influence of the second star. The more massive star transfers matter through the inner Lagrangian point to the other star.

This mass-transfer process is unstable and results in rapid changes in the stars. To see this, recall the nature of the Roche lobes. The more massive star has the larger lobe. The star evolves, fills its lobe, and begins to lose mass. As the star loses mass, the star has a smaller region of influence, so its Roche lobe shrinks, as illustrated in Figure 3.2. Matter otherwise safely attached to the star finds itself cast adrift because the Roche lobe is smaller. That causes the loss of even more mass, resulting in an even smaller Roche lobe. A positive feedback operates in the sense that the more mass the star loses, the more it is forced to lose. The more massive star only approaches the condition of mass loss on the relatively slow timescale dictated by the contracting of the core. After the mass loss starts, however, it continues at a rapid pace, independent of any internal changes in the structure of the star.

This rapid phase continues until the stars have equal mass – the bigger one having lost mass, and the smaller one having gained it. Up to this point, the stars have been spiraling closer together as the star transferred mass. This is due, in large part, to the conservation of angular momentum. Mass is being added to the less massive star that moves with a higher velocity. Higher mass at a higher velocity would mean excess angular momentum. The stars correct this problem by moving together, since a smaller-size orbit has less angular momentum.

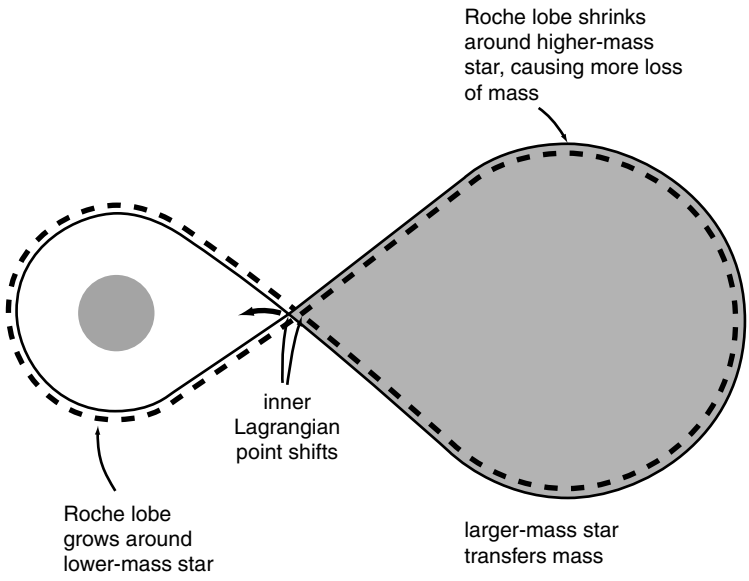


Figure 3.2 When the more massive star in a binary system loses mass, the process is unstable. As the more massive star loses mass, its Roche lobe becomes smaller, thus biting more deeply into the mass-losing star and causing even more mass loss. This effect is exacerbated because the requirement for angular momentum to be conserved also forces the stars to spiral closer together, making both Roche lobes smaller. As mass is transferred, the location of the inner Lagrangian point shifts to reflect the changing balance of the mass.

That the stars get closer together during the rapid phase of mass transfer only enhances the rate of transfer because the Roche lobes of both stars, particularly of the star losing mass, get smaller as the stars move together (Figure 3.1).

Although it does slow down, the mass transfer does not halt after the stars attain the condition of equal mass. Now conservation of angular momentum works to make the stars spiral apart. As the star continues to lose mass, it is now the smaller-mass, higher-velocity star. Angular momentum would decrease if the star did not move to a larger-size orbit as mass moved from the more quickly moving star to the slower. The tendency for the stars to move apart once the mass-losing star becomes the less massive means that, as the star loses mass, its Roche lobe gets bigger, not smaller. In order for the mass transfer to continue, the star must expand to fill its new larger Roche lobe. This expansion occurs, but only on the longer timescale of the

internal changes of the structure as the core contracts. The mass transfer no longer involves a positive feedback, and it is thus slower; but mass transfer will continue until the star ceases its attempt to become a red giant. The Algol system is presumably in this slow mass-transfer phase.

3.6 LARGE SEPARATION

When the two stars are of relatively large separation, but still close enough to qualify as a ''close'' binary, mass transfer does not begin until the more massive star has become nearly a full-fledged red giant. In this case, the mass-losing star will have a large envelope and a tiny core. The mass transfer continues until virtually the whole envelope vanishes and only the core remains.

If the original star was not too massive (less than about 8 solar masses), the core left behind will be dense and supported by the quantum pressure. It will just cool to become a white dwarf. The result will be a tiny white dwarf orbiting around a more massive main sequence star. The main-sequence star will have grown in mass because it is the repository of much of the envelope matter that originally shrouded the white dwarf.

A more massive star (one originally more than about 8 solar masses) can leave a larger core behind. Such a core will be supported by thermal pressure. It can continue to evolve without the envelope by contracting and heating until new nuclear fuels ignite in its center. The likely outcome for such a core will be to develop an inner iron core that is susceptible to the inevitable collapse. The situation is then similar to that for single stars. The collapse could create an explosion that would leave a neutron star behind. Alternatively, the collapse could be complete, resulting in the formation of a black hole. The result is that we could reasonably envisage the creation of binary systems with a normal star orbiting any of the types of compact stellar remnants we have discussed: white dwarfs, neutron stars, or black holes. We will discuss these cases in Chapters 5, 8 and 10.

3.7 SMALL SEPARATION

If the two stars are too close together, the stars evolve in a very different way. Stars swell a bit in size as they consume their hydrogen on the main sequence. This is because the helium that builds up in the center occupies less volume than the hydrogen did. When the helium

contracts, the gravitational energy transfers to the outer parts of the star, causing those parts to gain energy, expand, and cool slightly. The process is very similar to that which causes a star to become a red giant, but on a much smaller scale. If the stars are very close together, even this gentle swelling on the main sequence can cause the more massive star to fill its Roche lobe.

The twist comes after the rapid phase of transfer halts, when the two stars have equal masses. Ordinarily, the mass-losing star is a red giant and is evolving internally so rapidly that the mass-receiving star, which is still on the main sequence, is a totally passive partner. In the present case, however, we end up with both stars still on the main sequence. The mass-losing star is evolving slowly, continuing to push mass onto its companion. The evolution of the companion speeds up as it gains mass. Normally, the speed-up is insignificant, but for the case of close stars, the second star also swells to fill its Roche lobe. Each star then tries to transfer mass to the other simultaneously. The situation gets quite messy.

One thing that surely happens with both stars shoving mass beyond their Roche lobes is that material escapes to the region where it surrounds both stars. This matter will orbit in a disk that is in the same plane as the orbit of the two stars and that surrounds both stars. Matter flows outward into this disk, so such configurations have been dubbed *excretion disks* to distinguish this flow from *accretion disks*, where material settles inward. Accretion disks will be the topic of Chapter 4. The system probably ejects some material completely into the surrounding space.

Computer calculations show another interesting possibility. With both stars trying to move mass onto the other, only one can win. The calculations show that the star that had the smaller mass may win this contest, or lose it, depending on your point of view, in the sense that it transfers all its mass to the larger one. The big star consumes the little one! The net outcome is not some exotic binary, but a single star, perhaps surrounded by an excretion disk, the sole evidence of the cannibalism.

3.8 Evolution of the second star

In the standard picture where the star of initially smaller mass remains patiently on the main sequence until the other star completes its evolution, the second star eventually gets its turn. The second star will consume the hydrogen in its center, including

perhaps some of that added by the other star. Then the second star will begin to swell as its core contracts, and it, too, will eventually fill its Roche lobe.

At this point, the second star will begin to lose mass to the first. The second star does not particularly care what form its companion is in; it will just proceed to push mass over onto it. From an astronomer's point of view, the results can be quite exciting because the star receiving the mass is a white dwarf, a neutron star, or a black hole – a compact star with a large gravitational field. The effect can be quite spectacular. Astronomers have observed many systems where a star is transferring mass to a compact star. Some of these binary systems with compact stars may have evolved in the rather clean way described in the previous paragraph, with the second star simply swelling to fill its Roche lobe. In other cases, we will see that the actual evolution is probably more complex.

3.9 COMMON-ENVELOPE PHASE

The principal factor that can spoil the simple picture of one star filling its Roche lobe and transferring matter to the other star that passively accepts the mass is that the second star is unlikely to be a completely passive partner. The mass-gaining star can resist the process, as happened for two main-sequence stars very close together. The issue is, if neither star wants the mass, where does it go?

This issue arises more critically for stars that are more compact. For a star of a given mass, whether it is a main-sequence star, a white dwarf, or a neutron star, the strength of the gravitational pull depends only on the distance from the center of the star. The gravity does get stronger, the closer one gets to the center of the star. For this reason, the gravity at the surface of a white dwarf is much greater than the gravity at the surface of a normal star of the same mass, and the gravity at the surface of a neutron star of the same mass is greater even yet. The implication is that, if matter falls from a mass-transferring star at a given rate onto a normal star, the impact of the matter with the stellar surface will liberate energy and create luminosity at a certain rate. If the same star transfers mass to a white dwarf at the same rate, the energy liberated when the matter strikes the white-dwarf surface will be much greater, thus generating much more heat and a much larger luminosity. The case of a neutron star will be even more extreme. Although a black hole does not have a surface, matter can still respond to the effects of the strong gravity

very near the black hole. The result can again be the generation of a large luminosity.

The luminosity generated by the matter that falls in can serve to resist that very infall. The luminosity flooding outward can exert a pressure. In the extreme case, and this case arises in common circumstances for neutron stars, the luminosity can exceed the Eddington limit (described in Chapter 2). This means that the infalling matter is creating a luminosity so great that the resulting pressure is sufficient to prevent the infall! Even in less extreme circumstances, the energy of infall can inhibit the infall. Faced with this resistance, some of the matter will not collect on the mass-gaining star but will go in orbit around both stars.

When this process gets extreme, the matter lost from one star goes predominantly into orbit around both stars, interacts with itself, and bloats to become an approximately spherical (in the imagination of theorists, anyway) bag of gas in which the core of the mass-losing star and the mass-gaining star orbit. The resulting configuration is known as a *common envelope* because the envelope of matter surrounds both stars.

This situation can profoundly affect the orbits of the stars. Now they are not orbiting in the vacuum of space but in a bag of gas. The gas resists their motion, the stars feel friction and drag, and their motion heats the gas. The drag will tend to slow the forward velocity. In the ever-present grip of gravity, the result will be that the stars spiral toward one another and end up orbiting even faster. This will create more friction, heat, and drag and cause the orbits to shrink even faster. The energy and angular momentum lost from the stars goes into the common envelope at an ever-increasing rate.

The details of this process are not well understood, but the principle of conservation of energy gives insight into the general nature of the subsequent events. The gravitational energy from the decaying orbits eventually becomes equal to the gravitational energy that binds the common envelope to the two stars. At this point, the energy injected into the envelope by the motion of the stars will be sufficient to blow the envelope away. This process is not an explosion but something more like the ejection of a red-giant envelope to make a planetary nebula. The common envelope will be ejected and the two stars, the core of the mass-losing star and the mass-gaining star – whatever configuration they may be in – will again orbit in the vacuum of space, but now they will be very close together. Astronomers think this process produces pairs of white dwarfs, neutron

stars, and perhaps black holes, in addition to various combinations of these stars and normal stars. We will explore these combinations in Chapters 5 (see especially Section 5.3), 8, and 10.

3.10 GRAVITATIONAL RADIATION

Suppose two stars have survived as compact stars, white dwarfs, neutron stars, or black holes that have weathered mass transfer from first the originally more massive star, then the originally less massive star and any intermediate common-envelope phase. Now they are orbiting quietly in space. Is this the end of the story? The answer is no!

An important prediction of the general theory of relativity is that gravitational waves spread like ripples through curved space. If a wiggle occurs in the curvature of space, waves will propagate outward carrying off energy and momentum. Imagine an elastic rubber sheet on which you grab a pinch and shake it up and down, or the act of poking your finger in the surface of a still pond. Ripples will move outward across the sheet or pond. Ripples in space–time will propagate in the elastic curved space described by general relativity.

Two stars moving in orbit cause a rhythmic change in the curvature of the space around themselves as they circle. The effect is as if you were to twirl a small paddle on the surface of a pond. Ripples spread out across the pond, and gravitational waves spread out through space away from the orbiting stars. The waves carry energy and angular momentum away from the stellar orbits and cause the stars to spiral closer together in the grip of gravity. Eventually, they must collide in some way. In some very special, but important, cases, this loss of energy can determine the life and death of stars. We will discuss these issues further in Chapter 6.

4
Accretion disks: flat stars

4.1 THE THIRD OBJECT

One of the major developments of mid-twentieth-century stellar astrophysics was the understanding that there is often a third ''object'' in a binary star system, especially in a system undergoing mass transfer. Matter from one star swirls around the other forming a configuration known as an *accretion disk*. Such disks were first recognized in the study of white dwarfs in binary systems. With the advent of X-ray astronomy, it became particularly clear that accretion disks play a prominent role in binary systems containing neutron stars and black holes. In many circumstances, the accretion disk is the primary source of visible light; in others, the disk is also the primary source of X-ray radiation and, in yet others, the disk channels matter into streams of outgoing material and energy. One dramatic fact is that, without accretion disks, we would not yet have discovered any stellar-mass black holes.

One star in a binary system must undergo mass transfer to feed the disk with the matter needed for the disk to exist at all. The disk forms around the star receiving the transferred mass. An accretion disk thus also depends on a more ordinary star (considering black holes to be ''ordinary'' in this context!) for the gravity to hold the disk together. Given this support from the two stars in the binary system – one to provide matter, one to provide gravity – the accretion disk then effectively has a life of its own. The accretion disk has a structure and evolution that depends only incidentally on the properties of the star at its center or the one providing it mass. The disk is almost like a separate star, a flat star. The disk generates its own heat and light and can have eruptions that have nothing directly to do with either of the stars.

4.2 HOW A DISK FORMS

In common situations, the matter that feeds the disk flows from the companion star through the inner Lagrangian point that connects the two Roche lobes in the binary system. The structure of the inner Lagrangian point makes it act like a nozzle. The matter thus leaves the mass-losing star in a rather thin stream in the orbital plane of the two stars. In reality, the matter may spray in a messier fashion, but most of the matter remains in the orbital plane. If the two stars were stationary, this matter would flow from one star directly along the line connecting the centers of the two stars and strike the mass-gaining star. In a binary star system, however, the stars are constantly moving in orbit, so the mass-gaining star is a moving target. The matter may leave the mass-losing star headed for the other star, but because the other star moves along in its orbit, the transferred matter cannot fall directly onto the mass-gaining star.

If the mass-gaining star is small in radius, and white dwarfs, neutron stars, and black holes all qualify in this regard, then when the mass flow first starts, the stream of matter will miss the mass-gaining star entirely, passing behind the star as the star moves along its orbit. The gravitational domain of the mass-gaining star captures the matter, however, so the stream circles around and collides with the incoming stream. As this process continues, the flow of self-interacting matter will form first a ring and then a disk. From that point on, the transferred matter will collide with the outer portions of the disk and become incorporated into the disk.

The process by which the self-colliding stream of matter becomes an accretion disk involves the angular momentum of the matter in a crucial way. When the stream of matter first circles around the mass-gaining star, it has a certain angular momentum with respect to the star it orbits. Conservation of angular momentum forces the matter to move in a circular path of a certain size. The size of this path depends on the motion of the two stars. If the matter just stayed in this path, it would form a ring, somewhat like the rings around Saturn. To form a filled-in accretion disk that extends all the way down to the surface of the star, the material must settle to ever-smaller orbits. Matter in a smaller orbit will have a higher velocity, but the net effect is still to have a smaller angular momentum. Only if the orbiting matter loses some of its angular momentum can the matter move inward and settle onto the central star. The angular momentum must be conserved in the whole binary system, but the

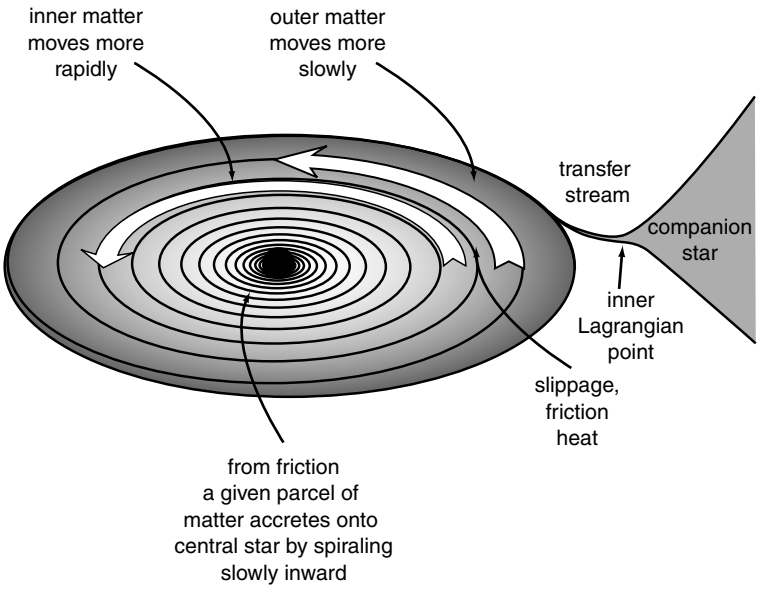


Figure 4.1 The orbiting of matter in an accretion disk naturally makes the matter that is farther from the central object move more slowly than matter that is nearer to the center. This creates a constant ''rubbing'' of the streams of matter. The rubbing results in friction and heating of the matter so that it radiates. The friction also causes the matter to slowly spiral down onto the central object.

matter in the disk must transfer some of its angular momentum elsewhere. Without this loss of angular momentum by the disk matter, the matter would stay in a ring. With a loss of angular momentum, the matter can settle inward, forming a full-fledged accretion disk.

One of the remarkable things about accretion disks is that they are structured in just such a way to provide for this transfer of angular momentum elsewhere, as illustrated in Figure 4.1. Kepler's third law tells us that because the matter in the disk that is closer to the central star must have a smaller orbit where the gravity is higher, matter in a smaller orbit must move faster. Thus each piece of material in the disk finds the material just beyond moving a little slower, and the material just within its orbit moving a little faster. The result is an inevitable rubbing of all the orbiting streams of material on all the adjacent streams. Each stream is slowed down by the slower, outer, adjacent stream and is thus forced to spiral inward. The result, ironically, is for the matter to end up moving faster because the material picks up

energy from the gravity of the central star. This process is fundamentally the same one that caused a star to heat up as it lost energy, as we discussed in Chapter 1. The effect of conservation of energy in the presence of gravity is to gain speed (or temperature) when some energy is taken away from the gravitating matter. The result of the rubbing and slipping inward is that the matter gradually settles onto the surface of the star. This process of gradual addition is known by the general term *accretion*, and hence the resulting flat structures are known as accretion disks. The angular momentum that is lost from the disk is gained by the orbiting stars or perhaps blown from the system by winds. The total angular momentum is, in any case, conserved.

4.3 LET THERE BE LIGHT – AND X-RAYS

The other important aspect of the inescapable friction that causes the matter to spiral in and accrete on the star is that friction heats the matter in the process. The heat escapes as radiation that astronomers can study. Because the orbital velocities are lower in the outer portions of the disk, the amount of slipping, friction, and heat are relatively low. The outer portions of the disk are typically about as hot as the surface of the Sun and emit much of their energy in the optical portion of the spectrum. In the middle of the disk, the velocities are higher, the friction and heat are greater, and the energy characteristically emerges in the ultraviolet portion of the spectrum. This is the end of the story if the mass-gaining star is a white dwarf, because the matter spiraling inward in the accretion disk collides with the white dwarf before the matter gains substantially more energy. For neutron stars and black holes, however, conditions can get even more extreme. The velocity of the spiraling matter can approach the speed of light. The frictional heating is immense. The matter gets so hot that the radiation emerges as X-rays, as shown in Figure 4.2. This is one reason that the search for neutron stars and black holes in binary stars requires X-ray instrumentation. Those instruments work best on satellites above the absorbing atmosphere of the Earth, so the astronomy of neutron stars and black holes has been primarily one of the space age. We will tell this story in Chapters 8 and 10.

4.4 A SOURCE OF FRICTION

The study of these flat stars called accretion disks has been a major undertaking in astronomy over the last three decades. The

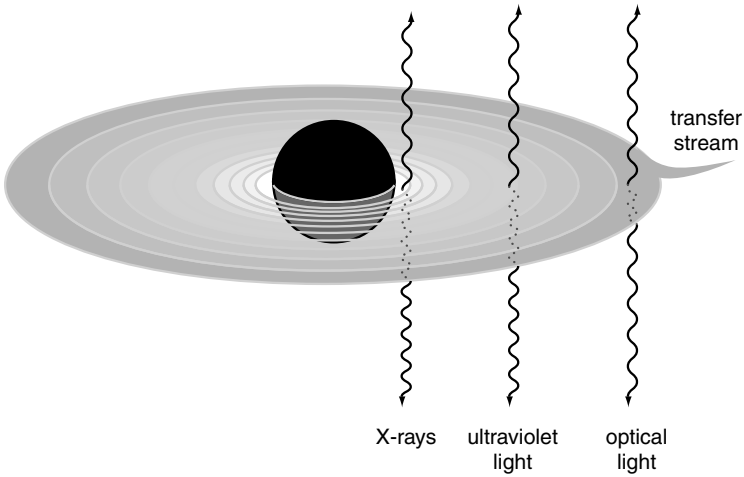


Figure 4.2 Because the orbital velocity of the matter in an accretion disk increases inward, the resulting friction and heat increase, and the resulting temperature of the orbiting matter rises. The outer parts of an accretion disk typically radiate in the optical and the middle parts in the ultraviolet; the innermost parts, if they exist, radiate X-rays.

understanding of accretion disks is still in a somewhat crude state. The situation is analogous to the early days of stellar evolution when there was an understanding of the balance between pressure and gravity, but the power source of stars was not known. The problem was that nuclear physics had not been invented. For accretion disks, the physics that determines the heating of the disk is known in principle, but its application is very complex in practice. The net effect is much the same. The drawback for accretion disk theory is that we do not know the nature of the friction, and so the mechanism to generate heat in the disk remains an important unknown.

We know that the normal microscopic rubbing of molecules in a gas is vastly insufficient to provide the friction and heat in observed accretion disks. Rather, the friction must come from large-scale roiling in the disk. Work of the last few years has provided evidence that magnetic fields must play a role in this process to generate the turbulent roiling motions and to couple one eddy in the complicated flow to another to make the interaction and the friction effective.

One compelling theory advanced by Steve Balbus and John Hawley of the University of Virginia is that any magnetic field in the disk becomes naturally and unavoidably stretched, twisted, and amplified in the orbiting matter in the disk. A simple analog of the

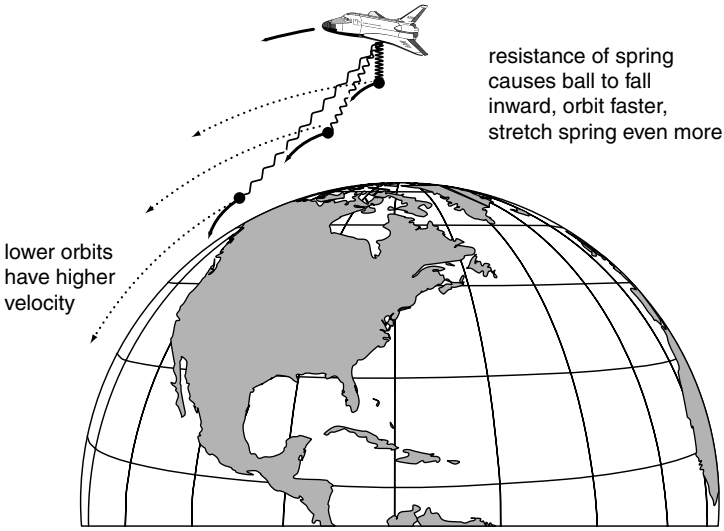


Figure 4.3 A satellite in a lower orbit than the space shuttle would orbit more rapidly. If the satellite is coupled to the shuttle by a spring, the spring will add some drag, causing the satellite to settle inward and to orbit even faster. The process will run away until the satellite burns up or the spring breaks.

process is to imagine a satellite connected to the space shuttle by a stretchy spring, as shown in Figure 4.3. If the satellite travels in a slightly lower orbit, the satellite will move faster than the shuttle. This will increase the tension in the spring and result in a ''drag'' on the satellite. Normally, if there is drag on a moving object, it will slow down. In the case of an orbiting satellite, however, the drag of the spring that slows the satellite leaves it with too little speed to maintain its orbit. The satellite must settle into a lower orbit where gravity is stronger and things orbit with even higher velocity. The net effect of the drag by the spring is to make the satellite settle into a lower orbit, closer to the Earth, where it moves even faster! This is yet another example of the working of conservation of energy (and angular momentum) when gravity is present. When a gravitating system loses energy, it heats up (like a star) or moves faster (like the satellite). When the satellite settles inward, it gains an even larger relative velocity with respect to the shuttle. The satellite will thus move even farther from the shuttle, increasing the tension in the spring and increasing the drag even more. The process clearly runs away, until the satellite burns up or crashes into the Earth or the

spring breaks. In accretion disks, the shuttle and the satellite are represented by two blobs of matter in different orbits, and the spring depicts a line of magnetic force connecting them, as illustrated in Figure 4.4. Any attempt to connect the blobs by means of the magnetic field will cause them to orbit even farther apart and increase the tension in the magnetic field until it snaps. The snapping magnetic field can put energy into the roiling matter and drive the turbulent motions that make the friction and heat. This general process is called the *magneto-rotational instability*. We will see it again in Chapter 6 on supernovae.

This magnetic coupling process must exist in accretion disks and play a role in their friction. It may not be the whole story because this theory does not seem to account for the full variability of the friction deduced from observations of accretion disks. Other theories propose that dynamos that generate magnetic fields spontaneously arise in the disk. Energy from the orbiting stars powers the dynamos. Eventual understanding will probably combine both of these ideas and more.

4.5 A LIFE OF ITS OWN

One of the most compelling pieces of evidence that an accretion disk can have its own behavior is when a disk flares with increased brightness. In most systems, the matter flows from the companion star so rapidly that the accretion disk is kept hot and ionized, and the disk radiates steadily. In other systems, however, the flow of matter being transferred is not sufficient to keep the disk in the hot, bright state, and the disk flares only occasionally. Astronomers observe this behavior in disks around white dwarfs, neutron stars, and black holes. There may be a variety of phenomena involved in this flaring, but there is one process that certainly happens in common circumstances. Under certain conditions, the flow of matter in the disk cannot be steady. Rather, the matter stores and then flushes from the disk. The flushing stage is especially bright and causes the flare of radiation. This process is rather independent of the two stars that feed the disk and hold it together. The timing of the flare events and their specific observational features do depend on the central star. If the central star is a white dwarf, astronomers call the flaring a *dwarf nova* (Chapter 5). If the central star is a neutron star or black hole, the flushing of the disk results in an *X-ray transient* (Chapters 8 and 10).

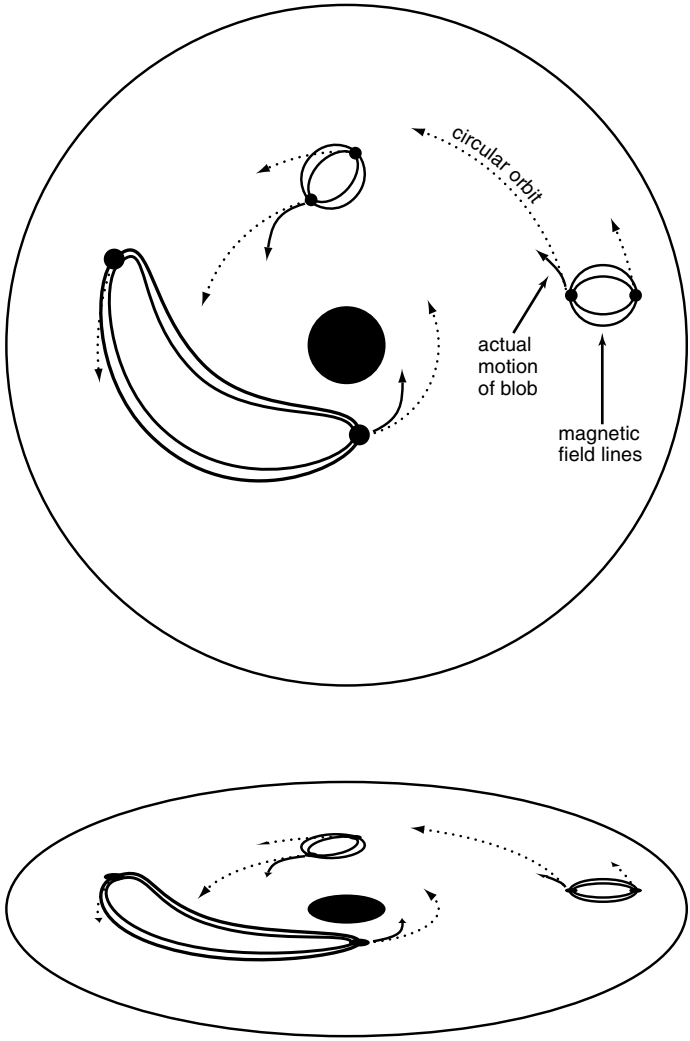


Figure 4.4 Separate blobs of matter orbiting in an accretion disk and linked by magnetic fields behave in a manner that is analogous to the shuttle, satellite, and spring combination shown in Figure 4.3, with the magnetic field playing the role of the spring. The pull of the magnetic field on the inner, more rapidly orbiting blob, will make it settle inward, stretching the magnetic field and causing even more drag on the inner blob and more settling. The stretched magnetic field will eventually ''snap,'' and the energy released will cause the matter to roil, to heat, and to radiate.

The theory behind this behavior is that the generation of the friction and heating in the disk depends on the temperature in the disk. When the disk is at a low temperature, less than that at the surface of the Sun, the matter in the disk is rather transparent. Any heat generated by the low friction can easily escape as radiation, thus maintaining the low-temperature state. In this low-friction state, there is little tendency for the matter to settle inward, but new matter flows from the companion star. The addition of matter increases the density of the material in the disk. As the density increases, however, the matter becomes more opaque, radiation cannot escape so easily, and the temperature must rise. This leads to a runaway process. The reason is that, as the matter heats, it becomes even more opaque to radiation. This traps more heat, leading to a greater opacity and an even greater trapping of the heat.

The result is that the disk can exist in a cool, barely accreting state, with low luminosity, until enough density accumulates to trigger this runaway heating. The beginning of such an outburst is illustrated in the top two panels of Figure 4.5. A wave of heating runs through the disk. The wave can begin on the outside of the disk, as shown in the second panel of Figure 4.5, or deeper down in the disk, depending on circumstances. The disk suddenly becomes very hot and very bright. The disk reaches maximum brightness when the heating fully envelopes the disk, as shown in the middle panel of Figure 4.5. The friction increases dramatically in the hot state, and so material that had accumulated in the outer parts of the disk rapidly moves inward. Ironically, this motion of the matter in the disk shuts the process off. As the outer portions of the disk thin out, they become more transparent again. They can radiate more easily, lose their heat, and lower the temperature. Now the inverse process sets in. As the temperature drops, the material becomes less opaque and more transparent, and this leads to a greater loss of heat, lower temperature, more transparency, and even greater loss of heat. A wave of cooling sets in from the outer parts of the disk that thin out first. This is illustrated in the fourth panel of Figure 4.5. The cool front sweeps inward, causing the majority of the matter in the disk to settle back into the cool storage state, as shown in the last panel of Figure 4.5. After an interval of storage, the cycle will then repeat.

The net effect is that the disk can exist in its cool storage state for a considerable time. The amount of time depends on circumstances, but the interval can vary from weeks to decades. The disk may be essentially undetectable during this phase. Then the eruption

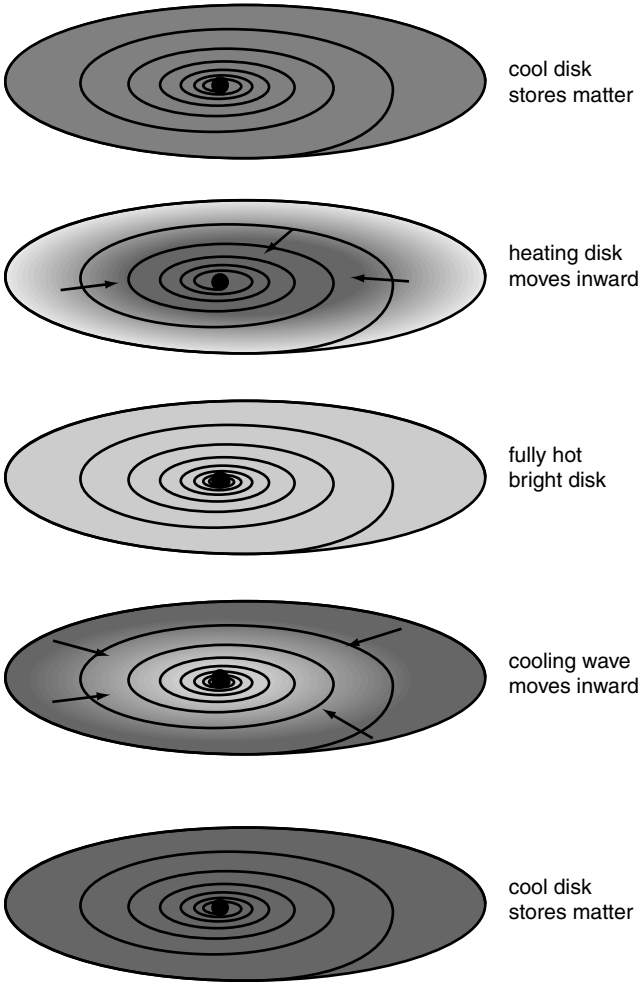


Figure 4.5 When mass is fed into an accretion disk at a rather slow rate, the disk goes through a cycle of cool, dim storage and hot, bright flushing of matter. (Top) Most of the time the disk matter is cool and rather transparent, so little of the matter added from the companion star flows through the disk. (Second) As matter accumulates, the density rises, and the disk turns more opaque, trapping the heat. This leads to a heating instability and a heating wave that propagates through the disk. (Middle) When the disk is fully heated, it is temporarily very hot and bright, the peak of the flare. (Fourth) As matter settles inward, the outer parts thin out, turn more transparent, and cool. A cooling wave moves inward through the disk. (Bottom) After the whole disk has cooled, the storage process begins again.

occurs, and the disk becomes very hot and bright for a short time, typically one-tenth the time the disk was dim, and is readily visible to astronomers. No sooner has the eruption occurred, however, than the disk starts to cool. Astronomers who want to study this transient bright phase must scramble!

An important aspect of this cycle of quiescence and eruption is that the process can be quite independent of the stars in the system. During the whole process, the mass-losing star can be pumping matter in at a perfectly constant rate. The star around which the accretion disk swirls provides a constant gravity. The flaring activity is a feature of the disk alone. In more complex systems, the mass can flow from the mass-losing star at a variable rate. The mass-gaining star can have a hard surface or strong magnetic field of its own (in the case of either neutron stars or white dwarfs). Either of these situations can lead to interesting variations.

4.6 FAT CENTERS? THE DAF ZOO

Another important idea has emerged in the last few years. The inner parts of accretion disks may not be so flat. Under certain circumstances, as the disk cools after its heating episode, the density can get so low that interactions among the particles are rare, and the efficiency of radiation can drop. This again leads to a retention of heat. The excess heat leads to pressure that causes the disk to swell up and become fatter, as shown in Figure 4.6. If this happens, this portion of the disk can become so hot that matter and antimatter, electrons and positrons, are created. The disk assumes a more nearly spherical configuration, and matter falls inward on the central star almost uniformly from all directions.

Under these circumstances, the matter can fall in so rapidly that the flow of matter carries the heat generated into the central star before the energy radiates away. This is especially true if the central star is a black hole. The heat energy disappears into the black hole just as the matter itself does. The technical term for carrying some property along with the flow is *advection*. In this case, the supposition is that a substantial part of the energy generated in the flow is advected into the black hole, rather than being radiated away as would be case with a thin accretion disk. The resulting structure has been termed an *advection-dominated accretion flow*, or *ADAF*, to discriminate this structure from the *disk-dominated accretion flow* (which could be called *DDAF*, but is usually just called the disk) that was the subject of the bulk of

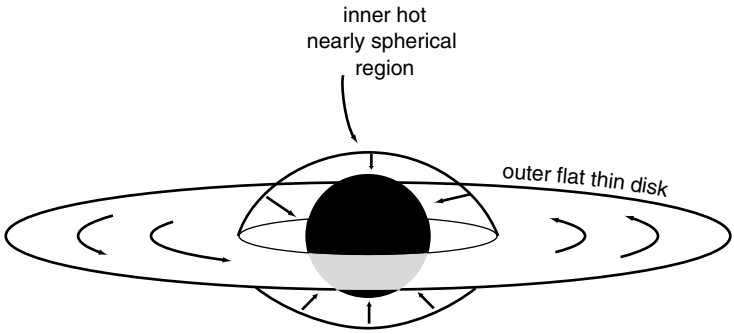


Figure 4.6 The inner portions of accretion disks, especially those surrounding black holes, can retain their heat and swell to become a fat, nearly spherical region. In the outer, thin disk, the matter orbits in a single plane, but in the inner, fat portion, the matter can flow nearly radially inward, can circulate in turbulent convection, or can be blown out in a wind. Radial inflow can sweep heat down into the black hole before it can be radiated away, so the inner fat regions can be relatively dim.

this chapter. The ADAF model for the physics of this region was popularized by Ramesh Narayan at Harvard and Insu Yi, one of the brightest young people to get his Ph.D. from our department in Austin, but who has apparently disappeared into the world of import/export commerce.

According to ADAF models, the result of advecting heat energy down the black hole rather than radiating it away is that this fat, inner portion of the accretion flow is especially dim. What little energy leaks out corresponds to especially high energy radiation – high-energy X-rays and gamma rays. There is some evidence that such regions do form in the centers of unstable accretion disks as they settle back into their storage state (Section 4.5) and that they may form around supermassive black holes in the centers of galaxies. One of the outstanding issues, the subject of current research, is when, why, and how a disk makes the transition from the relatively cool, flat configuration of a standard accretion disk to the very hot, fat configuration. Understanding this transition may give new clues for how to find and study black holes.

Picking up on this theme, other researchers have argued that the fat inner ball will not just sit there, slowly sinking inward. Marek Abramowicz of Sweden's University of Göteburg and his colleagues have argued that this inner structure must be roiled by turbulent

boiling or convection. The resulting structure will have some of the same, low luminosity, fat geometry properties of the ADAF model, but also some important conceptual, quantitative, and observable differences. This alternative structure has been called a *convection-dominated accretion flow*, or *CDAF*. Roger Blandford of Stanford and Mitch Begelman of the University of Colorado are convinced that any such structure must blow a wind from the surface (Chapter 2, Section 2.2). Thinking of the salute this outflowing matter might give, Blandford and Begelman named their model (with tongue more than slightly in cheek) *advection-dominated inflow–outflow solutions*, or *ADIOS*. More recently David Meier of the Jet Propulsion Laboratory has invoked the notion that, with the magneto-rotational instability and other dynamo effects, magnetic fields will be an important and generic part of the problem. Meier draws on the power of twisted magnetic fields to drive not just generic outflow, but jets from black holes to describe a general *magnetically-dominated accretion flow*, or *MDAF*. The true structure of the inner parts of accretion disks around black holes probably involves aspects of all these ideas. Once again, understanding of the nature of the accretion flow near black holes may help us understand the existence and nature of black holes. We will return to these topics in Chapter 10.

5
White dwarfs: quantum dots

5.1 SINGLE WHITE DWARFS

White dwarfs are certainly the most common stellar ''corpses'' in the Galaxy. There may be more white dwarfs than all the other stars combined. The reason is that low-mass stars are born more frequently, and low-mass stars create white dwarfs. In addition, after a white dwarf forms, it sticks around, slowly cooling off, supported by the quantum pressure of its electrons. This means that the vast majority of the white dwarfs ever created in the Galaxy are still there. The exceptions are a few that explode or collapse because of the presence of a binary companion. There are probably ten billion and maybe a hundred billion white dwarfs in the Galaxy. Most white dwarfs have a mass very nearly 0.6 times the mass of the Sun. A few have smaller mass, and a few have larger mass. Exactly why the distribution of the masses is this way is not totally understood.

White dwarfs provide clues to the evolution of the stars that gave them birth. To fully reveal the story, astronomers need to probe the insides of the white dwarf. Ed Nather and Don Winget at the University of Texas invented a very effective technique to do this. The technique uses the seismology of the white dwarfs to reveal their interior structure, just as geologists use earthquakes to probe the inner Earth. Under special circumstances, depending on their temperature, white dwarfs naturally oscillate in response to the flow of radiation from their insides. The oscillations cause small variations in the light output. To do white-dwarf seismology, careful observations must be made over extended times, days to weeks. The problem is that the Sun rises every day, and that makes observations difficult. Nather and Winget thus invented the ''Whole Earth Telescope,'' in which a network of small telescopes in various sites around the world

is coordinated by telephone and the World Wide Web. The trick is that as the Sun rises and the target white dwarf sets in one part of the world, the Sun is setting on the opposite side of the world, and the target white dwarf is rising. With careful planning, the white dwarf can be observed constantly from somewhere on the globe for weeks at a time.

The results have been striking. The Whole Earth Telescope has measured the masses of some white dwarfs with exquisite accuracy. The team has measured the rotation of some of the stars and probed the inner layers of carbon and oxygen. The outer layers, thin shells of hydrogen and helium, have provided clues to the birth of the white dwarfs. By these techniques and others, measurements of the ages of some of the white dwarfs are possible.

Measuring the ages of the white dwarfs is especially interesting because the ages reveal the history of the Galaxy. Because essentially all the white dwarfs ever born are still around, they can tell the story of when the first white dwarfs formed when the Galaxy itself was young. The white dwarfs cool steadily, but they cool slowly. The oldest, coolest white dwarfs are dim and difficult, but not impossible, to see. Studies of the oldest white dwarfs reveal that the first formed about 10 billion years ago. The Galaxy itself presumably formed only a few billion years before that. This argument leads to the conclusion that the Galaxy is relatively young compared to some estimates. The exact age of the Galaxy remains uncertain, but estimating its age with white dwarfs is now an established method.

5.2 CATACLYSMIC VARIABLES

A significant number of the white dwarfs in the Galaxy are not alone, but in binary systems. These white dwarfs are especially interesting in the context of this book because they share properties with more exotic objects like neutron stars and black holes in binary systems. Most of the white-dwarf binaries are the result of the first stage of mass transfer, when the originally most massive star forms a white-dwarf core and transfers the remainder of its mass to a stellar companion. In some cases, both stars have undergone mass transfer, leaving two white dwarfs in orbit.

Some of the most common and interesting examples of the second stage of mass transfer are the *cataclysmic variables*. These variable ''stars'' are all binary systems in which mass flows from one star, first into an accretion disk and then onto a white dwarf. The basic

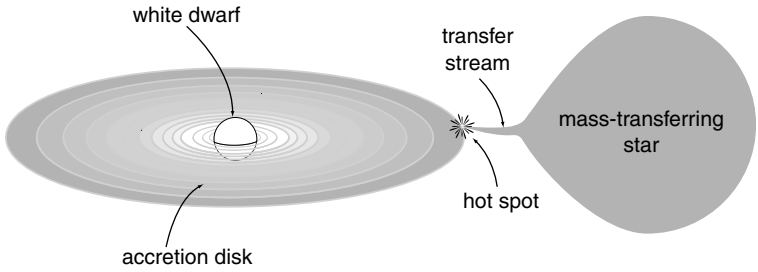


Figure 5.1 Schematic illustration of a cataclysmic variable. The basic components are a star that fills its Roche lobe and transfers mass through an accretion stream, the bright hot spot where the stream strikes the outer rim of the accretion disk, an accretion disk, and a central white dwarf.

components of a cataclysmic-variable system are illustrated in Figure 5.1. The star losing the mass is often a small main sequence star that sometimes has less mass than the companion white dwarf. Emitted radiation tracks the stream of material passing through the inner Lagrangian point and merging with the disk. Most of the light from a cataclysmic variable comes from neither the white dwarf nor the mass-losing star but from the so-called *hot spot* where the transfer stream collides with the outer edge of the accretion disk. This collision is very energetic and so produces a great deal of heat and light. Some light also comes from the friction and heating in the inner reaches of the accretion disk itself, as described in Chapter 4.

Several types of cataclysmic variables exist. The types are differentiated by their specific observational properties and the mechanisms thought to cause their variability. Cataclysmic variables all fall under the general category of the novae, or new stars. This is because historically the brightest flares would cause a ‘‘new’’ star to appear where none had been seen before. The star system is not new, of course, merely below the threshold of detectability until the system flares. The phrase ‘‘supernova’’ is an offshoot. For a long time, all suddenly flaring events that caused a new star to appear were classified with the same general term, ‘‘nova.’’ With the discovery that some events were in distant galaxies, and hence intrinsically very much brighter, shining over great distances, the term ‘‘supernova’’ was applied. We now know that novae and supernovae involve very different phenomena, although they are not completely unrelated. Novae might eventually turn into supernovae, and some novae involve thermonuclear explosions.

Dwarf novae are the most gentle of the cataclysmic variables. Dwarf novae flare up irregularly to be about ten times brighter than they usually are. The flares occur with intervals of weeks to months and last for days to weeks at a time. This interval is too short to build up any reservoir of thermonuclear fuel. The energy involved comes from heating as material from the mass-losing star settles in the gravitational field of the white dwarf. There are two competing ideas of how the flare occurs. One is that the mass-losing star undergoes surges that throw over extra mass from time to time. The problem with this picture is that one would expect the hot spot to flare first, before the disk, but this is not observed. The alternative is that matter piles up in the accretion disk until some instability causes the matter to suddenly spiral down toward the white dwarf, leading to an increase in the frictional heating and the light output in the process.

Detailed studies suggest that the disk-heating instability described in Chapter 4 (see Figure 4.5) is the primary cause of dwarf novae. Matter piles up in the disk in a cool, dim storage phase until the disk becomes opaque and traps the heat. This very heating causes an increase in the opacity, yielding more heating, more friction, and yet more opacity. The result is a rapid transition of the disk to a hot bright state. When the central star is a white dwarf, the observed result is a dwarf-nova outburst. During the outburst, the extra luminosity will heat the surface of the companion star and may cause the companion to transfer more mass. Both suggested mechanisms may thus play some role in the dwarf-nova outburst mechanism.

Recurrent novae flare to become about a thousand times brighter than the conditions prior to the outburst. These flares occur every 10–100 years. The mechanism of the outburst is unknown. Although both kinds of systems involve mass transfer through an accretion disk onto a white dwarf, dwarf novae do not have recurrent nova outbursts, nor vice versa. The difference may follow from the rate of mass transfer. If the rate is fast enough, the disk will steadily channel all the mass to the white dwarf. The disk will not have the luxury of waiting until enough matter has collected to begin to drop the matter onto the white dwarf. A faster mass-transfer rate might explain why a recurrent nova does not undergo dwarf-nova outbursts, but that does not explain the nature of the recurrent nova outbursts.

Classical novae, or in casual terms, novae, flare from ten thousand to a hundred thousand times brighter than their normal state. None has ever been seen to recur. The suspicion that classical novae repeat at intervals of about 10 000 years has been around for decades. There

is, however, little direct evidence for that particular timescale, which is too long for the brief recorded history of astronomy. The established evidence, both observational and theoretical, is that the mechanism of the classical nova outburst is a thermonuclear explosion. The idea is that as matter flows from the companion star, the matter settles onto the white dwarf in a dense layer supported, as is the white dwarf, by the quantum pressure, as shown in Figure 5.2. The inner white dwarf is probably composed of carbon and oxygen that require extreme conditions to ignite and burn. The material collecting on the outside is hydrogen, which burns more easily. As the hydrogen collects, the density and temperature increase until the hydrogen ignites. Because the hydrogen is supported by the quantum pressure, the thermonuclear burning does not increase the pressure and hence cannot at first cause expansion and cooling. Rather, the burning is unregulated, and an explosion ensues. The explosion does not involve the whole star like a supernova, only the outer layers. Nevertheless, the result is spectacular, giving a great flare of light and blowing matter off the surface of the white dwarf at high velocities. If the current theories are correct, the white dwarf will then begin to accumulate more hydrogen from its obliging companion until the conditions are yet again ripe for an explosion.

5.3 THE ORIGIN OF CATACLYSMIC VARIABLES

"Careful readers" (to which class the author never belonged) may have noticed that they were sandbagged earlier in the first general description of cataclysmic variables. The sleeper was the comment that in most cataclysmic variables the star losing mass is a small main-sequence star. Let us think that through. If a small main-sequence star is losing mass, the star must be filling its Roche lobe. Because the star is not large, the lobe must be small, which means that the stars – the main-sequence star and the white dwarf – must be very close together, almost touching. How then did the white dwarf form in the first place? The separation must have been large so that the progenitor of the white dwarf could form a well-developed core–envelope structure and become a red giant before mass transfer began. If the stars were very close together originally, the big star would eat the little one (Chapter 3, Section 3.7). No cataclysmic-variable system could evolve. The conclusion is that the two stars must have been far apart initially, even though they are very close together now.

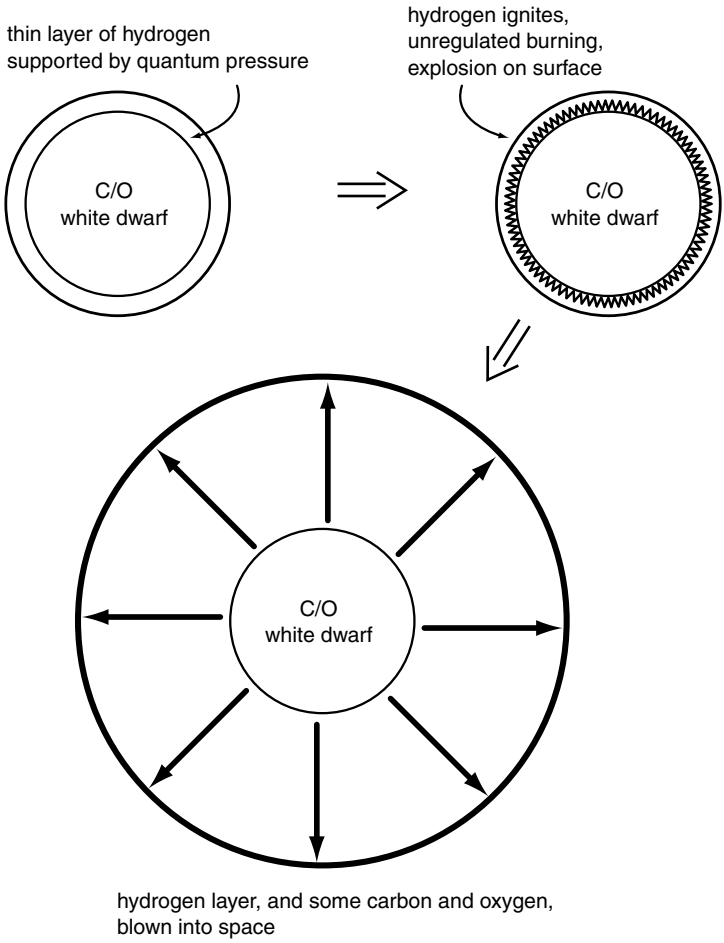


Figure 5.2 The mechanism of a classical nova explosion. (Top left) Hydrogen from the companion star passes through the accretion disk and accumulates in a thin, quantum-pressure-supported layer on the surface of a white dwarf, often composed of carbon and oxygen. (Top right) When the density and temperature in the hydrogen layer get large enough, the hydrogen will begin thermonuclear burning. Because the hydrogen is supported by the quantum pressure, there will at first be no mechanical response, the shell will just get hotter, and the burning will be unregulated. This will result in an explosion. (Bottom) The explosion will blow the hydrogen layer into space, along with some of the carbon and oxygen from the central white dwarf.

What is necessary to perform this bit of stellar legerdemain is to find a way to drag the stars together. The mechanism proposed to accomplish this is the *common envelope*, described in Chapter 3 (Section 3.9). We have discussed that matter can spill outward to orbit around both stars in a binary system. There is a strong suspicion that when a red giant goes into the first stage of rapid mass transfer, mass flows at such a rate that the second star is glutted. The matter falling on the star causes heat and extra radiation, and the pressure of that radiation will prevent the rapid flow of matter onto the star.

With a red giant pouring forth mass in copious amounts and the companion refusing to accept it, the matter will enshroud both stars, as shown in Figure 5.3. Unlike the case of an excretion disk where the matter orbits both stars in the orbital plane, this great amount of matter will form an approximately spherical red-giant-like envelope around both stars. Both the tiny white-dwarf core of the original red giant and the innocent main-sequence companion will orbit around inside this envelope. The result is a common-envelope or ''double-core'' system. The main-sequence star and the white dwarf are not orbiting in the vacuum of space now but in the frictional medium of their common gaseous shroud. The friction causes the two stars to spiral together.

The developments that follow then are particularly unclear, but speculation goes as follows. The white dwarf and the main-sequence star finally get very close together, so close that the Roche lobe of the main-sequence star gets smaller than the star. Notice that the star does not evolve and expand to fill the lobe; the lobe shrinks along with the orbit to fit the star. At this point (perhaps from the heat of theoretical astrophysicists waving their arms), a burst of energy blows away the common envelope. As the ejected matter floats away, a fully formed cataclysmic variable emerges, as illustrated in Figure 5.3. In this view, the system is ''born'' within the common envelope as a main-sequence star already filling its Roche lobe and transferring mass to a white dwarf. The beginning of the transfer of mass from the main-sequence star to the white dwarf may be the energy source that ejects the common envelope.

The simplest, cleanest, mass-transfer process to imagine is that the red-giant envelope flows from one star to the other and thus bares the white-dwarf core. The second star subsequently expands to fill its Roche lobe and transfers mass back to the white dwarf to form a cataclysmic variable. This simple picture is probably relatively rare in practice. Even though many details must yet be understood, the

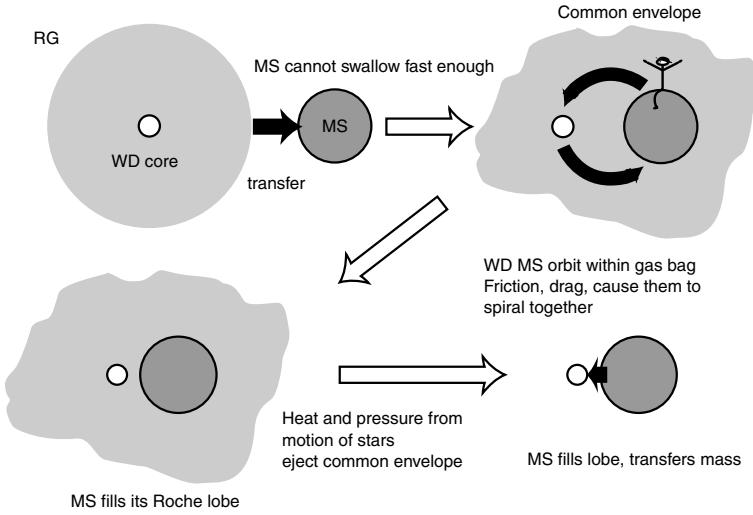


Figure 5.3 When a red giant (RG) in a binary system transfers mass to a main-sequence (MS) companion faster than the main-sequence star can assimilate that matter (upper left), a common envelope will form, engulfing both the white-dwarf (WD) core of the original red giant and the main-sequence star. Friction and drag will cause the white dwarf and main-sequence star to spiral together (upper right). As the two inner stars get very close together, the main-sequence star can nearly fill its decreasing Roche lobe and the heat and pressure of motion of the two stars in the bag of gas can expel the common envelope, much like the formation of a planetary nebula (lower left). The outcome can be a main-sequence star filling its Roche lobe and transferring mass to a white dwarf, a common form of cataclysmic variable (lower right).

formation of most cataclysmic variables probably involves the more complicated common-envelope process.

5.4 THE FINAL EVOLUTION OF CATACLYSMIC VARIABLES

The ultimate fate of cataclysmic variables is very uncertain. There are two general possibilities. These systems could just fizzle out. The ordinary star could eject its envelope and leave behind a second white dwarf so that mass transfer stopped. Alternatively, cataclysmic variables could end in a cataclysmic implosion or explosion. Even the fizzle could be interesting, involving some fascinating contortions. Let us examine the catastrophic possibilities first, then return to the fizzle.

In some observed cataclysmic variables, the mass of the white dwarf is within about 10 percent of the Chandrasekhar limit, and the mass is increasing steadily. This situation immediately invokes speculation concerning the outcome if the white dwarf reaches the limiting mass. One possibility is that the nuclear fuel of which the white dwarf is composed – for instance, carbon and oxygen – ignites. For a white dwarf near the Chandrasekhar mass limit, the density is very high. With these conditions, the quantum energy of the carbon nuclei can trigger nuclear reactions, even if the temperature and the thermal energy are at absolute zero. As we have described many times now, nuclear ignition under conditions where the star is supported by the quantum pressure is very unstable. Ignition of carbon under these conditions would lead to a violent explosion. This explosion would occur in a star devoid of hydrogen, save perhaps for a negligibly thin layer on the surface. Such a picture is the most probable origin of one kind of supernova, as we will explore in Chapter 6.

The white dwarf could possibly be made of iron that disintegrates upon compression, or, more likely, of oxygen, neon, and magnesium, elements that can absorb electrons rapidly. In these circumstances, when the Chandrasekhar limit is approached, the white dwarf may collapse rather than explode. This process will leave a neutron star in orbit around the main-sequence star. This collapse may result in the ejection of little or no mass. The energy of the collapse might come out almost entirely in the form of neutrinos, so that there would be little or no optical display. A process this violent, however, is likely to be bright as well.

All these potential catastrophes depend on the mass getting very close to the limiting value of the Chandrasekhar mass, within a percent or so. One interesting open question is whether the mass ever gets that high. Nova explosions certainly blow off matter that has accumulated on the surface of the white dwarf. If all the matter that has accumulated is ejected in the outburst, the mass of the white dwarf will not increase. The situation could be even worse. Nova explosions are observed to expel an excess of carbon and other heavy elements. This strongly suggests that a nova explosion expels not just the outer layer of accumulated hydrogen but also some of the guts of the white dwarf itself. This would mean that the mass of the white dwarf shrinks as a result of nova explosions. If this is the case, the white dwarf will remain in the binary system until the companion star evolves and forms a white dwarf of its own. Thus nova explosions

might lead to circumstances where the final fate of the cataclysmic variable is a fizzle rather than a catastrophe.

The cataclysmic variable might fizzle, but the story is not over just because the system produces two white dwarfs. We need to inquire about the ultimate fate of two orbiting white dwarfs. They can no longer evolve on their own. Supported by the quantum pressure, they will just cool off if left to their own devices. The white dwarfs do not remain unmolested, however. As they revolve about, their orbital motion generates gravitational waves. The gravitational waves carry off energy and angular momentum, and the orbit must shrink. As described in Chapter 3, gravitational radiation affects all stellar orbits. Gravitational radiation is a very small effect, so that any other normal interaction between the stars is more important. Only when the two white dwarfs reach a state of total quiescence can the small effect of gravitational radiation become important. This will inevitably happen to two white dwarfs, however, and they must spiral together. The outcome depends on the specific properties of the white dwarfs as stars supported by the quantum pressure.

For a normal star supported by thermal pressure, the addition of mass causes the star to attain a larger radius. Remove mass, and the star shrinks. For a white dwarf supported by the quantum pressure, the opposite situation holds. The addition of mass causes an overall compaction of the star. The star thus shrinks in radius as the mass increases. Removal of mass from a white dwarf allows the star to expand in the smaller gravity and attain a larger radius. This behavior has crucial implications for the ultimate fate of one of the white dwarfs.

The two white dwarfs will spiral together until the separation and hence the Roche lobes become small enough so that one of the white dwarfs fills its lobe. Which one will that be? The one with the smaller mass has a smaller Roche lobe but a larger radius. The smaller-mass white dwarf will fill its lobe first and begin to lose mass to the larger-mass dwarf. This is not good news for people rooting for the underdog! As the smaller-mass dwarf loses mass, its Roche lobe shrinks, but its radius gets even larger! The white dwarf will lose mass at an ever more rapid pace. The only outcome can be the disappearance of the small-mass dwarf. The larger-mass dwarf will simply gobble up the smaller-mass one. Some mass may slop out into space, but this will be little consolation to the disappearing dwarf.

The smaller-mass star may not disappear entirely. When the mass of the object gets down to the size of a planet – less than a

thousandth the mass of the Sun – its structure may rearrange. If the material becomes rock-like, like the Earth, then the remains of the little white dwarf may cease expanding. The result could be one white dwarf orbited by a desolate rocky chunk. Given sufficient time for gravitational radiation to act, even that chunk could spiral down to the surface of the remaining white dwarf and be consumed.

Alternatively, the process of disrupting the smaller-mass white dwarf may not end gently at all. As the larger-mass white dwarf consumes the smaller-mass one, the larger-mass white dwarf gets more mass, shrinks to a smaller volume, and hence develops a higher density. This increase in density could result in the ignition of carbon burning in the more massive white dwarf. The resulting catastrophic burning in the more massive white dwarf would blow the star apart. This is yet another proposed mechanism to create a certain type of supernova from a white dwarf. We will see this in more detail in Chapter 6.

White dwarfs may just be small quantum-pressure-supported dots, but they can do very interesting things. They may hold the key to understanding the fate of the Universe. We will see that in Chapter 11.

6

Supernovae: stellar catastrophes

6.1 OBSERVATIONS

Which stars explode? Which collapse? Which outwit the villain gravity and settle down to a quiet old age as a white dwarf? Astrophysicists are beginning to block out answers to these questions. We know that a quiet death eludes some stars. Astronomers observe some stars exploding as supernovae, a sudden brightening by which a single star becomes as bright as an entire galaxy. Estimates of the energy involved in such a process reveal that a major portion of the star, if not the entire star, must be blown to smithereens.

Historical records, particularly the careful data recorded by the Chinese, show that seven or eight supernovae have exploded over the last 2000 years in our portion of the Galaxy. The supernova of 1006 was the brightest ever recorded. One could read by this supernova at night. Astronomers throughout the Middle and Far East observed this event.

The supernova of 1054 is by far the most famous, although this event is clearly not the only so-called ''Chinese guest star.'' This explosion produced the rapidly expanding shell of gas that modern astronomers identify as the Crab nebula. The supernova of 1054 was apparently recorded first by the Japanese and was also clearly mentioned by the Koreans, although the Chinese have the most careful records. There is a suspicion that Native Americans recorded the event in rock paintings and perhaps on pottery, but other evidence is that the symbols are generic. An entertaining mystery surrounds the question of why there is no mention of the event in European history. One line of thought is that the church had such a grip on people in the Middle Ages that no one having seen the supernova would have dared voice a difference with the dogma of the immutability of the heavens.

One historian, the wife of one of my colleagues, has an interesting alternative viewpoint. She argues that the people who made careful records of goings-on in medieval Europe were the monks in scattered monasteries. Some of these monks were renowned for their drunken revelries and orgies, in total disregard for their official vows of abstinence and celibacy. Would such people have shied from making mention of a bright light in the sky when they kept otherwise excellent records? (Never put it in writing?) The truth may be more mundane, having to do with weather or mountains blocking the view. A report of a few years ago called attention to a reputed light in the sky at the time of the appointment of Pope Leo, but this has not been widely accepted. In any case, there is no confirmed record of the supernova of 1054 in European history.

Five hundred years later, the Europeans made up for lost time. The supernova of 1572 was observed by the most famous astronomer of the time, the Danish nobleman Tycho Brahe. Tycho made the careful measurements of planetary motions that allowed his student, Johannes Kepler, to deduce his famous laws of planetary motion. Tycho also carefully recorded the supernova of 1572. His data on the rate at which the supernova brightened and then dimmed in comparison to other stars gives a strong indication of the kind of explosion that occurred. The heavens favored Kepler in his turn with the explosion of a supernova in 1604. Kepler also took careful data, by which we deduce that he witnessed the same kind of explosion as his master. Although there are counterarguments and some controversy, both Tycho's and Kepler's supernovae are widely regarded to be the kind of event modern astronomers label Type Ia.

Shortly after Kepler came Galileo and his telescope, and then Newton with his new understanding of the laws of mechanics and gravity. This epoch represented the birth of modern astronomy. Astronomers now have large telescopes, the ability to observe in wavelengths from the radio to gamma rays, and the keen desire to study a supernova close up. Ironically, however, Kepler's was the last supernova to be observed in our Galaxy. Supernovae go off rarely and at random, so a long interval with none is not particularly surprising, just disappointing. We do observe a young expanding gaseous remnant of an exploded star, a powerful emitter of radio radiation known as Cassiopeia A. From the present size and rate of expansion of the remnant, we deduce that the explosion that gave rise to Cas A occurred in about 1667. By rights, this should have been Newton's supernova, but no bright optical outburst was seen. Evidently, this

explosion was underluminous. There are reports that Cas A was seen faintly by John Flamsteed, who was appointed the first Astronomer Royal of England by King Charles II in 1675, but there are questions concerning the timing and whether or not that sighting was in the same position as the remnant observed today. Astrophysicists have calculated that supernovae are brighter if they explode within large red-giant envelopes (see Section 6.6). The suspicion is that the star that exploded in about 1667 may have ejected a major portion of its envelope before exploding or that the star was otherwise relatively small and compact. That condition, in turn, may have prevented Cas A from reaching the peak brightness characteristic of most supernovae. We will see in Chapter 7 that supernova 1987A, the best-studied supernova of all time, had this property of being intrinsically dimmer than usual.

A new chapter in this story was written by the *Chandra X-ray Observatory* launched on July 23, 1999. After astronomers had searched for decades with other instruments, the *Chandra Observatory* found the compact object that was demanded to exist in the remnant of this massive star. Ironically, the very first image obtained by *Chandra* for publicity purposes was of Cas A, since everyone knew an image of Cas A would be spectacular. To everyone's surprise and delight, there was a small dot of X-ray emission right in the dead center of the expanding cloud of supernova ejecta. This central source is putting out X-rays with a luminosity of only about one-tenth that of the total light of our Sun. This explains why it was not seen before. The compact object in Cas A is much fainter than the neutron star in the Crab nebula and, as of this writing, we have still not figured out if it is a neutron star operating under its own power or a black hole with a disk feebly emitting while accreting matter from its surroundings. The *Chandra* website asked for readers to vote between these choices. That is an amusing exercise, perhaps, but it is not the way science is done. One give-away might be a regular pulse of emission, a frequent clue to a rotating neutron star (Chapter 8), but not expected for emission from a disk around a black hole. So far, no such pulsed emission has been seen.

There is a theme that runs through this discussion of historical supernovae and their currently observed remnants, both compact and extended, but that is not immediately obvious. That is that there are two kinds of explosions, ones that leave behind compact remnants and ones that do not. Among the latter are SN 1006 and Tycho's supernova of 1572. Among the former are the Crab nebula and Cas A.

As we will explore in this chapter, there are two fundamental explosion mechanisms, one associated with the collapse of the core of a massive star that must leave behind some sort of compact remnant, either a neutron star or black hole, and the other that blows the star, believed to be a white dwarf, to smithereens, leaving no compact object. A more subtle, but significant, clue is that when a compact object is observed, there is evidence for some sort of elongated appearance or even directed, jet-like flow. This is true for both the Crab and Cas A and for other historical supernovae, events recorded by the Chinese in 386 and 1181, and a nearby, well-studied remnant in the constellation Vela that is thought to have exploded about 10 000 years ago. This correlation also applies to SN 1987A (Chapter 7). The opposite seems also to be true; that when no compact remnant is seen, there is no substantial elongation. SN 1006 and Tycho are examples of this. An exception is the remnant of Kepler's supernova of 1604, about which arguments still rage. *Chandra* X-ray images of Kepler show some elongation and so I will bet it is of the core collapse variety, although there is, as yet, no sign of a compact remnant. We will explore the significance of the association of compact objects and jet-like structure in Section 6.5.

All supernovae directly observed since 1604 (with the possible exception of Cas A), and hence all supernovae seen by modern astronomers, have been in other galaxies. Any single galaxy hosts a supernova only rarely. Supernovae occur roughly once per 100 years for spiral galaxies like the Milky Way. Astronomers do, however, observe a huge number of galaxies at great distances. The chance that some of these galaxies will have supernovae go off in them is appreciable. Before supernova 1987A, about thirty supernovae were recorded every year. Closer attention was paid to discovering supernovae after supernova 1987A, and the current rate of discovery is about 100 per year. Many of these supernovae are so distant and so faint that scant useful data are obtained from them, but special programs have yielded good data on very distant supernovae. This will be discussed in Chapter 12.

From the studies of supernovae in other galaxies, astronomers have come to recognize that there are two basic types called, cleverly enough, Type I and Type II. This differentiation was first made in the 1930s when Fritz Zwicky began systematic searches for supernovae at Caltech. The categories of supernovae are traditionally defined by the *spectrum* that reveals the composition of the ejected matter. Complementary information is obtained from the *light curve*, the pattern of

rapid brightening and slower dimming followed by each event. As more supernovae have been discovered, the dividing lines of this taxonomy have been blurred by events that share some properties of Type I and some of Type II. As for any developing science, one begins with categories and then seeks to replace mere categories with a solid base of physical understanding.

The spectra of *Type I supernovae* are peculiar in that they reveal no detectable hydrogen, the most common element in the Universe. Some Type I supernovae, called Type Ia, appear in all kinds of galaxies – elliptical, spiral, and irregular. Type Ia tend to avoid the arms of spiral galaxies. Because the spiral arms are the site of new star formation, the suggestion is that Type Ia supernovae explode in older, longer-lived stars. This implies that the progenitor stars of Type Ia supernovae are not particularly massive because massive stars live only a short time. Just how low the mass of these Type Ia supernovae may be is a question of current debate. The light curve for Type Ia supernovae is very identifiable. There is an initial rise to a peak that takes about two weeks, and then a long slower period of gradual decay over timescales of months that is very similar for all these events. The data recorded by Tycho and Kepler suggest that they both witnessed Type Ia supernovae. No other galactic supernova has sufficient records to make an identification by type. For decades, all Type Ia supernovae were thought to be virtually identical, but more recent careful observations have revealed small, but real, variations among them.

Near the peak of their light output, *Type II supernovae* show normal abundances in their ejected material, including a normal complement of hydrogen. The material observed at this phase is very similar to the outer layers of the Sun. These supernovae have never appeared in elliptical galaxies. Type II supernovae occur occasionally in irregular galaxies, but mostly in spiral galaxies and then within the confines of the spiral arms. The reasonable interpretation is that the stars that make Type II supernovae are born within the spiral arms and live an insufficient time to wander from the site of their birth. Because they are short-lived, the stars that make Type II supernovae must be massive. The light curve of a typical Type II supernova shows a rise to peak brightness in a week or two and then a period of a month or two when the light output is nearly constant. After this time, the luminosity will drop suddenly and then less rapidly with a timescale of months. This pattern of light emission with time is consistent with an explosion in the core of a star with a massive, extended red-giant envelope, as will be explained in Section 6.6.

To confuse the issue, one and maybe two other varieties of hydrogen-deficient supernovae were identified in the 1980s. These are called, with a further flight of imagination, *Type Ib* and *Type Ic*. The two types are probably closely related. Unlike Type Ia, but like Type II, Types Ib and Ic only seem to explode in the arms of spiral galaxies. Therefore, Types Ib and Ic are also associated with massive stars. Type Ib show evidence for helium in the spectrum near maximum light. Type Ic show little or no such evidence for helium. On the other hand, both types show evidence for oxygen, magnesium, and calcium at later times. This is the strongest argument that Types Ib and Ic are closely related. They show little or no evidence for the strong line of silicon that is a major characteristic of the spectra of Type Ia. Type Ia supernovae show essentially only iron at later times, another factor emphasizing their difference from Types Ib and Ic. The composition revealed by Types Ib and Ic is similar to that expected in the core of a massive star that has been stripped of its hydrogen. In the case of Type Ic, most of the helium is gone as well. This suggests an origin in a star much like a Wolf–Rayet star, but a direct connection to this class of stars has not yet been established. The light curves of Types Ib and Ic are somewhat similar to those of Type Ia, but are dimmer at maximum light.

A bright supernova observed in 1993, SN 1993J, gave yet more clues to the diversity of processes that lead to exploding stars. SN 1993J revealed hydrogen in its spectrum, so this event was a variety of Type II. As the explosion proceeded, however, the strength of the hydrogen features diminished, and strong evidence for helium emerged. In this phase, SN 1993J looked much like a Type Ib. There were a few events like this known before, and several have been seen since. Apparently this star had most, but not all, of its hydrogen envelope removed, probably in a binary mass-transfer process. In other cases, the removal of hydrogen is more nearly complete, and in yet others, for the Type Ic, the helium is removed, too. There is yet no direct observational proof for binary companions in Types Ib or Ic or the transition events like SN 1993J, but computer models suggest this is the case for SN 1993J, at least. Strong winds from massive stars could play a role for the Types Ib and Ic, and the relative importance of winds versus binary mass transfer has not been resolved.

6.2 THE FATE OF MASSIVE STARS

The evidence indicates that Types II and Ib/c supernovae represent the explosion of massive stars. These stars have presumably evolved from

the main sequence to red giants and have had a series of nuclear-burning stages producing ever heavier elements in the core. Just which massive stars participate in this process is still debated.

One way to deduce the masses of the stars that make supernovae is to examine the rate at which the events occur in various galaxies. The death rate can then be compared to the rate at which stars are born with various masses. We know that there are many low-mass stars born every year in a galaxy like ours and rather few massive stars (*why* this should be true is a question under active investigation). If we consider stars with mass in excess of about 20 solar masses, we find such stars are born, and hence die, too infrequently to account for the rate at which Type II supernovae explode. If we consider stars with less than about 8 solar masses, we find that such stars die in excess profusion. Stars with mass between about 8 and about 20 solar masses are born and die at the rate of about once per 100 years in our Galaxy. This is also the approximate rate at which we deduce Type II supernovae occur. Type II supernovae probably come from stars of this mass range. Many of these stars, particularly on the upper end of this mass range, are thought to form iron cores that collapse to form neutron stars. There is thus a strong suspicion that Type II supernovae leave neutron stars as compact remnants of the explosion, and that the gravitational energy liberated in forming the neutron star is the driving force of the explosion.

The rate of explosion of Types Ib and Ic supernovae is not well known because relatively few of them have been discovered. Their rate is roughly the same as that for Type II. This suggests that Types Ib and Ic come from roughly the same mass range as Type II. One possibility is that Types Ib and Ic come only from Wolf–Rayet stars that formed by the action of strong stellar winds in stars more massive than 30 solar masses (Chapter 2, Section 2.2). This is probably not the only source of Type Ib and Ic events. Because very massive stars are rare, there would probably be too few of them to explain the rate of explosions. This suggests that Types Ib and Ic also come from stars that were born with less than 30 solar masses. A binary companion would then be necessary to help strip away the hydrogen envelope. Nevertheless, the basic arguments that pertain to Type II supernovae hold also for Types Ib and Ic. If Types Ib and Ic come from massive stars, to account for their rate of occurrence and their sites in spiral arms, Types Ib and Ic are also very likely to be associated with core collapse to form neutron stars.

At the lower end of the mass range suspected to contribute to Type II supernovae, the evolution may be slightly different. The outcome, core collapse, is basically the same. Computer calculations show that for stars with original mass between about 8 and 12 solar masses the core will be supported by the thermal pressure when carbon is burned. This stage of carbon burning is then regulated and gentle in the standard way. The carbon burns to produce neon and magnesium, but the oxygen that typically coexists with the carbon after helium burning does not get hot enough to burn. As the core, now composed of oxygen, neon, and magnesium, contracts, the quantum pressure comes into play before any other fuel can ignite. The stars in the mass range 8–12 solar masses will therefore form cores supported by the quantum pressure and consisting of oxygen, neon, and magnesium. The atomic nuclei of neon and magnesium are capable of absorbing an electron, thus turning one proton into a neutron, and transmuting themselves into an element of lower proton number. This process reduces the electrons that are responsible for the quantum pressure that is supporting the core. The result is that the core collapses before any of the elements in the core begin thermonuclear burning. During the collapse, the remaining nuclear fuels – oxygen, neon, and magnesium – are converted to iron. The net result is a collapsing iron core, just as for the more massive stars where the iron core forms before the collapse ensues. These two processes of iron-core collapse may give identical results, or there may be some subtle difference between collapse triggered by absorbing electrons rather than by heating and disintegrating the iron. These differences could affect the explosive outcome. There is some evidence that stars in the lower-mass range with the collapsing oxygen/neon/magnesium cores may be especially efficient in producing some of the rare heavy elements like platinum.

A different way of addressing the question of which stars explode is to ask which stars do not explode because they cast off their envelopes gently and leave white-dwarf remnants. This question has been addressed by counting the number of white dwarfs in stellar clusters of various ages and then estimating what stars must have produced those white dwarfs. Such estimates are roughly consistent with the statement that all stars below about 8 solar masses make white dwarfs, and hence do not make supernovae, at least not right away.

Estimates of the rate of formation of neutron stars in the Galaxy are similar to estimates of the rate of formation of Type II supernovae.

This does not prove that Type II supernovae produce neutron stars, but the notion that the two processes are directly related is a nearly universal working hypothesis. The problem with this hypothesis is that no calculations have been able to show satisfactorily that the energy liberated in forming a neutron star can routinely cause an explosion. Despite rather gross changes in the physics over the last three decades, many calculations keep stubbornly predicting no explosion, but total collapse. This does not necessarily mean that such explosions do not occur in nature. The calculations may leave out some important piece of physics. That physics might be presently unknown to us, or the process might be too complex to calculate effectively, like the effects of rotation or magnetic fields. Alternatively, we may find that not all stars that develop collapsing cores do form explosions. Some may leave black holes with no explosion at all.

6.3 ELEMENT FACTORIES

Stars with an initial mass larger than 20 solar masses should form iron cores that collapse. There are so few of these stars that whether they explode or not will not change the total supernova rate appreciably. Some other way must be devised to determine whether or not they explode. The observation that suggests that some of the massive stars must explode is the simple but profound one that says that about 1 percent of the material in stars is composed of elements heavier than helium. These elements cannot be produced in the big bang. On the other hand, we know from theoretical calculations that heavy elements in reasonable proportions are produced naturally in the massive stars in the process of forming an iron core. The conclusion is that at least some of the most massive stars must explode in order to eject their complement of heavy elements into space to be incorporated in new stars.

Calculations show that stars with mass between 8 and about 15 solar masses contain too little in heavy elements outside the collapsing core to contribute substantially to the production of elements like carbon, oxygen, and calcium that are abundant in stars, as well as in our bodies. Thus the stars that are presumed to account for most, if not all, of the Type II supernovae are not significant contributors to synthesis of the heavy elements. Stars with mass between about 15 and 100 solar masses produce substantial amounts of heavy elements. If these stars explode and eject their heavy elements, this freshly

synthesized material will mix with the hydrogen in the interstellar gas. New stars form from this enriched mixture. If all stars from 15 to 100 solar masses explode, the new stars will have about the right amount of all the abundant heavy elements.

This picture has led to the widespread belief that the most massive stars must explode and produce the heavy elements. There is probably a great deal of truth in this notion. As observations get more accurate, however, there are hints that the broad picture must be reassessed. Detailed stellar spectra of both young and old stars have allowed new accurate measurements to be made of the way that various elements have been produced throughout the history of the Galaxy. There is a suggestion that if all the massive stars from 15 to 100 solar masses explode, many of the basic heavy elements like carbon, oxygen, and iron will be produced in greater quantity than is observed in the stars in the vicinity of the Sun. A possible resolution of this dilemma is that some of the massive stars collapse completely. In this picture, some massive stars would explode, ejecting heavy elements and leaving neutron stars behind as compact remnants. Others would produce no explosion and would leave behind black holes as the only remnant of their previous stellar existence.

The pattern that seems to best satisfy all our present knowledge would have stars from about 8 to about 30 solar masses exploding and those from 30 to 100 solar masses collapsing and swallowing all their heavy elements. The most reasonable position is probably to conclude that we do not yet know enough about the nuclear and evolutionary processes in stars to conclude with any certainty which stars explode and eject the heavy elements we see.

6.4 Collapse and Explosion

In the collapse of an iron core, the protons capture electrons and convert to neutrons. Each reaction creates a neutrino. This is the process by which the composition is converted to neutrons, the necessary step to make a neutron star. For every neutron formed, there must also be a neutrino. The result is a *lot* of neutrinos.

When the collapse reaches the density of atomic nuclei, the strong nuclear force has a repulsive component. This provides a strong outward pressure. In addition, the quantum pressure of the neutrons plays a role. The result of the increased pressure is that the collapse halts (temporarily, at least). The basic processes as they are thought to occur in a massive star are shown in Figure 6.1.

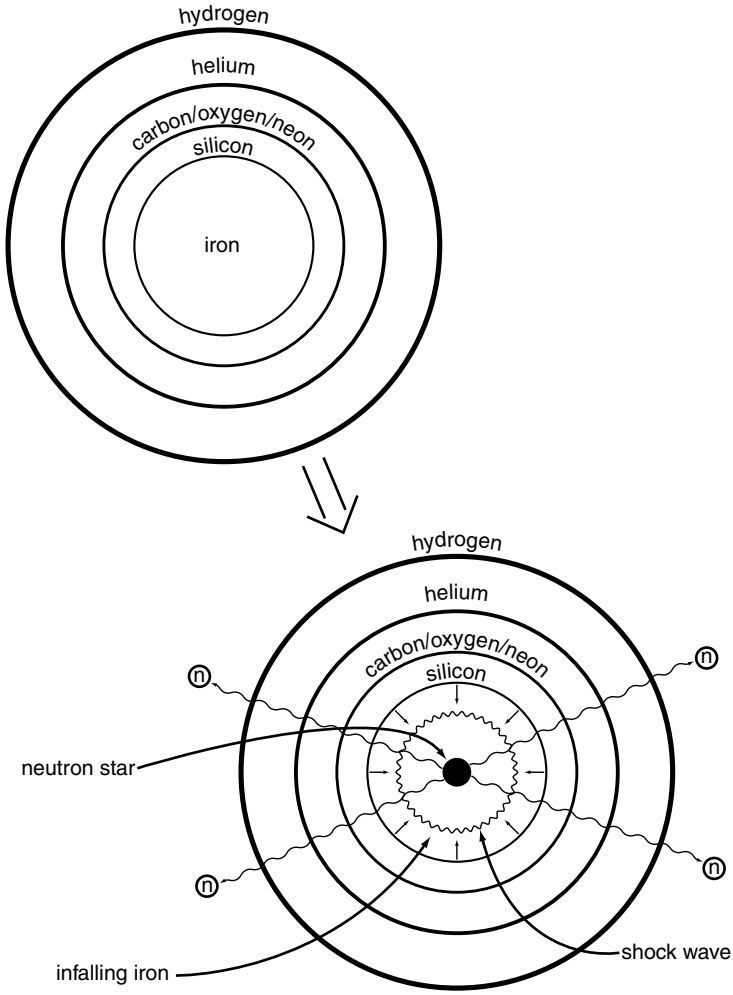


Figure 6.1 The collapse of the iron core of a massive star to form a neutron star. (Top) The star passes through many phases of regulated nuclear burning and forms an iron core. (Bottom) The iron core collapses to form a neutron star, momentarily leaving the outer layers hovering. The creation of the neutron star creates a huge flood of neutrinos. The rebound of the neutron star produces a shock wave that propagates outward into the infalling matter. If the shock wave is strong and the explosion is successful, the outer layers will be blown off in a supernova explosion, and the elements produced in the star will be spread into space. If the explosion is not strong enough, the outer layers will also fall in and crush the neutron star into a black hole.

If you drop something heavy, like a bowling ball, appreciable energy is released when it lands. The more massive the object, the greater the energy released. The farther the object falls, the greater will be the energy released. Imagine dropping the bowling ball from the top of a tall building. Imagine dropping a sports utility vehicle from the top of a tall building. Now imagine the gigantic release of energy when a star with the mass of the Sun collapses to the tiny size of a neutron star, only a few kilometers across. A huge energy is released when the neutron star forms. This energy is several hundred times more than is necessary to blow off the outer layers, those containing calcium, oxygen, carbon, and helium, and any outer envelope of never-burned hydrogen. The problem is that most of the energy produced in the collapse is lost to the neutrinos that can easily stream out of the newly born neutron star and through the infalling matter. If 99 percent of the energy is lost, 1 percent can remain. That is enough to cause an explosion. If 99.9 percent is lost, however, that is too much. The explosion will fail, and the outer matter will continue to rain in and crush the neutron star into a black hole.

The exact treatment of this problem has proven to be very difficult. The requirement is to determine whether 99 or 99.9 percent of the energy is lost to neutrinos, or whether it is some fraction in between. The energy lost to neutrinos must be determined to about one part in a thousand. Uncertainties in the complex physics involved in core collapse have been larger than this critical difference. A related problem is that the explosion process tends to be self-limiting. If more of the energy is trapped, then the rate of infall of new matter from the outer parts of the star is slowed. This, however, decreases the rate at which the collapse produces energy that can power the explosion. The result has been that for decades computer calculations have tended to give results that teeter on the edge of success, some giving explosions, many giving complete collapse to form black holes with no explosion. No completely clear, accepted, reproducible result has emerged. The stars know how to produce these explosions, but astrophysicists are struggling to figure it out.

Over the last couple of decades, research on this topic has involved two basic mechanisms by which the collapse of an iron core might be partially reversed to make a supernova explosion. One is called *core bounce*. When the neutron star first forms, the new star overshoots its equilibrium configuration giving a large compression to the neutron core. There is then a rebound. This rebound sends a strong supersonic shock wave back out through the infalling matter.

The core takes about 1 second to collapse after instability sets in. The core bounce creates the shock in about 0.01 second. If everything works, in this short time a huge explosion should be generated.

If the shock wave is sufficiently strong, the outer matter is ejected, and the neutron star is left behind; however, the shock must run uphill into the infalling matter. Some of the energy of the shock is dissipated by the production and loss of neutrinos. The shock also must do the work of breaking down the infalling iron into lighter elements, protons and neutrons, to form the neutron star. The shock wave can thus stall with insufficient energy to reach the outer layers of the star. Matter can continue to rain down on the stalled shock front, as illustrated in the bottom part of Figure 6.2. The shock front hangs in mid-flow, much as a bow wave stands off a rock in the middle of a stream, as shown in the top of Figure 6.2. The matter will continue to be shocked as material hits this front, but the shocked matter will settle onto the neutron star, just as the water will be slowed, but not stopped, by the rock in the stream. When enough matter lands on the neutron star, the neutron star will be crushed into a black hole. Most calculations currently show that the core bounce alone is not sufficient to cause an explosion.

The other mechanism that has been actively considered takes advantage of the tremendous stream of neutrinos leaving the neutron star. Normal matter, the Sun, is essentially transparent to neutrinos because neutrinos interact only through the weak nuclear force. The only exception to this is neutron star matter. This matter, nearly as dense as an atomic nucleus, is so dense that it can be opaque or at least semitransparent to the neutrinos. Although most of the neutrinos will get out into space, a small fraction will be trapped in the hot matter that lies just behind the shock front created by the core bounce. The slow accumulation of neutrino heat may provide the pressure to reinvigorate the shock, driving the shock outward and causing the explosion. Slow in this case means about a second.

The mechanism for depositing a small fraction of the neutrino energy behind the shock may be related to the boiling of the newly formed neutron star, as shown in Figure 6.3. When the collapse is first halted and the neutron star rebounds, the neutron star is very hot. This heat can cause the neutron star to boil much like a pan of water boils on the stove. The boiling provides a mechanism for carrying the heat upward, in the case of the pan, or neutrinos outward, in the case of the neutron star, by mechanical motion that bodily carries the heat or neutrinos. Under the right circumstances, this boiling process can

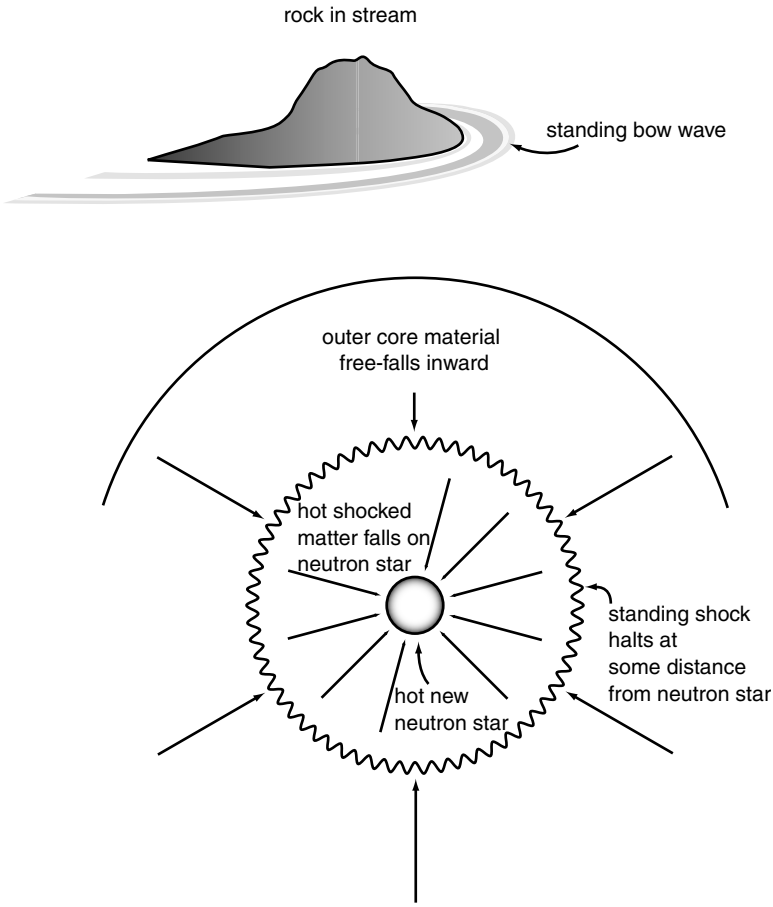


Figure 6.2 A rock in a stream will cause a standing bow wave to form in front of it. Because the water, not the rock, is moving, the wave can also stand still. In the collapse of a stellar core, the shock wave formed by the rebound of the neutron star will move outward into the infalling matter. It can reach a position where the pressure of the hot gas inside the shock (the analog of the rock) supports the shock as the outer matter of the star continues to rain downward. As the matter flows inward, the shock can hover at one radius as a ''standing shock.''

be much more efficient in transporting neutrinos than a slower process of leaking radiation, or neutrinos. Calculations of this process in neutron stars are very challenging because the motion is complex. All modern calculations that can follow motion in more than one (radial) dimension show that neutron stars do boil. There is a consensus that explosions will not occur without this boiling. There is still debate

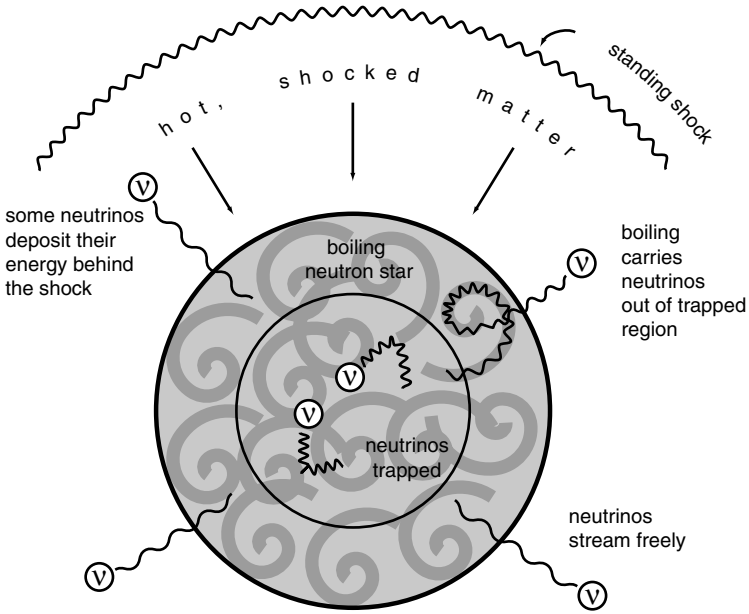


Figure 6.3 Deep within a newly formed neutron star, the matter is so dense that even the neutrinos are trapped. The neutrinos can bounce around, but they cannot escape directly. If the neutron star is hot and boiling as computer calculations show, some of the matter containing neutrinos will boil to the surface, where the trapped neutrinos can escape. This can enhance the rate of loss of neutrinos from the neutron star. Some fraction of these neutrinos can interact with matter beyond the neutron star, but behind the standing shock. If the flood of neutrinos enhanced by the boiling of the neutron star is large enough, sufficient neutrino energy might be deposited behind the standing shock to reinvigorate it and send it all the way out of the star, leading to a successful supernova explosion.

about whether this process of boiling neutrinos is sufficient to cause an explosion.

6.5 Polarization and Jets: New Observations and New Concepts

For the past thirty years, most calculations of core collapse and subsequent events treated the configuration as spherically symmetric. Even if the neutron star boils, the structure of the neutron star may, on average, be spherically symmetric. There are a number of lines of

evidence, however, that the explosions that result from the core collapse process are intrinsically nonspherical. Matter may be ejected more intensely in some directions than others.

Some hints of this perspective have been with us for a long time. As we will see in Chapter 8, we observe hundreds of neutron stars as rotating, magnetic *pulsars*. If we look at the supernova remnants, the expanding clouds of gas that have produced neutron stars in supernova explosions, there is evidence for nonspherical behavior. The famous Crab nebula is hardly round. X-ray images obtained by the *Chandra Observatory* show a torus of matter shed by the neutron star and jets of high velocity matter being spurted out in opposite directions along the axis of the torus. The neutron star is even running away in space directly along this jet direction. Cassiopeiae A shows evidence of a jet-like flow in one direction and a somewhat more diffuse, but distinct flow in the opposite direction. Unlike the case for the Crab pulsar and a couple of other examples, the compact object in Cas A seems to be running away perpendicular to the orientation of the jets, not along them. That must be a clue, but we do not yet know what it is telling us about the explosion process and compact object in Cas A.

Thus, the situation was, until recently, that we knew that the left-overs of core collapse were frequently rotating, magnetic neutron stars. What we did not know was whether the rotation and magnetic fields were crucial to the process, or present but incidental to the explosion. Similar arguments applied to the supernova remnants that showed evidence for asymmetries of various kinds. Were these aspects of a few peculiar supernovae, or was something systematic going on?

The technique of measuring the *polarization* of the light from supernovae provided a new window of observations and major new insights into the explosion process. Electromagnetic radiation consists of an electric component oscillating in a fashion that is perpendicular to the magnetic component, with both perpendicular to the direction of motion of the electromagnetic wave (or photon of light in the quantum description), as illustrated in Figure 6.4. The process of measuring the polarization of the light is one of determining the direction in which the electric field is oscillating. In supernovae, the light scatters off electrons in the outer material of the supernovae before proceeding to astronomers' telescopes, millions of light years away. This scattering gives an average net orientation of the electric component that is perpendicular to the surface of the

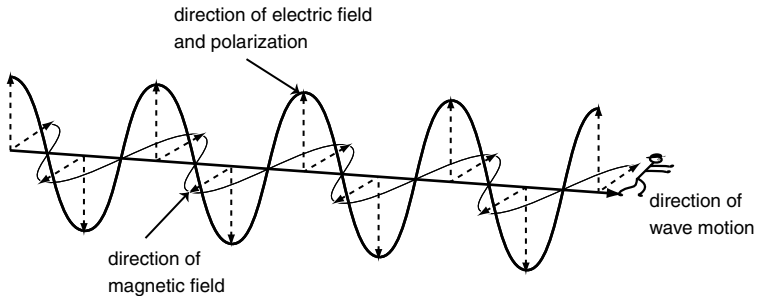


Figure 6.4 Electromagnetic radiation consists of an electric-field component oscillating in a fashion that is perpendicular to the magnetic-field component. In this illustration, the electric field oscillates up and down, as shown by the sinusoidal line and the arrows representing the electric field at its maximum amplitude; the magnetic field is oscillating back and forth as shown by the corresponding line and arrows. The wave itself moves off at the speed of light in a third direction that is perpendicular to the electric and magnetic fields. The technique of polarimetry consists of measuring the intensity and orientation of the electric-field oscillation, up and down in this illustraton. Since the orientation of this electric field can be different at different parts of an exploding supernova, polarimetry gives a method for learning about the shape of a supernova that is too distant to obtain a direct image.

supernova. If the supernova is perfectly spherical (or if it is at least round in the aspect it presents to us) all directions will be represented in the light and there will be no net direction to the electric component. If, however, the supernova matter is asymmetric in some fashion, then some parts of the matter will provide more light, and more heavily represent the orientation of the surface they represent, than others. The net effect will be to impart a net orientation to the electric component of all the light from the supernova and this will give a net polarization for astronomers to measure. The basic nature of this effect is illustrated in Figure 6.4. The bottom line is that if a supernova reveals a net polarization it *cannot* be spherically symmetric; it might be pancake shaped, or cigar shaped, or, much more likely, some more complicated shape, but it cannot be round.

Starting about a decade ago, our group at Texas, led by my colleague Lifan Wang who was then a Hubble Postdoctoral Fellow here, began to collect polarization data on every supernova that was accessible to us. The early days were hard. Lifan used a small telescope and had to add up data from several nights to get enough signal. This

was reminiscent of the heroic days of astronomy early in the twentieth century when a single night's data was simply not enough. We learned that lesson and have migrated our program, now led by Dietrich Baade at the European Southern Observatory, to the magnificent Very Large Telescope (VLT) array in Chile, where similar observations on their eight-meter telescopes can be done in a half hour!

The first thing Lifan noticed as the data began to come in was that there was a distinct difference between Type Ia supernovae and all the core-collapse supernovae. Near and after peak light, Type Ia were barely polarized, if at all. They were essentially round. All of the core-collapse supernovae showed significant polarization. They were definitely not round! As even more data came in over the last few years, we realized that the strength of the polarization got larger as the supernova aged, thinned out, and allowed us to see deeper into its depths. This meant that the cause of the asymmetry was not some incidental aspect of its environment, but that the inner depths were asymmetric; the very machine driving the explosion was severely out of round. We also realized that Type Ic were highly polarized. These supernovae have lost their hydrogen and helium envelopes allowing us to see deeper into the explosion, even at early times. The lesson is the same. The inner depths, driven by the explosion process, are highly non-spherical.

Another important lesson was that in many of the cases, the polarization was not random. The net orientation of the electric field always pointed in the same direction independent of time or even of the wavelength observed. This meant that the supernova ejecta were somehow driven along a special direction during the explosion. Even more data has shown that this behavior is not universal. Sometimes more than one direction is indicated by different ejected elements and sometimes the data seem to indicate truly random directions in space. Still, this tendency for the ejected supernovae matter to ''point'' in a special direction is a powerful aspect in many cases and a strong clue to what is going on.

If, in common circumstances, the supernova is somehow ''pointing'' to a certain direction in space, how can that happen? What would tell an exploding star that one direction was somehow special? The obvious answer seems to be rotation. A sphere at rest will have no special orientation, but a rotating sphere, or a planet like the Earth, or a star like the Sun, or the Galaxy (which is not spherical) for that matter, have a special direction, the direction aligned with the axis of rotation. Rotation automatically selects a special direction.

What that specific direction is depends on the accident of birth and maybe subsequent jostling, but that direction is an intrinsic characteristic of a rotating object. There are, however, ways of setting up special directions that do not require rotation. One is that the newborn neutron star may end up oscillating with respect to the outer stellar material: neutron star to the left, star to the right; then vice versa. We have to keep such possibilities in mind as we go forward.

As the polarization data first began to accumulate, the first thing we thought of were jets. Jets blowing along special directions are a ubiquitous aspect of gravitating accreting systems. Protostars blow jets. We see jets of matter coming from the centers of galaxies and from black holes in binary stellar systems (see Chapter 10). The infall of the iron core to form a neutron star is an extreme case of a gravitating, accreting system. Perhaps, we thought, a similar thing was happening in the core collapse supernovae.

Another important ingredient in this context is magnetic fields. As outlined above (and will be explored in detail in Chapter 8), pulsars are neutron stars that both rotate and are magnetic. Most of the theories of how to produce jets depend on tangling up magnetic fields. Perhaps, then, magnetic fields are also important to the actual process of the explosion of the supernova. This is hard to prove, but I think my student and colleague Shizuka Akiyama and I have taken an important step in this direction. We have examined the physics of the magneto-rotational instability that was first discussed in Chapter 4 (Section 4.4) in the context of accretion disks. Amplifying magnetic field by this mechanism requires a gravitating system with *shear*, the process by which some matter slides past other matter. The flow in accretion disks intrinsically involves shear; the matter closer to the central star naturally moves faster than the matter further out. The same thing is true in core collapse. As the iron core collapses to form a neutron star, it is like a skater pulling in his arms (Figure 1.2); the neutron star will spin much faster than the original iron core. The difference naturally forms a shear and drives the magneto-rotational instability that will rapidly grow any feeble magnetic field that might be present in the original iron core. The implication is that the magnetic field will naturally grow in this environment. It is not consistent to consider only rotation and ignore the magnetic field. Rotation and magnetic fields will come hand-in-hand in the core collapse environment. The important issue is just how big is the magnetic field and just what it will do to the matter. This is a tough problem, but, to my mind, the polarization is telling us that rotation

and magnetic fields are intrinsically coupled to the explosion process, shaping the explosion if not actually causing it.

The polarization then points to an important role for rotation and magnetic fields in the very explosion process itself. If this is the case, then the current numerical calculations may be missing a major ingredient necessary to yield an explosion. The most obvious mechanism for breaking the spherical symmetry by singling out a specific direction is rotation, because rotation defines a rotation axis. Proper treatment of rotation, abetted by magnetic fields, may be necessary in order to understand fully when and how collapse leads to explosions. All the energy of collapse is provided by gravity. This energy temporarily goes into two components: the hot bath of neutrinos that will slowly leak out of the neutron star and the tremendous fly-wheel of the rotating neutron star itself. Tapping the energy of that fly-wheel and sending it up the rotation axis may be just the process that explodes and shapes core-collapse supernovae. Adding the effects of rotation and magnetic fields is even more of a computational challenge, but computer power grows steadily, and progress will be made in this area in the next few years. Other suggestions that rotation and magnetic fields are important to the core-collapse process are presented in Chapter 11.

To pursue the question of the role of jets in supernovae, my colleague Alexei Khokhlov, then at the Naval Research Laboratory and now at the University of Chicago, explored what jets might do to supernovae. This calculation glossed over a number of complications that need to be investigated more deeply, but addressed fundamental issues by assuming that a newly formed neutron star could launch jets along the rotation axes in about a second, while the outer parts of the star hovered, waiting to be blasted into space or to collapse into a black hole depending on the outcome of the collapse. To correspond to a Type Ib or Ic supernova, the hydrogen envelope of a massive star model was omitted, and only the core of helium and heavier elements was retained (Khokhlov and my Texas colleague Peter Höflich have since done calculations covering more general conditions). The jets penetrated to the surface of the helium core in about six seconds. As they propagated, the jets drove bow shocks that blow sideways as well as forward, much as a motor boat creates a bow wave as it powers across a lake (see also Figure 6.2). Unlike a lake, a star is basically spherical and the bow waves blown away from the jet open up away from the jet like a flower petal and wrap around the star. If the jets are basically symmetrical in the ‘‘up’’ and ‘‘down’’ direction, the

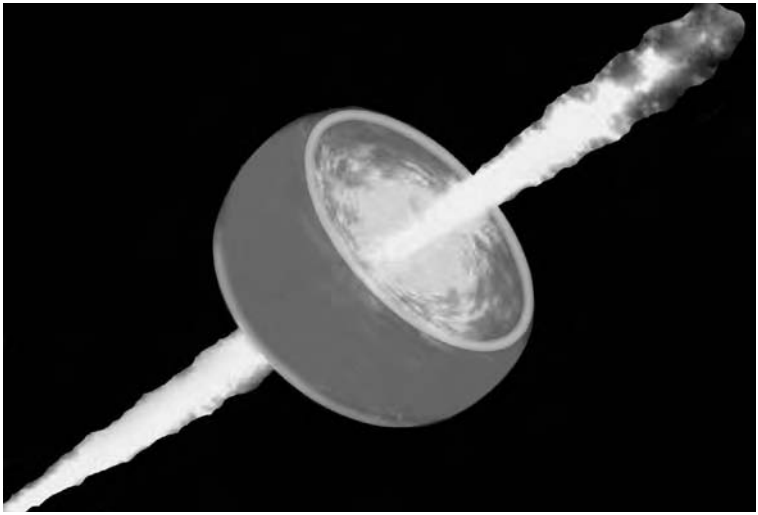


Figure 6.5 The collapse of a rotating iron core to form a rotating magnetic neutron star may yield strong jets. Computer simulations show that sufficiently strong jets can explode the star and leave a typical shape to the ejecta, as illustrated here. Twin jets will blow matter out along the rotation axis. As the jets plow out through the star, their bow waves drive circular shock wave patterns that propagate away from the jets, as illustrated by the lighter rings along the top and bottom perimeters. These shock waves will collide along the equator. That causes matter to be expelled in an expanding torus or doughnut of matter along the equator. The result is a canonical ''bagel and breadstick'' shape that could account for the shape of core-collapse supernovae as seen in images and as measured by polarization (adapted from a NASA illustration).

down-going bow shock from the ''up'' jet will collide with the up-going bow shock from the ''down'' jet at the equator. The result is that shortly after the jets penetrate the surface, the sideways bow shocks converge and eject the matter out along the equator. If the star has no hydrogen envelope, as assumed by Khokhlov, then the final result is two jets of matter along the axes and a strongly asymmetric, doughnut-like explosion in the equatorial direction, as illustrated schematically in Figure 6.5. This is the generic shape predicted for a jet-induced supernova. Although the polarization observations cannot uniquely prove this is the shape, the data are consistent with this shape as the source of the observed polarization in many cases.

Although I have used various props to illustrate this generic shape of jet-induced supernovae (once a carrot and chocolate doughnut were the only supplies available; very messy), my favorite is a breadstick and a bagel because it alliterates. The breadstick threaded through the hole in the bagel represents the matter ejected in the jets. The bagel represents the matter blown out along the equator by the converging bow shocks. This concoction captures some of the sense of the nature of the explosion, but one must recall that it is a static image; in reality matter will be rushing outward in both the ''bagel'' and the ''breadstick'' directions.

The explosion computed by Khokhlov was driven entirely by the jets. The stalled shock and the neutrinos described in Section 6.4 played no role. This trial calculation does not prove that jets alone explode supernovae, but it does show that sufficiently strong jets can do so in principle. Further study will probably show that both jets and neutrinos are necessary in varying degrees. If jets are a critical part of the explosion in many, if not all, core collapse events, then many issues such as nucleosynthesis and the production of black holes must be reconsidered.

These developments leave open the issue of how jets are formed in supernovae if, indeed, they are. One aspect of the problem is that the magnetic fields probably do not represent the strongest force during the core collapse process; the magnetic forces are intrinsically weaker than the pressure forces in the neutron star. On the other hand, the pressure and gravity basically push along a radial direction and cancel one another. The magnetic field has the special property that it can push laterally where ordinary pressure and gravity offer little resistance. Magnetic fields may help to direct matter and energy toward the rotation axes. By catalyzing the motion of energy in that direction, magnetic fields may help to tap the rotational energy to flow into axial jets without contributing to the brute energy of the flow itself.

One aspect of this problem that Shizuka Akiyama and I have recently emphasized is the somewhat counterintuitive notion that the final spin and magnetic field of the neutron star, will be an irregular function of the original spin of the iron core. If the iron core spins slowly, the neutron star will spin slowly and generate only a weak magnetic field. If the iron core spins a bit faster, the neutron star will spin a bit faster and generate a stronger magnetic field. If, however, the iron core spins faster than a certain amount, then the centrifugal force of rotation will tend to give an extra source of support to the

neutron star, in addition to its normal pressure. That means that the neutron star will not collapse quite as far or achieve quite so high a density. That, in turn, means that the neutron star will rotate a bit more *slowly*, like a skater who has only pulled her arms in part way. It also follows that the magnetic field generated will be less strong for this faster iron core rotation.

The rotation of the iron core may thus be an important determinant of the final outcome of the collapse. It is conceivable, for instance, that very slowly rotating iron cores will fail to trigger an explosion (as many of the most sophisticated computer calculations today show!). Somewhat faster rotation of the iron core will generate more rotation of the neutron star and stronger magnetic fields, perhaps triggering successful jet-induced (and neutrino-boosted) supernovae. With even higher rotation of the iron core, however, the neutron star rotates less fast, generates weaker magnetic fields and perhaps there ensues in this situation total collapse to form a black hole. This is only a hypothesis, but it illustrates how thinking about the core-collapse problem might change, once rotation and magnetic fields are brought into the picture.

What would make one star have a slower rotating iron core and another a faster rotating core? This is also a difficult problem that is the subject of current active research. The evolution of stars from the main sequence to the iron-core phase will tend to be accompanied by a migration of angular momentum outward from the faster inner core to the slower outer envelope, thus slowing the spin of the iron core that ultimately forms. The rate at which the core is spun down may also be a sensitive function of the magnetic field that exists in the star, another focus of current research. In addition, the outcome is probably influenced by whether the star has a binary companion. Two stars in orbit can induce a mutual *torque* on one another, thus pumping some of their orbital energy and angular momentum into the spin of the cores of the stars, yielding, other things being equal, faster-spinning cores. In other circumstances, the stars could form a common envelope (Chapter 3, Section 3.9). The two stars might eject the common envelope and form a new compact binary system, but it might be even more likely that the two stars (or an immersed star and the core of the star that formed the common envelope) merge to form one exceedingly rapidly rotating core that could, if the circumstances are right, proceed to form an iron core. The issue of the success of a supernova and whether a given star yields a neutron star or a black

hole might then depend on whether or not the star was born in a multiple star system.

A subject that is developing as this second edition goes to print is our recent recognition that rotating neutron stars will be subject to forming shapes that not only depart from spherically symmetry, but even from axial symmetry, shapes like spiral arms and other, more complex geometries. Most of the work showing this behavior has ignored both the fact that a new-born neutron star will still be immersed in the supernova environment with matter raining down on it, and that the neutron star will be magnetic. In this case, we again hypothesize, the nonaxially symmetric motion will rattle the magnetic field, generating magnetohydrodynamic waves that will sap the energy of the rotation and carry that energy somewhere else, maybe up the axes in jets. It will take some effort to explore these ideas thoroughly, but we again see the expanding range of possibilities once rotation and magnetic fields are considered.

6.6 TYPE IA SUPERNOVAE: THE PECULIAR BREED

The principal peculiarity of Type I supernovae is that such events have no hydrogen in their ejected material. The hydrogen envelope that surrounds most stars has either been ejected or consumed to make helium or heavier elements. As noted in Section 6.1, there are two rather different observed categories of Type I. Some of them, the Types Ib and Ic, like Type II, occur only in spiral or irregular galaxies. The Type Ia supernovae occur in all types of galaxies. This makes Type Ia events different in some fundamental way and worthy of special attention.

In particular, Type Ia supernovae occur in elliptical galaxies, whereas Types II, Ib, and Ic do not. Elliptical galaxies have converted essentially all their gas into stars long ago and to a great extent have ceased the making of stars. Thus elliptical galaxies are thought to consist only of old, low-mass, long-lived stars. The high-mass stars born long ago should be long dead. This has given rise, in turn, to the idea that Type Ia supernovae must come somehow from low-mass stars. Because spiral galaxies contain a mix of high-mass and low-mass stars, that spirals produce both Type Ia and Type II supernovae is not surprising.

Another aspect that has driven thinking about Type Ia supernovae is that their observed properties are remarkably uniform. Type Ia events tend to follow the same light curve. In addition, as Type Ia

brighten and decline, the alterations in their spectra follow a very predictable course. Because white dwarfs of the Chandrasekhar mass would be essentially identical and hence undergo nearly identical explosions, the observed homogeneity of Type Ia has pointed to an origin in exploding white dwarfs. We now know that all Type Ia supernovae are not exactly identical. The reasons for this are the subject of active current research, as will be discussed later.

The most popular notion for how to turn a low-mass star into a supernova is thus to rejuvenate a white dwarf. The idea is that the more massive star in an orbiting pair could evolve and form a white dwarf. The low-mass companion could then take a long time to evolve, but it would eventually swell up as a red giant and dump mass onto the white dwarf. If the total mass accumulated by the white dwarf approaches the Chandrasekhar mass of about 1.4 solar masses, the white dwarf might then explode. A variation on this theme is that the white dwarf could grow in mass in a cataclysmic-variable system where the mass flows from a main-sequence star (Chapter 5). This process is slow, and the system could still last a long time before exploding. Yet another possibility is that Type Ia supernovae arise from systems of two white dwarfs that slowly merge due to the emission of gravitational waves generated by their orbital dance (Chapter 5, Section 5.4).

Careful studies of the observed properties of Type Ia supernovae are completely consistent with the general picture that the explosion occurs in a white dwarf. Near peak light, the spectra of Type Ia supernovae show elements such as oxygen, magnesium, silicon, sulfur, and calcium. These are just the elements expected if a mixture of carbon and oxygen burns to produce somewhat heavier elements consisting of differing numbers of ''helium nuclei.'' As a Type Ia supernova evolves, the spectrum becomes dominated by iron and other similarly heavy elements. These elements can be produced by burning carbon and oxygen all the way to iron. The nuclear binding energy of iron is at the bottom of the ''nuclear valley,'' where the neutrons and protons in the nucleus are most compressed (Chapter 2, Section 2.4).

In the process of expanding and thinning out, the outer, more tenuous portions of a supernova are seen first, and the inner, denser, more opaque portions are only seen later. The information revealed by the evolution of the spectra is then consistent with a configuration in which the denser inner portions of the exploding star burn all the way to iron and iron-like elements, and the outer parts are composed

of matter that results from carbon burning, but that is not so thoroughly processed. Computer models of exploding white dwarfs give results that match this pattern rather well. The exact nature of the combustion is still being explored, but the most successful models adopt a progenitor that is a carbon/oxygen white dwarf with a mass very near to, but less than, the Chandrasekhar mass.

At this point, I must correct a long-standing and erroneous view of the nature of Type Ia supernovae. This view is shared by many wise experts and neophytes alike because they have not followed this research closely. A casual view that permeated the astronomical community and the popular astronomical literature decades ago, and that is very difficult to root out, is that to make a Type Ia supernova, matter is added to a white dwarf until the Chandrasekhar mass is exceeded and the white dwarf collapses. *This is wrong!* The reason this notion is so persistent, I suspect, is that the idea of exceeding the mass limit and collapsing is simple and visceral. In addition, the ''other'' means of making supernovae does involve core collapse, and so it is easy to confuse the two mechanisms. There are also circumstances where some white dwarfs might collapse, but if so, the process does not yield the events we observe as Type Ia supernovae. Rather, mass is added, we believe, increasing the density in the center of the white dwarf until finally carbon can ignite. This condition of carbon ignition and subsequent unregulated thermonuclear runaway happens when the white dwarf has a mass about one percent less, not more, than the Chandrasekhar mass, and it blows the white dwarf up completely, so there is no collapse. This is a somewhat more complicated and perhaps less intuitive process (think dynamite!), and this may be why it has not permeated all corners of the community of interested people. Nevertheless, the supernova community stopped talking about exceeding the Chandrasekhar limit and collapse in the 1960s, and it is rather dismaying to find experts in related areas, never mind popular astronomy enthusiasts, still referring to this outmoded physical picture. The overwhelming observational evidence is that Type Ia supernovae arise from carbon/oxygen white dwarfs of mass a little less than the Chandrasekhar limit that do not collapse, but blow up completely by a process of thermonuclear explosion.

Type Ia supernovae explode because the white dwarf is supported by the quantum pressure, and any burning under those circumstances is unregulated, as we discussed in Chapters 7 and 5. For Type Ia supernovae, burning is unregulated in the extreme. As a white dwarf approaches the Chandrasekhar limiting mass, the central

density gets very high. Formally, the density would go to infinity just at the Chandrasekhar limit, but in practice other physics, in this case carbon burning, will come into play. The high density triggers the ignition of carbon but also ensures that, under these circumstances, the quantum pressure will be exceedingly large. The white dwarf will have a finite temperature that will help to promote the carbon burning, but the thermal pressure is negligible. The story of unregulated burning we have told before will then play out in the most dramatic way. The carbon begins to burn and to release energy. The quantum pressure does not budge. There is no mechanical response to expand and cool the star and damp the burning. The burning goes even faster, raising the temperature even more and producing ever faster burning. Under the extreme conditions at the center of a white dwarf with a little less than the Chandrasekhar mass, the burning cannot be controlled, the oxygen also ignites, and all the fuel is consumed to iron-peak elements in a flash. The result is a violent thermonuclear explosion.

There are two different ways of propagating a thermonuclear explosion in a white dwarf. One is a subsonic burning like a flame, a process called a *deflagration*. The other is a supersonic burning that is preceded by a shock front, very much like a stick of dynamite. This process is known as a *detonation*. We have known since the 1970s that Type Ia explosions cannot be the result of pure detonation. The supersonic burning rips through the model white dwarf before it can expand and adjust, and essentially the whole star is converted to iron-like matter. That is not what we see! We must account for the oxygen, silicon, sulfur, and calcium in the outer layers. The most sophisticated current models, those that best match the data, have the unregulated carbon burning begin as a boiling, turbulent deflagration and then make a transition to a supersonic detonation, as illustrated in Figure 6.6. These are known as *deflagration-to-detonation models*.

Both deflagration and deflagration-to-detonation models naturally create iron-like matter in the center, and intermediate elements like magnesium, silicon, sulfur, and calcium on the outside. These models also predict that the white dwarf is completely destroyed, leaving no compact remnant like a neutron star or a black hole. This comparison of theory and observation thus strongly points to an interpretation of Type Ia supernovae as the explosion of a carbon/oxygen white dwarf at just less than the Chandrasekhar limit.

There are ways to distinguish white dwarfs that explode only by subsonic deflagration and those that explode in the more complex

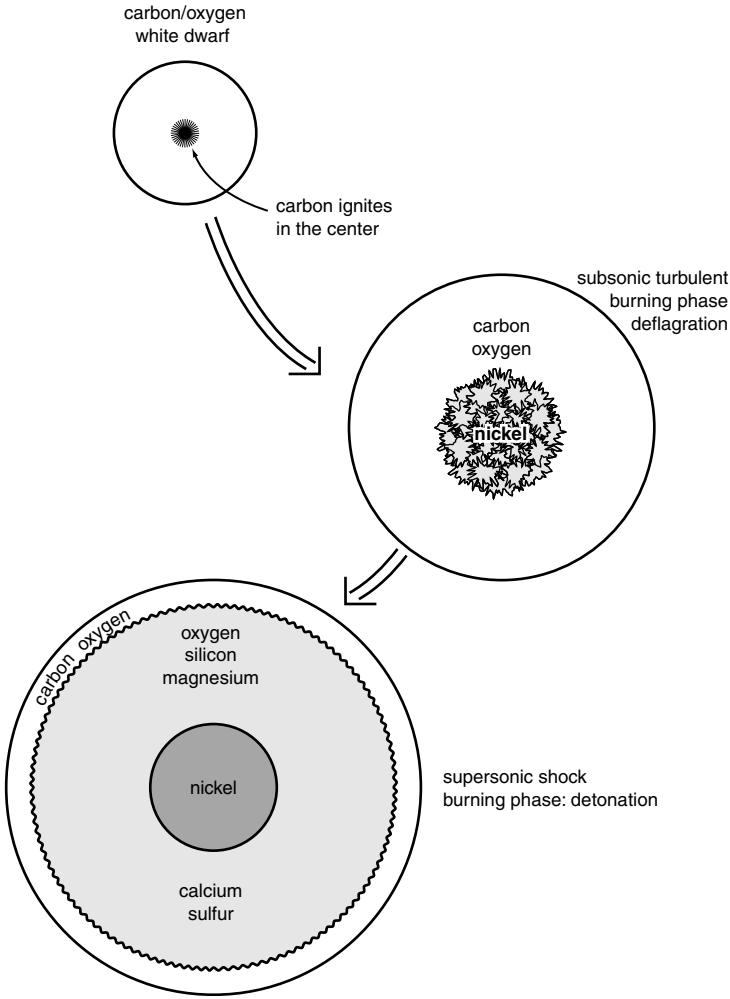


Figure 6.6 (Top) A Type Ia supernova explosion begins with the ignition of carbon near the center of the white dwarf. (Middle) A turbulent, roiling, burning front that moves less rapidly than the speed of sound spreads out from the center, at first converting all the burning matter to radioactive nickel. The pressure waves from this burning cause matter beyond the burning regions to expand before the burning reaches them. (Bottom) At some point, the burning front begins to propagate supersonically, producing a shock wave that triggers the burning. This detonation wave moves so rapidly that the outer portions of the star cannot expand substantially farther before they are overtaken by the burning. The detonation burning leaves behind oxygen, magnesium, silicon, sulfur, and calcium, the elements seen in the outer layers of Type Ia supernovae. A thin layer of unburned carbon and oxygen on the outside of the white dwarf might survive the explosion.

deflagration-to-detonation picture. The deflagration pushes matter out ahead of it at nearly the speed of sound, but the burning proceeds at intrinsically less than the speed of sound so it cannot catch up with, and burn, all the expanding matter. This means that models that rely purely on deflagration to explode the supernova must leave some unburned matter, still composed of carbon and oxygen, in the outer, fast-moving layers. The deflagration models also tend to leave ''fingers'' of unburned carbon and oxygen extending down to the center of the explosion. My colleagues at the Naval Research Laboratory, Vadim Gamezo and Elaine Oran working with Alexei Khokhlov, have shown that deflagration-to-detonation models drive a detonation through the ''fingers'' of unburned matter left by the deflagration phase and through the outer layers. The result is to scour the unburned carbon from the ejected matter. Observations in the infrared are a powerful way to look for carbon. Observations and analysis by my colleagues here in Texas, Howie Marion and Peter Höflich, have shown that carbon seems to exist neither in central ''fingers'' nor in the outer layers of normal Type Ia supernovae. Some outer, high-velocity carbon is seen in some ''sub-luminous'' events, but this is naturally accounted for in deflagration-to-detonation models by triggering the detonation somewhat later. At this writing, the evidence seems to strongly favor some version of the deflagration-to-detonation models. The physics of when and why the explosion makes the transition from deflagration to detonation remains to be solved satisfactorily.

Convergence on deflagration-to-detonation models for the explosion does not, however, answer all the mysteries about the nature of Type Ia supernovae. For Type II supernovae, we think we understand the broad outlines of the evolution of massive stars to form collapsing iron cores. We do not understand how the collapsing core results in an explosion. For Type Ia supernovae, the situation is just the opposite. There is nearly unanimous agreement that the mechanism of Type Ia supernovae is a violent thermonuclear explosion that obliterates the star. Despite this convergence of opinion on the mechanism, there is no generally accepted picture of the evolutionary origin of these peculiar events. The question of how the white dwarfs grow to the Chandrasekhar mass is still a knotty, unsolved problem. There has been no direct evidence that Type Ia supernovae arise in binary systems. Despite this lack of direct evidence, all the circumstantial evidence points to evolution in double-star systems, and there are few credible ways of making a white dwarf explode

without invoking a binary companion. The challenge is to figure out what binary evolution leads to a Type Ia explosion.

New perspectives on the nature of Type Ia supernovae came with evidence produced in the 1990s that confirmed a long-standing suspicion. Type Ia supernovae are not all identical. They show interesting variations that are mostly subtle, but real. In some cases, the variations are not even so subtle. The general trend is that Type Ia supernovae that are brighter than average decline from maximum brightness a bit slower than average. The events that are a bit dimmer than average (some by as much as a factor of two) decline more rapidly. Models of exploding Chandrasekhar-mass white dwarfs can account for this behavior if the explosion in some stars makes the transition from a subsonic deflagration to a supersonic detonation a little earlier than in others. Why this should be so is the object of current research.

The observed variety of Type Ia behavior seems to correlate with the nature of the host galaxy. Elliptical galaxies seem to produce selectively Type Ia supernovae that are of the dimmer, more rapidly declining variety. Within spiral galaxies, the inner portions seem to produce the full range of behavior, but the outer parts of the galaxy produce especially homogeneous explosions. We do not yet understand all the variables, but there is probably a variety of ways of making white dwarfs explode, and the progenitor systems can display a range in ages. Some Type Ia supernovae may come from mass transfer in ''normal'' binary systems, from some variation on a cataclysmic variable. Others may come from merging white dwarfs. Some may come from stars near 8 solar masses that have relatively short lifetimes and others may come from stars with closer to 1 solar mass that have lifetimes approaching that of the Universe itself.

The task of figuring out the prior evolution of Type Ia supernovae is made harder if one accepts that the supernovae arise in white dwarfs of the Chandrasekhar mass. Recall from Chapter 5 that the average white dwarf has a mass of only 0.6 solar masses. This means that the mass must more than double if the process starts with one of these white dwarfs. The task might be made easier if the white dwarfs born in binary systems are systematically more massive. There is some evidence that this may be the case. Note that if the white dwarf is in a system that undergoes a classical nova explosion every 10 000 years or so, the mass of the white dwarf could actually decrease! This is not an easy problem.

For this reason, there has been considerable attention paid to mechanisms that would lead a white dwarf to explode, even though it had less than a Chandrasekhar mass. The most likely such model is one where a white dwarf accretes mass rapidly enough that the accreted hydrogen remains hot and supported by its own thermal pressure. The hydrogen then burns on the surface of the white dwarf in a regulated manner, and a nova explosion is avoided. Under these circumstances, however, a thick layer of helium can build up surrounding the inner carbon/oxygen core. The helium layer can be supported by the quantum pressure. If this helium ignites, computer models show that a violent explosion occurs. The explosion not only burns the helium but can send a shock wave inward that causes the inner carbon/oxygen white-dwarf core to burn as well. All this happens very quickly, a matter of seconds, so the result is a single powerful explosion. This is a very plausible mechanism to produce an explosion. The problem is that this mechanism does not produce results that are in good agreement with the observations. The helium burns to iron-like material on the outside that should be seen first and produces only thin layers of intermediate elements like silicon and calcium that are ejected with the wrong velocities. The ejecta tend to be too hot as well. Despite the appeal of these models, nature seems to prefer exploding white dwarfs of nearly the Chandrasekhar mass.

There are currently two ''best bets'' for how to generate Type Ia supernovae. Both involve mass transfer onto a white dwarf in a binary system. One invokes transfer of hydrogen from a red giant at just the right rate. The mass transfer must be rapid enough that the collected hydrogen does not undergo a nova explosion that ejects the hydrogen along with part of the white dwarf. Apparently, the mass transfer must be rapid enough that even the helium remains hot, supported by the thermal and not the quantum pressure, so that igniting the helium does not cause an explosion with the wrong properties. If the mass transfer is too rapid, however, a common envelope of hydrogen will engulf the white dwarf. The hydrogen should show up in the explosion. That would be a violation of the basic observational definition of a Type I supernova. There may be binary configurations where the mass transfer is ''just right.'' The hydrogen will burn gently to helium, the helium will burn gently to carbon and oxygen, and that carbon and oxygen will settle onto the core to cause the core to grow toward the Chandrasekhar mass. Candidate systems have even been identified among a special class of X-ray sources called *supersoft X-ray sources*.

An interesting clue to this problem was provided by the discovery by Mario Hamuy of Carnegie Observatories of a supernova that had obvious evidence for hydrogen, but, when one looked, an underlying spectrum that was that of a Type Ia. The hydrogen was apparently transferred from an ordinary red-giant companion. Polarization observations by our group showed that the hydrogen was distributed in an extended disk, as one might think appropriate for a strong mass transfer that slopped matter out of the binary system as well as onto the white dwarf. It is not clear how common this sort of explosion is, although a few other candidates have been identified. It is also true that while the hydrogen was totally obvious in this event, careful searches for wisps of hydrogen have failed to produce any evidence in normal Type Ia.

Another line evidence concerning the binary nature of Type Ia has been found by the recent discovery of high-velocity shells containing calcium that are somehow detached from the supernova ejecta. Where data has been obtained, these shells show polarization and hence some breakdown in spherical symmetry. My Texas colleagues Chris Gerardy (now at University College London) and Peter Höflich have argued that the calcium is in a shell otherwise composed of hydrogen (or perhaps helium) that preexisted in the binary system and was compacted and ejected by the supernova explosion. In models, the calcium radiates efficiently in the compacted shell and the hydrogen (or helium) radiates more feebly and remains invisible. This high-velocity calcium thus may be a clue to the nature of the binary system and hints that the system contains a hydrogen-rich star, even though the hydrogen is not directly detected. The swept-up matter revealed by its calcium emission may come from an accretion disk, from the companion star, or perhaps from matter that was previously part of a common envelope that still lingers nearby.

The other popular model for producing a Type Ia supernova is by the merging of two white dwarfs in a binary system (Chapter 5, Section 5.4). This merging must happen sometimes. Some binary white dwarfs are seen. There is still controversy concerning whether there are enough binary white-dwarf systems with total mass exceeding the Chandrasekhar mass to produce Type Ia supernovae at the observed rate. In addition, the process by which the smaller-mass white dwarf fills its Roche lobe and comes apart, dumping its mass on the larger-mass white dwarf as described in Chapter 5, is complex and not well understood. The disrupted matter will swirl around the

larger-mass white dwarf in a thick disk. How that matter will settle onto the remaining white dwarf is not completely clear.

Yet another way to pursue evidence that Type Ia explode in binary systems is to look for the left-overs; not a compact remnant, but the companion star that would be left behind if the explosion occurs in a mass-transferring binary system. The matter in stars is rather concentrated toward their centers and that makes them tough. A nearby supernova could strip off some matter from the outside, but a companion star will easily survive the explosion. On the other hand, the companion star will be released from its orbit when the binding gravity of its companion disappears in the explosion. The companion should thus be slung out of the site of the explosion. The companion might be a rather normal little red main sequence star as observed in many cataclysmic variable systems, but there are billions of them in the Galaxy, so identifying the companion is not a simple thing to do. Pilar Ruiz-Lapuente from the University of Barcelona and her colleagues focused on the remnant of Tycho's supernova. Using images from the *Hubble Space Telescope*, they did not find any red giants that could be the companion. But they did identify a yellow star much like our Sun that is moving out of the vicinity of the explosion at about three times the average speed of other nearby stars. They suggest that this star is the surviving companion of Tycho's supernova.

The accumulating clues thus suggest that Type Ia do arise in binary systems and that the most common configuration involves mass transfer from a relatively normal companion star. White-dwarf mergers might contribute to some small fraction of Type Ia explosions, but there is no firm evidence for that at this time.

6.7 LIGHT CURVES: RADIOACTIVE NICKEL

Supernovae display a variety of shapes to their light curves. Type Ia supernovae are the brightest. They decay fairly rapidly in the first two weeks after peak light and then more slowly for months. Some Type II supernovae have an extended plateau and some drop rather quickly from maximum light. Both types seem to have a very slow decay at very late times, several months after the explosion. Types Ib and Ic supernovae are typically fainter than Type Ia by about a factor of two, but they have similar shapes near peak light and show evidence for a slow decay at later times. These patterns tell us something about the star that exploded and about a fundamental process that is probably taking place in all of them: radioactive decay.

When a supernova first explodes, the matter is compact, dense, and opaque. To reach maximum brightness, the ejected matter must expand until the material becomes more tenuous and semi-transparent. The size the ejecta must reach is typically 10 000 times the size of the Sun. This is 100 times the size of a red giant and 100 times the size of the Earth's orbit. As the matter expands, however, it cools. If the matter must expand too far before heat leaks out as radiation, the material may have cooled off so that there is no more heat to radiate.

Most Type II supernova explosions are thought to occur in red-giant envelopes. These are very large structures. After the explosion, large envelopes do not have very far to expand before they become sufficiently transparent to leak their heat as light. As they begin to radiate, Type II supernovae still retain a large proportion of the heat that was deposited by the shock wave that accompanied the supernova. Near maximum light and on the typical plateau that lasts for months, Type II supernovae shine by the shock energy originally deposited in the star. The deposited energy presumably arises in the core-collapse process.

For a Type I supernova, however, the story is different. Whether the exploding star is a white dwarf, as suspected for a Type Ia, or the bare core of a more massive star, as suspected for Types Ib and Ic, the exploding object is very small. The expected sizes range from one-tenth to one-thousandth of the size of the Sun. These bare cores are vastly smaller than the size to which they must expand before they can leak their shock energy. The result is that the expansion strongly cools the ejected matter, and by the time the matter reaches the point where it could radiate the heat, the heat from the original shock is all gone. This kind of supernova requires another source of heat to shine at all. All the light from Type I supernovae comes from radioactive decay.

The nature of a thermonuclear explosion is to burn very rapidly. If the explosion starts with a fuel built from multiples of helium nuclei – carbon, oxygen, or silicon – that has equal numbers of protons and neutrons, then the immediate product of the burning will also have equal numbers of protons and neutrons. This is because the rapid burning takes place on the timescale of the strong nuclear reactions. To change the ratio of protons to neutrons requires the weak force and thus a longer time. Nature, however, does not leave the burned matter with equal numbers of protons and neutrons. Rather, Nature prefers to form the element with the most tightly compacted nucleus, that of iron, which has 26 protons and 30 neutrons.

Nature manages to make iron in a thermonuclear explosion in a three-step process. The first step is to forge an element that is close to iron but that has equal numbers of protons and neutrons. This element, like iron, has a nucleus that is tightly bound by the nuclear force and has the same total number of protons plus neutrons, 56, but with 28 protons and 28 neutrons. This is the element that will form first, before the slower weak interactions come into play. This condition singles out one element, nickel-56. The unregulated burning of carbon or oxygen or silicon will naturally first produce nickel-56.

Nickel-56 is, however, unstable and therefore undergoes radioactive decay. The radioactive decay is induced by the weak force. One of the protons in the nickel converts to a neutron. The result is the formation of the element cobalt-56 with $28 - 1 = 27$ protons and $28 + 1 = 29$ neutrons. In the process, an electron is absorbed to conserve charge, and a neutrino is given off to balance the number of leptons. Excess energy comes off as gamma rays, high-energy photons. The gamma rays can be stopped by collision with the matter being ejected from the supernova and their energy used to heat the matter. The hot matter shines as the light we observe on Earth. The power of the light falls off as the nickel decays away and as the matter expands, so that it is less efficient in trapping the gamma rays. The neutrino always just leaves the star and plays no role in this heating.

The cobalt-56 that forms is also unstable. Again, the weak force induces a proton to convert to a neutron. The result has $27 - 1 = 26$ protons and $29 + 1 = 30$ neutrons. This is just good old iron-56, Nature's ultimate end point. This decay again produces a neutrino and gamma-ray energy. In this case, charge is conserved by emitting an antielectron, or positron. The positron will quickly collide with one of the electrons that are floating around normally, one for every proton. The annihilation of the electron will produce another source of gamma rays. Iron-56, with 26 protons and 30 neutrons, sits at the bottom of the nuclear energy valley, and so it is stable. This radioactive decay scheme, nickel to cobalt to iron, is just one of nature's ways of rolling things down the nuclear hillside to become iron.

The radioactive decay of these elements is controlled by a quantum uncertainty. One does not know what atom will decay, but on the average half will decay in a given time. For nickel-56, the time for half to decay is 6.1 days. After another interval of 6.1 days, half of the remaining half will decay, so that after 12.2 days only one-quarter of the original nickel remains. After 18.3 days, only one-eighth of the

original nickel will survive. This timescale, about a week, is the time for the gamma rays from the radioactive decay to pump energy into the exploding matter. Likewise, the cobalt-56 decays with a half-life of about 77 days, roughly 2 months. These times are long compared with the times for the basic explosion to ensue, a matter of seconds. That is why the nickel-56 forms first in this type of explosion and the iron forms only later, over several months. The observed light curves of Type I supernovae decay somewhat faster than the decay of nickel-56 in the early phase and of cobalt-56 in the later phases. The reason is that not all the gamma rays produced in the decay are trapped and converted to heat and light. Some of the gamma rays escape directly into space.

For Types Ib and Ic, the amount of nickel required to power the light curve is about one-tenth of the mass of the Sun. This amount of nickel is consistent with many computations of iron core collapse. The nickel is produced when the shock wave, of whatever origin, impacts the layer of silicon surrounding the iron core. Type Ia supernovae are generally brighter and must produce more nickel, of order 0.5–1 solar mass. The dimmest Type Ia events require only 0.1–0.2 solar mass of nickel. The models of Type Ia supernovae based on thermonuclear explosions in carbon/oxygen white dwarfs of the Chandrasekhar mass produce this amount of nickel rather naturally in the explosion. The amount can vary depending on, for instance, the density at which the explosion makes the transition from a deflagration to a detonation, so the variety of ejected nickel mass can also be understood, at least at a rudimentary level.

If Types Ib and Ic are related to the cores of massive stars, as the circumstantial evidence dictates, then their explosion mechanism should be similar to that of Type II supernovae. This suggests that Type II should also eject about 0.1 solar mass of nickel-56. This is not enough to compete with the heat and light from the shock near maximum light, but as the ejected matter continues to expand and cool, the shock energy dissipates, and the supernova gets dimmer. At this phase, the dimmer but steady source of radioactive decay should take over. The evidence from fading Type II supernovae shows that this is the case. Once again, not all the gamma rays are trapped. Some must radiate directly into space. A properly designed gamma-ray detector flown in orbit should see these missing gamma rays and directly confirm the validity of this picture. As we will see in Chapter 7, this was the case for SN 1987A.

When Betelgeuse blows

For years, every time I gave a popular lecture on supernovae, someone would ask, ''What will happen to the Earth when a nearby supernovae explodes.'' Each time I would say, ''I thought about that a little a long time ago, but I really need to work that out, so I know how to answer this question.'' Then after the lecture, I would return to work-a-day issues and forget until the next popular lecture. To get a record down on paper that I can use in the next lecture, here is a sketch of what will happen when the most likely nearby star explodes.

Betelgeuse is a red-giant star that marks the upper-leftmost shoulder of the constellation of Orion as we look at it from Earth. You can see it easily from anywhere in the northern hemisphere on a winter or spring evening. We do not know the precise mass of Betelgeuse, but we can make an intelligent guess. That will give us a good idea as to its fate and what will happen at the Earth.

Thanks to careful measurement by triangulation we know quite accurately how far away Betelgeuse is. It is 427 light years away. That is long by human standards, but right next door in a Galaxy that is 100 000 light years across. There are closer stars, but none that are likely to explode. At this distance, Betelgeuse presents little threat to the Earth, but we will sure notice it when it goes off. It is a good example of the low-level impact that will contribute to the stochastic history of bombardment of the Solar System by astronomical events over its 5-billion-year history. Such events should occur roughly once every million years.

From the power received at Earth over all wavelength bands and its distance, we can estimate that Betelgeuse emits a luminosity of about 50 000 to 100 000 times that of the Sun. From computer models, we can further estimate that this luminosity in a red giant requires a star of original main sequence mass of about 15–20 solar masses. This mass is such that, in the absence of a stellar companion, and Betelgeuse seems to have none, there will be little mass loss to winds, so this is probably a pretty good estimate. Stars in this mass range are predicted to evolve iron cores and undergo core collapse to form a neutron star and an explosion. Betelgeuse is nearly a canonical candidate for a Type II supernova explosion. We do not know exactly when it will explode. The final stages after a star of this mass becomes an extended red giant are

typically no more than 10 000 years. We do not know when in the next 10 000 years it will explode (it may be tomorrow!), but we can estimate the progression of events when it does.

Upon core collapse, Betelgeuse will emit 10^{53} ergs of neutrinos, each with an energy characteristic of a nuclear reaction. This burst of neutrinos will take about an hour to pass through the hydrogen envelope and into space. They will arrive in the Solar System 427 years later and be the first indication that Betelgeuse has erupted. These neutrinos will deliver about 2×10^8 recoils in the body of a 100-pound woman. This effective level of radiation exposure is far less than a lethal dose (by a factor in excess of 1000, depending on how the energy is actually deposited) but might cause some chromosomal damage. The shock wave generated by the collapsing core and the formation of a neutron star will require about a day to reach the surface. The breakout of that shock will generate a flash of ultraviolet light for about an hour that will be about 100 billion times brighter than the total luminosity of the Sun. This burst may not exceed the ultraviolet light from the Sun at the Earth, but could affect life on outer satellites if there is any, or any explorers from Earth, if we have ventured far from the Sun by the time this happens. This blast of ultraviolet light might cause some disruption of atmospheric chemistry. The ejecta of the supernova will expand and cool after shock breakout, and the total luminosity will first dim and then rise to maximum in about 2 weeks as the supernova material expands to about 100 times the Earth's orbit, and the photon diffusion time through the expanding matter becomes comparable to the time required for appreciable expansion of the matter. The total luminosity will then be about a billion times that of the Sun. At its distance, Betelgeuse will be a factor of about one million dimmer than the Sun, magnitude −12, about the same as a quarter Moon. This phase will last during the ''plateau'' phase of the light curve, 2 or 3 months. The observed surface of the supernova during this interval will be roughly constant at an effective temperature of about 6000 K, slightly hotter than the Sun. After the hydrogen envelope has expanded and electrons and protons have all recombined to make neutral hydrogen atoms, the envelope will be nearly transparent, and the light curve will begin a rapid decline.

In a typical supernova of this type, the emission is dominated for the next year or so by radioactive decay of cobalt to iron (nickel will have already decayed away). The expanding

envelope of hydrogen is likely to remain opaque to these gamma rays until substantial decay has occurred, so such an event is unlikely to provide a substantial source of gamma rays. If Betelgeuse produces a bright pulsar (Chapter 8), it might be a substantial source of gamma rays for thousands of years.

The ejecta from Betelgeuse will freely expand for about 1000 years and span about 20 light years in that time. During this time, the ejecta will be cold and dim. The supernova material will then start to pile up appreciable mass in interstellar matter and enter the supernova remnant phase. The supernova remnant will turn on as an X-ray source and begin to produce cosmic rays by acceleration of particles at the shock front. The supernova material will slow down, but a shock will race ahead into the interstellar matter, decelerating as it sweeps up ever more mass. The shock wave in the interstellar matter will be fully developed in about 20 000 years when it has expanded to about 30 light years. The shocked matter will begin to radiate substantially and cool off when it has expanded to about 100 light years, about 100 000 years after the explosion. The remnant will plow on through the interstellar matter. The shock from Betelgeuse will be very mild by the time it reaches the Solar System and will probably be easily deflected by the solar wind and magnetopause. The exception might be if there is a low-density, interstellar ''tunnel'' between us and Betelgeuse that would channel some of the energetic matter to us before it slowed down.

All these effects would be much stronger if the supernova were only 30 light years from the Earth. There are no candidate stars around us now, but on its galactic journey, such nearby explosions have probably happened several times in the 5-billion-year life of the Earth. Such events could be dangerous by triggering harmful mutations, but they might also be helpful because evolutionary ''shocks'' can also single out healthy mutations and drive biocomplexity. The Earth is coupled to this complex galactic environment, and the story of life on Earth will not be fully known until such long-term, sporadic effects are understood.

7
Supernova 1987A: lessons and enigmas

7.1 THE LARGE MAGELLANIC CLOUD AWAKES

The first supernova discovered in 1987 turned out to be the most spectacular supernova since the invention of the telescope. SN 1987A was the first supernova easily observable with the naked eye since the one recorded by Kepler in 1604. This event also brought the first direct confirmation that our basic picture of the exotic processes that mark the death of a massive star is correct. SN 1987A is the best-studied supernova ever, but the story is still unfolding, and there is much to learn.

SN 1987A did not explode in our Galaxy, but in a nearby satellite galaxy to our own Milky Way galaxy. This satellite galaxy cannot be seen from the northern hemisphere. The first European to record it was Magellan during his epic attempt to sail around the world. In English, it carries the name of the Large Magellanic Cloud for this reason. People native to the southern hemisphere were undoubtedly familiar with it before that. The Aborigines living around Sydney had long had another name for it: Calgalleon, which had to do with a woolly sheep. The Large Magellanic Cloud has a somewhat smaller companion that has picked up the unimaginative name, Small Magellanic Cloud. In the same Aboriginal dialect, it was rendered Gnarrangalleon. There is poetry!

The Large Magellanic Cloud is only 150 000 light years away, as shown in Figure 7.1. This is not much farther than the span across the Milky Way itself, about 50 000 light years. By contrast, the Andromeda galaxy, Messier 31, the great sister spiral galaxy to the Milky Way in our local group of galaxies, is about 2 million light years away. The nearest rich cluster of galaxies that has provided many well-studied supernovae in the last several decades is about 50 million light years away. The most distant supernovae ever found are more than a billion

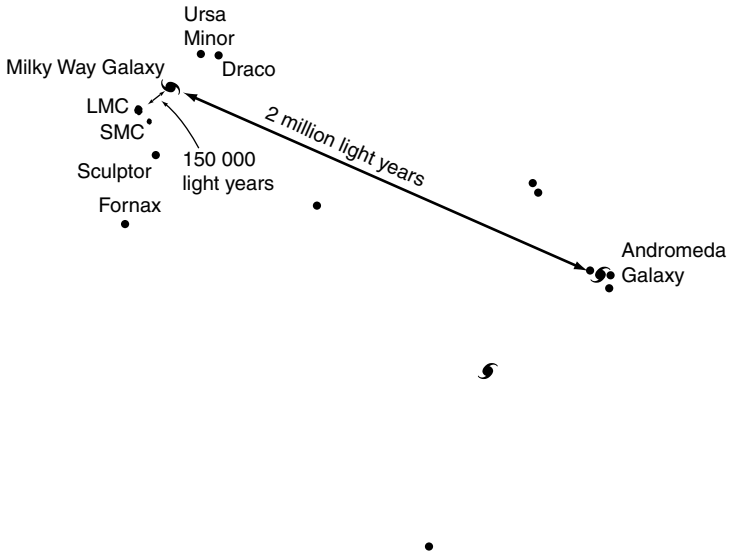


Figure 7.1 A schematic sketch of some of the 21 galaxies known to exist in the local group. These galaxies are distributed in three dimensions. This perspective corresponds to looking approximately along the plane of our Galaxy. The great Andromeda spiral galaxy is about 2 million light years away. By contrast, the Large and Small Magellanic Clouds are very close. The 150,000 light years to SN 1987A in the Large Magellanic Cloud was not much farther than one end of our Galaxy is from the other.

light years away. The nearness of the Magellanic Cloud was responsible for the great apparent brightness of SN 1987A. Intrinsically, it was relatively dim as supernovae go.

The known distance to the Large Magellanic Cloud gives us another perspective. The supernova actually exploded about 150 000 years ago, before modern *Homo sapiens* walked the Earth. By an incredible piece of luck, the light arrived at Earth just as our science had developed to the point where we could read many of its most important messages. We had to crawl out of our caves, invent fire and the wheel, develop agriculture and writing, and witness the flowering of Greece, the Middle Ages, the Renaissance, and the Industrial Revolution. We had to develop modern science, quantum theory, Einstein's theory, an understanding of the way stars work, and the techniques for detecting neutrinos and get all this done before the light arrived! Whew!

On the other hand, if the supernova had been a mere 100 light years farther away, technology would have advanced, and we might

have learned vastly more from it. On a personal note, if I had known that the light from the supernova were encroaching on the orbit of Pluto in September of 1986, I might not have agreed to be the Chair of my department that fall. By the next spring, I felt as if I were trying to drink from two fire hoses at once.

The Large Magellanic Cloud is neither a spiral nor an elliptical galaxy. Rather it is classified as an irregular galaxy. It has a large central band of rather young, newly formed stars, but then a more distended array of older stars. Off to one side of the central band, there is a region of especially intense recent star formation. The highlight of this region is called 30 Doradus by astronomers, or the Tarantula nebula by star gazers for the ''hairy'' arms of gas that extend from the center. The 30 Doradus region contains a very young cluster of very massive stars, perhaps 100 solar masses apiece. Surrounding the middle of 30 Doradus are large patches of gas and dust and other young massive stars, somewhat older than the core cluster of 30 Doradus. By careful study of the stellar ages, astronomers have been able to track propagating swaths of star formation in the region. One of the stars left behind in a prior wave of star formation became SN 1987A. Despite the obvious evidence for ongoing star formation, the Large Magellanic Cloud is relatively immature, in the sense that it has not processed as much of its gas through stars as has the Milky Way. The amount of heavy elements in the Large Magellanic Cloud is only about one-quarter of that in our Sun.

7.2 THE ONSET

SN 1987A was discovered and first formally reported on February 23, 1987, by Ian Shelton, a graduate student from the University of Toronto who was using a small telescope at the Las Campanas Observatory, high in the Chilean Andes. The first person to notice it may have been one of the night assistants, Oscar Duhalde, a Chilean of Basque extraction (Figure 7.2). Oscar had worked on the mountain for years and was justifiably proud of his familiarity with the southern sky. He stepped out of the dome for a cigarette and looked at the Large Magellanic Cloud. He noticed that there was a new light in 30 Doradus but did not remark to anyone at the time about it. The supernova was still faint at the time, only hours old, and Duhalde's note of it remains one of the remarkable parts of the story. Half a world away in Australia, Rob McNaught was working on his routine survey of the sky for asteroids. He was especially tired that evening and went to bed

Figure 7.2 Photo of the author and Oscar Duhalde at the site of Ian Shelton's original discovery at Las Campanas Observatory at the time of the tenth anniversary of the discovery of SN 1987A. (Photo courtesy of the author.)

without developing his plates. He awoke the next day with the astronomical world full of news of Shelton's announcement and found, when he did develop his image, that he had the first permanent recording of the light from the supernova. Who knows how many other people might have seen something and not mentioned it. There were rumors, but none were confirmed. Figure 7.3 shows a series of photos taken by McNaught with his patrol camera as SN 1987A appeared, brightened, and dimmed over the course of several months.

(a)

(b)

(c)

(d)

(e)

Figure 7.3 Series of photos of SN 1987A taken by Rob McNaught. The first was taken on February 22, 1987, the day before the supernova. This photo shows the broad central band of newly formed stars in the Large Magellanic Cloud. The entire galaxy is much bigger than the scale encompassed by this photo. In the upper middle is the Tarantula nebula or 30 Doradus, and to the lower right of that the supernova is in the final stages of silicon burning and near to undergoing core collapse. The second photo was taken on February 23, when the supernova was only hours old. The neutrinos were long gone, but the shock wave had only recently broken through the outer layers of the star, and the supernova was brightening rapidly. This was when Oscar Duhalde noticed it. The next photo is from February 24, when the supernova was a day old. By this time, Ian Shelton had made his discovery, and the world was awakening to the amazing event. The next photo was taken on May 20 when the supernova was near maximum brightness. The image of SN 1987A does not look much brighter than the other photos because the exposure was shorter. Note that the main bar of stars and 30 Doradus look fainter in contrast and the supernova stands out clearly. The last photo was taken on August 23, as the supernova was fading. This is the time when I saw the supernova (see box) and received this precious set of slides from Rob McNaught. (Photos by Rob McNaught.)

Seeing SN 1987A

I was one of the first people to hear about the supernova in the northern hemisphere. One of our ex-graduate students, Marshall McCall, was at the University of Toronto when the news came in from Ian Shelton. Marshall promptly called me. I called Nino Panagia who had used the *International Ultraviolet Explorer* satellite to study previous supernovae. Then I called Bob Kirshner at Harvard, perhaps the preeminent supernova observer of the time. I think Bob has never quite forgiven me for calling him second. Bob was also suspicious because I had been around at a meeting in Sicily in 1978 when a wonderful prank was played on him, pretending to bring news of a supernova in Andromeda. I was completely uninvolved in that prank, but guilty by association. Bob's first reaction was that I was pulling his leg. After my call, he went down the hall to the Center for Astronomical Telegrams and found their teletype spewing news of the supernova, although no one had bothered to tell him. I think he was irritated at that, too.

One of my first reactions to the supernova was to try to think of a way to go see it. This was reinforced by one of my colleagues, Don Winget, who said, ''Craig, you will die a bitter old man if you don't see this supernova for yourself.'' Upon more reflection, I decided that I could be of more use by staying in Austin and trying to contact as many people as possible in the southern hemisphere to alert them to the event and helping to guide observations. I am not an observer myself. I did have some experience in trying to coordinate observations of supernovae at McDonald Observatory and few observatories at the time had any experience in observing supernovae.

One of the first things I did was to consult with Brian Warner, an astronomer visiting Austin from South Africa. We communicated with his colleagues who were beginning to make observations. One of the things I had learned was that if one looked at crude data when it first comes off the telescope, there was some danger of mistaking the strong spectral line of hydrogen that is prominent in Type II supernovae with the strong silicon line that is characteristic of Type Ia supernovae. Some people had mistaken Type Ia for Type II on this basis. I tried to issue this caution to my South African colleagues. They had data showing excess emission in this tricky region of the spectrum. I

merely meant to be careful in the identification when they said they thought it was hydrogen. Somehow this came across in the tense rush of those first few hours as a statement that their feature was not hydrogen, but silicon, and that they were looking at a Type Ia. They announced that. Meanwhile other astronomers had done a quick and dirty analysis and recognized that they were, indeed, looking at hydrogen and announced, correctly, that SN 1987A was a variety of Type II supernova. I think some of the South Africans still hold a mild grudge against me for that.

I also thought that the supernova might emit X-rays. A few supernovae had done so, but there was no clear understanding of the mechanisms and timing of the X-rays. It did seem that if there were going to be X-rays, it was important to look very early in the explosion when the ejected matter was hot and bright. I called Walter Lewin, an X-ray astronomer at MIT. Walter pointed out that the Japanese had just launched a new X-ray satellite called *Ginga*, meaning galaxy in Japanese. Walter said that I should call Professor Minoru Oda, the scientist who was the head of the *Ginga* team. I looked at my watch and we did a quick calculation. It was one in the morning in Tokyo. Walter said, ''If I were you, I would call him.'' I noticed that Walter did not volunteer himself to make the call. I decided, what the heck, once in 400 years, it was worth the disruption. I got Oda's home number from Walter and rang him up. His wife answered, very sleepy, but very polite. I have the feeling she had handled emergencies before, if not one quite like this. She put Professor Oda on the phone, and I tried to explain the circumstances as best I could. No one could be sure the supernova was producing X-rays, but looking at it with *Ginga* was the only way to find out. Professor Oda thanked me and hung up. I heard years later that Professor Oda had his own version of this story of ''some crazy American calling him in the middle of the night.'' Fortunately, he did not remember who it was. As it turned out, there were no X-rays to be seen in those first few days, so I could have waited until it was a civilized time in Tokyo to call. *Ginga* did see X-rays a few months later, a detection that revolutionized some of our ideas about the supernova.

I did get a chance to see the supernova myself. Our Japanese colleagues added the topic of SN 1987A to a previously scheduled meeting in Tokyo in August of 1987, which was six months after the discovery. The reasonable thing to do seemed to be to go to Tokyo by way of Australia. I went with my colleague, Robert

Harkness, an expert on the theoretical supercomputer calculations of radiation from supernovae. Robert is also an expert on airplanes. He knew all about the Qantas stretch 747 that we flew from Los Angeles to Sydney. He had also learned from Brian Warner that Brian had been able to see SN 1987A from the window of the upper-level, first-class lounge for which 747s were so famous.

On the other hand, Robert cannot sleep on airplanes. I can. I had a nap while Robert sat in his seat. I woke up for a meal and then slept again. Robert ate little and sat some more. I awoke feeling great while we were in the middle of our 14-hour flight to Sydney. Although Robert was a bit out of sorts by this time, I asked the flight attendant if we could venture into the upstairs lounge to try to get a peek at the supernova. She asked the captain and he, in turn, invited us, not into the lounge, but onto the flight deck.

So up we scrambled to meet the crew of relatively young Australians, the pilot Jeff Chandler, the copilot, and the navigator. I'm sure this would not have happened on an American airline, and I'm not sure it was strictly legal on Qantas. In any case, the crew were fairly bored from the long flight and keen on the distraction we provided. We asked whether they knew where the Large Magellanic Cloud was. The navigator laughed and replied he had no idea. They flew by computer and never looked at the stars. Robert, no observational astronomer himself, then leaned down and peeked out the window next to Captain Chandler and announced, ''There it is!''

Indeed, our flight path was such that the Large Magellanic Cloud was at about 10 o'clock from the nose of the aircraft, easily seen out the captain's left window. It was not trivial to see the supernova. Although it was still fairly bright, it had faded from maximum. My admiration for Oscar Duhalde and what he noted in those first few hours went up. I had brought along some binoculars. With them, I could make out the bright dot of light next to 30 Doradus.

Then Captain Chandler had an idea. He said that fresh oxygen helps visual acuity. He pulled his oxygen mask from its holder. This was not a full-face mask, but tubing that was more reminiscent of the oxygen lines for patients in hospitals. There was a framework that supported the thing over your ears. We spent the next 10 minutes passing around the mask and

binoculars. The drill was to take the mask, snort a few deep drafts of oxygen, then rip off the mask (and in my case eye glasses), hold up the binoculars, and peer at the supernova. Frankly, I could not tell that it made any difference, but it sure was amusing! These were not, perhaps, ideal circumstances, but I can say that a few optical photons from the degraded gamma rays from the radioactive decay of supernova-created cobalt made it into my very own retinas. I may die a bitter old man, but it won't be for lack of seeing this remarkable event.

Robert and I spent a couple of days in Sydney among the city lights where viewing the supernova was not practical. We then proceeded to Canberra, site of Mount Stromlo Observatory and the location of the small meeting that was our excuse for this Australian junket. I gave a public talk that first night. I mentioned my curiosity about the native names for the Magellanic Clouds and the next day got a call from a gentleman by the name of Edward Wheeler, no relation that we could identify. He provided me with the names for the Large and Small Magellanic Clouds according to one of the dialects spoken around Sydney when the first British settlers arrived in 1798. The Aborigines speak some 500 languages, so possibilities for other wonderful names like Calgalleon and Gnarrangalleon are enticing. Afterward, there was a clear night, but Robert and I were still exhausted from our trip (and a couple of late nights in Sydney), so we made no attempt to see the supernova that evening. That would have required staying awake until two a.m. We had a beer with our host, Mike Dopita, and went to bed.

It clouded up that night. The patch of clouds did not cover all of Australia, but only that fraction we were destined to visit: Canberra, Sydney, and the other major observatory, the Anglo-Australian Observatory at Coonabarabran in the north. By the time we got to Coonabarabran, we were aware that our chances were slipping away. Both Robert and I awoke on the mountain top and watched fog blow over, opening occasional ''sucker holes,'' but never giving a good view of the sky, never mind the Large Magellanic Cloud. We talked a little desperately of getting a car and driving down off the mountain because there was some thinking that the fog might be a localized, mountain-top phenomenon. The bottom line was that we left Australia the next day, having never seen the supernova from the ground. Thank goodness for that Qantas crew.

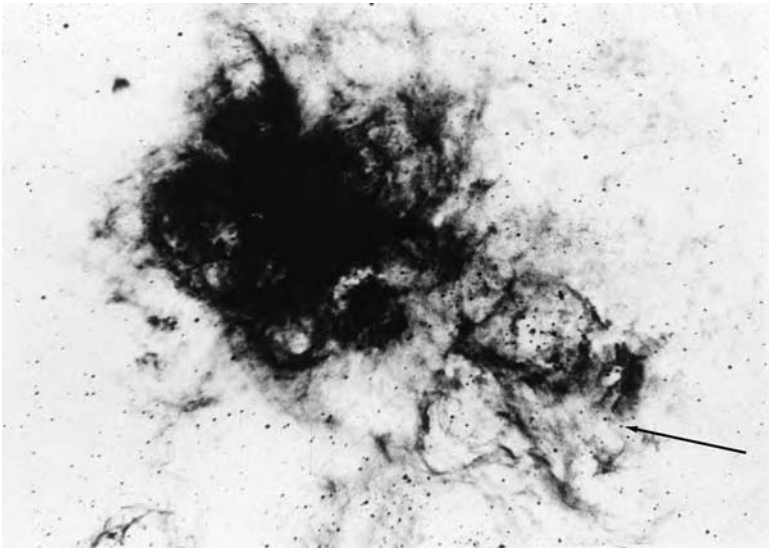


Figure 7.4 Photographic negative of 30 Doradus and Sk-69 202. The black dot at the tip of the arrow is Sk-69 202, soon to become SN 1987A. (Photo by You-Hua Chu.)

7.3 LESSONS FROM THE PROGENITOR

SN 1987A is one of a very few supernovae for which there is any evidence of the star that existed before it exploded. The star was seen in photographs taken for other purposes. It was listed in a catalog of hot stars in the Magellanic Clouds compiled by Norman Sanduleak. The star that exploded was listed by its position in the sky and known as Sk-69 202. You can make it out if you know where to look in Figure 7.4.

Sk-69 202 was not well studied. It was on a list of stars that German astronomer Rolf Kudritzki was investigating intensively, one by one, but it blew up just before Rolf got to it. There is some scientific import to the lack of attention drawn to the star. As Peter Conti, a hot-star expert from the University of Colorado, remarked, there was nothing special about Sk-69 202. It did not vary in light output. It did not have any anomalous emission lines. It did not seem to be shedding mass at an especially noticeable rate or in a special way. There was simply no hint at all that Sk-69 202 was special until it disappeared in a violent flash of light. We still do not know why that was so.

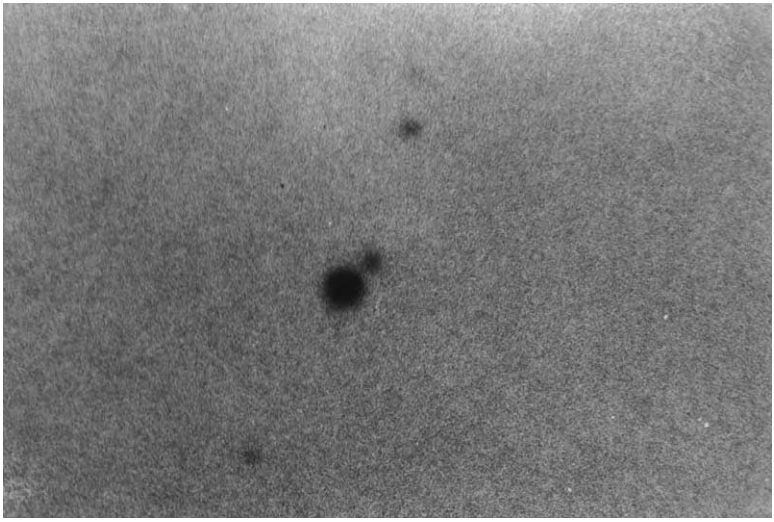


Figure 7.5 Image of Sk-69 202, the progenitor of SN 1987A. Note Star 2 at the upper right, about 2 o'clock, less than one diameter away from the main, dark spot in this negative image. Star 3 is revealed as a slight blurring of the image of Sk-69 202 in the lower left, about 7 o'clock in this orientation. (Photo by You-Hua Chu.)

A blown-up photographic image of Sk-69 202 is shown in Figure 7.5. The original, larger-scale photo was taken for other reasons, part of a study of star formation in the vicinity of the 30 Doradus nebula, by You-Hua Chu of the University of Illinois. The big dark patch in the center of Figure 7.5 is Sk-69 202. It is just a point of light, but it looks big because the photographic process smears out the image. The brighter the star, the more intense and the larger the image. This also became known as Star 1, the star that blew up. To the upper right in this image is what is known as Star 2. This is another star in the Large Magellanic Cloud. It is somewhat less massive than Sk-69 202 was. It is not physically or gravitationally close to Sk-69 202 – it is several light years away – but it was probably born in the same burst of star formation that gave rise to Sk-69 202 and other fainter stars in this image. Dr. Chu gave me this slide when I went to Champaign-Urbana to present an already-scheduled colloquium on another topic about a week after SN 1987A erupted. She saw something in the photo that was part of a story that played out over the next few months.

When SN 1987A first went off, the vicinity of the supernova shown in Figure 7.5 was lost in the intense glare of the explosion. SN 1987A faded first in the ultraviolet. As it did, Star 2 in Figure 7.5 could

be identified. The surprise was that something was also left behind at the location of Star 1 in the images from the *International Ultraviolet Explorer* satellite, the only ultraviolet instrument available at the time of the explosion. The lingering ultraviolet image left some people wondering whether the wrong progenitor star had been identified. What You-Hua Chu had recognized was that the lower left part of the image in Figure 7.5 was somewhat blurry. She was sure there was a third star there, Star 3, that was obscured by the brighter, smeared image of Sk-69 202 in Figure 7.5. As SN 1987A continued to fade, careful positions were measured, and it was determined that the lingering image was not at the location of Star 1, but slightly offset. There was, indeed, a third star, Star 3. Both Star 2 and Star 3 show up clearly in later images taken with the *Hubble Space Telescope* after the supernova faded (see Figure 7.6). Other people got more credit for resolving this mystery at the time, but there is no question in my mind that Chu knew of the existence of Star 3 within days of the explosion. She scored another coup a decade later, at a meeting in Chile to celebrate the tenth anniversary of the discovery of the supernova, when she reported that she had discovered the first star to have rings around it, like the progenitor of SN 1987A (Section 7.8).

From preexplosion observations such as Figure 7.5, we know that Sk-69 202 had a mass of about 20 solar masses. This follows from knowing the luminosity. The luminosity is a clue to the mass of the evolved helium core, even though that core was buried in a surrounding hydrogen layer. From our knowledge of stellar structure and evolution, we can then estimate the mass that the star originally must have had to make such a massive core. The luminosity suggests that the core was about 6 solar masses, and such a core arises in a main sequence star of about 20 solar masses. The star shed some mass while evolving. The best estimates are that the star retained about 15–18 solar masses by the time it exploded.

Somewhat surprisingly, the star that exploded was not a red supergiant, as might have been expected given the basic theory of stellar evolution and the observation that there are many red giants of 20 solar masses in the Large Magellanic Cloud. Instead, the star was relatively compact and blue, a blue supergiant. The reasons for this are still not fully understood. The relatively small size produced an unorthodox and somewhat dim light curve. The light curve is by now well understood, given the starting conditions of the star when the explosion erupted. A legion of computer models based on single stars has been calculated in the attempt to understand the compact starting

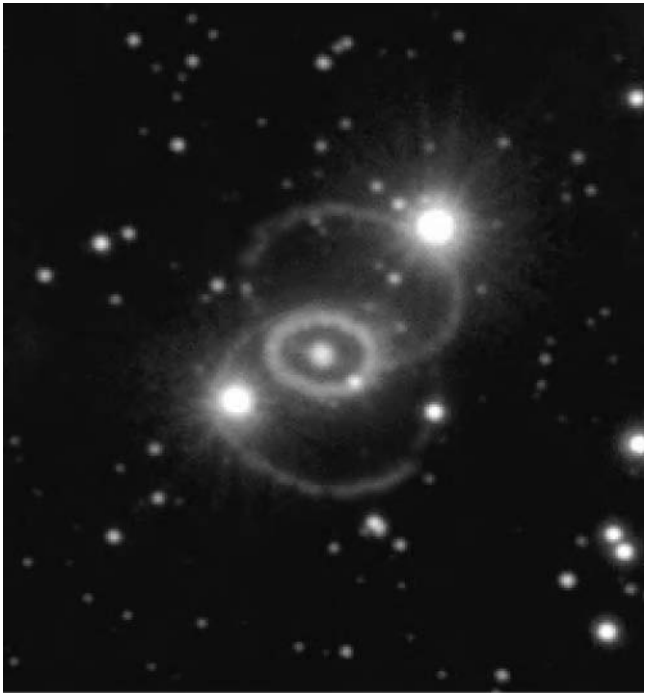


Figure 7.6 The rings of SN 1987A. These three rings are thought to compose parts of an ''hourglass'' shape with the smaller, brighter central ring the ''neck'' of the hourglass and the two larger rings the upper and lower ''rims,'' all seen at a tilt of about 40 degrees. The matter that forms these rings (and other structures not shown here) was shed from the progenitor star before it exploded and was illuminated by the light from the explosion. Note that the rings are not exactly colinear; the edge of the upper ring passes across the central ejecta, whereas the lower ring circumscribes the ejecta and the inner ring. Star 2 in Figure is just beyond the upper ring to the upper right. Star 3 is the image to the lower left just inside the lower ring. The smaller bright dot just opposite Star 3 directly on the rim of the lower ring is yet another star in the Large Magellanic Cloud. (*Hubble Space Telescope* photo by NASA.)

conditions, none of them entirely satisfactory. These models based on single stars may be wrong. The current hot idea is that Sk-69 202 might have been in a binary in which the companion was engulfed in a common envelope and dissolved, leaving only one star to explode. This process might have caused the envelope of the progenitor to contract to a smaller radius and produced some of the other special features of SN 1987A that we will discuss in Section 7.8. There is no

definite sign of any current companion, but that is consistent with none ever having existed, or the companion having been consumed by the supernova progenitor.

7.4 NEUTRINOS!

SN 1987A brought us a wealth of new understanding, but the single most important aspect was the burst of neutrinos that were detected from Earth. SN 1987A generated about 10^{57} neutrinos. Most of these went off in directions away from the Earth. Only a tiny fraction arrived at the Earth, and of this number only a tiny fraction interacted with the detectors so that their presence could be recorded. In the case of the neutrinos, the fact that the ''observatories'' were in the northern hemisphere was irrelevant. The neutrinos, with their ability to interact weakly and hence penetrate matter easily, raced up through the Earth. The same property meant that most of the neutrinos that passed through the detectors also did so without any interaction. Of the original 10^{57}, only nineteen neutrinos interacted with atoms of water in the detectors generating recorded flashes of light. Neutrinos were first detected by the Kamioka experiment in Japan, mentioned in Chapter 1 in the context of the solar neutrinos. Some neutrinos were also seen by a similar experiment in a salt mine near Cleveland and in a special site under a mountain in the Caucasus, what was then the Soviet Union. Those nineteen detected neutrinos were sufficient, however, to show that the basic picture of core collapse was correct. SN 1987A gave birth to extragalactic neutrino astronomy. Unfortunately, with the scant evidence of the nineteen neutrinos, we cannot determine whether the mechanism of the explosion was a core bounce, neutrino heating, or some other related process.

Putting the story together after the fact, astronomers realized that the neutrinos arrived at the Earth before the light. The reason is that the neutrinos are generated in the core collapse, or shortly thereafter, for about 10 seconds. The neutrinos that escape from the newly formed neutron star race outward at very nearly the speed of light. If neutrinos have a small mass as current theories suggest, then they will not travel at quite the speed of light, but so close to it that the difference is negligible. The shock wave that causes the star to explode propagates very rapidly, about one-thirtieth the speed of light. This is faster than the speed of sound in the star, but not at the speed of the departing neutrinos. It took the shock wave about an

hour to propagate to the edge of the blue supergiant and generate the first intense burst of light seen by Oscar Duhalde and recorded by Ian Shelton and Rob McNaught. Those first photons were thus a light hour behind the neutrinos, a lag of about 10 million kilometers, about the radius of Jupiter's orbit. The pulse of neutrinos and that first pulse of light raced each other for 150 000 years, but the light could not catch up. The neutrinos arrived an hour ahead of the optical photons. At this moment, almost 20 light years beyond the Earth, the pulse of neutrinos is still ahead of the leading edge of the pulse of light.

7.5 NEUTRON STAR?

The detection of the neutrinos was dramatic confirmation that a very compact object formed in SN 1987A by the process of core collapse. This result is completely consistent with stellar evolution theory for a star of initial mass about 20 times that of the Sun. The icing on the cake would be the direct detection of the neutron star.

We know that the supernova of 1054 that made the Crab nebula did leave behind a neutron star. This knowledge does not help us to reach general conclusions about how stars explode and make neutron stars because the Crab nebula is peculiar in many respects. It has a large helium content and slower expansion motions than are characteristic of most supernova remnants. Despite the useful observations of the Chinese, we do not know whether it was a Type I of some flavor, a Type II, or perhaps a transition event like SN 1993J (Chapter 6, Section 6.1). Astronomers of that era could not obtain spectra. Nevertheless, the Crab supernova and its left-over neutron star give us one distinct case with which to compare.

SN 1987A is the best-studied supernova ever, and we know it underwent core collapse, so the potential to learn about neutron star formation is great. As of this writing, however, SN 1987A is nearly 19 years old, and there is still no concrete evidence for a neutron star. This is important because there remains the possibility that the collapse could have generated an explosion and the observed neutrinos, but ultimately have crushed the nascent neutron star to make a black hole. SN 1987A seems to be a close cousin to the supernova that produced Cas A in about 1667. Both were dimmer than usual, both seem to have occurred in massive stars, and until very recently, neither had obvious evidence for a compact object. We now know that Cas A has a dim X-ray source associated with the compact object it left

behind (Chapter 6, Section 6.5; Chapter 8, Section 8.3). If the same sort of object exists in SN 1987A, we would not see it even with the *Chandra Observatory* at the distance of the Large Magellanic Cloud.

Current evidence does not prove that a neutron star is absent in SN 1987A. The neutron star in SN 1987A could be slowly rotating or not very magnetized and therefore not radiating very much. There is also a question of whether the neutron star could be ''beaming'' its radiation away from Earth as some pulsars are known to do (see Chapter 8). The argument against that is based on the fact that the expanding gas of SN 1987A must surround any pulsar. This gas should absorb any emitted pulsar energy and re-emit the energy in all directions. Whether the compact remnant is a neutron star or a black hole, it cannot be accreting much matter from its immediate environment or it would be bright enough to see. Recent observations with the *Hubble Space Telescope* by Jenny Graves of Harvard and her colleagues show that any optical source associated with this compact star must not be much brighter than our Sun. What is certain is that, if there is a neutron star in SN 1987A, the two-decade-old neutron star is pumping out energy at a rate that is less, by a factor of ten thousand or more, than the nearly 1000-year-old Crab nebula.

7.6 THE LIGHT CURVE

SN 1987A also provided the most direct evidence that radioactive decay of nickel-56 and cobalt-56 can power supernova light curves. Because it was a relatively compact star, Sk-69 202 had to expand farther before it could leak the heat from the original shock. It did not have to expand as far as a Type I, but about ten times farther than a normal Type II exploding in a red-giant state. Thus SN 1987A cooled more than a normal Type II and had less shock heat to radiate by the time it could radiate. This made it dimmer than a normal Type II supernovae (Chapter 6, Section 6.6). The fact that the star that exploded was a blue supergiant with a smaller initial radius made SN 1987A naturally dimmer than a normal Type II explosion in a red giant.

Models of the explosion of SN 1987A show that the shock energy dissipated in the expansion about a week after the explosion, yet the supernova did not attain maximum light for two months more. That power came from radioactive decay of nickel to cobalt to iron. Models show that the peak light in SN 1987A is produced solely by decay of nickel and cobalt. After the peak light, the light curve

declined at a well-defined rate, showing the precise half-life of decay of cobalt-56. From the brightness of the tail, one can read off precisely how much nickel was originally ejected and how much iron will eventually expand into space. The answer is 0.07 solar mass. This is a little on the low side compared to prior expectations but in the range expected for a star of 20 solar mass. In addition, there is direct spectroscopic evidence for the cobalt, and satellites rigged to measure gamma rays detected the gamma rays that were predicted to come from the decay of cobalt. The direct evidence for nickel and cobalt decay in SN 1987A gives us increased confidence that the same process accompanies core collapse in other explosions in massive stars. Understanding these processes in SN 1987A also gives us more confidence to use them in the rather different environment of the thermonuclear explosions of Type Ia supernovae.

7.7 THIS COW'S NOT SPHERICAL

There is an old joke, one version of which has a scientist hired to study the efficiency of a dairy. He begins his report with the statement, ''First we assume all cows are spherically symmetric.'' This is an in-joke that carries a lot of weight with astronomers. Stars are almost perfectly spherically symmetric because gravity pulls in on them in all directions. Stars are not exactly spherically symmetric, however, if they rotate rapidly or have a strong magnetic field. Still, to make headway in understanding new phenomena, physicists and astronomers have learned that it is often fruitful to make simplifying assumptions to block out the rough truth. Details, out-of-roundness, can be added later as needed. For SN 1987A, it was needed.

The first computer models of SN 1987A assumed that the cow was spherically symmetric. That simplifies the analysis, making minimal computational demands on what are already complex computer calculations. Such simplified models were the obvious place to start. The first clue that they were substantially wrong came from the detection of X-rays. At a meeting in Tokyo (see box) six months after the first detection of the explosion, in August of 1987, several theorists presented their predictions that the expansion should lead to the free streaming of X-rays and gamma rays from the radioactive decay in about another year. Japanese astronomers had recently launched a new X-ray satellite. They calmly stood up and reported that they had already detected the X-rays!

The reason for the early onset of X-rays was that SN 1987A was not expanding as a uniform sphere with the hydrogen on the outside, a helium layer deeper in, and the nickel, cobalt, and iron down in the deepest, slowest-moving layers. SN 1987A was instead a roiling, turbulent mess that stirred the elements it ejected. Further thought and subsequent computer models showed that fingers of radioactive nickel should, and did, reach out into the outer layers. Streams of hydrogen and helium should plunge inward. The outward mixing of nickel allowed the X-rays and gamma rays to emerge earlier than predicted from the simple models. We learn from our mistakes. By now, the understanding of the complicated structure of SN 1987A and how those lessons apply to other types of supernovae has reached a fairly sophisticated level (Chapter 6, Section 6.5).

7.8 RINGS AND JETS

The most dramatic direct evidence that something about SN 1987A was not sedately spherically symmetric is from the amazing pictures of the rings around the supernova. These were first discovered from the ground but were widely illustrated by images from the *Hubble Space Telescope*. As the epic of SN 1987A unfolded, the *Hubble Telescope* was launched, found to be out of focus, and repaired in a dramatic space walk. The focused *Hubble* images revealed a central ring around the supernova that is tilted in its aspect to us. There are also two fainter rings, nearly but not quite concentric with the first. These preexisting ring structures and the central, expanding supernova ejecta are shown in Figure 7.6. The *Hubble* images also show that the ejected matter is not round in profile, but elongated. This can be seen in Figure 7.7.

The origin of these rings is still debated. They must have formed by matter shed by the progenitor star before it exploded. One popular model is that the star blew a slowly moving wind from its equator while it was a red giant and then a faster wind after it contracted to become the blue supergiant that eventually exploded. The fast wind is supposed to have shaped the slow wind to form the bright ring and to have expanded outward to form the other two rings. Unfortunately, computer models show that the inner, bright ring often does not survive the interaction in the form observed. Another hypothesis is that the rings were shed when the progenitor of SN 1987A consumed a smaller-mass binary companion.

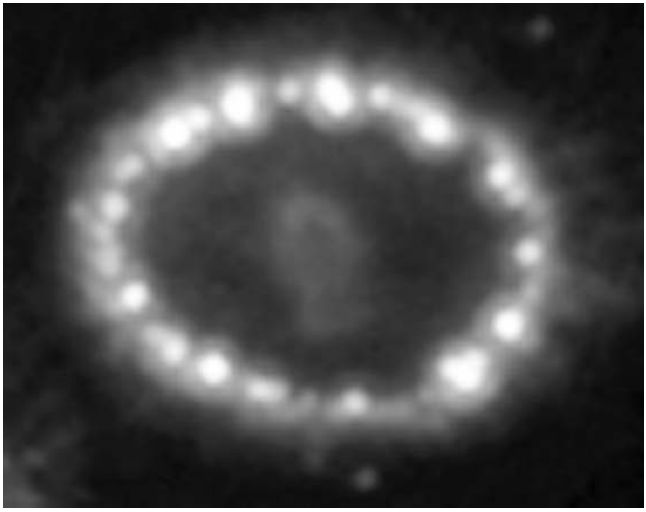


Figure 7.7 An image of SN 1987A taken on November 28, 2003, shows the result of the collision of the most rapidly expanding supernova material (not visible) with relatively dense concentrations of matter in the inner ring. The result is a string of bright spots resembling jewels on a necklace. The larger spot at about 11 o'clock first began to brighten in 1997, heralding the onset of the long-awaited collision. The central image is the glow of somewhat slower moving ejected material that is heated by radioactive decay. Note that this matter is distinctly out of round, showing that the explosion was distinctly aspherical. This portion of the ejected matter, thought to come from deep within the explosion, is elongated in a direction roughly perpendicular to the major axis of the inner ring. (*Hubble Space Telescope* photo by NASA.)

What has been clear all along is that the inner ring is only a few light years across. The most rapidly moving outer portions of the exploding star are moving at a substantial fraction of the speed of light, at least 10 percent. This implied that in a few years, or perhaps a couple of decades, the ejecta should smash into the ring. The expected result was a renaissance for SN 1987A. Astronomers predicted a new brightening in the optical, the radio, and the X-rays from the gas heated by the collision. The ring is formed of bits and clumps of gas. Each of those was predicted to light up when the shock wave hit it, making the ring sparkle like fireworks over timescales of months to years.

The first estimates of when the collision should occur were based on the notion that there was no material between the supernova

and the inner ring to slow the ejecta down. The answer was about 10 years, or, roughly, 1999. More study showed that the space between the supernova and the ring did contain matter. The time for the collision was put off to about 2005. That is not long in the big scheme of things; however, it is long in the life of an astronomer waiting to check a theory.

Given this new timescale, there was thus a little surprise when the *Hubble Space Telescope* revealed that a small portion of the ring had brightened in 1997. Many people thought that the collision had begun. Others worried that there might be some other unexpected anomaly. Ground-based observations from March 1998 showed that many more clumps were lighting up. The collision had indeed begun. Why it occurred faster than the revised estimates is a puzzle. By 2002, the ring was alight with glowing dots, as shown in Figure 7.7. Studies with *Hubble* and *Chandra* are continuing to probe these regions to learn about the nature of the ring and the response of gas to high-velocity shock waves.

The image of the ejecta in Figure 7.7 also shows the shape of the inner, slower, moving ejecta. This is the matter that is still heated by radioactive decay; no longer cobalt-56, but other longer-lived trace elements. Note that this glowing region is decidedly out of round. The axis of this region points nearly along the axis of the outer rings. In addition, this is the direction defined by the polarization measurements of SN 1987A, by an apparent directed ejection of nickel-56 indicated by Doppler shift measurements, and by an ejection of energy, still not understood, called the *mystery spot* that gave a flash of X-ray and optical radiation for a brief time about two months after SN 1987A exploded. Lifan Wang and collaborators (including me) pointed out that all this evidence was consistent with a single special direction, an axis, that ran from the outermost gas, shed by the star before it exploded, right down into the heart of the supernova. SN 1987A is completely consistent with, even the prime example for, the explosion of a core collapse supernova that explodes in a preferential direction, as outlined in Section 6.5 of Chapter 6.

There was another amusing wrinkle to this story. We predicted on the basis of all this evidence for spatial orientation that a spectrum of the ejected matter in SN 1987A would show a Doppler red shift, that matter was moving away from us in the top of the image in Figure 7.7. Here was the chain of reasoning. Other work had proved that the inner ring is nearly a perfect circle; it only looks like an ellipse because it is tilted at 45 degrees. Just looking at the image in

Figure 7.7, you cannot tell whether the ring is tilted 45 degrees ''up'' or 45 degrees ''down'' as it is projected on the sky. More work had shown that the ring was expanding and that the ''top'' part of the ring was moving toward us and hence was the part nearest to us on the Earth. The ring is tilted ''up'' so that the top is the part nearest to us and the bottom is the part of the ring on the far side. That means that if the elongated ejecta shown in Figure 7.7 were aligned with the axis of the ring, perpendicular to the plane of the ring, then the ''top'' part of the ejecta should be moving away from us, hence showing a red shift in its spectra. So, of course, we obtained a spectrum of the top part of the ejecta with a challenging observation with *Hubble* and found a blue shift!

This caused a brief consternation, but led to deeper insight. What we realized was that the ejecta you can see in Figure 7.7 was primarily composed of iron and iron-like elements. What we had measured, however, was not an atomic feature of iron, but of the element calcium, because the latter is especially distinct and easy to measure (look for the lost coin under the street light, because the light is better!). According to the jet-induced models for supernovae described in Section 6.5 of Chapter 6, iron should be blown out along the jet, the breadstick, but calcium should be preferentially blown out along the equator, in the bagel! Material in the bagel should have the same orientation as the plane of the ring, so calcium on the ''top'' should be moving toward us, just as the ring itself is, and so should have a blue shift, as observed! After those twists and turns, we concluded that SN 1987A is qualitatively consistent with a supernova that was blown up by a jet.

7.9 OTHER FIRSTS

Further observations revealed two other ''firsts'' for SN 1987A. Both were expected at some level, but never before seen. One was the formation of molecules. Molecules of varying complexity fill the interstellar medium. If the density is high enough, single atoms can bind together to form molecules. This apparently happened in SN 1987A. After about 200 days, SN 1987A showed evidence for at least carbon monoxide (CO) and silicon monoxide (SiO). There are other ways of forming molecules, but one cannot help thinking that the first steps toward molecular complexity that lead to life might begin in supernovae like SN 1987A.

The other interesting observation was to see "dust." The interstellar medium is also full of tiny bits of grit that astronomers call dust. Astronomical dust is interstellar dirt, formed of clumps of graphite (carbon) or sand (silicon oxides) or rust (iron oxides). Theories had predicted that the carbon, oxygen, silicon, and iron in supernovae might in some circumstances coalesce into dust. SN 1987A gave the first firm observational evidence for this process when the light curve got dimmer after about 500 days, as it became shaded in a cloud of its own dust. Studies of this process showed that the dust formed in dense patches, again emphasizing that the ejecta of the supernova were not uniform, but very clumpy.

Astronomers will continue to follow this piece of astronomical history as it evolves. This amazing event has much more to teach us.

8

Neutron stars: atoms with attitude

8.1 HISTORY – THEORY LEADS, FOR ONCE

In 1932, the brilliant Russian physicist Lev Landau argued on general grounds that the newly discovered quantum pressure could not support a mass much in excess of 1 solar mass. He addressed his discussion to electrons, but the type of particle did not matter. In 1933, the neutron was discovered, after Landau's paper had been submitted. In retrospect, Landau's arguments applied to the quantum pressure of neutrons as well. An object supported by the quantum pressure of neutrons should be smaller and denser than a white dwarf, but it should have nearly the same maximum mass, about 1 solar mass.

Fritz Zwicky of Caltech was one of the world's first active supernova observers. Quick on the pickup, Zwicky suggested in 1934 that supernovae result from the energy liberated in forming a neutron star. Not until a year later, in 1935, did the precocious young Indian physicist, Subramanyan Chandrasekhar, present his rigorous derivation of the nature of the quantum pressure and the mass limit to white dwarfs that bears his name.

Robert Oppenheimer made history with his leadership of the Manhattan Project, but among his most widely known papers are two published with students in 1939. The first of these papers used the complete theory of general relativity for the first time to estimate the upper mass limit of neutron stars to be 0.7 solar mass. The second paper explored the result of violating that limit with the resulting production of a black hole. The upper limit to the neutron star is now commonly referred to as the Oppenheimer – Volkoff limit, after the authors. In the 1960s, repulsive nuclear forces between the neutrons were added to the purely quantum effects. As a result, the estimates of the maximum mass of neutron stars rose to between 1.5 and 2.5 solar masses.

In 1964, John Archibald Wheeler suggested that the power radiated by the Crab nebula could plausibly be provided by the rate of loss of rotational energy of a neutron star. This proved to be a prescient guess. At about the same time, Rudolph Minkowski, an old cohort of Fritz Zwicky, was studying the Crab nebula. He pointed out that, although most of the stars seen in a photograph were foreground or background stars, one, apparently buried in the heart of the nebula, had a peculiar spectrum and an abnormally blue color. Minkowski could not prove that this peculiar star was in the nebula. There was not a shred of rational evidence relating Wheeler's speculation to Minkowski's observations, but the relation turned out to be true.

Theoretical astrophysicists often find themselves dragging along behind the observations, trying to explain some exciting new phenomenon *ex post facto* (quasars represent a superb example). In the case of neutron stars, however, the theorists were way out in front. More than three decades passed from the first theoretical discussions of neutron stars until some confirming evidence came in.

In 1967, Jocelyn Bell was a graduate student working with Anthony Hewish on a peculiar radio telescope at the University of Cambridge in England. The telescope was a series of wires run helter skelter, designed to look for rapid modulation of radio signals by the solar wind. What Ms. Bell noticed among the reams of data was a source of regularly pulsed radio emission. The pulses lasted 0.016 seconds and recurred quite regularly, every 1.337 301 15 seconds, with astounding accuracy.

The investigators were mystified at first and then, after some contemplation, petrified. There had been a long-standing expectation that any extraterrestrial civilization would signal its existence with some regularly modulated mechanism. The strange signals were dubbed LGMs, short for little green men, and a strong air of secrecy cloaked the lab. This conclusion was too significant to be blabbed about, while further checks ensued.

Soon, other such sources were discovered. Significantly, and much to the relief of the researchers, they found the pulse periods were gradually increasing. The fantastically accurate period was not locked in as it would be with an artificial mechanism, but slowly drifted. Whatever these things were, they represented a natural phenomenon. The discovery of *pulsars*, pulsating radio sources, was announced to the world. Anthony Hewish won the Nobel Prize for Physics for the discovery of neutron stars as pulsars in 1974. To the

discomfit of some, Jocelyn Bell, whose perspicacity revealed the unexpected signal, did not share in the award. Dr. Bell, a gracious woman, went on to a fruitful career as an X-ray astronomer.

8.2 THE NATURE OF PULSARS – NOT LITTLE GREEN MEN

What were these pulsars? They could not be ordinary stars. Even the light travel time across the Sun is a few seconds, and the pulses in these objects lasted only a fraction of a second. More practically, the fastest motion the Sun could withstand would be if it changed substantially in about a half hour. This is the Sun's dynamical timescale, the time it requires to respond to an imbalance between gravity and pressure. Any global motion of the whole Sun on a faster timescale, whether by rotation, oscillation, or any other mechanism, would mean that the Sun would tear apart.

White dwarfs are more compact and able to withstand rapid movement. One second – a characteristic time between pulsar pulses – is just about the natural timescale for a white dwarf. Just after the discovery of pulsars there was a great flurry of activity exploring white dwarf models for pulsars. The white dwarfs were pictured to be rotating or oscillating. Some people even considered neutron stars. Because neutron stars were even more compact, they would have no trouble responding quickly enough. The natural dynamical timescale for an oscillating neutron star is about 1 millisecond, or 0.001 second, so there was some question why a neutron star should respond as slowly as 1 second. At first, neutron stars were considered a radical, though not impossible, explanation for pulsars.

The studies that showed that the periods of pulsars lengthened with time continued as the theorists thrashed around for a consistent explanation of pulsars. The gradual lengthening of the time between pulses turned out to be a key, if subtle, clue. Studies of oscillating stars show that they tend to respond more rapidly as they lose energy. The reason is that the oscillations themselves tend to make the star somewhat more bloated and unresponsive. As the oscillations die away, the star gets more compact and bounces more quickly. A rough analogy is to drop a ball and listen for the bounces; they become closer together as the ball bounces less and less high. The lengthening of time between pulses suggested that the pulsar phenomenon had nothing to do with oscillations. As a rotating object loses energy, it spins more slowly, and so the time to make one revolution lengthens. This is in accord with the behavior of pulsars, so some

rotational phenomenon was considered the most likely explanation for pulsars.

The next major breakthrough came from studies of the Crab nebula. Ten or twenty pulsars had been discovered, all with periods of about 1 second. Then astronomers focused on the strange star Minkowski had pointed out years before. The star turned out to be a pulsar! The period of the pulses was much faster than had been seen in any other pulsar. The time between pulses was only 0.033 seconds. This time is so short that no white dwarf could oscillate or rotate that fast without being torn apart. The pulsar in the Crab nebula had to be a neutron star, and so, presumably, did all the others! Only rotating neutron stars could account for the whole range in periods, from fast to slow. A big star cannot rotate rapidly, but a compact star like a neutron star can rotate rapidly or slowly, depending on circumstances.

The pulsar in the Crab nebula rotates relatively rapidly because it was born only a short time ago and has not had time to lose much rotational energy. The pulsars with spin periods of about a second are deduced to be 1 million to 10 million years old. The Crab pulsar is so energetic that it emits pulses of optical light as well as radio radiation.

We still do not understand clearly why the radiation comes from the pulsars in pulses. That radiation comes from the pulsars at all is, however, a clue to another important property. The neutron stars must contain strong magnetic fields to generate radiation. Fundamentally, radiation is caused by wiggling a magnetic field. This causes a wiggling electric field, which in turn causes a wiggling magnetic field, which causes a... Coupled wiggling electric and magnetic fields are at the heart of the process of electromagnetic radiation. Without a magnetic field, the rotating neutron star could not emit the kind of radio radiation observed. Thus pulsars must be *rotating, magnetized neutron stars*. That the pulsars are magnetic is not too surprising. Ordinary stars like the Sun generate magnetic fields. If such a star were compressed to the size of a neutron star, the magnetic field would be amplified by a factor of about 10 billion. The resulting magnetic field would be just about what is required to generate the radiation in pulsars. Whether squeezing the field of the star that collapsed to form it is the origin of the magnetic fields of pulsars is still not clear. The newly born neutron stars may act like dynamos and make their own magnetic fields.

The simplest magnetic field a neutron star could have is a so-called dipole field like a bar magnet, with a north pole and a south

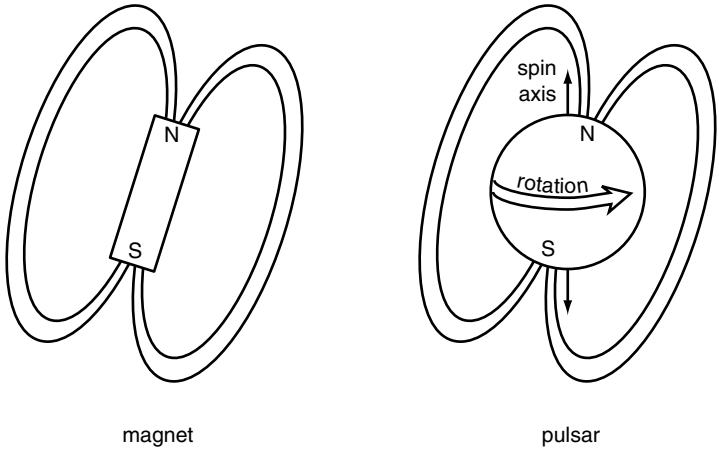


Figure 8.1 The simplest configuration of a magnetic field in a neutron star is a dipole field like a bar magnet, with a north pole and a south pole (left). The lines of magnetic force link the poles. To emit radiation, the magnetic axis of the neutron star must be tilted with respect to the rotation axis (right).

pole, as shown in Figure 8.1. The lines of magnetic force for such a field are arching loops, out one pole and into the other, exactly like the pattern of iron filings around a bar magnet. If the magnetic field is perfectly aligned with the axis of rotation, there will be no radiation, at least no pulsed radiation. The reason is that the magnetic configuration is too symmetric. If the magnetic field is perfectly aligned, there is no effective change in the magnetic field as the neutron star rotates. A wiggling magnetic field is required to generate radiation, and a perfectly aligned magnetic field causes no wiggles as the neutron star rotates.

Radiation will occur if the axis of the magnetic field is tipped with respect to the rotation axis. Then as the neutron star rotates, the magnetic field points in different directions, and the magnetic force at any given point in space varies continuously. This misalignment is not so special a requirement when one considers that the magnetic poles of the Earth are not lined up exactly with the rotation axis and that the magnetic poles even occasionally swap ends.

If pulsar radiation comes from the magnetic poles, we can even understand the pulses because the magnetic poles sweep around like beams from a lighthouse. A pulse would be detected every time a radio ''lighthouse beacon'' pointed at the Earth. This is the most popular view of the origin of the pulses. Theories have been

constructed in which the rotating magnetic fields generate huge electrical fields right at the magnetic poles. The energy in the electric field is so great that it can rip electrons from the neutron star surface or create electron/positron pairs. The particles cause a gigantic spark as they flow along the electric and magnetic fields toward the neutron star or out into space. The spark, like a bolt of lightning at the pole, emits a burst of radio static. This is the particular mechanism envisaged by which the magnetic field ''wiggles'' and gives rise to radiation.

There is still debate as to exactly where and how this spark forms. As the pulsar rotates, the magnetic lines of force are carried around with it. Any charged particles caught in the magnetic field are forced to spiral along the field, but they cannot move across the field. The result is that as the neutron star rotates, the particles must rotate as well. All the particles locked to the rotation of the neutron star make a complete circle in the same time, but to accomplish this, the more distant particles, with a greater circumference to travel, are forced to move at tremendous velocities. At not too great a distance from a neutron star, the particles would be whipped around at the speed of light. The path on which particles locked to the neutron star's rotation would move at this limiting speed is known as the *speed-of-light circle*. The distance would be a thousand miles in the case of the Crab pulsar and 30 000 miles – roughly the Earth's diameter – for a pulsar with a period of 1 second. Because particles cannot move at the speed of light, the particles must be ripped off the magnetic field lines at the speed-of-light circle. The wrenching process involved would generate radiation. Some theories argue that the great forces generate electron/positron pairs and accelerate them near the speed-of-light circle so that the ''spark'' occurs there. Other theories argue that the particles to be accelerated are those pulled from the neutron star so that the spark arises closer to the neutron star surface.

By now, some 600 pulsars have been discovered. Most of these are nearby in the Galaxy because their radiation is relatively feeble and cannot be detected from great distances. Extrapolation from the known number of pulsars leads to the estimate that as much as 1 percent of the mass of the Galaxy may be in the form of neutron stars, about one billion of them all told. Most of these would be ''dead'' pulsars, which could no longer radiate. Pulsars live about 1 million to 10 million years before their magnetic fields decay away or become aligned with the rotation axis, so that no pulses of radiation are possible.

8.3 PULSARS AND SUPERNOVAE – A GAME OF HIDE AND SEEK

When supernovae explode, they inject a large amount of matter and energy into the surrounding gas of the interstellar medium. An explosive ''cloud'' plows out into the interstellar gas, much like a mushroom cloud rises from a hydrogen bomb on the Earth. For a bomb on Earth, the ''cloud'' rises upward from the ground; for a supernova, the cloud expands outward in all directions. The resulting expanding remnant of a supernova is marked by radiation in the radio that occurs when the shock wave from the supernova compresses and heats the interstellar gas and sends electrons spiraling around the interstellar magnetic field at nearly the speed of light. Interior to the shock wave that marks the point of collision, the shocked gas is so hot it emits X-rays. A *supernova remnant* can span several light years.

These extended supernova remnants live only about 100 000 years before they fade into the general interstellar gas. Pulsars ''live'' for 1 million to 10 million years. After that time, the neutron star is still around, but it no longer emits radio pulses. Thus pulsars live for about ten times longer than the extended remnants. One expects most pulsars not to be associated with an extended remnant, but that every extended remnant in which a pulsar was born should still surround that pulsar. Most pulsars are not associated with extended supernova remnants, as expected. Strangely enough, the converse is also true. Most extended remnants show no sign of a pulsar. The Crab nebula is a conspicuous exception to this rule. This negative conclusion has been strongly reinforced by searches for pulsars with X-ray satellites.

This is a puzzling observation. Either no neutron stars are formed in many supernova explosions, or they are not rotating or magnetic so that they cannot emit radio pulses or related traces in the X-ray band, or the pulsars pick up such a high velocity that they escape out of the gaseous remnant. It is possible that in many cases the radio radiation from pulsars is ''beamed'' so that it does not shine toward the Earth. On the other hand, the X-ray radiation, similar to that emitted strongly from the Crab nebula, shines in all directions, so it would be difficult to hide. This raises yet another question. If pulsars are born at the same rate as supernovae explode, but many supernovae do not make pulsars, then apparently there is a way of making pulsars without the associated explosion and optical outburst that identify a supernova. No one knows how this is accomplished, if, indeed, it must be.

This is the context in which one considers the situation with Cas A and SN 1987A. All the evidence is that Cas A represents the explosion of a star of about 20 solar masses. Such a star is predicted to make a neutron star, but until recently (Chapter 6, Section 6.1), no compact remnant had been seen. The same arguments apply to SN 1987A in a somewhat different context because that supernova is still so young. SN 1987A came from a star of about 20 solar masses. It emitted neutrinos, so we know it had a gravitational collapse, yet any neutron star must be much dimmer than the 1000-year-old pulsar in the Crab nebula. Does this mean neutron stars exist in Cas A and SN 1987A but are especially dim? Does this dimness apply to the lack of observed neutron stars in older supernova remnants? Or did Cas A or SN 1987A ultimately create a black hole, and, if so, does this apply to the older supernova remnants? These questions remain central to the study of the final evolution of massive stars.

The point of X-ray light in the center of Cas A will continue to be the subject of intense investigation, but a few conclusions are immediately clear. The source is ten thousand times dimmer than the pulsar in the Crab nebula. If it is a neutron star, it is clearly not putting forth the effort to radiate that it might. Even just the heat energy stored in a newly formed neutron star could generate more light than this, never mind any pulsar radiation. On the other hand, a small rate of accretion could make either a neutron star or a black hole shine in X-rays like this, so either could be powered by the fall-back of some supernova ejecta that did not quite make it. This discovery also sheds light on the situation with SN 1987A. If a compact object this dim resides in the center of SN 1987A, then it is no wonder that it has not yet been detected. Progress on the study of the point of light in Cas A will undoubtedly also help us to understand whether SN 1987A left behind a neutron star or black hole.

8.4 Neutron Star Structure – Iron Skin and Superfluid Guts

Neutron stars are sometimes referred to as giant atomic nuclei because, like nuclei, they are composed essentially entirely of baryons, neutrons. Because they are so massive and bound by gravity, neutron stars have a ''personality'' beyond that of any atomic nucleus.

Neutron stars have about as much mass as the Sun, but, because of their very high densities, they are only 10 to 20 kilometers in radius. Their very outermost layers are of nearly normal composition.

There are still protons and electrons. The material is probably mostly iron because all thermonuclear processes should have gone to completion. The topmost material is probably gaseous, an atmosphere hanging above the solid surface, just as on the Earth. One major difference is that in the huge gravitational field of the neutron star, the atmosphere would be only a few meters thick. The solid surface can support mountains and other rugged terrain. Mount Everest dropped onto a neutron star surface would be crushed to a foot or so in height. Typical hills and valleys on the surface of a neutron star would range up to several inches in height.

The outer solid crust of iron-like material on a neutron star would be a few kilometers thick. An important difference in the structure of this material is that the crust is permeated by the huge magnetic field. This magnetic field alters the structure of atoms. Electrons can move along a magnetic field line but cannot move across field lines. This rule applies even to the electrons in atoms if the magnetic field is strong enough. The result is the deformation of atoms into long skinny strings, with the electron clouds elongated along the magnetic field lines and confined in transverse directions. These atoms can in turn be linked to form new kinds of long skinny molecules, which could only exist in the extreme conditions of the crust of a neutron star.

Deeper into the neutron star, electrons are squeezed tightly by the exclusion principle, and the quantum energy they acquire forces them to combine with a proton to form a neutron. The nuclear forces cannot hold a large excess of neutrons into a nucleus, so neutrons begin to leak out of specific nuclei and move around freely in the material. This process is known as *neutron drip*. The densities at which it occurs are higher than the highest density of any white dwarf, but these conditions are still found only a few kilometers deep in the neutron star.

Upon reaching depths where the density is comparable to the density of normal atomic nuclei, nothing resembling a normal atom can exist. The material is essentially all neutrons, although there is a scattering of protons and electrons. The few electrons can still exist because they are so sparsely spread that the effects of exclusion are small, and their quantum energy is not appreciable. There is one proton for every surviving electron to balance charge. The densities are so high that the exclusion effect on the neutrons is dominant and their quantum energy, moderated by effects of nuclear forces, provides the pressure to support the neutron star. The quantum uncertainty

in the ''cloud'' that represents a massive neutron is smaller than that for the cloud of the smaller-mass electron. This is why electrons feel squeezed first, and neutrons must be raised to much higher densities before the exclusion of one neutron by another has an appreciable effect.

A remarkable transition in the nature of neutron-star material is made at higher densities. The nuclear forces between neutrons have another important role besides just altering the pressure. The nuclear forces cause the quantum waves that represent the neutrons to line up in a special way that minimizes the repulsive nuclear forces. The result is that the neutrons are thought to form what is called a *superfluid*. A superfluid is a special state of matter in which all the particles flow in consonance and the result is absolutely zero viscosity, no resistance to motion. Water has much less viscosity than molasses, but a superfluid has none at all! Physicists have created superfluids in the laboratory by cooling liquid helium to near absolute zero. This reduces the thermal energy in the helium, and helium has no interfering chemical reactions because it is a noble gas. The result is that the quantum properties dominate, and the quantum waves of the helium atoms can line up in such a way as to form a superfluid. The resulting material flows so easily that if care is not taken, it will flow up the side of the beaker and out of the experiment! Lev Landau, with whom we introduced this chapter, won the Nobel Prize in Physics for his work on liquid helium in 1962.

At the highest densities in the center of a massive neutron star, the quantum effects among the neutrons can cause yet another arrangement of the structure. Theories predict that the neutrons will clump together into a rock-like solid. This material would be somewhat akin to the solid crust. In the crust, the solidification is due to electrical forces on the electrons, whereas in the core the solidification is due to nuclear forces among the neutrons. At these most extreme densities, the huge gravitational energy can be converted into mass. Exotic particles that do not normally exist in nature could spring spontaneously into existence, but there is no proof that such processes occur.

This picture of the interior of a neutron star just sketched follows from the theoretical extrapolation of known physics to extreme conditions. Fortunately, there is some evidence that the picture is at least qualitatively correct. This evidence comes from ''glitches'' observed in the rate of pulses from pulsars. As we have said, pulsars generally slow down with time, in the sense that their pulses slowly

get farther and farther apart. This effect is quite gradual, of order of one part in a million per year, and it is only due to the exceedingly accurate rate of pulses that the slowdown can even be detected. Occasionally, however, a pulsar will speed up for a short while, and the time between pulses will become shorter. After some time, the pulses will settle back into their old pattern of gradual slowing. This behavior is known as a ''glitch,'' which means, in general, an unexpected interruption or change in behavior. Glitches have been observed in a few of the youngest pulsars. Apparently, the older pulsars have settled down into a state where they do not glitch anymore. The Crab pulsar has been observed to glitch. There is another supernova remnant in the direction of the constellation of Vela. This supernova remnant is only about 10 000 years old. It also contains a pulsar that has been observed to glitch.

No one has seen a pulsar in the process of glitching. Rather, the pulsar is observed at one time and then a little later, and the period is found to be slightly shorter. From such observations a few days apart, one can conclude that the glitches happen on a time that is shorter than a few days (possibly much shorter), but no more accurate statement can be made. The thing that is of particular interest is that after a glitch, the pulsar requires a considerable time, of order a month, to return to its original period and resume the same gradual lengthening of the period. That the time to return to normalcy is so long seems to strongly suggest that the inner portions of the neutron star are superfluid.

Glitches are thought to occur as a neutron star adjusts itself to the loss of rotational energy as it slows down. The understanding of how that adjustment occurs has evolved over the decades since glitches were discovered. An early model envisaged the spinning neutron star to form an equatorial ''bulge'' that was frozen in when the neutron star cooled and its outer layers solidified. As the neutron star spun more slowly, the bulge would settle by cracking and breaking. Conservation of angular momentum would cause the neutron star crust to rotate slightly more rapidly when the crust broke and settled into a smaller radius. This was thought to represent the formation of the glitch. The slow healing time was then thought to represent the long time necessary for the outer solid crust to bring the inner, zero viscosity, superfluid core into a common spin rate, after which the whole neutron star would begin to lose rotational energy and once again begin to spin ever more slowly. The reason to mention this picture is that it is a simple physical one that was reasonably easy to

describe in lectures. I used it for decades, and it appears in other books. It is also wrong. More careful study showed that the mechanism of glitches is more interesting and subtle. The idea of crust cracking has survived in another context that will be described in Section 8.10.

The current model for glitches is based on considerations of exactly how the magnetic field that is such an obvious part of the external aspects of a pulsar threads the inner superfluid core. It turns out that a magnetic field cannot penetrate the superfluid, but only normal matter. For the magnetic field to thread the superfluid core, there must be ''vortices'' of normal matter that extend through the superfluid core, roughly parallel to the spin axis of the neutron star. The spinning vortices of normal matter are the repository of the angular momentum of the material in the inner core. The vortices of normal matter also provide the path for the magnetic field to pass from the north to the south pole within the neutron star. The vortices that allow normal matter and the magnetic field to thread the superfluid are ''pinned'' to irregularities in the normal matter of the outer crust. In this picture, a glitch occurs because the vortices have a memory of the past when the outer crust was spinning faster. At intervals, some of the vortices unpin from the crust and coalesce, allowing the whole neutron star to adjust to its slower rotating, lower angular momentum state. Although the whole neutron star adjusts to the lower rotational state, this unpinning causes the outer crust to temporarily rotate more rapidly, giving rise to the glitch. As the neutron star attains its new equilibrium rotational state, the vortices again pin to the crust and slow it down so that the gradual slowing of the whole neutron star can continue. The bottom line is still that the glitch phenomenon cannot be explained without invoking a superfluid core.

8.5 BINARY PULSARS – ''TANGO POR DOS''

The accurate periods of pulsars make excellent clocks. If the clock were to move, the *frequency* of the pulses would be changed by the *Doppler shift*. The frequency of the radio emission would also be changed, but the radio radiation is continuum radiation, which, without spectral ''lines'' – specific identifiable frequencies – gives no detectable Doppler shift. The pulses themselves are a marvelous substitute. With this clock, astronomers can look for periodic changes in the velocity of a pulsar that would indicate that the neutron star

was in orbit. The evidence shows that to a high degree of accuracy the vast majority of pulsars are not in binary star orbits. Astronomers were very excited when in 1975 careful searches paid off, and a radio pulsar was discovered to be in a binary orbit. Since then, eleven more binary radio pulsars have been discovered. They are the exception that proves the rule; the vast majority of the known pulsars are single stars.

The discovery of the first binary pulsar led to a host of interesting results. The orbit was worked out from the Doppler shift of the pulsar period, and the prediction was made that any companion star of ordinary size would cause the eclipse of the neutron star once each orbit. No eclipse was seen. The lack of an eclipse implies that the companion star is itself a compact star, probably a white dwarf or neutron star.

Nature has been kind to put neutron stars in binary orbits. Study of the binary orbits allows the determination of the neutron star masses, a fundamental property that cannot be accurately measured by any present techniques for the multitude of single radio pulsars. The period of the orbit gives information about the masses of the stars, using Kepler's third law. The mass of the first binary pulsar is one of the few known neutron star masses. Both stars seem to have a mass of very nearly 1.4 solar masses. Other binary neutron stars have also had their masses weighed in this manner, and they also appear to have very nearly this mass. The coincidence of this number with the Chandrasekhar limit requires some comment. If a white dwarf attained the Chandrasekhar limit and collapsed to form a neutron star, the neutron star would be somewhat lower in mass. This is because some energy is inevitably ejected in the process of forming the neutron star, if only in the form of neutrinos. A great deal of energy must be ejected and the mass equivalent, in terms of $E = mc^2$, of the minimum energy loss is about 0.2 solar mass. To make a neutron star of 1.4 solar mass, the initially collapsing object would have to be 10 or 20 percent more massive, and hence somewhat greater than the Chandrasekhar mass. Just why neutron stars should form from cores of a precise mass that somewhat exceeds the Chandrasekhar mass is not clear.

The accurate orbital timing of the first binary pulsar showed that the orbit was decaying. The two stars are slowly spiraling together. Recall the final evolution of two white dwarfs from Chapter 5. They are imagined to spiral together as they give off gravitational radiation. In the binary pulsar system, the change in the

orbit is precisely what would be predicted as the result of gravitational radiation. With one stroke, this observation confirms, indirectly but strongly, the predicted existence of gravitational radiation by Einstein's general theory and shows that gravitational radiation works in binary systems to draw stars together, just as the astrophysicists had predicted. Whatever the companion of this binary pulsar, white dwarf, or neutron star, gravitational radiation will eventually cause them to collide and merge. The discovery and analysis of the binary pulsar and the remarkable proof of gravitational radiation led to the award of the Nobel Prize to Joe Taylor and Russell Hulse, the radio astronomers at the University of Massachusetts (Taylor is now at Princeton, Hulse at the University of Texas in Dallas) who made the discovery and analysis of the first binary pulsar. For this second Nobel Prize for work on neutron stars, the important contribution of the graduate student (Dr. Hulse) was recognized.

The binary pulsars, by being the exception to the rule, also lead us to ask why the strong majority of pulsars are not in binary systems. The binary pulsars provide a clue to the answer. One possibility is that neutron stars are commonly ejected from binary orbits by the explosion that creates them. Arguments based on conservation of energy and angular momentum show that if half the total mass of a binary system is ejected in an explosion, the system will be disrupted, with the two stars flying off in opposite directions. In addition, pulsars are observed to sail through space at rather high velocities. There are a number of reasons to think that pulsars are given a ''kick'' by the process of violent gravitational collapse that creates them. Such kicks will also help to tear neutron stars away from any binary companion. Ejecting matter in the explosion and kicking the pulsars probably account for most of the single pulsars. The exceptions can also be understood at some level. For one thing, the star that blows up will frequently be the less massive star because it will have transferred mass to the companion. If the exploding star contains less than half the total mass of the two stars combined, then it cannot eject more than half the total mass, and the binary system will not be disrupted. The kicks to newly formed neutron stars may not be delivered in random directions, but, inasmuch as they are, some of the kicks could help to keep the neutron star in orbit, despite the loss of mass and gravity from the binary system by the supernova process itself.

The circumstantial evidence that Types Ib and Ic supernovae arise from massive stars that have lost their outer envelopes by mass transfer suggests that they create neutron stars in binary systems.

Whether these neutron stars remain in the binary is not clear. There is a strong suspicion that, for systems in which the neutron star is still in a binary, the neutron star was born in some version of a Type Ib or Type Ic supernova explosion.

There may be another reason why the radio pulsars, in particular, are mostly single. An important feature of the first binary pulsar is that the companion star is known to be compact. No mass is being transferred in the system. As we will see in the next section, neutron stars are known to exist in binary systems in which the neutron star is not a radio pulsar. These systems are transferring mass. One reasonable hypothesis is that mass transfer prevents the emission of radio pulses by blocking the radio emission or by shorting out the sparking mechanism and preventing the radio radiation in the first place. With this picture, one would say that the binary pulsar is special, not because the neutron star remained bound in a binary system, but because the companion star is unable to transfer mass and spoil the radio pulses. Those neutron stars that were always single stars or that were ejected from binary systems have no problem because they have no companion to interfere. Most neutron stars left in binary systems are not radio pulsars because they have the misfortune to be neighbors to a living star that insists on sharing some of its matter.

An amazing new chapter in this story came with the discovery by Andrew Lyne of the University of Manchester and his colleagues of two pulsars in orbit, known as J037–3039 A and B; B with a rotation period of 2.8 seconds and A with a rotation period of 23 milliseconds (see Section 8.9). The most surprising aspect of this discovery, aside from the fact that both compact objects are active pulsars, is that the plane of the binary orbit is oriented almost directly at the Earth so that the pulsars eclipse one another. This means that one object no more than a few miles across is getting between the Earth and another tiny object only a few miles across!

The opportunity to observe the eclipses has opened a whole new gold mine of information about neutron stars and pulsars, including an in-depth exploration of the magnetic field surrounding the pulsars. Detailed timing of the orbit gives the mass of each neutron star and the rate of decay of the orbit by gravitational radiation. The masses in turn give new information on the inner structure of the neutron star. There are indications that a wind from the fast pulsar is mussing up the magnetosphere of the slow pulsar. There is also information in the shape and evolution of the nearly circular orbit. The latter means

that there could not have been a huge kick in the explosion when either neutron star formed and may have some implications for jet-induced supernovae. There is some speculation that the second neutron star, at least, was formed by the collapse of an oxygen/neon/magnesium core of a star of initial mass of 8 to 12 solar masses (Chapter 6, Section 6.2), rather than of the iron core of a more massive star.

8.6 X-rays from neutron stars – hints of a violent universe

X-ray observations have been mentioned where appropriate throughout this book. The next subject owes its very existence to the advent of X-ray astronomy, however, and so a word of history is in order. In the last three decades, the science of X-ray astronomy has matured to become a major independent branch of astronomy. X-rays must be collected above the absorbing shield of the Earth's atmosphere. The first observations were made with brief rocket flights that only tantalized the scientists that launched them. There were glimpses of intense sources of high-energy X-rays.

The revolution in X-ray astronomy began with the launch of a small astronomical satellite dedicated to the detection of X-rays in 1972. The satellite was launched from a site in Kenya and was called *Uhuru*, the Swahili word for freedom. This first satellite could not locate the source of any X-ray emission very accurately, and, although better than rockets, it was not tremendously sensitive. *Uhuru* was on station for a long time compared to a rocket at perigee, however, and it could look for X-rays for orbit after orbit. The result was stupendous. The whole sky was alight with X-rays. It was like Galileo's invention of the telescope: to look with a new tool and to find that previously unknown or inconspicuous objects glared forth when examined properly. X-rays were seen from stars, from galaxies, from every direction! Above the protective layer of the atmosphere, the Universe was a far more violent place than astronomers had suspected. For opening this new perspective, Riccardo Giacconi, the leader of the *Uhuru* team, was awarded the Nobel Prize in Physics in 2002.

Many X-ray satellites have been flown in the last 30 years. Several have been launched by the United States, others by European countries. Japan has had a very successful series of satellites and nearly took over the field when the U.S. support for X-ray astronomy lagged in the 1980s. Russia has also had a number of successful experiments. A major step of this first burst of activity in a new field

was the launching by NASA of a large satellite in 1978, bearing the name *Einstein*, because it was the centennial year of his birth. This satellite contained a device that could focus X-rays like a proper telescope. It could measure details in an X-ray picture with an accuracy of one arcsecond, equivalent to that of ground-based optical telescopes. In six years, the science of X-ray astronomy made an advance in sensitivity and detail equivalent to the leap from Galileo's first telescope to the giant modern reflectors. The new *Chandra Observatory* mentioned in Chapter 1 is the latest step in this progression, and there are more and better projects under construction and on the drawing boards.

One of the subjects to benefit most from the new science of X-ray astronomy was the study of neutron stars. This is because the great gravity of these objects causes tremendous heating of any matter that falls upon them. The matter becomes so hot that the maximum intensity of radiation comes in the X-ray portion of the spectrum. Under proper circumstances, neutron stars are just natural X-ray emitters.

Some of the first X-ray sources examined with *Uhuru* showed a peculiar behavior. The intensity of the X-rays was not constant, but faded away at regular intervals, typically every few days. Most of the scientists who worked on the early X-ray experiments building the detectors were physicists, not astronomers. The erratic behavior in the signal puzzled them. Astronomers – at least many amateurs who delight in such things, if not the professionals who specialized elsewhere – would have immediately identified the cause. The problem was that the X-rays were being eclipsed. The X-ray source was in a binary star orbit and was simply disappearing behind the other normal star periodically. This companion star was the source of matter that fell onto the neutron star and produced the X-rays.

This understanding led to a rapid series of identifications of orbiting neutron stars. A major new branch of astronomy was born almost overnight as the new sources were identified and characterized, and theorists rushed to understand their properties. The X-ray observations provided an exciting new way to probe the nature of mass transfer, accretion disks, and the structure and behavior of the neutron stars themselves. Although the existence of accretion disks had been demonstrated in the cataclysmic variables, it was the exciting new realm of neutron-star X-ray sources that resulted in the sudden growth of interest and developments in the understanding of accretion disks.

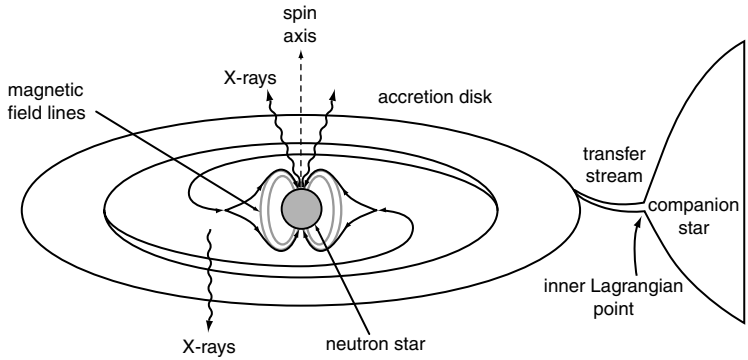


Figure 8.2 Binary X-ray sources consisting of a neutron star with a low-mass companion, like Hercules X-1, are very similar to cataclysmic variables, but with the white dwarf replaced with a neutron star. The companion star, often a main-sequence star, transfers mass from its Roche lobe through a transfer stream that collides with an accretion disk. The matter joins the disk and spirals slowly down toward the neutron star. When the magnetic force of the neutron star exceeds the pressure forces in the disk, the matter is diverted to follow lines of constant magnetic force. These paths lead to the magnetic poles of the neutron star. X-rays can be emitted from the inner, hot portions of the accretion disk and from the magnetic poles where matter actually strikes the neutron star surface.

Over the next few years after the launch of *Uhuru*, X-ray astronomers realized that there were two basic classes of binary neutron-star X-ray sources (and a handful of oddballs that resist categorization). The first class consists of a neutron star in orbit about a normal, fairly low mass star. The other class consists of neutron stars in orbit around high-mass normal stars. In this case, the normally evolving star typically has a mass in excess of 10 solar masses.

The classic example of the first type is the first X-ray source discovered by *Uhuru* in the direction of the constellation Hercules, the system named Hercules X-1. Detailed studies over decades have shown that Her X-1 is a nearly textbook example of mass transfer to a neutron star in a binary system, as shown schematically in Figure 8.2. A star of about 2 solar masses, slightly evolved on the main sequence, is filling its Roche lobe and transferring mass. The mass settles into an accretion disk. As friction operates in the disk, the matter spirals down toward the neutron star and gets heated. In the inner portions of the accretion disk, the orbital velocities are very high, so the frictional heating is strong, and the material in the disk itself emits

X-rays. When the spiraling matter gets near the neutron star, the magnetic field of the neutron star channels the matter toward the magnetic poles. When the material finally lands on the surface of the neutron star, the impact causes more heating and further X-rays.

Although X-ray satellites are crucial to the discovery of X-ray sources, one should not forget that the astronomy advances most efficiently where standard earthbound optical techniques can be brought to bear in complementary studies. This is because, as a matter of practice, there is a tremendous amount of information available in the photons emitted in the optical band. This is, after all, where most stars emit the majority of their radiation. Most of our practical knowledge of the Universe is obtained in the optical, so X-ray (or radio, infrared, ultraviolet, or gamma ray) information must be integrated into the realm of classical optical astronomy to come to full fruition.

As an example, studies of Her X-1 would be woefully incomplete without the optical studies of the companion star. It is the optical studies that tell us the type of star, its evolutionary state, and the fact that it is filling its Roche lobe. Coupled optical and X-ray studies were used to completely characterize the orbits of the two stars and to obtain a direct measure of their masses using Kepler's law. The mass of the neutron star comes out to be very nearly 1 solar mass. This mass seems to be significantly less than the 1.4 solar masses that has been measured so precisely for several of the binary pulsars, as mentioned in Section 8.5. There is no understanding of why this should be so. It is presumably an accident of birth of an especially low mass progenitor core, but it might have involved an especially large ejection of the mass from the collapsing core. In this game, even ''typical'' objects are not so typical.

The observations of Her X-1 suggest that a star of initial mass between 10 and 15 solar masses evolved and shed its envelope. The bare core probably evolved on its own for a while and then collapsed. Like cataclysmic variables, there is a strong hint in Her X-1 that the original evolution was not just a simple case of one star losing mass to the other. For one thing, the two stars are too close together now for the first star to have developed a dense core and red-giant envelope. Also, the relatively low mass of the companion star suggests that it did not accept all the mass that the first star lost. Her X-1 is probably another example of common-envelope evolution in which the 2-solar-mass star was engulfed in the envelope of the more massive star. Much of the first star's envelope was presumably lost out of the

system, and the core of the massive star and the smaller-mass companion spiraled together. Perhaps the smaller star filled its Roche lobe while still enshrouded in the envelope of the other. Whether any of this helps to explain the relatively low mass of the neutron star is not clear.

The other kind of binary X-ray-source systems, those with high-mass normal companions, is typified by the third X-ray source *Uhuru* discovered in the direction of the constellation of Centaurus, Centaurus X-3. The basic difference between Her X-1 and Cen X-3 is that the mass-losing star in the latter is fairly massive, about 20 solar masses. This turns out to make an important modification to the mass transfer process, if not the ultimate outcome. When Cen X-3 was first discovered and the companion optical star identified, attempts were made to work out the orbits. According to the standard picture, the assumption was made that the companion filled its Roche lobe in order to transfer mass to the neutron star. The answers that emerged did not make sense. The mass of the neutron star was derived to be so low, about 0.1 solar mass, that the gravity should be so weak that any neutron star should expand to be a white dwarf instead.

The problem was that the companion star does not fill its Roche lobe! Rather, such a massive star blows an appreciable stellar wind. It loses mass through this wind whether it has a companion star or not. In this case, however, there is a neutron star, the gravity of which reaches out and ensnares some of the passing wind. The matter from the wind then settles into an accretion disk. With this picture, things make more sense. The orbital information from Cen X-3 is not as accurate as that from Her X-1, never mind the binary pulsars. The best estimate for the mass of the neutron star comes out to be a little more than a solar mass, but a mass of 1.5 solar masses cannot be excluded. This is a reasonable result.

The disproportionate mass between the neutron star and the massive normal companion in Cen X-3 has one interesting consequence. The neutron star raises tides on the surface of the companion, just as the Moon does on the Earth. Energy is expended in dragging those tides around, and the energy comes out of the orbit, causing the neutron star to spiral toward the other star. If the companion is not too massive, the tidal drag causes it to spin faster until the companion rotates at exactly the speed that the neutron star orbits. Then the tide just sits in one place on the surface of the star, and there is no drag. For a massive companion, however, there is too much inertia. The central star and the tides always lag behind the

orbital motion, dragging the neutron star down. There is no limit to this process, and eventually the neutron star should collide with and disappear into the companion star. The neutron star could spiral to the center, swallow matter from the star, collapse to make a black hole, and then eat the whole star! This may be the fate in store for Cen X-3.

Her X-1 and Cen X-3 share another very important feature. The X-rays they emit come in pulses, 1.2 seconds apart for Her X-1 and 4.8 seconds for Cen X-3. The behavior is very reminiscent of the pulses from radio pulsars, but the energy is coming in the X-ray portion of the spectrum. In addition, for extended periods of time the pulses get steadily more rapid, whereas, except for glitches, the radio pulses slow down.

Despite the exotic nature of the radiation, the X-ray pulses are easier to explain than the radio pulses. Much of the explanation borrows heavily from the knowledge gained by studying radio pulsars. The neutron stars are presumed to be magnetized and rotating. The crucial difference is that, whereas a pulsar must generate radio radiation by its own devices, the X-rays are caused by an external agent, the dumping of mass upon the neutron star.

With the presence of the magnetic field, the matter arrives at the neutron star in a special way that promotes pulses. The matter spirals down in the accretion disk until it encounters the outer reaches of the magnetic field. At that point, the matter finds that it cannot continue in orbit because it cannot move across the lines of magnetic force. Rather, the matter falls along the lines of force, as shown in Figure 8.2. These lead naturally to the north and south magnetic poles of the neutron star. The matter is channeled so that it falls selectively on the magnetic poles, not at random on the surface of the neutron star. The intense X-radiation then comes from the magnetic poles, as if there were two bright spots on an otherwise dark surface. If the magnetic axis is misaligned with the axis of rotation, then, as the neutron star spins around, first one then the other bright spot points at the Earth, just like a lighthouse. The observer detects a pulse of X-rays as the pole is swept into view by the rotation. With mass transfer, one can understand fairly easily why the radiation comes from the poles and hence why there are pulses.

The influence of mass transfer also explains why the pulses tend to speed up rather than slow down. There are two competing effects. The loss of energy in the radiation tries to slow the neutron star down. The matter arriving from the accretion disk, however, carries with it

the angular momentum of its orbit. As the matter lands on the neutron star, the spin is transferred to the neutron star. This turns out to be the dominant effect in many circumstances, and the neutron star rotates faster and faster until the mass transfer stops or the neutron star is rotating as fast as the accreting matter where it begins to interact with the magnetic field. If the neutron star tries to rotate too fast, its magnetic field acts like a paddle to splash matter out of the accretion disk, which slows the neutron star down. Both Her X-1 and Cen X-3 have gone through episodes lasting a couple of years where they have stopped speeding up (Cen X-3) or have even tended to spin more slowly (Her X-1). This is presumably because they have ejected matter or the rate of mass transfer has declined so the accretion disk has retreated, allowing the neutron star rotation to slow. Even though the spin-up by accretion makes good sense, the slow-down process must be rather prevalent because many X-ray pulsars have rather long periods, some as long as 800 seconds.

8.7 X-RAY FLARES – A STORY RETOLD

Recall from Chapter 5 that there were two basic classes of flaring binary white-dwarf systems: the dwarf novae where the accretion disk is the source of the activity and classical novae caused by thermonuclear explosions on the surface of the white dwarf. Suppose the white dwarf were replaced by a neutron star. Similar phenomena will occur.

X-ray astronomers see several accreting neutron stars in the Galaxy that are labeled as *X-ray transients*. In this context, the general word ''transient'' refers to a particular phenomenology, implying a particular physical cause. Every few years, these X-ray transients emit a flare of X-rays that lasts for about a month or so. At least two of these systems are well studied and are known to be in binary systems. There is a strong suspicion that the process causing this outburst is similar to that in dwarf novae, an instability in the flow in the accretion disk. The accretion disk instability described in Chapter 4 does not depend sensitively on the nature of the object around which the disk circles. If matter flows into the disk from a companion star at an appropriate rate, the disk will go into the storing and flushing mode that characterizes the dwarf novae. If the object receiving the mass is a neutron star, however, then in the flushing phase, matter from the disk is spiraling down onto a neutron star. The matter gets intensely hot and emits X-rays. The timescales are somewhat longer in the X-ray

transients than in dwarf novae, and there are no quantitative models, but the disk instability is a plausible picture for the origin of the X-ray transients.

There is also a neutron star analog of classical novae. In 1978, a fascinating new class of X-ray sources was discovered. Russian scientists first noticed the phenomena. Some X-ray sources show an occasional brief, strong burst. The power rises in about a second and then decays over the course of the next minute or so. The bursts recur every few hours more or less randomly. After the Russians reported these bursts, a search of old *Uhuru* data also showed the effect. The American astronomers just had not noticed it at first in the welter of data with which they had to deal.

The display in the X-ray bursts is not like the rather demure pulses from Her X-1 and Cen X-3 or like the occasional flares of the X-ray transients. The bursts are very energetic compared to the pulses of Her X-1 or Cen X-3. They are comparable in power to the X-ray transients but much shorter in duration. They call for a completely different physical explanation.

Of the more than 100 X-ray sources in the Galaxy with low-mass companions, about 40 are *X-ray bursters*. None of the binary X-ray sources with high-mass companions display this behavior, and neither do the few low-mass systems that display X-ray pulses like Her X-1. Like the general population with low-mass companions, the X-ray bursters tend to cluster toward the center of the Galaxy, as do the oldest stars in the Galaxy. At least nine of the X-ray bursters are seen to be in globular clusters that are also old assemblages of stars. Most X-ray bursters show no evidence for binary motion, but evidence has been reported for orbital motion in at least one X-ray burst source. The guess is that all these systems are in binary systems, but nature conspires to hide the fact. If the systems are seen edge-on, it is most easy to determine the Doppler motion due to their orbit, but in this case the neutron star and its X-rays can be obscured by the accretion disk. If the system is nearly face-on, the X-rays can be seen, but the orbit is difficult to determine because all the motion is almost at right angles to the observer. The Doppler shift only registers the component of motion directly toward or away from the observer. The X-ray bursters do not show any sign of X-ray pulses (an exception will be described later). The interpretation is that the neutron stars in these systems have very low magnetic fields, so matter is not focused on the magnetic poles, and there is no X-ray ''lighthouse'' effect.

The theory for the burst sources is based on thermonuclear explosions on the surface of the neutron stars. Calculations have shown that as hydrogen accretes onto the surface of a neutron star, it is heated and burns in a regulated fashion. Under proper circumstances, the resulting helium, however, piles up in a layer supported by the quantum pressure. As we have seen in several instances, this condition leads to unstable burning when the helium finally gets hot and dense enough to ignite. The X-ray bursts are thus thermonuclear explosions on the surfaces of the neutron stars. There is therefore a direct parallel for this explanation of the X-ray bursts and the explanation of the outbursts in the classical novae, the basic differences being in the nature of the compact object doing the accreting. Because of the high gravity of neutron stars, relatively little, if any, matter is ejected from the neutron star in an X-ray burst. The high gravity also causes the very short timescale of the explosion on the surface of the neutron star, as compared to the effects in a classical nova that can linger for a year or more.

The theory of these nuclear outbursts shows that they only occur if the rate of accretion of matter onto the neutron star is relatively sedate. This allows the layer of helium to build up, supported by the quantum pressure. At high accretion rates, the helium stays hot, is supported by the thermal pressure, and burns in a regulated, nonflaring way. One of the implications of this theory is that if the neutron star is strongly magnetic, then even a sedate rate of accretion will be focused onto the magnetic poles, giving an effective high rate of accretion at those two spots. That will provide the circumstances for hot magnetic poles and X-ray pulses, but it will mean that the rate of the accretion at the poles is high enough that the helium will ignite and burn in a regulated way. This is another reason to argue that neutron stars that show X-ray pulses have large magnetic fields and no X-ray bursts, and neutron stars that show X-ray bursts have small magnetic fields and do not display X-ray pulses.

The Eddington limit discussed in Chapter 2 plays an interesting role in the neutron-star accretion process associated with these X-ray burst sources. Recall that the Eddington limit is a limit to how bright an object can be without blowing away matter by the sheer pressure of the outflowing radiation. The Eddington limit depends on the gravity of the object, and so the limiting luminosity scales with the mass. For accreting neutron stars, there is a close coupling between the mass and the luminosity because the luminosity is caused by the infalling matter. This means that if the matter falls in at too high a

rate, intense radiation will be generated. The infalling matter will be blown away rather than accreting. If too much of the infalling matter is blown away, however, then there is not enough radiation to blow the matter away, and the infall can take place. The result can be to balance things so that some matter is blown away and some accretes. The luminosity adjusts so that the Eddington limit is not violated. Many of the binary neutron-star X-ray sources have luminosities somewhat below the Eddington limit, as if they had made their accommodation with the limiting luminosity. In the observed X-ray bursts, the luminosity rises until it bumps right into the ceiling of the expected Eddington limit for an object of the mass of a neutron star, about one solar mass.

At least one binary neutron star system, Centaurus X-4, displays both X-ray transient outbursts and X-ray bursts. As the X-ray flux from Centaurus X-4 declined from one month-long flare of the X-ray transient variety, it showed another brief flare of the X-ray burst variety before proceeding to decline. Presumably an accretion-disk instability flushed matter down toward the neutron star creating the X-ray transient. As matter accumulated on the neutron star, it underwent a thermonuclear outburst. Then the disk went into its storage mode; there was no fresh mass added to the neutron star, so no repeated X-ray burst.

8.8 THE RAPID BURSTER – NONE OF THE ABOVE

One particular source, the *Rapid Burster*, displays behavior that falls in the ''none of the above'' category. This system, known to intimates as MXB 1730–335 (for MIT X-ray Burst), was discovered about 20 years ago. When active, it bursts about four thousand times a day. The Rapid Burster is located in a globular cluster. It also occasionally has the more prominent bursts associated with the thermonuclear ignition of helium. Like the other thermonuclear burst sources, the Rapid Burster shows no sign of X-ray pulses that would indicate the rotation of the underlying neutron star. The presumption is that the magnetic field of this neutron star is relatively weak, so matter falls more uniformly on the surface and is not focused at the magnetic poles. The repetitive bursts that define the Rapid Burster are thought to be neither a thermonuclear burst on the surface of the neutron star nor the type of accretion-disk-heating instability similar to that of dwarf novae. The observations suggest that the matter rains down on the neutron star in blobs, like a rapidly dripping faucet, rather than in a steady gush.

There is no well-established theory for this behavior, but the suspicion is that it involves an instability of the matter on the inner edge of the accretion disk that may be due to a condition where the pressure of radiation becomes excessively large, larger than the pressure of the hot gas in the disk. For 20 years, the Rapid Burster was alone, but now it has some company.

In 1990, NASA launched another of its great observatories to complement the *Hubble Space Telescope*. This was the *Compton Gamma Ray Observatory*. We will talk about it more in Chapter 11. In December 1995, this satellite discovered a system known as the *bursting pulsar*, or, more technically according to its discovery instrument and coordinates, GRO J1744–28. Follow-up work on it was done by another NASA satellite, the *Rossi X-ray Timing Explorer*. This relatively modest satellite was named after Bruno Rossi, an MIT pioneer of X-ray astronomy, and was designed to follow X-ray behavior with very accurate timing. Observations with *RXTE* of the bursting pulsar showed an incredible array of behavior that indicate that this system may be an important link between systems like the Rapid Burster and the other X-ray burst sources.

As its name implies, the bursting pulsar is an X-ray pulsar. From the frequency of the pulses, one can deduce that the neutron star rotates about twice a second. Its orbital motion has also been detected. The neutron star is in a 12-day orbit around a small red giant that has lost almost all of its hydrogen envelope and now has a mass of about one-quarter the mass of the Sun. From January through May of 1996, the bursting pulsar showed large bursts, lasting about 10 seconds apiece every 2 hours or so. These bursts displayed characteristics of the staccato bursts of the Rapid Burster rather than the helium ignition flares of the X-ray bursters. The presumption is that the bursting pulsar has a stronger magnetic field than the Rapid Burster and hence can both generate ''lighthouse'' pulses of X-rays from the magnetic poles and can suppress nuclear flares by the focused, hot accretion at the magnetic poles. The fact that it still manages to show the instability of the inner disk means that the magnetic field is not so strong that it cuts out the inner part of the disk where that instability happens. The bursting pulsar is thus an interesting intermediate case that promises to teach us more about the conditions under which neutron stars evolve in binary systems. After May of 1996, the system got so dim that *RXTE* could no longer detect it, so for now, the bursting pulsar is keeping any further secrets it may have to reveal.

8.9 MILLISECOND PULSARS

In the last decade, a new variety of radio pulsars have been found that have generated great excitement because they link so many aspects of the formation and evolution of neutron stars. Theory predicts that neutron stars cannot rotate faster than about one thousand times per second without flinging themselves apart with the excessive centrifugal force. That limiting rotation rate corresponds to a rotational period of 0.001 second, or 1 millisecond. Thus one expects that the fastest pulses that could be discovered from a pulsar would be about 1 millisecond, and that a 1 millisecond pulsar would be on the verge of tearing itself apart. Realistically, one would expect that pulsars would rotate a little slower than this fastest possible limit and, hence, to have pulses of a few milliseconds. By this standard, the pulse period of the Crab nebula pulsar is dawdling along at a mere 33 milliseconds.

Special search techniques were developed to search for pulsars near this period limit, and they have been successful. Over two dozen *millisecond pulsars* have been found. In contrast to their longer period kin, about half of the millisecond pulsars are in binaries. The most rapidly rotating has a remarkably well-defined period, 0.001 557 806 448 85 seconds, or about 1.6 milliseconds. This neutron star is whipping around 642 times per second.

The next step is to account for the origin of the millisecond pulsars. Pulsars must be magnetic neutron stars. The Crab pulsar rotates 30 times per second; normal pulsars, about once per second. This is because the Crab pulsar is only 1000 years old. When it is several million years old, the Crab pulsar will have slowed down, and it will presumably also have a period of about 1 second. This suggests that millisecond pulsars might be very young, newly born neutron stars. More thought, and appropriate observation, shows just the opposite is the case. With a normal-strength magnetic field, a pulsar with a period of 1 millisecond would be losing energy so fast that it could not maintain its rapid rotation. By this argument, the millisecond pulsars should be slowing down very rapidly, but they are observed to be slowing scarcely at all. The millisecond pulsars must therefore have a smaller magnetic field than normal so that they lose little rotational energy into radiation. This in turn suggests that they are old, so that there has been time for their magnetic fields to decay away or otherwise disappear. If they are old, however, why have they not lost more of their rotational energy when they were younger with a more robust magnetic field?

The proposed resolution to this query is that the neutron stars were born in binary systems and that transfer of mass and associated angular momentum from a companion kept the neutron star spinning fast, even as the field decayed. Thus all millisecond pulsars should be in binary systems, but a significant fraction of them are not. This is another dilemma. If there were a binary companion, where did it go?

One possible answer to this further dilemma was suggested by the discovery of a particular millisecond pulsar in a binary system. This pulsar orbited a companion, a more or less normal star. It appeared as if the pulsar were killing the normal star because the star was losing mass at a high rate. The rapidly rotating neutron star produces a great flux of high-energy radiation, X-rays and gamma rays. It was first thought that this intense radiation was literally blasting away the companion star. Some astronomers termed this system the Black Widow star because the neutron star was perceived to be killing its mate. Subsequent observations showed that the star was probably losing mass on its own. In any case, the implication is that the companion will soon be gone, leaving a millisecond pulsar to spin alone in space. Roughly half of the millisecond pulsars are in binary systems with a companion star to transfer mass and keep them spun up. Presumably the other half of the observed millisecond pulsars have already dispensed with their companions in one way or another.

Another interesting millisecond pulsar revealed that it had objects of planetary mass orbiting it. These objects were discovered only by the exquisite timing that is possible with these pulsars. Tiny rhythmic oscillations in the pulse period revealed that the pulsar was being slightly tugged around in space by several small objects of mass about that of Jupiter. Whether these are true planets, left over from some ill-fated solar system that orbited the star before it exploded, or whether the ''planets'' are themselves left-over lumps of blasted star-stuff is not clear.

To put the millisecond pulsars in perspective, we need to take a step back in the evolutionary story. What sort of system gave rise to the original system of a neutron star orbiting an ordinary star? The explosion of a supernova in a binary system ejects a great deal of mass and hence decreases the gravity that holds a binary system together. That is why we think most ordinary pulsars are alone in space. They have not murdered their companions, but they may have unbound and ejected them from orbit. To prevent this, we need a fairly gentle way to make a neutron star. After the neutron star is born, it must

have a weak magnetic field or lose an originally strong magnetic field and then be spun up by accretion to become a millisecond pulsar.

If this is the evolution of the neutron stars that become millisecond pulsars, then such systems should pass through a phase in which the companion adds mass to the neutron star to spin it up. The result should be the production of X-rays. The natural conclusion is that the systems we see now as X-ray sources with neutron stars orbiting low-mass companion stars will evolve to become the millisecond pulsars. The problem is that if you work out the rate at which X-ray systems with neutron stars and low-mass companions are born and the rate at which millisecond pulsars are born, they disagree substantially. There do not seem to be enough low-mass X-ray systems to account for the number of millisecond pulsars. Either there is another way to make millisecond pulsars, or there is something we do not understand about the evolution of the stellar systems in the X-ray phase. If that phase lasted a shorter time than we think, there would have to be a higher production rate to account for the number we see at this epoch in galactic history. That would help close the gap.

Another mechanism that might avoid the phase of being an ordinary X-ray source during the spin-up phase has been suggested to produce millisecond pulsars. That mechanism involves the accretion of matter onto the O/Ne/Mg core of a star of original mass of about 10 solar masses. When such a core reaches its maximum mass, it will undergo electron capture and collapse to form a neutron star, but essentially all the core will collapse to make the neutron star, and very little is expected to be ejected (Chapter 6). This gives the maximum probability of maintaining a companion in binary orbit. This general process is called *accretion-induced collapse*, to distinguish it from core collapse brought on by the normal process of core collapse of a single evolving star as fuel is burned to heavier elements. This process is plausible in general, but it does not necessarily predict that the resulting neutron star will be rapidly spinning with a low magnetic field, the conditions required to be a millisecond pulsar.

The low magnetic fields required to explain the millisecond pulsars have raised a different conundrum. All radio pulsars are observed to fall on the short-period side of a limiting value of the period that depends on the strength of the magnetic field. The implication is that as pulsars age and rotate slower and slower, their magnetic fields decay away, so that for very old slow pulsars the combination of rotation and magnetic field is no longer able to generate the thunderstorms at the magnetic poles that are required to

make a radio pulsar. In a plot of magnetic field versus spin period, this limiting period is known as the ''death line.'' Taking a somewhat more pragmatic approach, Mal Ruderman of Columbia University argues that the cutoff may be different for different magnetic field configurations and hence the boundary may be a ''death valley.'' In any case, the notion persisted for two decades that the magnetic field of pulsars decays away with a timescale of perhaps 100 million years. Continued consideration of the numbers of pulsars with different field strengths and spin periods and the existence of the millisecond pulsars with very low magnetic fields has inspired reconsideration of this issue. There are suggestions that the field may not decay or that it is the accretion process itself that kills the field in the case of the millisecond pulsars. The origin and evolution of neutron-star magnetic fields is still a subject of active investigation.

8.10 Soft Gamma-Ray Repeaters – Reach Out and Touch Someone

Although the Sun occasionally belches a flare of particles that reach the Earth and affect radio communications, we are used to the stars being quietly remote in their isolated magnificence against the backdrop of dark space. Imagine our surprise, therefore, when one of them reached out and touched us in August of 1998 and another, in spades, in 2004! As the Earth sails around the Sun and follows the Sun around the Galaxy for billions of years, it is not isolated from the violent Universe around us.

A class of bursting events called *soft gamma-ray repeaters* has been defined over the last 20 years. At first, these events were confused and intermingled with the events known as *gamma-ray bursts*, the story of which we will learn in Chapter 11. The difference between ''hard,'' high-energy X-rays and ''soft,'' low-energy gamma rays is a matter of operational definition, and the dividing line is somewhat arbitrary. As the names imply, however, soft gamma-ray repeaters and gamma-ray bursts radiate most of their energy in the gamma-ray range. The soft gamma-ray repeaters emit somewhat less energetic photons than the gamma-ray bursts, a difference an expert can love. As we shall see in Chapter 11, no gamma-ray burst has ever been known to repeat. As data accumulated, however, it became clear that the sources that gave out the softer gamma rays could and did repeat their outbursts, if at irregular intervals. The question was, what were they? Gamma rays of any sort require high energies, and that suggests high gravity, so one

might think about white dwarfs, neutron stars, or black holes. Round up the usual suspects! An important clue was that all the soft gamma-ray repeaters turned out to be in supernova remnants.

The current most widely accepted theory for the soft gamma-ray repeaters was developed by Rob Duncan at the University of Texas and Chris Thompson, now at the University of Toronto. They were originally seeking an explanation for gamma-ray bursts, not soft gamma-ray repeaters. Their investigations led them to consider neutron stars with very strong magnetic fields. They developed a theory that, under certain circumstances involving, among other things, very rapid rotation, neutron stars could develop immensely strong magnetic fields. Whereas millisecond pulsars have magnetic fields about ten thousand times less strong than ''normal'' pulsars, Duncan and Thompson argued for magnetic fields thousands of times stronger than ''normal.'' The force of such magnetic fields could rival the gravity of the neutron star – strong indeed. Duncan and Thompson needed a name to distinguish their intellectual baby from the ''normal'' pulsars and millisecond pulsars, so they coined the name *magnetar* for a neutron star where the magnetic field rivaled gravity and pressure.

As they investigated the properties of magnetars, Duncan and Thompson realized that they should have a special activity. When they are first born, the magnetars would assume an equilibrium, balancing the magnetic fields, pressure, gravity, and the centrifugal force of their rapid rotation. The latter would cause the neutron star to bulge along the equator, and that bulge would tend to be frozen into place in the outer rocky crust of the neutron star. As the neutron star lost energy and slowed, the bulge would be too big for the slower rotation, and it would eventually crack and settle. This picture is very similar to the original explanation for glitches in pulsar rotation rates, which has now been supplanted, as mentioned in Section 8.4. In the context of the magnetar theory, however, Duncan and Thompson realized that such a crust cracking would send powerful waves into the magnetic field that looped above the neutron star surface. The magnetic field would have to readjust to the new structure of the neutron star, and the magnetic field would convert some energy into hot plasma. That hot plasma would radiate the gamma-ray energy for the timescales observed in soft gamma-ray repeaters. Duncan and Thompson proposed that soft gamma-ray repeaters were, in fact, magnetars, a variety of super-magnetized neutron star not previously recognized. They also recognized that after the first, major,

crust-cracking star quake, there could be more localized shifts in the crust as it adjusted to the rearranged magnetic field. This would give a smaller, dimmer source of soft gamma rays, but if the spot were carried around by the rotation of the neutron star, then one might see a ''lighthouse'' effect so that the gamma rays would be seen to ''pulse'' at the rotation rate of the neutron star.

This suggestion that soft gamma-ray repeaters were magnetars attracted some positive, some negative, and some bewildered reactions. To make progress, observational confirmation was needed, and that came in 1998 in a rapid succession of events. Careful observations with the *Rossi X-ray Timing Explorer* revealed the rotation rate and rate of slowing down of one of the soft gamma-ray repeaters. The observations were consistent with a neutron star with a magnetic field one thousand times stronger than ''normal.''

In August of 1998, Nature made sure we understood this lesson. One of the soft gamma-ray repeaters went off with a burst that was so strong that it affected the Earth! The gamma rays from this soft gamma-ray repeater affected the ionization of the upper atmosphere and interfered with radio communications worldwide. A wonderful contribution to the Op/Ed page of the *New York Times* described the awe-inspiring, incredibly intense, and widespread aurora witnessed by a bunch of guys on a fishing expedition above the Arctic Circle. This was one of the very few known events when a star in our Galaxy, but far beyond the Solar System, physically affected the Earth. There was no harm done, but this cannot have been the first time such a thing happened, and it was not the last. There is at least one record of a gamma-ray burst tickling the ionosphere; in this case the event happened not just in our Galaxy, but in a galaxy long, long ago and far, far away.

The event just described also brought evidence for a pulsar with a superstrong magnetic field. The eruption had the immensely strong burst that tickled the Earth's ionosphere, but then displayed a series of ''pulses'' just as Duncan and Thompson had predicted. They argued that hot spots should occur as the crust shifted in places. The rotation of the magnetar would give a lighthouse type effect as the hot spots were seen and then rotated out of sight. In hindsight, just this behavior had been seen in the first soft gamma-ray repeater observed in 1979 in the Large Magellanic Cloud. At the time, that outburst was strange and controversial. That misery is now comforted by the company of the nearly twin outburst of the nearby source that produced the August 1998 burst. One must be careful and continue to

seek evidence, but the magnetar theory is clearly the leading contender to account for the soft gamma-ray repeaters.

The latest chapter in this particular saga happened on December 27, 2004, while the new *Swift* satellite was still in its check-out phase. Another Galactic magnetar let off a huge burst of energy, 100 times brighter at its peak than the ones in 1979 and 1998. *Swift* detected this burst, but *Swift* was not needed: this burst ''pinned the needle'' on something like 15 other spacecraft ranging from Earth orbit to Mars. Once again this burst temporarily rattled the ionosphere of the Earth, even though it came from 50 thousand light years away, on the far side of the Galactic center from the Earth. Some of the radiation even reflected off the Moon. In this case, the theory demanded not just cracking of the neutron star crust and the production of hot spots, but the wholesale rearrangement of the huge magnetic field.

There are a handful of other objects that also seem to fit nicely into this scheme. These have been known as the *anomalous X-ray pulsars*. Like the soft gamma-ray repeaters, the anomalous X-ray pulsars are all found in supernova remnants. They show no evidence for binary companions. They all have rather long periods that fall in a restricted range of 6–11 seconds, very similar to the soft gamma-ray repeaters. They all seem to be spinning down, the spin periods getting longer and longer, as if the spinning source were simply losing energy. From the spin period and rate of decrease of spin, an indirect estimate can be made of the strength of the magnetic field and the result is a value comparable to magnetars: 100 to 1000 times stronger than normal radio pulsars.

A scheme that makes sense is that one neutron star in ten is born with an especially high rotation that allows the newly born neutron star to generate the high magnetic field. For the next 1000 years, that magnetar undergoes crust cracks and rearrangement and is active as a soft gamma-ray repeater. After that time, the neutron star rotates sufficiently slowly that it cannot generate strong gamma-ray outbursts, but for the next 40 000 years it can radiate enough to be seen as an anomalous X-ray pulsar. After that time, it will be cooler and slower and will be a ''dead'' magnetar. The nature of the supernova that gives rise to magnetars and the nature of dead magnetars are not clear. How often do we end topics on that note? Such a big Universe, so little time ...

8.11 GEMINGA

Yet another chapter in the neutron star story is told in the saga of the source known as *Geminga*. This source was first detected in 1973 by one of the early satellites with gamma-ray instruments. Two decades were required to figure out what it was. The name was given to it by an Italian X-ray astronomer, Giovanni Bignami. The name is nominally related to the fact that it is a gamma-ray source in the direction of the constellation of Gemini. More amusingly, it is an Italian double entendre related to the fact that the source could not be detected in the radio, one of the ongoing mysteries. In the dialect of Milan spoken by Bignami, *ghe'è minga* means it's not there.

Vision in gamma rays is blurry and there were lots of spots of light in the direction of Geminga. A long time was required to pin down the source. In the optical, stars, asteroids, and plate defects had to be ruled out. The *Einstein* satellite revealed an X-ray source that helped to narrow down the optical search. One thing became clear. Whatever the object was, it was damn dim in the optical. Suspicion that Geminga was a neutron star grew. In the late 1980s, a dim optical source was isolated. It turned out to be the real thing.

A major breakthrough came finally in the 1990s with observations from the *Compton Gamma Ray Observatory* and the German–US–British *Röntgen Satellite* or *ROSAT*, named for the discoverer of X-rays. Observations with these instruments showed that Geminga revealed both gamma-ray and X-ray pulses due to rotation with a period of 0.237 seconds. Geminga was a neutron star. Like the Crab nebula pulsar it emitted gamma rays, but unlike the Crab pulsar and so many others, it did not emit radio radiation. Various arguments suggested that it was very close to the Earth. That meant that, even though the gamma rays were detected, they were intrinsically feeble. That was why similar sources were not common. They would just be too hard to detect at greater distances. The small distance also explained why Geminga could be seen in the optical at all. Neutron stars have such a small radiating surface that one would have to be very close to be observed.

The close distance had another significant implication. There was a chance to detect the *proper motion* of the source, the motion across the sky due to its motion through space, and even the *parallax*, the apparent motion due to the Earth's orbit around the Sun. The former gives a hint of where Geminga arose; the latter, how far away it is. The parallax was measured in 1994 with the *Hubble Space Telescope*,

and Geminga is only 160 parsecs, about 500 light years away – right in our backyard! The proper motion was extrapolated backward, and Geminga's origin was traced to near a star in the Orion nebula. There is an expanding cloud of gas around a star there that might be the supernova that created Geminga. The time for Geminga to get from Orion to where it is now is about 350 000 years, which is consistent with the age measured from the rate of slowing of the spin and with the estimated age of the supernova remnant. There are other possible interpretations, but the strong implication is that Geminga arose in a supernova explosion rather nearby about 350 000 years ago. Early hominids were leaving the veldt then and beginning to explore the planet – not so long ago.

The interpretation of Geminga is that it is a neutron star with a rather normal magnetic field. In 350 000 years, it has spun down so that it can barely generate gamma rays by particle creation and acceleration near the magnetic poles. Its surface is still hot and glows in the optical, if dimly. The most likely reason why radio is not observed is that the radio is created, but that it is radiated away from the Earth by an accident of orientation. Overall, Geminga is very special because of its nearness to Earth, but it may represent a normal phase in the aging and evolution of normal neutron stars. In looking to the past of Geminga, we may also be looking to the future when Betelgeuse erupts at about the same distance, the story foretold in the box in Chapter 6.

9

Black holes in theory: into the abyss

9.1 WHY BLACK HOLES?

Black holes have become a cultural icon. Although few people understand the physical and mathematical innards of the black holes that Einstein's equations reveal, nearly everyone understands the symbolism of black holes as yawning maws that swallow everything and let nothing out. Can there be any compelling reason to understand more deeply a trivialized cultural metaphor? The answer, for anyone interested in the nature of the world around us, is an emphatic yes! Black holes represent far more than a simple metaphor for loss and despair. Although black holes may form from stars, they are not stars. They are objects of pure space and time that have transcended their stellar birthright. The first glimmers of the possibility of black holes arose in the eighteenth century. Two hundred years later, they are still on the forefront of science. In the domain of astronomy, there is virtual certainty that astronomers have detected black holes, that they are a reality in our Universe. In the domain of physics, black holes are on the vanguard of intellectual thought. They play a unique and central role in the quest to develop a ''theory of everything,'' a deeper comprehension of the essence of space and time, an understanding of the origin and fate of our Universe.

There is a certain inevitability to black holes in a gravitating Universe. Einstein's theory says that for sufficiently compressed matter, gravity will overwhelm all other forces. The reason lies in the fundamental equation, $E = mc^2$. Because mass and energy are interchangeable, one of the implications of this equation is that energy has weight. The very energy that is expended to provide the pressure to support a star against gravity increases the pull of the gravitational field. The more you resist gravity, the more you add to its strength.

The result is that if an object is compressed enough, gravity becomes overwhelming. Any force that tries to resist just makes the pull all the greater. When gravity exceeds all other forces, the object will collapse to form a black hole.

The first people to contemplate the notion that gravity could become an overwhelming influence were John Mitchell, a British physicist, and the Marquis de Laplace, a French mathematician. Mitchell in 1783 and Laplace in 1796 based their arguments on Newton's theory of gravity. They used the concept of an *escape velocity*. The notion is that to escape from the surface of a gravitating object, a sufficiently large velocity must be imposed to overcome the pull of gravity and ''escape'' into space. If the velocity is too small, the launch will fail. If it is just right, a launched vehicle will just coast to a halt as it gets far away from the gravitating object. With more velocity, a launched vehicle will still have a head of steam as it breaks free of gravity and it will continue to speed away. That is the whole idea behind tying two big, solid-fueled boosters and an external liquid fuel tank to the space shuttle when it goes up from Cape Canaveral. The shuttle must achieve escape velocity, or near it, to get into orbit, and that means lifting off the launch pad really fast!

Mitchell and Laplace used this idea of an escape velocity to argue that an object could be so compact that the escape velocity from the surface would exceed the speed of light. By some coincidence, an algebraic formulation of this escape velocity condition in the context of Newton's theory of gravity gives the correct result for the ''size'' of a black hole using the correct theory of gravity, general relativity. Mitchell did not, apparently, coin a zippy shorthand name for his intellectual creation. Laplace called his hypothetical compressed entities *corps obscurs*, or hidden bodies. (The modern French equivalent is *astres occlus*, or closed stars. The literal translation, *trous noirs*, has also gained acceptance after some initial resistance because of its suggestion of double entendre.)

With some hindsight, we can see that Newton's theory of gravity was flawed. This theory predicted that, if two masses got infinitesimally close together, the force would go to infinity. A general lesson of physics is that, when infinities arise, there is a problem with the mathematical formulation that reflects some omission in the physics. Another problem with Newton's law of gravity is that, although it prescribed how the strength of gravity scaled with the mass of a gravitating object (to the first power) and the distance between objects (inversely as the square of the distance), it did not say

how gravity varied in time. Consider two orbiting stars. A literal use of Newton's law of gravity says that, as one star moves, the other instantaneously responds to the fact that the motion has occurred. Thus according to Newton's law of gravity, the effect of gravity propagates infinitely fast. This second troublesome infinity violates the idea that nothing can move faster than the speed of light. Finally, and perhaps most compelling from a strictly practical point of view, Newton's gravity did not work.

Newton's law of gravity is spectacularly successful in most normal circumstances, when distances are large and speeds are slow. Astronomers still use it to great effect to predict the orbits of most stars. Rocket scientists use it to plot the paths of spacecraft even as they do complex orbits that carry them around planets, getting a boost from the interaction. The *Galileo* spacecraft went through a remarkable series of bank shots around the inner planets, picking up speed in the various encounters with Venus and Earth, before being flung to Jupiter. The recently launched *Cassini* spacecraft completed the first stage of its voyage to Saturn by first looping inward to circle Venus. *Cassini* received a kick from the orbital motion of Venus that gave it the momentum to sail out to Saturn. The success of gravitational multiple-bank shots shows that Newton's gravity works very well in this regime.

For very fine measurements, however, Newton gives the wrong answer! The predictions of Newton do not agree with observation, with the way Nature works. Classic examples are the rate of rotation of the perihelion of Mercury and the deflection of light by the Sun. In contrast, Einstein's theory of gravity has passed every test of observation. A modern example is the use of global positioning systems (GPS) in boating, camping, and driving, as well as military and industrial uses. This system works by timing the signals from an array of orbiting satellites. It is based on the mathematics of the curved space and warped time of Einstein, not the simple law of gravitation of Newton. If the silicon chips in the GPS detectors knew only about Newton, boaters in the fog and soldiers in the field would get lost!

As we shall see, giving up Newton for Einstein does not represent merely swapping one set of mathematics for another. Rather, Einstein brought with him a revolution in the fundamental concept underlying gravity. Newton crafted his mathematics in the language of a force of gravity as the underlying concept. Physicists and astronomers still use the notion of a gravitational force in casual terms, even though it has become outmoded in a fundamental way.

Einstein's view was radically different. For Einstein, there is no force of gravity. Instead, Einstein's theory represents gravity as a manifestation of curved space. A gravitating object curves the space around it. A second object then responds by moving as straight as it can in that curved space. The curved space results in deflections of motion that are manifested as gravity, even though the object is in free fall, sensing no force whatsoever. Much of this chapter will be devoted to exploring this conception of gravity.

The progress of our understanding of gravity is not over, however. We have come to understand that, even though it has passed every experimental test, Einstein's theory has flaws. It has its own nasty infinities that represent some omission in the physics. Ironically, the hints of a new, better theory are again cast in the language of force, but not the force of Newton. In notions being developed today, the force is quantum in nature and may play on a field of ten or eleven dimensions, not the three of space and one of time that sufficed for both Newton and Einstein. We will begin with an exploration of black holes, as portrayed in Einstein's theory, and see how deeper issues arise. Some of those issues will be explored in Chapter 12.

9.2 THE EVENT HORIZON

As described by general relativity, a black hole is a region of space–time bordered by a one-way membrane called an *event horizon*, as shown schematically in Figure 9.1. Matter or light can pass inward through the event horizon, but nothing that travels at or less than the speed of light, even light itself, can get back out. The term ''event horizon'' comes from the notion that if an ''event,'' like a firecracker exploding, occurs just outside the event horizon, the light can reach an observer, and the fact that the event occurred can be registered. If the firecracker goes off just inside the event horizon, however, no information that the event occurred can reach the observer. The event takes place beyond a horizon so that it cannot be seen. Once inside the event horizon, escape is impossible without traveling faster than the velocity of light. The location of the event horizon is thus intimately related to the fact that the speed of light is a speed limit for all normal stuff. The simple argument of Mitchell and Laplace concerning the formation of a *corps obscurs* relates to the size of the event horizon. The size of the event horizon scales with the mass of the black hole. For a black hole with ten times the mass of the Sun, it would have a radius of 30 kilometers, about 20 miles in radius. The

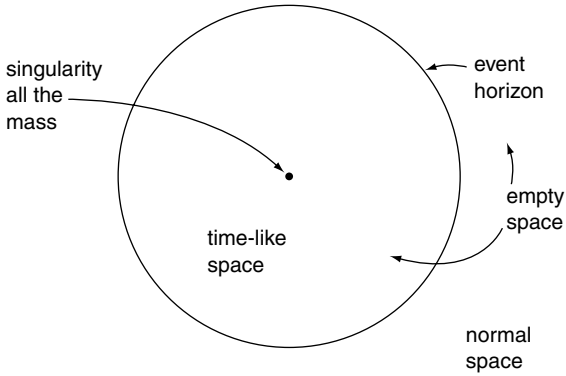


Figure 9.1 The simplest, nonrotating black hole has two basic elements – the event horizon, interior to which nothing can escape, and the singularity, where everything, including space and time, are crushed out of existence. Within the event horizon, space takes on a time-like aspect (Section).

nature of the event horizon in the context of curved space and time will be explored in more depth in Section 9.5.

9.3 SINGULARITY

When Newton was pondering the means by which apples bonked him on the head and, more particularly, how the Earth kept the Moon trapped in orbit, he intuited an important aspect of gravity. He realized that the gravity of the Earth must act from the center of the Earth, not, for instance, from its surface. This was not a trivial conclusion, and he needed to prove that it was true. Newton knocked off his gravity studies for a while and invented the mathematics of calculus in order to prove his conjecture. With his new mathematical tools, Newton was able to prove that, although the mass of the Earth is distributed throughout its volume, each little piece of the Earth acts in concert as if it were in the center. The result is that for any object beyond the Earth's surface, the gravitational attraction of the Earth will act as if all the mass of the Earth were concentrated at a point in the center. This is true for any spherical gravitating body. The gravitational attraction depends only on the distance from the center of the body, not on the radius or volume of that body. Armed with this mathematically proven conclusion, Newton went on to formulate his theory of gravity with a mathematical expression that said that the

force of gravity between two spherical objects depended only on the masses of the two objects and on the inverse square of the distance between their centers.

As an example to make this property concrete, imagine that the Sun were suddenly compacted to become a neutron star of the same mass. It would get cold and dark on the Earth, but the Earth would continue in exactly the same orbit because the gravitational pull it feels from the Sun depends only on the mass of the Sun, not on how big it is. Another implication is that we are in no danger of falling into a black hole. All the black holes we know or suspect are far away. The gravity would be frightful if we were to get near their centers, but at a large distance from their centers, the gravity gets weak as it does at a large distance from any object, and vanishingly small if the distance is very large. In this context, there is one interesting difference between normal stars of any kind – suns, white dwarfs, or neutron stars – and black holes. The former act as if all their mass were concentrated at a point in the center. For black holes, this is literally true.

Inside the event horizon, all mass that falls into a black hole is trapped. Even though there is no material surface at the event horizon, the matter within the black hole still signifies its presence by exerting a gravitational pull. The gravitational acceleration exists outside the event horizon and causes the formation of the event horizon. Although the black hole still exerts a gravitational pull, the matter itself is crushed out of all recognizable existence. General relativity predicts that the matter compacts into a region of zero volume and infinite density at the center of the black hole. Even more profound, space and time cease to exist at this point. Such a region is called a *singularity* and is illustrated schematically as a point in Figure 9.1. For a black hole, all the mass that creates the gravity is literally at this point in the center, at the singularity.

The infinities associated with the singularity are clues that Einstein's theory is not a complete theory of gravity, despite its great success. We know in principle what is lacking. Einstein's theory does not contain any aspects of the quantum theory. The uncertainty principle of the quantum theory tells us that it is not possible to specify the position of anything exactly, including the position of an infinitely small singularity. The notion of a singularity as it arises in Einstein's theory is thus an intrinsic violation of the quantum theory. With a theory of gravity that properly incorporated quantum effects, which general relativity does not, the singularity would probably be altered to be a region of exceedingly small volume and immense, but

not infinite, density. It is the nature of that exceedingly small volume, the singularity that forms inside a black hole, the singularity from which our Universe was born, that is the heart of the quest for a new, deeper understanding of physics.

9.4 BEING A TREATISE ON THE GENERAL NATURE OF DEATH WITHIN A BLACK HOLE

The manner in which a black hole crushes matter out of existence, save for its gravitational field, is rather graphic. Consider something falling into a black hole, say a human body – feet first. In this case, at every instant the feet are going to be closer to the center of the black hole than is the head. Gravity is thus going to be stronger at the feet and will pull the feet away from the head. The natural forces on an extended body tend to stretch it along the direction toward the center of the gravitation. At the same time, all parts of the body are trying to fall toward the center. The left shoulder is trying to fall toward the center. The right shoulder is trying to fall toward the center. As the body gets closer to the center, the distance between separate paths directed at the center gets ever smaller. The shoulders get shoved together, and whatever is in between must suffer the consequences. A body falling into a black hole will be stretched feet from head and crushed side to side. This is known jocularly as the ''noodle effect.'' Anything falling into a black hole will be noodlized, as shown in Figure 9.2.

The technical name for this simultaneous radial stretching and lateral crushing is the *tidal force*. It is precisely the same effect as causes the tides on the Earth. Here, the Moon pulls on the Earth and its oceans, pulling them toward the Moon and pushing them in sideways to form the tidal bulges in the oceans, the faintest form of noodle. As a body falls into a black hole, the tidal forces increase drastically. First the body stretches into a noodle and breaks apart. Then the individual cells stretch into noodles and are destroyed. Next gravity overcomes the electrical forces that bind matter into molecules and atoms. Atoms will be wrenched out of molecules and electrons pulled from atoms. As infall proceeds, the rising tidal forces will overcome the nuclear force, stretching out the atomic nuclei and breaking them apart into individual protons and neutrons. In their turn, the protons and neutrons will break up into quarks, and the quarks into whatever comprises them. These building blocks will in turn be subject to supernoodlization until the singularity is reached

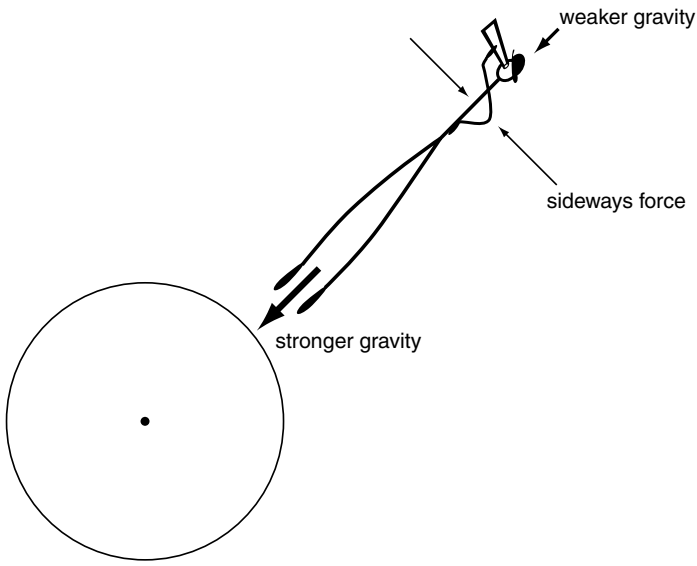


Figure 9.2 Any material body falling into a black hole will have its bottom pulled from its top and its sides crushed together in a tidal ''noodlizing'' effect.

and matter as we know it ceases to exist. Another way of characterizing the singularity in Einstein's theory is that the tidal forces become infinite. Physicists are gaining the first hints of what conditions may be in the singularity that will prevent that infinity. A discussion of this topic is postponed to Chapter 12.

9.5 BLACK HOLES IN SPACE AND TIME

9.5.1 *Curved space and black holes*

Black holes are in the most fundamental way a beast of curved space. Visualizing this curvature that occupies all of three dimensions is very difficult for creatures such as us who are limited to a three-dimensional perspective. Even the experts have difficulty picturing the immense complexity of curved space. They have invented tricks to help with the perception. We will describe these tricks because they help, but even they represent only a shadow, and a fairly complicated one, of the truth.

The notion of curved space raises a general question. How do we characterize it? A line inscribed in a wavy two-dimensional space may be straight in some sense from our three-dimensional perspective, but not truly straight at all. Likewise, a properly ''straight'' line in a curved two-dimensional space may look strangely curved from another perspective. The ability to define and construct straight lines in curved space is fundamental to understanding how curved space works.

What do we mean by a straight line in curved space? There is a rigorous way to decide which lines are straight in a given space, a way that is intuitively reasonable as well. To obtain a straight line in a curved space, start with a small portion of the space where it is, for all practical purposes, flat. Think of any measurement you would normally make on the surface of the Earth, ignoring the fact that the Earth is really a closed spherical surface. In this small, nearly flat portion, use two short straight sticks. Lie one stick down. Now extend the second stick so that it partially overlaps the first, so that you know it is pointed in the same direction as the first, but so that it also extends out a way. Now hold down the second stick and slide the first along, keeping it parallel to the second stick until it extends out a way. Continue in this manner, extending each stick in turn a little way, in such a manner that you are always assured that each extension goes in precisely the same direction as the last. As you proceed, draw a line using each stick in turn as a straight edge. Never look off at a distance to orient yourself. This technique depends on the fact that you are looking only at the local little patch of very nearly flat space in which you find yourself at any given instant. This method of drawing a straight line is called *parallel propagation* because each step consists of extending one of the sticks parallel to the other. One can prove mathematically that the line you draw as a result of this tedious operation is the shortest distance between any two points along it. What more could you want from a truly straight line? The operation of parallel propagation is what you approximate every time you sketch a freehand straight line. You do not make two marks on a paper and then try to make the distance between them as short as possible. Rather you start your pencil off in some direction and then, trying to keep your hand steady, continue the line parallel to itself. That is what makes parallel propagation so intuitive. It is what you really do to sketch a straight line.

In a flat space, parallel propagation will give the ordinary straight lines known and loved by tenth-grade geometry teachers.

Parallel lines constructed in this fashion will never cross. Triangles made of three such lines will have 180 degrees as the sum of their interior angles. This is the geometry of Euclid, the geometry of flat space. In an arbitrarily curved space, watch out! Viewed from above, lines drawn as straight as possible by the method of parallel propagation will appear wackily curved if the surface is curved; but parallel propagated lines are as straight as possible and will be the shortest distance between two points, even if the space is curved.

A particular trick the mathematicians have developed for picturing curved space is to project a three-dimensional curved space onto two dimensions in a special way, like casting a shadow. One dimension is suppressed, and the resulting two-dimensional figure is displayed as a two-dimensional surface in three-dimensional space. It becomes something we can look over, around, and under from our three-dimensional perspective and get a feel for the real thing. The technical name for the image that results from projecting the two-dimensional representation into ordinary flat, three-dimensional space is called an *embedding diagram*, because the two-dimensional ''shadow'' is embedded in the three-dimensional space.

To perform this trick for a black hole, one of the dimensions of rotation is suppressed. The resulting figure looks like a cone, or as if you were to poke your finger into a rubber sheet, as shown in Figure 9.3. The distant, still flat, parts of the sheet are the simple two-dimensional projection of flat, uncurved, three-dimensional space. The cone made with your finger is a technically proper representation of the curved space around a black hole (at least in qualitative shape, the mathematics of Einstein's theory tells the precise shape of the cone).

Full appreciation of the manner in which this cone represents the curved space of a black hole takes some time and quiet contemplation. One feature of the cone is immediately apparent and quite important. Consider the construction of a circle on the surface around the cone. This operation must be done in the confines of the two-dimensional surface. To go off this surface into three dimensions is cheating because that would be like going from the real three dimensions of a black hole into an unphysical honest-to-gosh fourth spatial dimension. To draw a circle, start at the center of the ''black hole,'' at the bottom of the depression of the cone. Draw a line out along the curved surface directly away from the center. This line is a radius line, despite the fact that, from our three-dimensional view of the operation, it follows the funny curved surface of the cone. Now

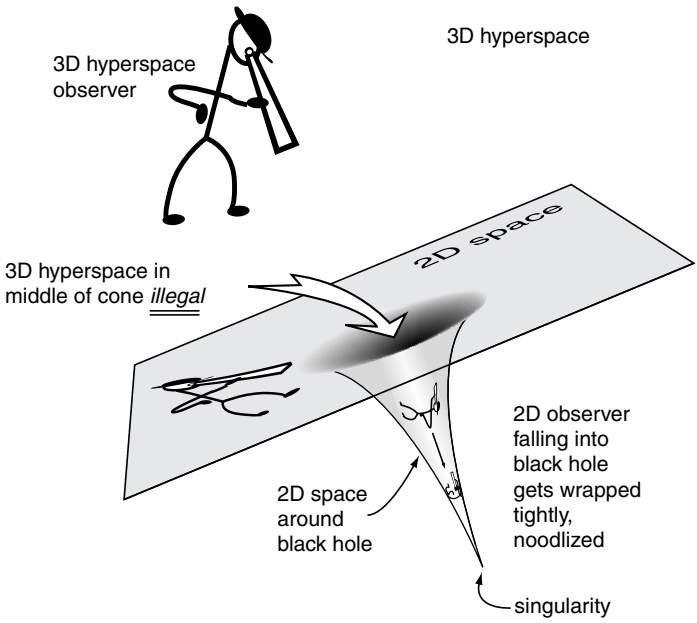


Figure 9.3 A schematic representation of the embedding diagram of the curved, two-dimensional space around a black hole. Far from the black hole the space is flat. Near the black hole, the space appears to be a ''cone'' to a three-dimensional hyperspace observer. A two-dimensional scientist falling into the black hole would be stretched toward the singularity, wrapped in the conical space, and crushed in the singularity. Note that in this view, the space corresponding to the two-dimensional black hole is on the cone. The region ''within'' the cone as perceived by the hyperspace observer is part of the higher, three-dimensional space that is imperceivable and inaccessible to a two-dimensional inhabitant of the two-dimensional space.

stop at some point along the surface of the cone and draw a circle, a line connecting all those points that are equally distant from the center.

Now imagine that you measure the length of the radius line and the circumference of the corresponding circle. Do you see that the radius in this curved surface must always be longer than normal? The ratio of the circumference to 2π times the radius is always less than one. The process of constructing the cone preserves this aspect of the original curved space, and the resulting embedding diagram lets it be seen graphically. In this curved space, the distance inward as represented by the radius is somehow stretched and lengthened. If you

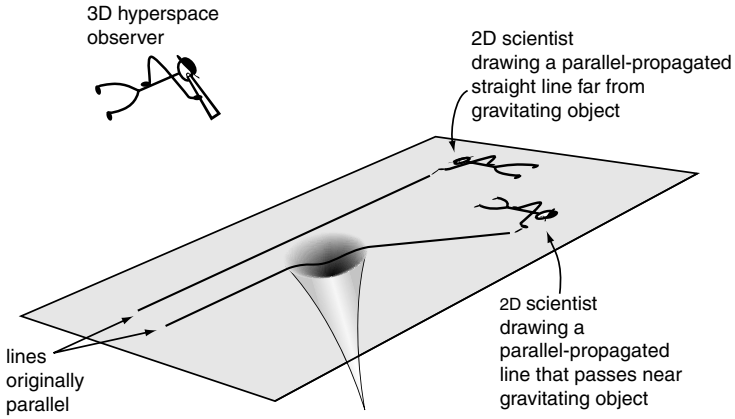


Figure 9.4 Two two-dimensional scientists draw parallel-propagated straight lines in their two-dimensional space. The lines begin parallel, but the one that responds to the curvature of the gravitating object will bend toward the center of curvature and emerge in a different direction. Both lines are legitimate straight lines in the two-dimensional space, even though one looks curved to a three-dimensional hyperspace observer.

were to go off to a flat portion of the rubber sheet and do the same operation, start at a point, go out a certain distance along a radius, make a circle, you would get the standard result – the circumference is 2π times the radius. That is the test for flat space.

Let us apply the technique of parallel propagation to the curved space around a black hole as portrayed by the projected two-dimensional cone, as illustrated in Figure 9.4. Figure 9.4 shows two scientists drawing lines by parallel propagation in the two-dimensional space they occupy. Both start at some distance out in the ''flat'' portion. One draws a parallel-propagated line that passes far from the black hole. This line looks straight to an imaginary three-dimensional hyperspace observer, the perspective we take whenever we look down from our three-dimensional hyperspace onto a two-dimensional embedding diagram. The other scientist draws a parallel-propagated straight line that skirts the deepest portion of the cone (we do not want anyone crushed by the infinite tidal forces!). As this line nears the lowest portion of the cone, think what happens. A small portion of the space surrounding this point is oriented differently than a small portion of the space out in the flat, away from the cone. The line drawn in this location is going around the axis of the cone, responding to the

''aroundness'' of the surface, despite the fact that it is going as straight as it can in the curved space of the cone. From this part of space, the line must head off in a direction different from the direction along which it originally aimed in flat space. As this line continues, it will eventually emerge into flat space once more, but in a different direction from the original line segment that started in flat space. This line is also a straight line in the two-dimensional curved space. From the superior three-dimensional position of the hyperspace observer the line looks curved. It is bent toward the center of the cone where the curvature is severe.

Looking from the point of view of the hyperspace observer is useful for perspective, but we must bear in mind that our reality is closer to that of the two-dimensional scientists. We must draw lines, do geometry, and figure out the curvature of space around gravitating objects as three-dimensional people in a three-dimensional space. We do not have the luxury of stepping out into some four-dimensional hyperspace and looking back to see how our space curves. We can determine that two initially parallel light rays passing by a star will diverge, just as the two scientists drawing the parallel-propagated lines in Figure 9.4 will determine a real divergence of initially parallel lines. The two-dimensional scientists cannot see the conical space around the gravitating object, as it is revealed to the hyperspace observer, but they can deduce its nature by doing careful geometry. They can, for instance, deduce that the radius of a circle in that part of space is long compared to its circumference.

We can explore the nature of space around a gravitating object a bit more. Think of an equilateral triangle composed of three straight lines surrounding the deepest point of the cone in Figure 9.4. Each line will look like an arc bowed outwards to a three-dimensional hyperspace observer. All observers will agree that the lines will not meet at 60-degree angles, and the sum of the interior angles will be greater than 180 degrees. How about parallel lines? Two lines drawn parallel initially will curve differently as they pass near the cone, and the one closer to the center will be bent more severely. The lines will not be parallel in the flat space to which they emerge. Lines drawn by parallel propagation will be the shortest distance between two points. A line that does not dip down in the cone must travel farther to reach a given point on the far side. Likewise, a line that goes too deeply within the cone will have wasted some motion and will have farther to climb out. There is a shortest distance between any two points, and the line that is shortest is straight, but there may be more than one

straight line between two given points. Think of a line that misses the bottom of the cone narrowly to the left. It will be bent to the right. A line that misses the bottom to the right will be bent to the left. These two lines will cross. From the point of beginning to the point of intersection, there will be two straight lines.

All this is rather abstract, but it applies to Einstein's theory of gravity in general, not just in the vicinity of black holes. Think of the straight line that just encircles the neck of the cone and closes on itself, as shown in Figure 9.5. A straight line cannot do that in flat space, but the cone shows that it is not just possible but demanded of certain straight lines in the curved space. That closed curved straight line in curved space is an orbit! In Einstein's theory, orbits are not caused by the action of a gravitational force as they are in Newton's theory. For Einstein, the gravitating body causes a curvature in space – of which our cone is a representation – and orbiting bodies are moving with no force as straight as they can in that curved space. The Moon is moving as straight as it can in the curved space around the Earth, and the Earth is moving as straight as it can in the curved space around the Sun. For such problems as planetary orbits, both Newton's theory and Einstein's give virtually the same numerical results, despite the vastly different concepts on which they are based. That Einstein's theory explains everything that Newton did in the regime of weak gravity is one of the powers of the theory. In addition, Einstein's theory predicts the nature of black holes that Newton's is powerless to describe.

Now, perhaps, you are prepared for the mind-bending exercise of attempting to picture the nature of curved space in its three-dimensional glory, with our toy two-dimensional cone as a guide. Figure 9.6 is an attempt to help do that. Draw a radial line out along the cone in the two-dimensional representation. At intervals, draw circles of constant radius, each with its own stretched-out radius. That will characterize the two-dimensional conelike surface as perceived by the three-dimensional hyperspace observer. What sort of three-dimensional curved space does the three-dimensional observer see in his own space? That's us! Imagine, if you can, rotating each of those circles in the two-dimensional space so that the swept-out locus of the rim of the circle is a two-dimensional sphere encompassing a three-dimensional volume. Now you have a set of nested spheres, but the distance from the center to the periphery of each sphere is ''stretched out.'' The distance to the center of each sphere in the empty space around a gravitating object is larger than it would have been in flat space.

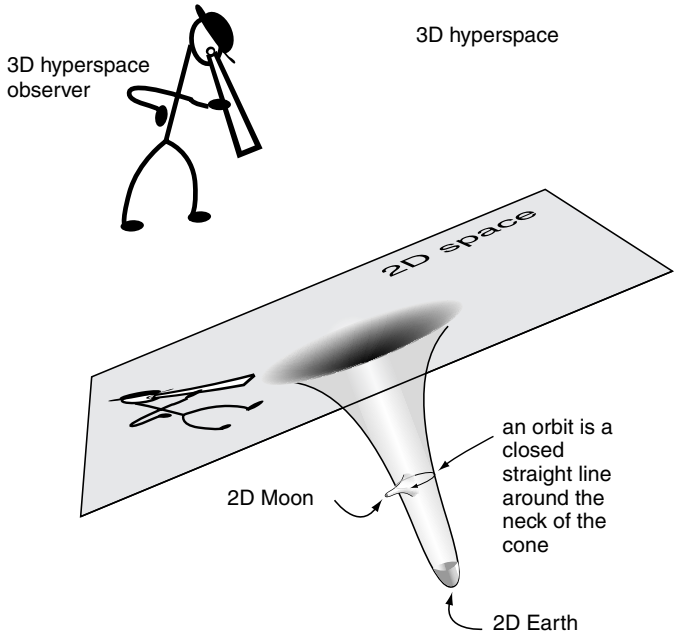


Figure 9.5 From the point of view of a hypothetical, three-dimensional, hyperspace observer, the space around the Earth would be a cone with the radius of a circle large compared to the corresponding circumference. The Moon moves as straight as it can by parallel propagating in the curved space around the Earth. In this cone-like space, one set of straight lines consists of those that close on themselves around the neck of the cone. This is Einstein's version of an orbit. The Moon, in turn, causes space to be cone-like in its immediate vicinity. This will cause rockets launched from the Earth to be deflected or to orbit even though they, also, are moving as straight as they can in the curved space. Note that the volumes of the Earth and Moon are reduced to areas in this two-dimensional representation.

This exercise is an attempt to represent the curvature of the three-dimensional gravitating space. Neither the three-dimensional observer in Figure 9.6 nor we can directly perceive this curvature as a cone or anything else. For that, we would have to be a denizen of some four-dimensional hyperspace to look down on our three-dimensional space. We simply cannot do that. We can do careful three-dimensional geometry in the confines of our own three-dimensional space and work out the nature of the curvature of our space without ever being outside of it. If you were to measure the circumference of a given sphere around a gravitating object and then

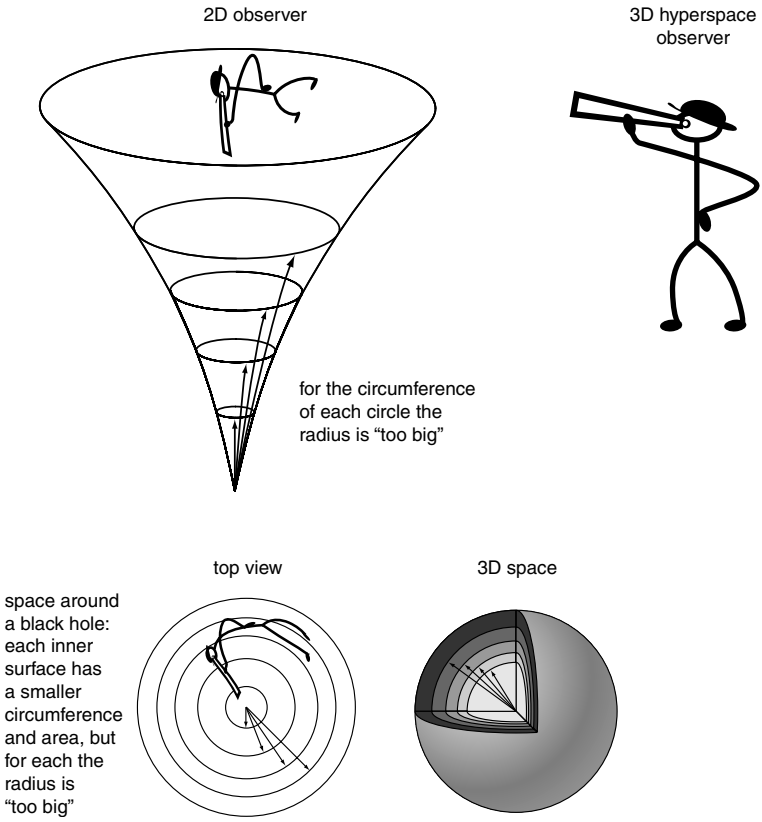


Figure 9.6 (Top) In the schematic two-dimensional curved space around a gravitating object, one can imagine circles of increasing radius and circumference. The circumference will always be smaller than 2π times the radius, and the discrepancy will be largest for the innermost circles. Both the two-dimensional resident of the two-dimensional space and the three-dimensional hyperspace observer will agree on that general property, but the hyperspace observer can see the cone-like space and the reason for the large radius is obvious. (Bottom) If the nested circles of the top diagram are rotated to map out a series of nested spheres, then one has a crude representation of the space around a three-dimensional gravitating object. Each of the spheres will have a circumference that is less than 2π times the radius. This is impossible to represent in three-dimensional space (never mind on a flat sheet of paper in this book!). A three-dimensional scientist could determine the curvature by doing careful geometry but could never ''see'' the curvature of three dimensions.

measure the distance to the center, you would find that the circumference was in every case less than 2π times the radius and that the smaller the sphere, the larger would be the discrepancy, just the property preserved in two dimensions and manifested in our cone representation. A three-dimensional scientist cannot, however, perceive where the extra length of the radius goes. All the scientist can or needs to know is that the radius is long compared to the circumference.

The important thing on which to concentrate is that such curvature exists in the space around the Earth, not just near a black hole. If you could draw a huge circle in the space around the Earth and then measure the radius of the circle, you would find that the radius was longer than you would expect if the space were flat. If you were to construct a triangle in the space around Earth consisting of three segments that are the shortest distances between the vertices, you would find that the angles added up to more than 180 degrees. All gravitating bodies curve the space around them! A black hole is only the most extreme example.

With this newfound perspective, let us return to the nature of black holes. Picture again a flat flexible sheet as a two-dimensional representation of flat, empty three-dimensional space. A star would cause a depression in the sheet. The star would be reduced to a two-dimensional spot of finite area (representing volume in the full three dimensions; check the Earth and Moon in Figure 9.5), and the depression representing curved space would extend beyond the star into the surrounding empty space. At no point within the star or beyond its surface is the curvature especially severe.

Suppose that the star were compacted to become a neutron star. This would be represented by making the spot smaller and the depression in the sheet much deeper. At rather large distances from the neutron star, the curvature of the sheet would be about the same. Near the neutron star, the walls of the depression will be nearly vertical (how one needs that three-dimensional, higher perspective to describe the goings-on!). As in the gravity of Newton, the strength of gravity depends on the distance to the center of the object. At the same relatively large distance, the gravity is the same. A neutron star has greater gravity than a normal star, not in the sense that it reaches out farther but in the sense that, because it is smaller in radius, one can approach much closer to the center of the gravitating star. A measure of the stronger gravity of the neutron star is the severity of the curvature of the flexible sheet at the bottom of the deep

depression. The sheet would change directions rapidly at the bottom, a measure of the large curvature.

When a black hole forms, all the matter is crushed into the singularity. The mass of the star is no longer represented by an area but by a point. The flexible sheet is stretched to extremes. The curvature undergoes a discontinuity at the bottom of the cone. The sheet changes directions by 180 degrees in an infinitesimal length. One can go around the neck of the cone in an infinitesimal distance (see Figure 9.3). This is a representation of the infinite tidal forces that accompany a real singularity. Somewhere down inside the depression of the cone, a circle represents the location of the event horizon. To get the full effect, you should picture the space as an escalator moving rapidly inward, flowing down toward the singularity. To move outward, you have to run up the down escalator. At the event horizon, the escalator moves inward at the speed of light. Because you cannot run faster than the speed of light in the piece of space you occupy, you are dragged down to the singularity once you cross within the event horizon.

The singularity is a region of mystery, where our present laws of physics break down. That does not mean black holes cannot exist. Einstein's theory is still quite valid at the event horizon, which is the only part of a black hole anyone will ever observe and live to tell about. The British mathematician Roger Penrose has proved what is called the *singularity theorem*. This theorem says that once an event horizon forms by any means, some singularity must form. The theorem does not prove that *all* matter must fall into the singularity once a black hole forms, but that conclusion seems somehow inevitable.

9.5.2 *Black holes and the nature of time*

Black holes cannot really be understood without a discussion of the nature of time in their vicinity. Like curved space, the flow of time is warped near and within a black hole. This makes temporal events difficult to picture in ordinary terms. One of the fundamental problems with a discussion of time in curved space is that everything depends on whose time you are discussing.

When two things are moving apart at a large relative velocity, the great Doppler shift means that all frequencies are observed to be lower. These frequencies include not only the frequency of light but also the tick of a clock, even the biological clock. Two people rocketing away from each other at great speeds will each see the other

aging more slowly than they themselves are. In the case of large gravity, there is a related effect. To an observer who is not in a large gravitational field, a clock that sits deep within the gravitational pull of some compact star will be seen to run more slowly. A person orbiting around the compact star will be seen to age more slowly. The photons that climb out of the region of highly curved space and strong gravity require some time, so that the rate of arrival of the photons at a distant observer is slow. There is a long gap between the arrival of one photon and the next. Each photon carries information concerning the ''age'' of the object that emitted it. Because the photons take longer to get out, they arrive when the outside observer has aged considerably. The outside observer detects the photons and sees the object in the gravitational field as younger.

Consider two investigators. One volunteers to fall down a black hole, giving her life for science. The other, the project scientist, volunteers to remain at a safe distance and monitor the proceedings. The first volunteer falls straight down into the black hole and by her own watch and biological clock passes through the event horizon, is noodlized, and dies in a few seconds. The project scientist, watching through his telescope, sees the watch of the falling volunteer running ever more slowly, and the volunteer herself aging more slowly. As the falling volunteer approaches the event horizon, time stops flowing from the vantage point of the distant observer, and he never sees the falling volunteer cross the event horizon. The reason is that the last photon emitted by the volunteer before crossing the event horizon takes a very long time to reach the distant observer. The distant observer can, in principle, always see some photons from the falling person, no matter how long he waits. When those laggard photons finally arrive, the distant observer sees the falling volunteer before she crossed the event horizon.

In practice, the photons that arrive at distant times in the future are highly red-shifted and difficult to detect. In addition, the time between their individual arrivals is very long. Most of the time the distant observer sees absolutely nothing. Because of the large red shift and the delay between arrival of photons, the actual perception is that anything falling into the black hole turns black very rapidly.

The term ''frozen star'' was invented to describe the mathematical solution of Einstein's theory that corresponded to the result of the absolute collapse of a star. This term focused on the fact that a distant observer can never see the surface of the star fall through the event horizon. There is thus a suggestion that the surface of the star

somehow lingers at the event horizon to be touched, and probed and explored. The term ''black hole'' was coined by John A. Wheeler in 1968 at a meeting in New York City on pulsars. Wheeler tried to come up with a graphic term to encourage his colleagues to contemplate even more extreme states of gravitational compaction than white dwarfs and neutron stars. The name ''black hole'' concentrates on the collapse and the fact that the star rapidly turns completely black, and on the fact that, after collapse ensues, no part of the star can ever be recovered. If you tried to fly down and grab some of this frozen star, you would find that the surface receded from your grasp as your time became its time and you could see it fall once more.

The term ''black hole'' is much more pertinent to the real situation because it directs attention to the actual collapse and to the interior of the black hole. The case is difficult to prove, but there is a sense that the term ''black hole'' itself spurred some of the marvelous work that followed. With this new term and new mode of thinking came complete mathematical solutions of the interior of black holes, where people's minds can reach, even if their bodies cannot.

9.6 Black-hole evaporation: Hawking radiation

As remarked earlier, Einstein's theory, for all its magnificence and success, is not complete. This theory is a so-called classical theory in that it incorporates none of the principles of the quantum theory. In Einstein's theory, as in Newton's, all motion and changes are smooth, and all positions can, in principle, be specified exactly. Einstein's theory is not compatible with our understanding of microscopic physics as described accurately by the quantum theory.

9.6.1 *Quantum event horizons*

The first successful attempt to include some of the principles of the quantum theory was done by the brilliant theoretical physicist from the University of Cambridge, Stephen Hawking. The process by which energy is converted into equal parts matter and antimatter is intrinsically a quantum mechanical process. Hawking's genius was to see how to add a little of the quantum process into the otherwise classical realm of Einstein's theory. He showed that the gravitational energy associated with the curved space in the vicinity of an event horizon will create particles and antiparticles. In principle, electrons and positrons, or even protons and antiprotons, could be generated. The

easiest particle to make, however, is the photon because it has no mass (technically speaking, a photon and an antiphoton are one and the same thing).

According to the quantum theory, no position can be specified exactly. This applies equally well to the position of the event horizon around a black hole. Because of the intrinsic quantum mechanical nature of things, you cannot say definitely whether something is inside or outside the event horizon, only whether something is probably inside or outside the event horizon. The location of the event horizon is then fuzzy. When two photons are created in the vicinity of the event horizon, there is a probability – purely quantum mechanical in nature – that one photon will be inside the event horizon and will disappear down toward the singularity, and the other will be outside the event horizon and fly off to great distances where it can be detected. Hawking's great discovery was that black holes are not truly black. They shine with their own radiance generated from pure gravitational curvature!

9.6.2 *A two-way street*

The physical implications of this discovery were immense and caused a wrenching turnabout in our view of black holes. The energy to create the radiation came from the gravitational field, but the gravitational field came from the mass of the matter that had collapsed to make the black hole. When the photons carry off energy, the energy of the black hole must decline. This can only happen if the mass of the black hole declines as well. As black holes emit Hawking radiation, they are shining away their very mass! Black holes are not completely one-way affairs after all. Even though it is still true that tidal forces will tear an object beyond recognition as it falls into the singularity, the mass is not gone forever. It will emerge later in the form of the Hawking radiation to permeate the Universe. A black hole is just nature's way of turning all that bothersome matter into pure random radiation. We will see that nature has yet other tricks with the same fate in mind. Gather ye rosebuds while ye may, a photon yet ye'll be!

Hawking discovered that the black hole radiation does not come out in an arbitrary fashion. The spectrum of the radiation corresponds exactly to a single temperature, when it might have been some odd, nonthermal shape. The temperature is determined in turn by the mass of the black hole. The variation with mass is inverse so that a massive black hole has a low temperature, and a low-mass black hole

has a higher temperature. For a black hole of stellar mass, the temperature is very low. Little radiation could be emitted in a time as short as the age of the Universe, and so the radiation is of little practical importance. Our standard picture of black holes as gaping one-way maws holds true.

9.6.3 *Mini black holes*

If the mass of the black hole should be less than that of an average asteroid, however, the situation is markedly different. Such small black holes would be very hot and would radiate prodigious amounts of radiation. As these small black holes radiate, their mass shrinks so they get hotter and radiate even faster. The process runs away faster and faster. In less than the age of the Universe, such small black holes could evaporate completely! The final stages of this process are so accelerated that the last energy would emerge in an explosion of high-energy gamma rays.

These so-called *mini black holes* could not be created in the collapse of an ordinary star. They might have arisen in the turbulence that may have marked the original state of the big bang. If this were the case, there could be swarms of mini black holes in the Universe, some of which would be explosively evaporating at any time. The properties of such explosions have been worked out theoretically, and the radiation has been sought, but so far unsuccessfully. The notion that such tiny black holes could exist persists, however, and we will touch on a modern view of the role they could play at the deepest levels of physics and cosmology in Chapter 14 (Section 14.5).

9.6.4 *White holes*

One can imagine (mathematically) the reverse of a black hole, or a *white hole*. A white hole is obtained by running time backward compared to the flow of events for a black hole. For a black hole, one starts with ordinary space. A star collapses to make a black hole, and then you have a black hole forever, gobbling up matter, but releasing nothing (forgetting for the moment Hawking radiation). Now run the movie backward in time. One must start with a white hole that has existed since the beginning of the Universe, spewing forth matter but swallowing nothing. At some time, the ‘‘last stuff’’ pours forth, and one is left with empty, flat space.

Black holes are regarded seriously because we can predict that they might well occur in the course of stellar evolution and because we think we have found them, as Chapter 10 will show. From the properties of known stars, the properties of the resulting black holes can be predicted. White holes are not regarded on the same footing because they must exist since the beginning of time. Their properties cannot be predicted because we cannot predict the beginning of the Universe. White holes could have any property – large mass or small. Because we cannot predict their properties, white holes have no firm place in the realm of ordinary pragmatic physics.

Hawking's discoveries may have been a first step toward putting the notion of white holes on a firmer basis. Hawking has blurred the distinction between white holes and black holes by introducing quantum mechanical properties to the event horizon. Now we see that a black hole can emit radiation, a property previously reserved for white holes. Likewise, a white hole should be able to swallow radiation. Hawking has argued that for very small objects the distinction between white holes and black holes may disappear.

9.7 Fundamental Properties of Black Holes

For all their exotic nature and the complexity of the theory that treats them, black holes can have only three fundamental intrinsic properties. These properties are their mass, their spin or angular momentum, and their electrical charge. These properties are distinguished because they can be measured from outside the black hole and, therefore, determined by ordinary techniques. The mass can be determined by putting an object in orbit around the black hole and seeing how fast it moves. The charge can be determined by holding a test charge and detecting the force of attraction or repulsion from the hole. In practice, one expects real black holes to be electrically neutral because they should rapidly attract enough opposite charge from their surroundings to neutralize any charge that might build up. Measurement of the spin of a black hole is a more subtle process. As the black hole rotates, it drags the nearby space around with it. This dragging can be measured, in principle, like the currents in the ocean. Once the mass, spin, and charge of a black hole are known, all its other intrinsic properties are set. For instance, for a noncharged, nonspinning black hole, the size given by the radius of the event horizon is strictly proportional to the mass. The temperature of the

Hawking radiation varies inversely with the mass. Other properties that a black hole might have, but cannot, are mountains like the Earth or sunspots and flares like a star. On a more fundamental level, black holes cannot have the property of a lepton number or a baryon number. The forces associated with leptons and baryons are short range and cannot extend outside the event horizon where they can be measured. Black holes do not so much violate the laws of conservation of lepton and baryon number as transcend them. In the realm of black holes, these fundamental physical laws of ordinary space are irrelevant. John A. Wheeler has coined an aphorism to describe this raw simplicity of black holes – he says ''black holes have no hair.''

To illustrate the power of this notion, consider two compact stars. Let one be made of neutrons, an ordinary neutron star. Let the other be made of antineutrons, an antineutron star! If these two stars were to collide, the neutrons and antineutrons would annihilate to produce pure energy and an explosion of unprecedented proportions. Suppose, however, we dump a few too many neutrons on the first star and it collapses into a black hole. Then we add some antineutrons to the second star so that it, too, collapses to make a black hole. Do we now have a black hole and an anti black hole? No, we have two identical black holes because the black holes transcend the law of baryon (neutron and antineutron) number. If the two black holes combine, the result is not an explosion but one larger black hole. The form of mass that originally collapsed to make a black hole becomes irrelevant after it has passed through the event horizon. Then only the total mass counts. While he was warming up, Stephen Hawking presented to the world the laws by which black holes combine to make larger ones, an exercise that alone would have assured his reputation as a brilliant physicist.

9.8 INSIDE BLACK HOLES

Just because black holes have only three fundamental properties does not mean that their nature, which derives entirely from specifying the values of those three properties, is not complex. Apart from quantum effects, the exterior of a black hole, the event horizon, is a model of simplicity: smooth, perfect, and unperturbed. The insides, however, as exposed by the powerful techniques of mathematics, are a wonder such as to strain one's credibility to the limits.

9.8.1 *Time-like space*

When we discussed the oddities of the flow of time near black holes (Section 9.5.2), we omitted the oddest twist of all. This aspect can never be observed directly, but it is the real factor that accounts for the existence of the event horizon that blocks our view. Inside the event horizon, space takes on the aspects of time (cf. Figure 9.1). No matter how rockets are fired or forces applied, any object must move inward toward the singularity (or outward, if we are dealing with a white hole) as it ages. There is no choice in the matter, just as you have no choice in the matter of your aging from eighteen to thirty-one. The same principle that drags you on into old age drags an object within the event horizon ever closer to the singularity. Within the event horizon, space is no longer the entity in which you can move around in three dimensions with impunity. There is only one direction, inward. The one-way nature of this space is intimately related to the one-way nature of time. Inside a black hole, space is time-like! The time-like nature of space is the reason that everything goes inward inside a black hole, and nothing can get out. It is the reason black holes are black.

9.8.2 *Schwarzschild black holes*

The simplest black hole is one with mass, but no charge or spin. This kind is called a *Schwarzschild black hole* after the physicist who first gave a mathematical description of such a beast, shortly after Einstein presented his general theory of relativity. There is a poetry to this name that is rendered as black shield from the German. This was the type of black hole illustrated schematically in Figure 9.1.

For a Schwarzschild black hole, the event horizon coincides exactly with what is called the *surface of infinite red shift*. A photon emitted from this surface will have an infinitely long wavelength by the time it escapes to great distances. The event horizon is round for a Schwarzschild black hole, and the singularity is a point at the center of the black hole.

Mathematical investigations have shown that even the lowly Schwarzschild black hole is not so simple. In the idealized case, where one assumes that all the mass is confined to the singularity and that a vacuum exists everywhere else, a black hole is really twain, two equal geometries sharing the same singularity. Each black hole has its own universe of empty flat space. These two universes exist at the same

instant but in different places. When moving at less than the speed of light, one cannot travel from one to the other but will instead fall into the singularity if passage between them is attempted. This idealized mathematical description does not apply to a black hole that has formed from the collapse of a star. Then the matter of the star introduces other changes in the geometry and curvature of space that are, as yet, too complicated for anyone to have been able to calculate. The ''other universe'' is undoubtedly just a mathematical fiction, but it gives a portent of the richness to come.

9.8.3 *Kerr black holes*

One has only to introduce some rotation to the black hole to complicate affairs in the most interesting fashion. The first basic mathematical solution corresponding to rotating black holes was discovered by the New Zealand physicist Roy Kerr, in 1963. Subsequently, the complete solution of the interior of a rotating black hole was worked out by others, but these black holes are still referred to as *Kerr black holes* to distinguish them from Schwarzschild black holes.

If a black hole rotates rapidly enough, the event horizon disappears completely. In this case, one could look directly into the fearsome maw of the singularity. Such a beast is known as a *naked singularity*, a singularity unclothed by an event horizon. There is no formal proof as yet, but there is a strong belief that no black hole can rotate fast enough to create a naked singularity. Certainly any star that rotated so fast would fling itself apart before it could collapse to make a black hole. Firing matter into a black hole tangentially would spin it up. Calculations show, however, that as the black hole nears the limit where the last veil might be dropped, gravitational radiation will become so intense as to carry away any increment in rotational energy. Perhaps there is some way to create a naked singularity, but it seems very difficult. Many researchers have adopted the as yet unproven doctrine that naked singularities cannot exist in the real world of astrophysics. This doctrine that nature denies freedom of expression to unclothed singularities is known informally as ''cosmic censorship.'' Stephen Hawking, a firm believer in cosmic censorship, bet Kip Thorne of Caltech that naked singularities cannot exist. He paid off on the bet when the carefully designed computer models of Matt Choptuik yielded naked singularities. No one has yet found one in their backyard.

Real rotating black holes may have matter swarming around inside the event horizon that will substantially alter the geometry of the inner reaches. The best we can do is to follow the mathematician's description of the idealized case where, once again, the assumption is made that all mass is confined to the singularity, and that all the rest of space is pure vacuum. The result is illustrated schematically in Figure 9.7. Welcome to Wonderland, Alice!

The first thing one discovers in the study of rotating black holes is that the singularity is not a point but a ring! One can imagine an intrepid explorer plunging through the center of the ring, avoiding the infinite tidal forces of the singularity itself. Retreating now to the outside, we find that for a rotating black hole the surface of infinite red shift separates from the event horizon. Both surfaces are oblate, flung out around the equator by centrifugal forces, but the surface of infinite red shift is more extended. There is a finite distance between the surface of infinite red shift and the event horizon at the equator. At the poles of the rotation axis, the two surfaces are still contiguous.

The surface of infinite red shift has another property. It is also the *stationary limit* with respect to sideways motion. The rotation of the black hole drags the local space around in the same sense as the hole rotates. The effect is stronger the closer one is to the black hole. At a moderate distance, one could fire rockets and overcome the effect in order to hover in one place. This requires some effort, like swimming upstream or walking up the down escalator. At the stationary limit, all efforts to remain still are fruitless. To resist moving around in the same sense as the black hole spins, one would have to fly backward in the local space faster than the speed of light. Inside the stationary limit, all material objects, including photons of light, are forced to rotate with the hole.

On the other hand, because the surface of infinite red shift is removed from the event horizon at the equator, one can, in principle (ignoring the huge tidal forces), fly inside the surface of infinite red shift and return. This can be done by moving with the rotation of the black hole, the path of least resistance. Some paths lead into the event horizon, and there will be no return; however, with a rotating black hole, the option exists to emerge from within the surface of infinite red shift.

The region between the surface of infinite red shift and the event horizon is called the *ergosphere*. This phrase was coined by Roger Penrose (of the singularity theorem) who investigated its properties. It derives from the Greek word *ergo*, meaning work or energy. Penrose

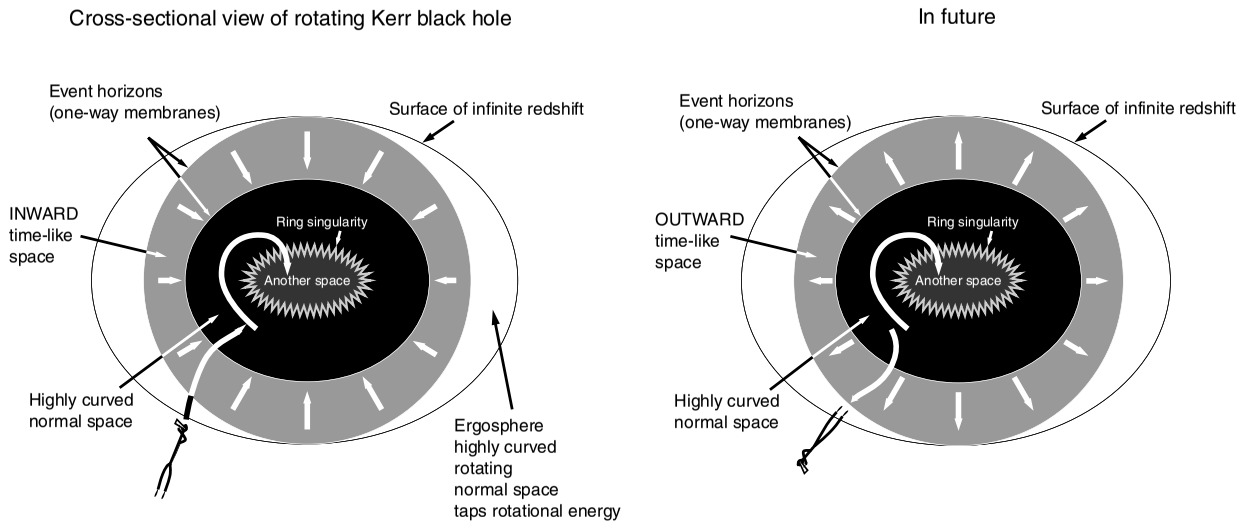


Figure 9.7 (Left) Schematic cutaway view of a rotating Kerr black hole illustrating the complex structure of the geometry. In this geometry one could, in principle, fly through the ingoing time-like space between the event horizons and survive in the highly curved, but "normal," space within that surrounds the ring singularity or the different space one would find by passing through the ring. (Right) In the future of the left-hand diagram is a portion of the geometry that would, again in principle, allow one to fly out through outgoing time-like space into a normal, low gravity space, but in another universe than the one from which one entered.

found that, under proper circumstances, energy could be extracted from the black hole. If one of a pair of particles is fired down the hole in a counterrotating sense from within the ergosphere, the recoil will throw the other particle out with more energy than both particles had originally, including their mass energy, $E = mc^2$. You do not get something for nothing. In this case, the excess energy in the ejected particle comes from the rotational energy of the black hole. After the particle is ejected, the black hole will be rotating less rapidly.

There is some question as to whether this *Penrose process* for tapping the energy of a rotating black hole can be of real astrophysical interest. The problem is that a considerable investment of energy must be made in firing the first particle into the event horizon in the proper fashion. A puny nuclear explosion would be far from sufficient; the particle must be moving at nearly the speed of light. Such reactions with massive particles may not occur spontaneously in nature with any reasonable probability. On the other hand, photons are already moving at the speed of light. There have been discussions of Penrose processes operating to swallow some photons and eject others at high energy. This process is also driven by the rotational energy of the black hole and is termed *superradiance*. There is some speculation that the gamma rays seen from quasars could be produced in this way, starting with photons in the more conventional X-ray or ultraviolet range that are produced in the inner edges of a hot accretion disk.

Let us now journey into the event horizon. As we pass within, we come to a region of time-like space in which we must move inward as we age. There is a crucial difference in the rotating case, however, for there is an inner boundary to this time-like region. At this inner boundary is another event horizon, which prevents a return to the space beyond. Within this second event horizon is a region of normal, if highly curved, space. This event horizon prevents a return to the time-like space, rather than preventing a return to normal space.

Within this inner volume of normal space is another surface of infinite red shift, but because one can move in and out of such a surface if appropriate moves are taken, it has no direct consequence. Around the equator of this inner surface of infinite red shift is the line we devoutly wish to avoid. That equatorial line is the location of the ring-shaped singularity. If we stumble against that, we are doomed by the infinite tidal forces.

The special property of this inner region of normal space is that we could elect to stay here forever. By careful choice of movement, we

can orbit around and never strike the singularity itself. This is very different from the case for a nonrotating black hole. There, the time-like space leads inexorably to the singularity.

Other options await if we continue our imaginary journey within the spinning black hole. At the same place, but in the future, there is a similar space–time structure. Here, however, the sense of the event horizons and time-like space are reversed. As one flies about, one could in principle elect to head outward, passing through an event horizon into a region of outgoing time-like space. This would be bounded by an outer event horizon, and beyond that would be an ergosphere, a surface of infinite red shift, and finally free space. Formally, mathematically, this is not the space from which we entered, but another, separate universe. The mathematical solution shows that in this new universe there will be another ingoing black hole like the original one we entered, so one can plunge down again and come out in yet a third universe. The idealized mathematical solution we are exploring has an infinite number of universes, all connected by rotating black holes!

Let us return to the central regions of the rotating black hole. We found there a more or less spherical region of normal space inside of which lay the ring singularity. Watch carefully now, Alice! The plane of the ring singularity divides the volume into two halves. You can maneuver from the top half, out through the inner surface of infinite red shift, and back in, to come to the bottom half. Alternatively, you could elect to plunge straight through the hole in the middle of the ring. In so doing, you would come to *a* bottom half, but not the one accessed by going out and around the ring. If from this new lower half you went out and around, you would be in a top half, but again not the one from which you started. The space through the ring is not the space you get to by going around the ring. If this is not passing through the looking glass, what is? You can imagine looking down through the ring and seeing another creature, perhaps a puce-colored eight-legged cat. If you go out around the singularity and look, you will not see the creature. Its space is only through the ring, not behind it.

If you join the creature through the ring, you can seek, in the future, a set of outgoing event horizons. These will again lead to an outer, flat universe, which is none of the ones we have discussed previously. As you leave this black hole, you will feel it pushing you. Unlike the others we have explored, this outgoing solution that exists through the ring antigravitates!

Having entertained ourselves thus, we must return to more sober reality. We do not diminish the wonder of the tale to point out again that what has just been described is an idealized mathematical solution. It is a marvelous, exact solution to the full set of equations describing general relativity. Nevertheless, a crucial assumption has been made in order to solve the equations at all. The assumption is that there is no mass anywhere except in the singularity. The presence of any matter or energy within the first set of event horizons would cause a change in the curvature and geometry, and the wonderful world of multiple universes would probably vanish. The solution to the equations with even a little matter present throughout the volume would not contain any of the extra spaces, in the future or through the ring. Even the presence of an explorer such as we imagined ourselves to be could change the whole situation.

Some research has been done to see what happens to the mathematical solution if the tiniest bit of extra matter is added inside the black hole. There is a strong suggestion that the whole geometry would begin to rattle and shake with the resultant generation of an intense flux of gravitational radiation. This radiation alone would alter the physical and mathematical situation, to eliminate the reality of the extra spaces and universes. At the very least, in the real Universe, photons of light will continue to flood down the black hole. As they plummet in, they are blue-shifted and attain incredible energies. This energy will build up at the event horizon in what has been termed a *blue sheet*. This sheet of energy would warp the geometry and wipe out any of the multiply-connected interior geometry.

The mathematical ''vacuum'' solution to the Kerr black hole is a marvelous, mind-stretching exercise. It probably has nothing to do with the guts of a real star-born black hole, rotating or not. The reality is fantastic enough, as we shall see in Chapters 10 and 11, and the mystery of the singularity remains. Black holes may form from stars, but they are vastly different from stars. One way to see this is to examine the intellectual frontiers to which research on black holes has led. There one finds mind-bending concepts of wormholes, time machines, multidimensional space, self-reproducing universes, and radical new notions of how to think of time and space under conditions where neither can exist. Those are the topics of Chapters 13 and 14.

10

Black holes in fact: exploring the reality

10.1 THE SEARCH FOR BLACK HOLES

Black holes, those made from stars, are really black! How can we hope to find them if they do exist? Some solitary massive stars may collapse to make isolated black holes drifting through the emptiness of space. There could be very many of these black holes. Estimates based on the number of massive stars that have died in the history of our Galaxy range from one to a hundred million black holes. The simple fact is that, until a space probe stumbles into one, we are likely never to detect this class of isolated, single black holes. We will certainly never see the black hole itself in any circumstances because no light emerges from it. Our only chance to detect the presence of a black hole is to find a situation where mass is plunging down a black hole, heats, and radiates. We can hope to detect the halo of radiation from such an accreting black hole, even if we never see the black hole itself. Black holes are so strange and so significant that the standard of proof must be exceedingly high. As we will see, the evidence is very strong, but still largely circumstantial.

Many astronomers search for giant black holes in the centers of galaxies. The evidence for those black holes has become rather strong in the last few years, but most of the evidence still involves matter moving far beyond the event horizon, and we know very little about the configuration of the accreting matter. There is no question that there are concentrations of gravitating mass in the centers of galaxies, including our own, that contain millions if not billions of solar masses, are small, and are not radiating anything like an equivalent amount of star light. One idea is that they could be a cluster of compact stars, neutron stars, or stellar-mass black holes, but the theory of such swarms of objects says they should quickly collide and

merge and make one large black hole. With some theoretical underpinning and compelling circumstantial evidence, the argument for these giant black holes is rather convincing. There are clues from the X-rays from some galactic cores that the space near the very center has just the character you would expect for that around a rotating, supermassive, Kerr black hole. More evidence of this kind may remove any ambiguity.

Another excellent hunting ground for black holes has proved to be in binary star systems, where mass transfer can feed the accretion and produce X-rays in the high gravity of a stellar-mass black hole. Here also the case has become very strong that we are observing black holes. This facet of black-hole research is closely connected to the topics covered in this book, so this story is worth telling in more detail.

Over thirty strong X-ray sources have been established to be in binary systems. Of these systems, about a dozen have some determination of the mass of the X-ray source itself. In most cases, the mass is in the range of one to two times the mass of the Sun. These are probably neutron stars. In some cases, pulsations are observed, and the case for rotating, magnetized neutron stars is clearly established. One should perhaps bear in mind, however, that, although a neutron star cannot have a large mass, there is no reason in principle why a black hole could not have a modest mass, particularly if it formed by adding a bit too much mass to a neutron star. We still have no unambiguous way of determining that we have a black hole with a mass less than the maximum mass of a neutron star, although there are some ideas for how to do this.

In the case of a black hole, there is no question of radiation from the surface of the object because there is no matter, only the ephemeral event horizon. All the X-rays must come from matter in the accretion flow. Within about three times the radius of the event horizon of a black hole, the gravity is so strong that the matter cannot spiral in a disk but must plunge headlong into the hole. In this state, the matter radiates much less because it is not subject to the friction of the accretion disk. In addition, the radiation emitted from this region is highly red-shifted, so it is difficult to detect with X-ray devices. Any X-rays detected from an accreting black hole will come from a halo in the disk, inside which there is only blackness. This particular way in which X-rays are emitted may prove sufficiently different from the X-ray emission mechanisms for neutron stars that black holes can be unambiguously identified, independent of their mass. For now, the story is a bit less certain.

10.2 CYGNUS X-1

One of the first binary X-ray sources discovered is a candidate black-hole system. This object was the first X-ray source discovered by the *Uhuru* satellite in the direction of the constellation Cygnus. Soon after its discovery, astronomers were describing Cygnus X-1 as a possible black hole. Absolute proof escapes us, but the net of circumstantial evidence has grown ever tighter. Cygnus X-1 is probably a black hole.

The chain of arguments proceeds like this. The fact that Cygnus X-1 emits a strong flux of energetic X-rays at all argues that it is a compact object with a large gravitational field. It could be a white dwarf, a neutron star, or a black hole. The intensity of the X-rays argues against the white dwarf possibility. Added evidence against a white dwarf is that the X-rays from Cygnus X-1 flicker on a timescale of milliseconds. We can use an argument based on how far light can go in a given time to say that the object must be smaller than the distance light can travel in 0.001 second. That distance is 300 kilometers, consistent with a neutron star or a black hole, but too small to be a white dwarf. A white dwarf would be too large and sluggish to vary rapidly. The conclusion that Cygnus X-1 is not a white dwarf, never mind an ordinary star, seems quite sound.

This leaves us with a neutron star or a black hole as the necessary object. There may be a foolproof way to tell the difference from the nature of the X-ray emission alone, but that argument is still under development and is difficult to apply cleanly to Cygnus X-1. Many feel that the millisecond fluctuations are themselves evidence of the nature of a black hole, but that has not been proven. The lack of regular pulsations is not sufficient because the object could be a slowly rotating or unmagnetized neutron star that could not produce detectable pulses. The only way we know to distinguish between a neutron star and a black hole is to argue that a black hole can exceed two or three solar masses, and, as discussed in Chapter 8, a neutron star cannot.

Careful study of the Cygnus X-1 system, both the X-ray source and its companion massive star, shows that the companion has a mass of about 30 solar masses, and the X-ray source a mass of about 10 solar masses. The latter is too much to be either a white dwarf or a neutron star. By a process of elimination, the reasonable conclusion seems to be that Cygnus X-1 is a black hole.

The presumption behind this chain of reasoning is that the massive star transfers mass to the black hole, and the infalling matter

emits X-rays before it plunges into the black hole, but all we really know for Cygnus X-1 is that a 10 solar mass ''thing'' is emitting X-rays. As an example, let us consider a way in which nature might be playing a trick on us. We know that triple-star systems are present in the Galaxy. We noted in Chapter 3 that the nearest star, Alpha Centauri, is in a triple system. Suppose that Cygnus X-1 consists of a neutron star of 1 solar mass orbiting an ordinary star of 9 solar masses, and that the pair of them are orbiting another ordinary star of 30 solar masses. If the 9-solar-mass star transfers mass to the neutron star causing the emission of X-rays, then we will have an X-ray source with total mass of 10 solar masses orbiting a 30-solar-mass star, just as the observations demand, yet there would be no black hole. This picture is unlikely, but not entirely impossible. The reason we can consider it at all is that the 30-solar-mass star would be considerably brighter than the 9-solar-mass star, so the latter could be lost in the glare. Attempts have been made to detect such a masquerading companion by searching for faint spectral lines that would shift around among the spectral lines of the brighter star as the Doppler shift responds to the orbital motion. No hint of such a secondary star has been forthcoming. It probably is not there, but a tiny doubt will always linger.

The massive companion to the X-ray source in Cygnus X-1 is blowing a stellar wind, as such stars do. The picture adopted for Cygnus X-1 is that the gravity of the black hole traps part of the wind. That matter then swirls into an accretion disk. The matter then spirals down, and the friction heats the gas to temperatures where the matter radiates X-rays. The companion is transferring mass at a sufficiently slow rate that it seems unlikely that the black hole in Cygnus X-1 could have started as a neutron star and then collapsed to a black hole, and subsequently grown to its present mass before the companion died. The presumption is that the black hole formed directly by the collapse of a 10-solar-mass object.

It does not follow that the black hole arose from a star whose initial mass was only 10 solar masses. A more likely prospect is that the progenitor star had a mass of around 35 solar masses. The other star, the normal companion that still exists, probably had about the same mass we see now, around 30 solar masses. Stars of 30–35 solar masses develop helium cores of about one-third their original mass. The originally more massive star thus probably grew a helium core of about 10 solar masses as it burned up the hydrogen in its center. At the same time, the star probably lost a great deal of mass due to its own stellar wind. The most likely time for this is when the originally

more massive star finally exhausted its central reserve of hydrogen and began to become a red giant. At this time, any mass remaining above the helium core probably flowed out of the binary system or onto the companion star. During this episode, the companion could have lost some mass to a wind and gained some from the more massive star, so it did not change appreciably.

Even though it has lost its hydrogen blanket, the now bare 10-solar-mass core of the first star is so massive that it is supported by the thermal pressure and continues to evolve with regulated nuclear burning. The core presumably burns a series of nuclear fuels until it forms an iron core. This core collapses, but instead of producing the explosion of a supernova, a black hole forms. All the matter in the core rains down through the event horizon. The net effect is that the 10-solar-mass black hole did not come from a 10-solar-mass star but more likely from one originally with somewhat more than 30 solar masses. The corollary implication is that this star did not explode but left a black hole instead. One is invited to think that all stars in this mass range, greater than 30 solar masses, leave black holes. A possible problem with this reasoning is that the very fact that the star was in close orbit with a massive companion may have altered the evolution in a way we do not understand. As we discussed in Chapter 6, there is little direct evidence concerning the end point of massive stars of a given initial mass. In any case, a common presumption is that stars of about 30 solar masses must explode to provide the heavy elements. Clues that stars of this mass make black holes means that there is no strong evidence to support this presumption.

10.3 OTHER SUSPECTS

Further observations showed that there are binary systems emitting X-rays that provide even better evidence for black holes than the famous Cygnus X-1.

One of these systems is LMC X-3. This object is the third X-ray source discovered in the nearby galaxy, the Large Magellanic Cloud, which also played host to Supernova 1987A. LMC X-3 is similar to Cygnus X-1 in that the X-ray source seems, from a study of orbital parameters, to have a mass of about 10 solar masses, and hence to be too massive to be a neutron star. In this case, however, the companion star is only about 10 solar masses as well. This means that it is much more difficult to hide a third star in the glare of the ordinary star than in the case of the more massive, and brighter, companion in the

Cygnus X-1 system. A three-body system with a neutron star orbiting an undetected normal star, with both orbiting the observed normal star, would be untenable. There would be obvious evidence of the third star. LMC X-3 may thus be a better candidate for a black hole than Cygnus X-1 because one cannot resort to the dodge of hiding some other source of mass and gravity in the system.

There is, however, a system in our Galaxy that is an even better candidate for containing a black hole in orbit. That is the system with the boring moniker AO620–00, named for its directional location in the Galaxy. This system seems to have a 5-solar-mass black hole orbiting a normal star that is not massive at all but about one-half the mass of the Sun. It is not clear how a star with original mass of about 30 solar masses, that could have a core of about 10 solar masses, which in turn could collapse to make a black hole, would come to have such a wimpy companion. Usually, massive stars seem to hang out with one another. On the other hand, nature may be tricking us here. If every 30-solar-mass star had a 0.5-solar-mass companion, the dinky star would be lost in the glare, and we would never know it. Nature may form stars in this way much more frequently than we realize, or there may be something else going on that is special to black-hole systems. One suggestion is that the little companion star forms from the matter spun off the star that forms the black hole. In any case, the small-mass, dim companion means that it is virtually impossible to hide another star in the system to trick us into thinking that an X-ray-emitting neutron star had a higher mass, therefore masquerading as a black hole.

Another argument adds to the case. AO620–00 underwent at least two outbursts that produced an excess light output, one in 1917 and one in 1975. The 1975 eruption produced a corresponding detected burst in X-rays. These bursts lasted for about a month and, in the optical at least, are rather reminiscent of dwarf-nova outbursts. Models of the behavior of accretion disks around black holes reproduce the properties of the optical and X-ray bursts with the same kind of physics that works for dwarf novae, as discussed in Chapters 4 and 5. The accretion disk collects matter until it undergoes an instability that dumps matter into the black hole at a greater rate, resulting in the outburst.

The arguments are still circumstantial. What we know is that AO620–00 contains an orbiting object with a large mass that emits X-rays but virtually no optical light. Nevertheless, it is very difficult to see how AO620–00 could be anything but a black hole.

There is a bit of a tendency to cry "black hole" whenever a strange new astrophysical phenomenon involving high energies turns up. That is one reason most astronomers are trying to be as conservative as possible about concluding that Cygnus X-1, LMC X-3, and AO620–00 are black holes. There is another danger: there are other black holes out there, and we are being too conservative to face the facts. The last few years have revealed that the Galaxy is full of systems like AO620–00.

10.4 BLACK-HOLE X-RAY NOVAE

One way to beef up our confidence that Cygnus X-1, LMC X-3, and AO620–00 are black holes is to find others. There is safety in numbers. The problem is that the combination is rare wherein a massive star makes a black hole, and we catch a comparably massive companion as it is transferring mass, but before the companion also dies. Only about one such pair should exist in the Galaxy at any one time. We may have discovered that one rare event in Cygnus X-1. It is possible that LMC X-3 is the only currently active black hole with a massive companion in that smaller galaxy, just as Cygnus X-1 may have that single merit in our Galaxy. The formation of black holes is associated with massive stars, and Cygnus X-1, the grandaddy of black-hole candidates, has a massive companion. The feeling lingered for a long time that all black-hole binaries, if they existed, would resemble Cygnus X-1. In the last decade or so, we have learned that the Galaxy is full of binary black-hole candidates, but, like AO620–00, they are wonderfully and surprisingly different from Cygnus X-1. These systems are even better candidates for black holes than the venerable Cygnus X-1, and they present better laboratories to explore the astrophysics of black holes.

Two basic characteristics distinguish the new class of black-hole candidates, of which AO620–00 is the prototype. They show a distinct transient behavior, and they have low-mass, relatively dim companions. These systems maintain a quiescent state for decades and then erupt in a sudden burst of energy. The energy output appears throughout the range of electromagnetic waves from radio to gamma rays. There is especially interesting behavior in the soft and hard X-ray bands. The outbursts last for about a year, and then the system fades to quiescence again. In the quiescent state, the only evidence of the system is the small-mass companion. Without an eruption to draw the attention of astronomers, these stars are lost among the billions of similar stars in the Galaxy. Without the ability to detect the associated

high-energy emission in X-rays and gamma rays, even the outburst may pass without special notice. Such eruptions may have been mistaken for classical novae in the past.

When AO620–00 underwent an outburst in 1917, before the invention of X-ray astronomy, it was taken for an ordinary nova. AO620–00 had a dramatic X-ray outburst in 1975, but it was several years before evidence came in that it might harbor a black hole. Only relatively recently has the realization dawned that the Galaxy contains many of these systems. The coverage of the sky with satellites that can monitor X-ray outbursts has been fairly thorough for the last decade. The result is that astronomers have discovered X-ray novae that are black-hole candidates at the rate of about one per year in the Galaxy for the last 10 years. Because these systems sit quietly undetected for perhaps 50 years for every year they are in outburst, then every one outburst may represent 50 sleeping systems. Our vigilance in watching the Galaxy is not perfect, and gas and dust could obscure some events. Allowing for such problems, one can guess that there could be 100 to 1000 such black-hole systems in the Galaxy. Thus they vastly outnumber systems like Cygnus X-1.

One of the principal goals in the study of these erupting systems is to find proof that they contain black holes, not neutron stars or some other configuration that can mimic the circumstantial evidence for a black hole. Currently the most reliable way to establish a black-hole candidate is to show that the compact object in a binary system has too much mass to be a neutron star.

Five or six black-hole novae are excellent black-hole candidates. These systems have at least a firm lower limit to the mass of the object emitting the X-rays that rules out a neutron star. Among these are AO620–00, V404 Cygni and Nova Muscae 1991. V404 Cygni is currently the best candidate for a black hole in a binary system. Many careful observations reveal that the mass of the compact star is about 12 solar masses, far more than is possible for a neutron star. Approximately another two dozen systems are good black-hole candidates based on the similarity of their optical and X-ray outburst behavior to the temporal and spectral behavior of the best-established candidates.

In most of the black-hole X-ray novae, the companion has a small mass. The companion stars are dim and hence difficult or impossible to detect, even when the system is at minimum light. In the systems where information is available about the mass of the compact object, there is also information about the mass of the

companion. In AO620–00 and Nova Muscae 1991, the normal star companion is substantially less than 1 solar mass. V404 Cygni is somewhat a special case. The companion has evolved past the main-sequence stage, but even then the remainder of the star has a mass of only about 4 solar masses. For most of the systems, the companions are low-mass, low-luminosity stars, with a mass considerably less than the mass of the putative black hole. There is no question of a third star masquerading in any of these systems, adding mass that would be mistakenly attributed to the compact object.

10.5 THE NATURE OF THE OUTBURST

To obtain a basic understanding of the behavior of these systems one of the most important questions to address is the reason for the outburst. The most promising model for the basic outburst is an instability not directly associated with either the black hole or the companion star, but within the accretion disk that passes matter between them. The companion star provides the reservoir of mass. If the mass flows too slowly from the companion, the accretion disk cannot remain in a hot, ionized state, and a steady rate of flow is not possible. These systems must undergo accretion-disk outbursts similar to those in dwarf novae and some neutron-star binary systems, as discussed in Chapters 5 and 8. In the simplest picture, the disk flares to make excess optical and X-ray radiation and then goes back into storage mode, accepting matter from the companion, but passing very little through itself and down the black hole. The disk emits little optical light and virtually no X-rays. The main thing observable in this state would be the companion star and perhaps the spot on the edge of the disk where matter rains in from the companion. The disk could develop a very hot, nearly spherical inner region, as discussed in Chapter 4 (Figure 4.6), which would alter this simple picture and give another source of luminosity in the ''off'' state. We will return to this topic in the next section.

This physical process of the disk instability does not depend on the exact nature of the compact object or of the star providing the mass. It can happen to accretion disks surrounding white dwarfs and neutron stars as well as black holes. The majority of the X-ray novae that display this outburst behavior show no explicit evidence for neutron stars and remain black-hole candidates.

The disk-outburst model can account for the decade-long periods of quiescence, which are set by the time for matter to collect or

ooze inward in the cold, low-viscosity disk. The rapid rise time of days can be associated with the timescale for heating waves to propagate in the inner disk. The year-long decline is governed by the more rapid viscous evolution in the hot state and the time for the cooling wave to propagate through the disk.

There are some explicit tests of this picture. The model predicts that in quiescence, the mass-transfer rate as determined from the luminosity of the ''hot spot'' where the accretion stream collides with the disk should be far greater than the flow into the black hole, as determined from the X-ray luminosity produced in the inner disk. These basic predictions are borne out by optical and ultraviolet observations of AO620–00 from the *Hubble Space Telescope* and X-ray observations with the *ROSAT* satellite. Other confirming evidence comes from the lack of helium emission lines. If the inner regions generated X-rays, the X-rays would excite the gas to produce fluorescent emission lines. The lack of those spectral features means that there cannot be many X-rays and hence little mass flow in the inner disk. These observations seem to show that the disk is storing matter.

One objection to the model is that the disk does not seem to cool in the decline phase as much as predicted. This may be due to the formation of a hot ''corona'' around the disk, much like the corona that surrounds the Sun. In that case, the observed surface temperature does not reflect the temperature of the body of the disk that the models predict. Another possibility is that the X-ray flux from the inner disk is not low because the mass-flow rate is low, but because the efficiency of emitting X-rays is low. We will discuss this in the next section.

10.6 LESSONS FROM THE X-RAYS

Near the maximum of the outburst, lower-energy X-rays from the black-hole novae show a component that seems to come from a hot, opaque, geometrically thin disk, as predicted by the disk instability models. The observations show no significant change in the inner radius of the disk as the systems cool after outburst. The only characteristic radius in the disk that could plausibly remain constant as the mass flux declines is the last stable circular orbit, within which matter must plummet straight into the black hole. Evidently, near the peak of the outburst, the accretion disk extends all the way down to the inner radius from which matter plunges directly down to the event horizon of the black hole and disappears. This conclusion

strongly affects considerations of the higher-energy X-rays that may contain direct clues of the existence and nature of the black hole, rather than the accretion disk.

The black-hole novae also show high-energy X-rays, ranging all the way up to gamma rays. A process known as *Compton scattering* can produce these high-energy X-rays when low-energy photons scatter from a hot plasma and pick up energy. Arthur Holly Compton won the Nobel Prize in 1927 for his discovery of this effect and was further honored by the naming of the *Compton Gamma Ray Observatory* (see Chapter 11, Section 11.2). Neutron star systems rarely display this kind of radiation, and then only in a truncated form. This high-energy radiation may be just the clue we need to clearly distinguish accreting black holes from accreting neutron stars without the need to invoke the mass limit of neutron stars. Some recent theories for this high-energy radiation have made the explicit argument that it can only exist as it is observed from systems with no hard surface. That argument, if confirmed, would rule out not only neutron stars but also some other bizarre suggestions that would nevertheless have a hard surface. The only small-radius, high-gravity objects we can now imagine that do not have hard surfaces are black holes.

This high-energy radiation is seen near the peak of the outburst of many of the black-hole X-ray novae. It probably comes from a hot corona surrounding the disk, although the exact nature of that corona remains elusive. The black-hole X-ray novae also commonly show radio outbursts that require an outflow of matter with very high energy electrons. This outflow could also be a source of high-energy radiation. The observed interplay between the high-energy radiation from a corona and the lower-energy X-ray radiation that is presumed to come from the accretion disk is complex and varies in time, but as the outburst decays, the high-energy radiation comes to dominate. This suggests a change in the structure of the accretion flow.

One possibility under active investigation is that, as the mass-flow rate declines due to the inward propagation of the cooling wave in the disk, the inner disk thins out and reaches a state where it cannot cool efficiently. Rather than dropping into the cold state of an accretion disk, this inner region can become very hot, and nearly spherical. Matter from this dilute, nearly spherical region then falls almost radially straight down the black hole. The basic notion of this sort of flow was outlined in Chapter 4 and is illustrated in Figure 10.1. This material does not radiate much, despite its high temperature, both because dilute gas does not radiate efficiently and because this

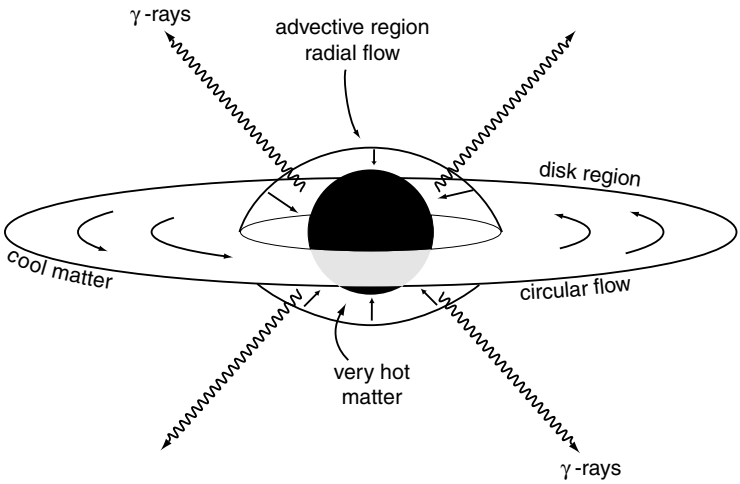


Figure 10.1 To account for the high-energy radiation observed from black-hole X-ray novae as they enter the low-luminosity state, a nearly spherical central advective region may form where the flow is nearly radial and the matter is very hot, but radiates inefficiently. Matter from the companion star spirals down through the accretion disk and then, perhaps by a process of evaporation, joins the hot advective flow before plunging down the black hole.

matter tends to plunge directly down the black hole, carrying its heat energy with it. In these circumstances, there is little time to radiate. This process is called *advective accretion flow*, to distinguish it from disk accretion flow. In a disk, most of the heat energy is radiated out through the face of the disk. In an advective flow, the heat is carried, or advected, down through the event horizon, so little heat is lost to radiation.

What little heat does radiate from an advective flow should, according to theoretical models, emerge as very high energy radiation, as observed. Because the radiation efficiency is low, a much higher mass flow rate must be sustained in order to produce even the feeble radiation that is seen. When applied to the black-hole X-ray novae, this theory suggests that a substantial amount of the mass transferred from the companion star does pass through the disk and down the black hole, even when the system is in its long-lived, low-luminosity state. Models based on this picture are rather successful in accounting for the feeble X-rays from the low-luminosity systems, even though the simple disk models say the disk should be cool and in a storage phase. This theory is on the cutting edge of research as this

book is being written and so there are a number of questions that have not been completely resolved. Among these are: how a cold disk can pass all the mass it must in order to feed the advective flow; how the advective region forms, perhaps by evaporation of disk matter; whether a substantial amount of matter transferred from the companion is blown away in a wind or other outflow before it can reach the black hole. All these issues are a sign of a vibrant and exciting research area.

One general notion has emerged. If the accreting object had a hard surface, photons from that surface would probably interfere with the matter in the advective region and prevent it from having the properties observed for the black-hole sources. This is one version of the argument that the black-hole X-ray novae cannot be neutron stars but must be objects with no surface. If this argument is right, they must be black holes, independent of the mass we measure for them.

10.7 SS 433

Another interesting class of objects in the astronomical zoo consisted for a very long time of a single entry. In 1980, Walter Cronkite brought this discovery to the attention of the world when he announced on CBS News that astronomers had found an object that was coming and going simultaneously! For those of you confused by that, read on.

The object was originally identified as being notable for its *emission lines*, excess power coming out at certain wavelengths of light. Normal stars show absorption by cool atoms, and emission is a sign of an energetic environment in some fashion. The object at issue is source number 433 in the catalog of objects with strong emission lines compiled by two astronomers, Stephenson and Sanduleak, so it is known as SS 433 (this is the same Sanduleak who cataloged the star destined to erupt as SN 1987A). Closer study showed that the emission lines in this object displayed a most peculiar behavior. There are two sets of emission lines, and they move around in frequency in opposite directions because of the Doppler effect. Each set of lines shows first a red shift and then a blue shift. The period of oscillation is 64 days. When one set of lines shows a red shift, the other set shows a blue shift, and vice versa. Thus when the gas causing one set of emission lines is moving toward us, the gas causing the other set is moving away from us, hence Cronkite's comment on the news. The actual interpretation that astronomers have given to this information is that

SS 433 is emitting jets of material in opposite directions, but somehow twisting around to throw the beams first in one direction, then in the other. Radio observations show an arcing series of blobs extending out beyond the object. Imagine that you are pointing a water hose overhead, but moving the nozzle in a circle. If you were to take a photograph at one instant, you would see blobs of water strung out along a widening helical path. That is what the radio astronomers see, confirming the picture of the oppositely directed rotating jets.

The real excitement came with the deduction of the velocity of the jet material. The jets are not directed at the Earth, but sideways, so normally one would not expect a Doppler shift. According to Einstein's special theory of relativity, however, even an object moving sideways shows a tiny Doppler effect. With ordinary velocities, the effect is undetectable. In order for there to be a measurable ''transverse'' Doppler effect in SS 433, the material in the twin beams must be moving at 80 percent the speed of light! SS 433 is ejecting opposing beams of material at nearly the speed of light. Active galaxies and quasars had shown similar jets, but this was the first time a star displayed such phenomena.

A further remarkable feature is that the material in the beams is not hot. SS 433 shows emission lines of neutral helium, but none from ionized helium so the matter cannot be tremendously hot. How the matter accelerates to the speed of light without getting heated in the process is a question that still plagues the theorists. One possibility is that radiation pressure can slowly accelerate the material and never push on it so hard that it gets hot.

SS 433 is surrounded by a radio source identified by the *synchrotron radiation* that arises when electrons spiral around in magnetic fields at nearly the speed of light. Some have identified this radio source as a supernova remnant left from the formation of SS 433. Others point out that if this is so, it is the largest supernova remnant in the Galaxy. A plausible alternative is that the remnant is a bubble blown in the interstellar gas by the relativistic particles ejected in the twin beams of SS 433 itself.

The actual nature of SS 433 still eludes satisfactory explanation. Clearly, the tremendous velocities require high energy and thus probably the high gravity of a compact star. One idea is that SS 433 contains a neutron star that is trying powerfully to emit radiation, perhaps because it is a young and energetic radio pulsar. If mass transfer has totally enshrouded it in a blanket of gas, a common envelope, however, the radio waves could not get out directly. The

energy then blasts out of two holes in the top and bottom of the envelope and makes the beams. This notion is given some support by other Doppler-shift measurements that indicate that besides the rotation of the beams, the whole object moves about with a period of 13.6 days. This probably represents a binary orbital period. The binary companion is presumably the source of the enshrouding envelope. Other theories attribute the energy to matter being swallowed by a black hole.

SS 433 remains an enigma in many regards, and the search for another object like it anywhere in the Universe went on for over a decade. Its close cousins, if not twins, were discovered only a few years ago.

10.8 MINIQUASARS

The black hole X-ray novae discussed in Sections 10.4 and 10.5 drew a lot of attention as evidence grew that they were black holes. The specifics were different in detail, but these objects had an inflow of matter, accretion disks, and, very probably, black holes. The same general description applies to the models for the energy sources of quasars and active galactic nuclei. The main difference is that the black holes in quasars are thought to be supermassive, up to a billion solar masses, and those in the black hole X-ray novae are 5–10 solar masses. The latter were clearly formed by the collapse of stars (although the details elude us). We do not know the origin of the supermassive variety.

One aspect of the supermassive black holes in galaxies is that they often emit beams of matter at nearly the speed of light. SS 433 was a hint in the direction that stellar-mass black holes could do the same thing, but ambiguity about its nature prevented a direct analogy from being drawn. That situation changed dramatically in the mid 1990s with the radio study of the outbursts of some of the black-hole X-ray novae.

Felix Mirabel is a radio astronomer of Argentine extraction who works in Paris. Luis Rodriguez is a Mexican radio astronomer. They began a project to monitor the radio emission of the black-hole X-ray novae. In 1994, they got data on an outburst in an otherwise obscure source that is hidden behind so much galactic dust that it cannot be seen with optical telescopes. The radio emission can penetrate the dust. Mirabel and Rodriguez discovered a remarkable behavior. They could identify discrete clouds of particles ejected from the X-ray

source that emitted radio radiation as they moved rapidly away from the central source. By watching these clouds from day to day, they could see how far apart they had moved in a given time interval. A simple calculation of their speed showed that they seemed to be moving at greater than the speed of light!

This apparently superluminal behavior had been seen before. It was first noticed 30 years ago when similar monitoring was done of quasars. This does not represent a breakdown of Einstein's theory, but a sort of relativistic optical illusion. The explanation for this phenomenon gave Sir Martin Rees, the eminent British astrophysicist, his first claim to scientific fame. The answer to this puzzling behavior is that the matter is ejected from the central source at nearly, but not quite, the speed of light. For the sources that appear superluminal, the jets of matter are pointed nearly toward us. In this case, the matter is chasing the radiation it emits and traveling at nearly the same speed. This foreshortens the apparent motion of a blob of emitting matter in such a way that it seems to be covering a large angle, and hence a large reach of space, in an impossibly short amount of time. The X-ray nova that Mirabel and Rodriguez observed was doing the same thing. The matter was being ejected in blobs that moved at nearly, but not more than, the speed of light, thus giving the appearance of superluminal motion.

At least one other black-hole X-ray nova has been discovered to display this superluminal motion. The second one has a measured mass for the compact object from the binary orbit that is more than 3 solar masses. This puts it firmly in the category of black hole candidate. The miniquasars have helped to put SS 433 in context. There are differences, but there are also obvious similarities. Even though there is still no firm proof that SS 433 is a black hole, we can deduce that if the jets of SS 433 were pointed more nearly directly at us, we would witness nearly, if not clearly, apparent superluminal motion.

The analogy between the black-hole X-ray novae and quasars as supermassive accreting black holes was already quite strong, but the discovery of the X-ray novae with apparent superluminal motion cemented the idea in many people's minds. The term ''miniquasars'' instantly became popular to describe the black-hole X-ray transients, especially those with the superluminal behavior. There is much to be learned about how black holes of either the stellar or supermassive variety launch the rapidly moving blobs of radio-emitting matter, but the discovery of the miniquasars is one more piece of evidence that black holes really exist on both the stellar and supermassive scales.

10.9 GIANTS AMONG US

The study of quasars has convinced astronomers that the only credible explanation for the immense luminosity, small size as indicated by the daily variability, and immense, sometimes superluminal, jets, is that they are powered by supermassive black holes. As described in Chapter 2, Section 2.2, accreting objects cannot have a luminosity brighter than the Eddington-limit luminosity, or they would blow the surrounding matter away with radiation pressure rather than accreting it, the very mechanism needed to produce the luminosity in the first place. The Eddington limit in turn depends on the mass of the accreting object; a larger mass with higher gravity can withstand a brighter, self-induced radiation, and still manage to draw matter inward. Accreting objects must then have a mass big enough that the Eddington limit to the possible luminosity is comfortably above the luminosity actually observed. This means the mass of the object must be big enough to withstand the observed luminosity. Estimates based on the Eddington luminosity argument as applied to the incredibly bright quasars yield estimates for the mass that range up to a billion solar masses for the very brightest.

Ironically, it has proven rather difficult to absolutely establish that quasars harbor these giant black holes. Velocities of gas believed to orbit near the black hole are consistent with the suspected large masses. In addition, recent observations with the *Chandra X-ray Observatory* and the *XMM-Newton X-ray Observatory* have revealed information from near the center that strongly suggests not just a black hole, but a Kerr black hole with rather specific rotational properties in some active galaxies. The assumption that quasars represent supermassive black holes is certainly consistent with all we know of quasars, and more specific data is promised. In the meantime, other evidence that such large black holes exist in the centers of galaxies has come from the study of more normal galaxies, such as our own Milky Way.

Investigations of giant black holes in ordinary galaxies were driven in part by the desire to understand what becomes of a quasar when it is no longer a quasar. In the standard picture, material from the surrounding galaxy must rain down on the central black hole so the luminosity can arise from the accreted mass, most likely from a large accretion disk. If that mass flow shuts off, the quasar activity will die out, but any black hole will still be there. Quasars are observed at large distances and from back in the past. The question is how many

current, quiet galaxies were once quasars and whether or not we can find evidence for their black holes.

Perhaps the most dramatic success in this field is the discovery and study of the supermassive black hole in the center of our own Milky Way Galaxy. The center of our Galaxy, in the direction of the constellation Sagittarius, is shrouded by the lanes of gas and dust in the disk of the Galaxy through which astronomers must peer to see the center. Ordinary optical astronomy is useless. Rather, astronomers have used longer-wavelength radiation, infrared and radio, to penetrate the murk. The target has long been a bright radio source known as Sagittarius A. The gas swirls around this region in a way suspiciously like gas falling into and swirling around a central source of gravity. A practical worry is that gas is subject to ephemeral forces of other sorts, other gas streams, the pressure of radiation, the guiding hand of magnetic lines of force. This gives caution about a literal interpretation of the swirling gas as caused only by a massive source of gravity, and yet that may prove the correct and simple interpretation. The most dramatic insights have come from studying the motions of stars near the Galactic center. Stars are like tough little nuggets. Their orbits are not swayed by streams of interstellar gas, magnetized or otherwise. They proceed like a bullet through a sandstorm, orbiting through the local gravitational field (the curved space!) caused by the collection of other stars and any giant single mass that might be present.

Unlike optical radiation, longer wavelength, infrared, and radio radiation can penetrate the murk between us and the center of the Galaxy. By observing the infrared radiation of stars, two teams of astronomers, one led by Reinhardt Genzel at the Max-Planck-Institut für Extraterrestriche Physik in Munich and one by Andrea Ghez at UCLA, have tracked the motions of individual stars near the center of the Galaxy, in a region smaller than the size of the orbit of Pluto, about 20 light days across. This technical tour de force has revealed not simply higher velocities of stars near the center, but with observations spanning a decade has shown the orbits of individual stars as they plunge, accelerating, toward the central source of gravity and then recede to outer, slower portions of the individual orbits. The result is unambiguous: there is a tiny, very dark, four-million solar mass concentration of gravity right at the dead center of our Galaxy. If this concentration of mass were a swarm of other dark objects, neutron stars or stellar mass black holes, they would quickly coalesce into a supermassive black hole anyway! The conclusion seems inescapable

that our Galaxy contains a four-million solar mass black hole. Astronomers are not resting on their laurels. What is needed next is an actual ''photograph'' of the dark spot, or other evidence of the strong Einsteinian curved space very near the event horizon. Such an observation may be possible in the near future with radio telescopes, and aggressive plans are afoot to do so.

In the meantime, other teams of astronomers have sought evidence for supermassive black holes in other galaxies scattered about the nearby Universe. My colleagues here at the University of Texas, John Kormendy and Karl Gebhardt, have been among the most ambitious and successful ''black-hole hunters.'' The search for supermassive black holes in normal galaxies proceeds not by looking for a large black dot, but by looking for evidence that stars orbiting near the center of the galaxy are caused to move more rapidly in the gravity of the black hole. This effort requires peeking with great sensitivity right in the heart of galaxies to see, on average, how fast the stars there move. One cannot see individual stars in these more distant galaxies, but the collective motion of the stars will broaden the spectral lines of light emitted by the stars. The average motion can be measured by the average Doppler shift. The *Hubble Space Telescope* with its great visual acuity played a key role in providing the needed data. The answer is that nearly all decent-size galaxies harbor black holes, and that many, if not most, galaxies could have been quasars in the past.

This work has provided an amazing new insight into the nature and import of these supermassive black holes, with Karl Gebhardt again playing a leading role. Decades ago (when my Texas colleague Greg Shields and I worked on this topic), it was thought that supermassive black holes were somewhat incidental to the host galaxy. The implicit assumption was that the black holes formed from matter that was left over from the formation of stars or shed by stars as they evolved, and that drained toward the center of the galaxy by uncertain processes. The size of the black hole could then be large or small, depending on the circumstances, but the assumption was that its presence was otherwise incidental to the galaxy as a whole. Instead, the new observations revealed that essentially every galaxy with a central bulge of stars, as possessed by our Milky Way and the nearby giant spiral galaxy Andromeda, contained a supermassive black hole. More dramatically, the mass of the black hole tracked in exact proportion to the mass of the bulge. Every bulge was about 800 times more massive than the central black hole. Galaxies that made more

massive bulges made more massive central black holes, or vice versa. To understand how remarkable this statement is, it is useful to note that the mass of the bulge is determined by measuring the average velocities of the stars that comprise it. This means that the velocities of the stars in the bulge are closely connected to the mass of the central black hole, even though the stars in the bulge are vastly too far away from the central black hole to feel its gravity now. Yet somehow these distant stars ''know'' about the presence of the black hole. How can this be?

The answer to this new profound question is not yet known. An idea that is gaining currency is that when the black hole first forms, the radiation from the accretion activity blows a strong wind that limits the mass that gathers in the black hole. Perhaps magnetic fields play a role in the feedback process. The general implications are clear. Somehow the mass of the central black hole is intimately connected to the basic processes of the formation and evolution of the galaxy as a whole. This revelation has spurred a great deal of theoretical activity and provided an even deeper rationale to search for black holes.

Another related area that is a current focus is the quest to find quasars at the greatest distances and hence in their most extreme youth. The youngest quasars found arise when the Universe was very young, only about 700 million years old. These quasars are seen shortly after the gas in the Universe was re-ionized after its cold hiatus in the Dark Ages that followed the big bang (Chapter 11, Section 11.1.6). Before that, the opacity of the gas was so high that it would be difficult to see things even as bright as quasars. Quasars probably do exist within the early murk. The problem is that astronomers are not at all sure how supermassive black holes could have grown so quickly. Mass can be thrown down their maws only as fast as the generated radiation pressure allows. If mass begins to flow in too quickly, so that the Eddington-limit luminosity (Chapter 2, Section 2.2) is exceeded, then the matter is instead blown away. This feedback limits how fast a black hole could grow by accretion alone. It may be that the first seed black holes formed from the collapse of massive stars were already pretty large, hundreds of solar masses, giving them a leg up. My colleague Volker Bromm argues that the first stars to form after the Dark Ages were massive, so this might fit together. Such black holes might settle into one another's gravity wells, spiral together by gravitational radiation and merge. Such a growth process would sidestep the Eddington limit and might be a very effective way to create supermassive black holes very quickly.

10.10 THE MIDDLE GROUND

Yet another hunting ground for black holes has arisen in an unexpected quarter. As noted in the previous section, the luminosity of an accreting object can help to guide an estimate of the mass. If the luminosity is greater than the Eddington limit, mass would be blown away by the radiation pressure from the star rather than accreting on it to provide the very luminosity observed.

With this understanding as background, X-ray astronomers have found sources of X-rays in nearby galaxies that are very bright, brighter than the gravity of a mere neutron star could hold together. These have been named *Ultra Luminous X-ray Sources* or *ULX*. At face value, the observed luminosity requires not only more mass than a neutron star can support, but more mass than binary black-hole candidate systems that are produced, as we suspect, from ''normal'' massive stars. In order to have the Eddington-limit luminosity meet or exceed the observed X-ray luminosity, the accreting object apparently must have more than 100 solar masses. To explain this new category of X-ray sources, astronomers began talking about ''intermediate mass black holes,'' black holes with considerably more mass than that suspected in Cygnus X-1 or those in black hole X-ray novae like A0620–00 or V404 Cygni, but far smaller than the million-to billion-solar-mass monsters that reside in the centers of galaxies. The ULX remain a topic of hot debate. Just as for quasars in the early days, it is difficult to prove that the source is a black hole. One has to rule out the possibility that the source is a cluster of smaller-mass objects that somehow mimic a single large mass. People are scrutinizing the spectrum of the X-rays to see if there are differences from ''normal'' binary X-ray sources that could be a clue to the nature of the gravitating object.

Suspicion that intermediate-mass black holes could exist, and account for the ULX, has been fed from another quarter, the search for black holes in the center of star clusters. The target has been the beautiful globular clusters, nearly spherical clusters of hundreds of thousands of small-mass stars that occupy the halo of the Milky Way and other galaxies. Globular clusters are thought to date from the epoch of formation of the galaxies themselves. Once again, the means to search for black holes in the centers of these clusters is similar to that for the search for supermassive black holes in the centers of galaxies; look for the motions of stars that point to a large dark mass in the center. Karl Gebhardt has again been a key player in this quest.

Such studies have revealed that at least a couple of globular clusters might have concentrations of dark gravitating mass in their centers. The cluster M15 in the Milky Way may have a central dark mass of 4000 solar masses. The cluster called G1 in our sister spiral, the Andromeda galaxy, may have a central dark mass of 20 000 solar masses. If either or both of these lines of evidence in globular clusters pans out, then yet another venue for black holes may have been discovered.

While the direct evidence for black holes in terms of a ''dark spot'' yet eludes us, there is a particular clue suggesting that these central knots of gravity in globular clusters may be black holes. The mass of the black-hole candidates seems to be the same ratio to the globular-cluster mass as does the galactic-bulge mass to supermassive black-hole mass; the candidate black holes have a mass about one thousandth that of the globular-cluster mass. Both galactic bulges and globular clusters are old. Both galactic bulges and globular clusters are roundish. Both galactic bulges and globular clusters appear to contain black holes that are a regulated fraction of the total mass. The physics that controls the formation of bulges and supermassive black holes may, then, apply all the way down in scale to the mass of globular clusters and their black holes. If this remarkable concordance proves true, then there is a hint that there is some powerful controlling physics at work.

Are the ULX black-hole candidates related to the globular cluster candidates? Globular cluster sources are not necessarily bright in X-rays nor are any ULX in globular clusters. The globular clusters require larger black holes than would the ULX, but there might be some continuum from stellar-mass black holes, to ULX black holes, to globular-cluster black holes and then on up to the largest found in the brightest quasars. The black holes in globular clusters might not be presently accreting a lot of matter and there might be intermediate-mass black holes in environments other than globular clusters. Astronomers have noted that starting with such large black holes might help to grow the supermassive variety more quickly through accretion or by merging them together to jump start the process in a way that would make the Eddington-limit luminosity irrelevant to the rapid growth. Certainly there is much more to learn about whether or not intermediate-mass black holes exist and, if so, their role in Nature.

11

Gamma-ray bursts, black holes and the Universe: long, long ago and far, far away

11.1 GAMMA-RAY BURSTS: YET ANOTHER COSMIC MYSTERY

There was a revolution in astronomy in the first few months of 1997. A major breakthrough occurred in one of the outstanding mysteries of modern astrophysics, the cosmic *gamma-ray bursts*. This story began in the 1960s. The United States launched a series of satellites that orbited the Earth at great distance, halfway to the Moon. They were called the *Vela* series, and they were designed to detect gamma rays and other high-energy photons and particles. If it strikes you that there must be something special about them to be so far from Earth, you are on the right track. They were not designed for astronomy, but primarily to detect terrestrial nuclear-bomb tests. They were also intended to study the background, other natural sources of high-energy photons and particles in the solar wind and the Earth's magnetosphere, to aid in the separation of bomb signals from natural signals.

Stirling Colgate was on the team in Geneva in 1959 working on the treaty to ban space, atmospheric, and underwater nuclear tests. He had done some calculations that suggested that when a supernova shock wave broke through the surface of the star there could be a pulse of gamma rays (see Section 11.4 in this chapter for an update of this topic). He was afraid that such an event would be misunderstood as a nuclear bomb and might trigger a serious miscalculation by one side or the other. He hassled both sides, the United States and the Soviets, concerning the need to understand potential astronomical sources of confusion, especially supernovae, lest they lead to disaster. In terms of giving credit, Colgate revealed the true father of modern supernova and gamma-ray burst research: ''Scratchy'' Tsarapkin. Anatoly Tsarapkin was the head of the Soviet delegation to the Geneva talks aimed at the Limited Test Ban Treaty. When Colgate said

supernovae might be confused with a test, Scratchy, not a scientist himself, fixed him with a steely glare and inquired, ''Who knows what a supernovae would look like?'' Colgate realized what thin ground he, and the U.S. delegation, were on. He returned to Livermore and made the case to Edward Teller that understanding supernovae must become a primary goal of the Lawrence Livermore Laboratory. The rest is history, much of it recounted in Chapters 6 and 7.

The *Vela* satellites were motivated, at least in part, by these concerns. Colgate found the Russians intractable. They would not do their own astrophysical background checks and feared satellites launched by the United States would be used for spying. The agreement to put the *Vela* satellites in high orbit was a response to the Russian demand for a guarantee that they not be used for spying. Both sides did launch spy satellites, of course, but this did not apply to the *Vela* series, the results of which were unclassified.

Perhaps the *Vela* series saw bombs, but they certainly detected outbursts of an extraterrestrial nature. One of the *Vela* series was instrumented to see X-rays and discovered the first X-ray burst (Chapter 8, Section 8.7). With the first extraterrestrial detections of gamma rays in 1967 (the *Vela 4* series), the scientists at Los Alamos could not convincingly rule out the Sun as the source. They had to wait until the launch of the next series (*Vela 5*), in 1969, before they were able to conclude rigorously that the gamma-ray signals were from neither the Earth nor the Sun but from elsewhere in outer space. The discovery was finally announced by Ray Klebesadel, Ian Strong, and Roy Olson in a paper in the *Astrophysical Journal* in 1973. This paper created a new scientific industry.

The bursts of gamma rays from beyond the Earth were seen at irregular intervals. These bursts lasted for 10–30 seconds and showed variations on times as short as a 0.001 second. Subsequent investigations showed that the gamma-ray bursts were primarily a gamma-ray phenomenon, with relatively little energy in the X-ray band, unlike other sources of gamma rays that emit abundantly at lower energies as well. That the dominant emission mode is gamma rays means that a high energy is involved. Gamma-ray bursts probably require high gravity and motion at nearly the speed of light.

The quest for an explanation of gamma-ray bursts was long handicapped by a lack of direct knowledge of the distance to the bursts. A debate raged as to whether they are in the Galaxy or at the farthest reaches of the Universe. This debate was brought into sharp focus by the immensely successful Burst and Transient Source

Experiment (BATSE) on the *Compton Gamma Ray Observatory*. The *Compton Gamma Ray Observatory*, named for Arthur Holly Compton (Chapter 10), was launched in 1991 as one of the series of Great Observatories planned by NASA. The *Hubble Observatory* was the first. Two others, the *Advanced X-ray Astronomy Facility (AXAF)* and the *Space Infrared Telescope Facility (SIRTF*), were downsized, descoped, and delayed for over a decade, but *AXAF* was finally launched as the successful *Chandra Observatory* in July 1999, and *SIRTF* was launched in August 2003 as the *Spitzer Space Telescope*. In the meantime, the *Compton Gamma Ray Observatory*, with BATSE aboard, was de-orbited in June, 2000. The dream of having all four Great Observatories in orbit at once was not realized, but the record is still fantastic, with, at this writing, *Hubble* in maturity, *Chandra* in ripe middle age, and *Spitzer* the active new kid on the block.

BATSE recorded 2704 new gamma-ray bursts in its active life, corresponding to about one per day. The surprising result was that the sources are, to great accuracy, distributed uniformly on the sky. There is no statistical evidence for any tendency to lie toward the plane of the disk of our Galaxy or toward the Galactic center. This contradicted any picture in which the sources were distributed throughout the Galaxy and viewed from the offset position of the Earth, 25 000 light years from the Galactic center. This result fueled increasing conviction that the sources of the gamma-ray bursts were in galaxies at cosmological distances because the distant galaxies are naturally distributed uniformly on the sky, on average. In addition, fainter sources are more abundant. The precise number of faint sources shows a pattern that is close to what one would expect if the bursts constituted a gamma-ray ''standard candle'' (see Chapter 12, Section 12.7) viewed in ever-larger volumes of space in an expanding Universe. There might, however, be other explanations for this pattern, and there is no particular reason to think that gamma-ray bursts are a standard gamma-ray candle.

The problem is that if the gamma-ray bursts are at cosmological distances, the intrinsic source of energy must be huge, comparable to or exceeding that of a supernova, but radiated essentially entirely in gamma rays. Everything about the cosmic gamma-ray bursts strains credibility, yet there they are.

One of the clearly defined problems in the study of gamma-ray bursts was the complete lack of counterpart events at other wavelengths, especially optical wavelengths. Without optical counterparts, the full weight of astronomical lore, much of it derived from optical

astronomy, could not be brought to bear on the issue. The problem was that the gamma-ray detectors could not provide sufficiently good locations. It is a difficult technical feat to bring gamma rays to focus. The gamma-ray sky has typically been ''fuzzy,'' a situation somewhat analogous to nearsighted people looking around with their glasses off. A given gamma-ray burst could be said to be ''over there,'' but ''there'' could not be precisely defined. The uncertainties in position were typically several to tens of degrees in radius (the full Moon subtends about 0.5 degree in angular diameter). In an area of the sky of that size, there can be thousands of stars. Finding the point of light that corresponds to a given 10-second-long gamma-ray burst was like seeking the proverbial needle in a haystack, a needle that was likely to vanish if you did not find it in less than a minute.

The nature of these events puzzled astrophysicists for nearly 30 years. Without the fetters of any relation to classical astronomy, theorists had a field day trying to explain the observations. The requirements for a theory in these circumstances are that it account for the observations and be self-consistent. Plausibility was not necessarily a constraint because gamma-ray bursts represented a new and unprecedented phenomenon. At a meeting shortly after their discovery, Mal Ruderman of Columbia University, who was giving the review talk on gamma-ray bursts, announced that it was easier to give a list of the people who had not presented a theory of gamma-ray bursts than it was to give a list of those who had. He showed a slide consisting of one name, Princeton's Jerry Ostriker who, for whatever reason, had not jumped on the gamma-ray-burst bandwagon.

Theories ranged from black hole collapse to ''relativistic bb's.'' The latter were supposed to be little grains of dust accelerated to near the speed of light and then arriving at the Solar System to crash energetically into the solar wind. Remember all the billion pulsars that have died in the Galaxy? One of the first theories, and one that generated more than a few chuckles, postulated that gamma-ray bursts were generated by comets falling onto those neutron stars. One of the little-known but supportive ideas of this hypothesis is that clouds of comets may very well spread nearly from one star to another. Space may be filled with comets, and the chance that one of them would occasionally fall onto one of those billions of neutron stars is not so low.

The argument that swayed some people into taking this comet idea more seriously is the problem of generating gamma rays at all with a neutron star. The problem is related to the Eddington limit

(Chapter 2). If energy is released on the surface of a neutron star, the material expands and cools in response to the radiation pressure. Under normal circumstances, such matter can get hot enough to emit X-rays, as we have seen in Chapter 8, but not hot enough to emit the more energetic gamma rays. The importance of the impact picture is that the material arrives in a lump and is compressed much more than would be either a dribble of gas or material just sitting on the surface. The effect might be enhanced if the infalling matter were a rock, so asteroids as well as comets have been considered. After a hiatus of a number of years, a similar idea was still around in 1998, although it sank under the weight of recent results.

There is a benefit to allowing the imagination of the theorists to run beyond the bounds of the known data. What was really needed were more data so that theory and observation could march hand in hand in some fruitful direction.

11.2 THE REVOLUTION

All this changed with the launch of a Dutch–Italian X-ray satellite, *BeppoSAX* on April 30, 1996. This wonderful name derives from the nickname of a pioneering Italian physicist and X-ray astronomer, Giuseppe Occhialini, known as Beppo to friends and colleagues, with the appendage for X-ray satellite in Italian, ''satellite per astronomia a raggi X.'' *BeppoSAX* was capable of looking everywhere on the sky for the weaker X-ray signal that characterizes gamma-ray bursts and to give a first coarse location, more accurate than BATSE provided. The key innovation for *BeppoSAX* was a second instrument that could be brought to focus by quickly slewing the satellite in an attempt to rapidly find the X-ray flare from the gamma-ray burst and to provide a much more accurate location, with an uncertainty of a few minutes of arc, an area on the sky several times smaller than BATSE provided. At that point, ground-based optical telescopes could be brought to bear to search the much smaller location to see if there were any optical component. All this was a bit of a gamble. If the whole gamma-ray-burst phenomenon in lower-energy X-rays and in the optical faded in the tens of seconds that characterized the gamma-ray bursts themselves, then there would be no time to slew the satellite, a process that would take at least hours, never mind time to obtain optical images, a process that might take a day (or night) even in the best of circumstances.

Another chapter of this story is worth telling if only to recognize the great effort and ingenuity that goes into the scientific enterprise that sometimes fails to pay off. At a meeting on gamma-ray bursts in Santa Cruz in 1981, the attendees recognized that studies of gamma-ray bursts were stymied by the lack of observations at other wavelengths. A project was born to design a satellite that would contain a gamma-ray detector, but also ultraviolet and optical detectors to look in the same direction and hence to get simultaneous information on the burst at other wavelengths. The project was named *HETE* for *High-Energy Transient Explorer* and the arduous process of design began. It won NASA competitions to build and launch and suffered the inevitable delays. *HETE* was finally scheduled to launch on November 4, 1996, a date that would have put it in competition with *BeppoSAX*. The Pegasus rocket carried *HETE* and an Argentine satellite to orbit, but a battery failed in the third stage. The shroud that held them could not be opened, and without its solar panels, *HETE* died in the darkened enclosure. That opened the way for *BeppoSAX*. To their credit, the *HETE* team regrouped, took the plans and spare parts, and built a new satellite. *HETE 2* was launched on October 9, 2000, and has been a valuable tool for the study of gamma-ray bursts, as will be outlined below. With the satellites still in its grip, the third stage of the Pegasus that carried *HETE 1* aloft burned up in the atmosphere on April 6, 2002, over the Indian Ocean,

BeppoSAX scored its coup on February 28, 1997, when it localized a burst sufficiently well that an optical follow-up was feasible. The result was the discovery of the first optical counterpart by a team led by Dutch astronomer Jan Van Paradijs. Van Paradijs saw the great flowering of gamma-ray burst research that followed from this identification, but was tragically struck down by cancer only two years later.

The fashion has been to label gamma-ray bursts by the year and day that they were discovered. Occasionally, two or more events have been discovered on the same day to mess up this scheme; then they get appendages of a, b, c, etc. With this convention, the breakthrough gamma-ray burst was thus named GRB 970228.

Two months later, in early May, *BeppoSAX* found another event, GRB 970508, enabling another optical identification. In this case, absorption lines of matter in front of this source proved that the source was at a cosmological distance, of order 1 billion light years or greater. In December of 1997, yet another optical counterpart was discovered associated with GRB 971214. After the gamma-ray burst

faded, a faint galaxy was revealed. The red shift of this galaxy was immense, with the wavelength of the detected light shifted by more than a factor of three from its natural wavelength. This galaxy was estimated to be 12 billion light years away. If GRB 971214 had radiated equally into all directions and hence followed the basic inverse-square law for apparent brightness (Chapter 12, Section 12.7; Chapter 14, Section 14.5), then estimating the distance from the red shift (and adopting specific values of the cosmological parameters) implied that the energy of this source was fantastically large. More energy would be required than the entire collapse and neutrino energy of a supernova, and more than even most exotic theories of colliding neutron stars and black holes could support.

Even GRB 971214 is not the record. That belongs to the first burst localized by *BeppoSAX* in 1999, GRB 990123. This burst brought in yet another interesting chapter in the saga. Many people realized that if an optical counterpart were ever to be seen, then an especially rapid response was needed. A special email notice system run by Scott Barthelmy and his colleagues at the NASA Goddard Space Flight Center in Maryland was set up. Even more extreme, some people began to wear beepers that were triggered electronically by a signal from a satellite, BATSE or *BeppoSAX*, so that they got buzzed the instant (allowing for the finite travel time of light and relay switches) a gamma-ray burst was detected. One of the things that this rapid response allowed was communication with automatically controlled robotic telescopes that would very quickly swivel to look for an optical counterpart, perhaps in the time frame of the original gamma-ray burst. This was the mission of *ROTSE*, the *Robotic Optical Transient Search Experiment*.

The first generation, *ROTSE I*, was a small telescope situated at the Los Alamos National Laboratory. It was designed and operated by Carl Akerlof and his associates at the University of Michigan, Los Alamos, and the Lawrence Livermore National Laboratory. *ROTSE I* was constructed to receive signals directly from the satellites that detect gamma-ray bursts and then to rapidly swivel and look at the location of a gamma-ray burst. *ROTSE I* was not very sensitive as telescopes go because it had only four wide-angle camera telephoto lenses, but it could see a fairly large portion of the sky at one time to look for variable sources. Another advantage is that it was quick! Quickness does not count if the weather does not cooperate or if the discovered gamma-ray burst is only visible from the southern hemisphere or if it is ''up'' in the north during daylight hours. This was the tale for the

first number of *BeppoSAX* bursts. *ROTSE I* did have a clean shot at some bursts, but it did not see anything.

Finally, on January 23, 1999, everything came together, and *ROTSE I* scored its first detection of a gamma-ray burst. *ROTSE I* detected the immediate optical counterpart of GRB 990123, the emission of light that occurs simultaneously with the burst itself. The results were dramatic. *ROTSE I* saw a flash of light that rose in about 10 seconds to ninth magnitude and then faded over the next minute or so. This peak apparent brightness was only about a factor of 10 dimmer than can be seen with the naked eye! Associated work on this gamma-ray burst revealed it to be at yet another immense distance. This makes GRB 990123 the intrinsically brightest optical event ever recorded in scientific history. Ho hum, another record for gamma-ray bursts. Actually, there is nothing to be blasé about here. If radiated uniformly in all directions, the implied peak optical luminosity of GRB 990123 was equivalent to ten million supernovae or ten thousand very bright quasars. This optical burst did not last long, but its intensity was very impressive.

Most of the energy emitted by GRB 990123 was in the gamma-ray range. Here again, GRB 990123 set a record. The detected gamma-ray intensity was among the strongest ever seen at the Earth. At the distance observed, the total energy in gamma rays was ten times higher than the previous record-setters like GRB 971214. If this gamma-ray energy poured out equally in all directions, the energy involved was equivalent to the complete annihilation of two solar masses of matter! One runs out of exclamation points.

These optical counterparts of the cosmic gamma-ray bursts thus revolutionized the field and proved the power of focusing optical astronomy on this decades-old problem. They opened a new era in the study of gamma-ray bursts that provided not only rapid progress in understanding the bursts themselves, but also promise of their use to explore the nature of the Universe at great distances.

The emission witnessed in the X-rays by *BeppoSAX*, in the optical by ground-based telescopes, and in the radio by radio telescopes, was discovered to last much longer than the original gamma-ray burst. Rather than tens of seconds, the X-rays last for days, and the optical and radio can stay above limits of detectability for weeks or months. This delayed emission of energy has been termed the *afterglow* of the gamma-ray burst. The general interpretation is that the process that energizes the event, whatever that process is, sends a powerful explosion out into the interstellar gas surrounding the event. The

explosion generates a strong shock wave that moves at very nearly the speed of light. The interaction of this shock wave with the interstellar gas can produce gamma rays, X-rays, optical emission, and radio emission in appropriate circumstances. The general process leading to this afterglow is called a *relativistic blast wave*. Models based on this process have been successful in accounting for many of the observations of the afterglow, including the spectrum of the radiation and the rate of decay that tends to drop off as one over the time since the original gamma-ray burst. If you wait twice as long, the glow is half as bright.

As remarked above, *HETE 2*, was launched in October of 2000. *BeppoSAX* continued to operate until April of 2002. *HETE 2* was not as effective as the most optimistic predictions for a variety of technical reasons, but it has provided data on key bursts that have driven progress in the field. The new kid on the block is the *Swift* satellite, launched on November 20, 2004. This satellite is just coming into full operation as this is being written. *Swift* is engineered to both discover gamma-ray bursts and to follow the optical afterglow with its own onboard telescope. There is also a global effort to respond to bursts with ground-based instruments, from small robotic telescopes to the giant telescopes that dot the planet: the Hobby-Eberly Telescope in Texas, the Keck telescopes in Hawaii, the Gemini telescopes in Chile and Hawaii, and the four Very Large Telescopes at the European Southern Observatory in Chile. There has been dramatic progress, but there is so much more to do.

The *ROTSE* story

I am involved in one of the robotic telescope projects, and there is a story there. The *ROTSE* team at the University of Michigan led by Carl Akerlof designed a second-generation robotic telescope with a larger aperture, but smaller field of view than *ROTSE I*. The idea was that one could afford a somewhat smaller field of view with more accurate first-cut satellite positions at the expense of being able to peer to fainter limits. *ROTSE II* was a bust for technical reasons. I am not sure what the problems were; Carl does not like to talk about it. In any case, the team pushed on to a third generation of telescopes, *ROTSE III*. These are small telescopes, with mirrors only about eighteen inches in diameter, but they can observe nearly two square degrees at a time and they are snake-fast. From receipt of an electronic command, a *ROTSE III* telescope

can be making fully robotic observations a mere six seconds later! This is the fastest response time of any similar instruments. One key goal is to search for the optical flash that is simultaneous with the gamma-ray burst itself, as was done by *ROTSE I* for GRB 990123. Even the *Swift* satellite itself will not routinely do that. *Swift* requires about a minute to train its optical telescope on any burst it discovers. The telescopes are housed in small enclosures reminiscent of, but somewhat larger than, a Porta Potty. They have tops that flip open automatically and are fully instrumented with weather stations to monitor conditions.

The chain of events involving me started at a meeting of the American Astronomical Society in June of 2001. Carl Akerlof gave a talk in which he outlined the success of *ROTSE I* and his proposed plan for four *ROTSE III* telescopes spaced around the world to provide maximal coverage. He mentioned that they were still exploring sites for the telescopes. As pure blind luck would have it, Carl sat down next to me after his talk. We had never met. I introduced myself and, with my typical, fools-rush-in naiveté, asked whether he might want to put one of the telescopes at McDonald Observatory. Carl was polite, but basically said ''I don't think so,'' and excused himself to rush off to the airport. I put the incident out of my mind.

I got a phone call from Carl about two months later, asking whether I might further consider the proposition of putting one of the *ROTSE III* instruments in Texas. Still having little idea what I was getting into, or exactly whose resources I was committing, I said, ''Sure.'' *ROTSE I* had been based at Los Alamos Laboratory, where gamma-ray bursts were discovered and where there was a long-standing interest and complementary projects for fast response telescopes. The presumption had been that one of the *ROTSE III* instruments would also go there; Texas was too close to provide the global geographical distribution that was desired. As it transpired, the lab administration gave indications of rather tepid support for *ROTSE*, among other things proposing to move the instruments from the lab grounds proper to a site 30 miles away in the mountains, where routine access and maintenance would be cumbersome. In addition, it was always difficult, and becoming more so, to get foreign associates onto the lab grounds, including that remote site. A Russian postdoctoral fellow was having such access problems and *ROTSE III* was designed to be an integrated foreign collaboration. That tipped the balance of a difficult

decision away from Los Alamos and to Texas. Off we went.

The first *ROTSE III* instrument, christened *ROTSE IIIa*, was installed in Australia. Texas got the second, *ROTSE IIIb*. Figure 11.1 shows *ROTSE IIIb* in the foreground of the Hobby–Eberly Telescope. The third, *ROTSE IIIc*, was installed in Namibia, where German scientists already had a radio-telescope site and another type of telescope to monitor the air showers formed by gamma rays. The fourth, *ROTSE IIId*, has been set up in Turkey. *ROTSE IIIa* and *b* have already done some interesting work with *HETE 2* bursts and are poised to be useful tools in the *Swift* era.

History played out in the background of these developments. The 9/11 attack came shortly after we decided to move the telescope to Texas. One of the minor, but significant, results was an even higher attention to security at Los Alamos. In addition, *ROTSE III* was installed at McDonald Observatory in February of 2003. A bunch of us were sitting in the Astronomer's Lodge at the observatory on the morning of February 3, having breakfast and planning the day's work, when one of the young scientists looked up from his laptop and reported that CNN was saying that radio contact had been lost from the Space Shuttle Columbia. That brave crew had died over our heads only moments earlier without our knowing it.

On a lighter note, we dedicated *ROTSE IIIb* with a quintessentially Texas tradition. While ships are dedicated by smashing a bottle of champagne over the bow, I felt it more appropriate to the West Texas environment and culture to stomp a jalepeño pepper. We had done this once before with the dedication of a special-purpose supernova search telescope. In this case, I again provided the jalepeños, and we have a nice little video of the team in fierce unison stomping the peppers into the grate work in front the enclosure door.

11.3 THE SHAPE OF THINGS

One of the issues that had to be confronted in the study of gamma-ray bursts was the manner in which the energy is released into the surroundings. There are a number of tightly intertwined issues here. Theoretical models of relativistic blast waves and the afterglow demand that a shock wave moves out from the source at speeds very close to the speed of light. To do this, the flow of energy must carry along with it very few ordinary particles, protons or, more generally,

Figure 11.1 The 0.45-meter Robotic Optical Transient Search Experiment telescope *ROTSE IIIb* in the foreground and the 9-meter Hobby–Eberly Telescope in the background at the McDonald Observatory in the Davis Mountains of West Texas. The four *ROTSE* telescopes were designed and implemented by a team from the University of Michigan headed by Carl Akerlof. McDonald Observatory is operated by the University of Texas. (Photo: Courtesy of Don Smith.)

baryons (Chapter 1). Too many of these particles of ordinary matter would slow the shock wave down so that it could not propagate with the deduced speeds. That is one thing that must distinguish an ordinary supernova and a gamma-ray burst. Both events have roughly the same amount of energy, but supernovae put their energy into moving a lot of ordinary matter at high, but not relativistic speeds. Gamma-ray bursts must put as much or more energy into a very small amount of mass.

Given the expansion at nearly the speed of light, a number of issues arise that come from Einstein's special theory of relativity. When motion with respect to an observer is high, lengths are foreshortened, and times are constricted. A gamma-ray burst that takes a minute as observed at the Earth may have spread over a region the

size of the Solar System at its origin. An event that takes several months to develop in the host galaxy of the gamma-ray burst may take only hours or days as observed at Earth. In particular, it may take many months for the relativistic shock wave to expand out from the source of energy, pile up mass in the interstellar medium, and slow to ordinary speeds. An observer on Earth would see all this playing out in a day or so. Turned around, when we see a gamma-ray-burst afterglow fading over a few days, it might have taken months in a far galaxy.

Another interesting effect is that, if a source of radiation moves toward an observer at a high speed, the radiation is thrown in the direction of the observer. This ''beaming'' can make the radiation seem brighter than it would otherwise be. In addition, if the source of energy is moving toward the observer, there is a very large blue shift, a ''boost'' of the energy of each photon that is detected. This can again make the source look brighter.

Such issues arise in trying to determine how bright a given gamma-ray burst really is and how much energy it emits. Even if the energy from a gamma-ray burst is emitted equally in all directions, it will be beamed and boosted and look brighter for a shorter time to an observer standing still on the Earth, compared to an observer at the same distance who moved with the velocity of the shock. Trying to figure out how bright a given gamma-ray burst ''really'' is in its own rest frame is a rather tricky business that requires an understanding of just how the boosting and beaming is working.

One can get a measure of the total energy emitted in the radiation independent of the beaming and boosting if the energy is emitted equally in all directions. The procedure is to add up all the energy received at Earth over the course of the burst event. That energy might have been emitted over a different time span in the frame of the explosion, but all the energy is all the energy, and it must all go somewhere eventually. If one assumes it goes off equally in all directions and corrects for the fact that things look dimmer by the inverse square of the distance (plus perhaps some corrections for cosmological warping), then the total energy in radiation of the explosion can be determined. For the first *BeppoSAX* events for which there was a measure of the red shift and hence the distance, the results were imposing, as mentioned earlier. For the event at the largest distance of the first few identified, GRB 971214 at 12 billion light years appeared to have emitted an energy comparable to the entire flow of neutrinos from a supernova, a huge amount of energy, and for GRB 990123 the corresponding amount would have been

ten times the neutrino energy of a supernova. In the early, heady, days of the afterglow revolution this was labeled by some as a result that threatened to challenge physics at a fundamental level. Challenges to the core of physics do arise from some astronomical observations as we will see in Chapter 12, but in this case the problem, while fascinating, had a more mundane yet far-reaching solution.

There is an important caveat to the method of measuring energy just outlined. If the flow of energy does not come out equally in all directions, if it is collimated in some way, if it flows out in a jet, then less total energy is required for a given observed burst, just in proportion to the amount of collimation, as shown in Figure 11.2. If the energy flows only into 10 percent of all available directions, then a given energy received on Earth requires only 10 percent as much total energy at the source. If the energy flows in a jet filling only 1 percent of the area around the source, then the energy at the source is only 1 percent of that deduced from the assumption that equal energy goes in all directions.

This collimation effect is not a fantasy. It is almost the rule rather than the exception. We see collimated flows from the Sun, protostars, planetary nebulae, binary black holes, and quasars. If the energy of a gamma-ray burst comes out in a collimated relativistic blast wave in only certain directions, then one must be careful in making estimates of luminosities and energies.

An example of this phenomenon is the ''blazars.'' Blazars are a certain subclass of quasars that are especially bright and highly variable. The common interpretation is that in these objects we happen to be looking right down the nozzle of a jet of matter ejected at nearly the speed of light. By the accident of the Earth's position in the beam, we see an especially bright source of radiation because of the beaming and boosting associated with the rapid motion toward us. We also see especially rapid time variability of the radiation that is thought to be associated with the shrinkage of time due to the relativistic motion. No one suggests that this energy is flowing out equally in all directions, thus requiring unprecedented amounts of energy, even for quasars. Rather it is assumed that, if we happened to observe the same object from the side, it would resemble an ''ordinary'' quasar. Understanding whether gamma-ray bursts are collimated and, if so, how and by how much became one of the key tasks facing the field.

My colleague, Lifan Wang, and I were among the first to point that this ''jetting'' or ''collimation'' might both be expected for gamma-ray bursts and important for their analysis, and that this

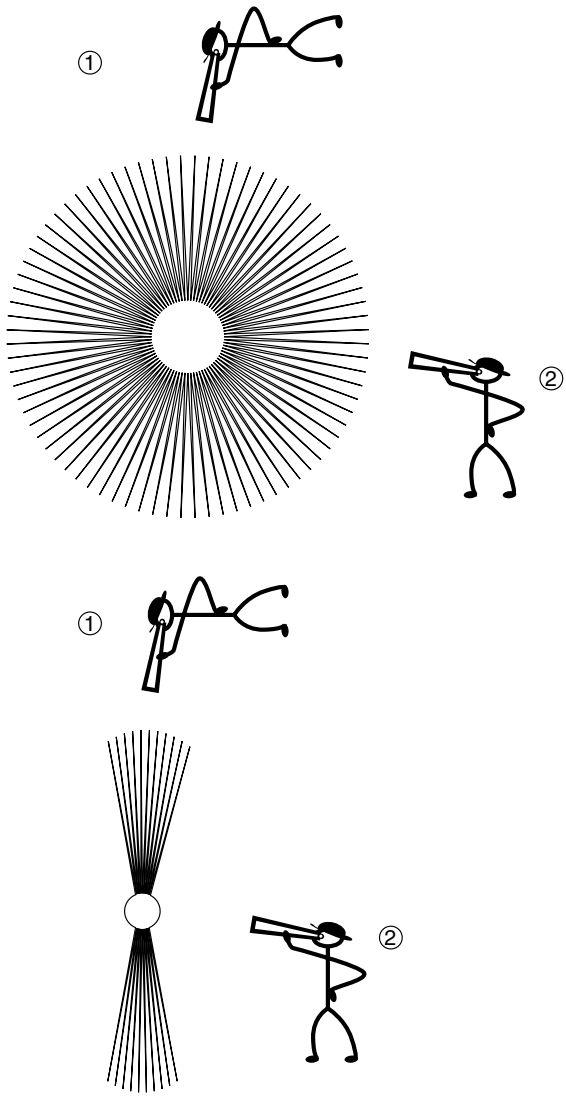


Figure 11.2 (Top) If the energy in a gamma-ray burst flows out equally in all directions, then it does not make any difference where the observer is. All observers at the same distance will see the same brightness and deduce the same energy. (Bottom) If the energy is collimated into a jet, however, the observer 1 who looks down the jet will see a much higher luminosity than the observer 2 looking from the side. If observer 1 assumes that the energy is emitted equally in all directions, he will deduce too large a total energy for the event.

property might link gamma-ray bursts to supernovae (next section). Our thinking was driven in part by our growing understanding from our polarization studies that core-collapse supernovae were asymmetric and often even ''jet-like,'' as outlined in Chapter 6. As things developed, it turned out we were on the right track. The proof that gamma-ray bursts involved jets and were related to supernovae came from different quarters, but we take some satisfaction that we had the correct basic ideas.

Our idea was to see how far one could go with using only relatively ordinary supernovae to produce gamma-ray bursts. The argument was that all gravitational-collapse events produce strong magnetic jets that punch out through the axes of the surrounding carbon/oxygen core. In ordinary Type II supernovae, the outer hydrogen layers would stop these jets. In Type Ic or Type Ib, the jet could escape into interstellar space making the gamma-ray burst.

In this picture, there are two components to the gamma-ray emission, one that radiates more or less equally in all directions with the energy about one thousand times less than a standard supernova expansion energy, and one component that is highly collimated in a relativistic jet containing perhaps 10 percent of the total supernova energy. The lower-energy component could be seen if the explosion occurred relatively nearby, 100 million light years or less, but would not be detectable with current instruments if the same event were at truly cosmological distances. The other gamma-ray component emerges in the jet so that all the gamma-ray energy contained in it is collimated to flow in a narrow angle. In this way, only some fraction of the supernova energy is required to be channeled into gamma rays.

With this picture in mind, Lifan and I were among the first to argue that the huge energies deduced for the very distant gamma-ray bursts was an artifact of assuming that equal energy is emitted in all directions, rather than being confined to the direction of the jet, as in the blazar picture described earlier. To reduce the required energy from the amount deduced in an ''all directions'' picture to some fraction of a supernova energy, the jet must be tightly collimated. The area of its cross section must be only one part in a thousand of the area surrounding the burst source. We noted that this is about the amount of collimation seen in typical jets from active galaxies, so it was not beyond the bounds of credibility. Whether it is produced in a real supernova is another story that is the subject of intense investigation, as outlined in Chapter 6.

If the jet moves at nearly the speed of light, the gamma rays will be blue-shifted and beamed strongly in one direction. This component could, in principle, be seen at cosmological distances if the jet happens to be pointed right at the Earth. Most of the jets will not be pointed at the Earth, so this picture requires many more gamma-ray-burst events that are not pointed at the Earth to account for the few that are. If the collimation is to one part in a thousand, then there must be one thousand jets not pointed at the Earth for every one that is. The required rate of bursts in that case would be roughly that for normal supernovae, approximately one per few hundred years per bright galaxy, giving a crude concordance to the argument.

While our reasoning was on the right track, the afterglows themselves produced the direct evidence that the energy flow is, indeed, strongly collimated, but probably not quite as much as we speculated. The important evidence is that, even though some of the afterglows fade roughly inversely with time as expected for spherical relativistic blast waves, a few were observed to decline more rapidly. The explanation for this behavior requires the invocation of a jet-like, rather than spherical, flow. A critical difference between a jet and a spherical blast wave is that, when it slows down, a jet can expand sideways. This sideways expansion can tap the energy of the jet and cause more rapid cooling and deceleration and hence a more rapid rate of decline of radiation output.

By now many burst afterglows have been analyzed and shown to reveal this behavior. Quantitative analysis by many people, including my colleague, Pawan Kumar, is consistent with them being collimated to only one percent of the sky, or even less. This is certainly well-collimated, but to a somewhat looser extent than what Lifan Wang and I guessed. This means that the energy is reduced by a factor of 100 or more, and that gamma-ray bursts must be 100 times more common than the actual rate of detection, about one per day, would imply. Even with this ''most are beamed away from us'' factor taken into account, gamma-ray bursts are deduced to be more rare than normal core collapse supernovae, and probably even more rare than the ''usual'' production of black holes. It is very unlikely that every core collapse supernova yields a gamma-ray burst, but even that conclusion is occasionally questioned (Section 11.6).

When this strong collimation was invoked for GRB 990123, the energy deduced for it was reduced from a mind-boggling level equivalent to the expansion energy of three thousand supernovae to about 10 percent of the total collapse energy of a neutron star, only

ten times the expansion energy of a normal supernova. A new phrase entered the literature, the ''isotropic equivalent'' energy. The idea was that this was the fictitious energy that would have been emitted if the burst radiated equally in all directions – isotropically. The isotropic-equivalent energy was a convenient measure of the *apparent* energy, but not to be confused with the *actual* energy emitted, the error made in the first blush.

Armed with this insight, people revisited the issue of the energy of the whole sample of gamma-rays bursts where adequate data was available, the best data involving the time behavior of all the radiation bands from radio to optical to X-ray. The remarkable result was that the rather wide spread in isotropic equivalent energy collapsed to a rather narrow distribution of actual energy emitted. It appears that all the bright gamma-ray bursts have an energy that falls within a rather narrow range (within a factor of a few). The energy deduced in this way is comparable to, but somewhat less than, the kinetic energy, the energy of motion, of a typical exploding supernova. This energy is 100 times less than the total energy released in neutrinos in core collapse, making it actually an interestingly small number, not a challengingly large one. The bottom line is that while gamma-ray bursts remain amazing and mysterious events, their energy is rather modest by supernova standards.

It is now generally accepted that many if not most gamma-rays bursts and their afterglows are jet-like. There are, however, other explanations for the rapid decline of the light of afterglows. If gamma-ray bursts arise in massive stars, as discussed in the next section, then they should be surrounded by the matter blown off in a stellar wind (Chapter 2, Section 2.2). Even a spherical blast wave would collide with this wind and slow more rapidly than if it only interacted with the dilute matter of the interstellar medium. This interaction can also account for the rapid declines seen in some afterglows. There is, of course, nothing to prevent a jet from colliding with a wind, and if the source of gamma rays pumps out energy for a prolonged time, the tendency for the power to decline can be overcome. There are lots of complications to be pursued and understood.

11.4 THE SUPERNOVA AND GAMMA-RAY-BURST CONNECTION

The third major achievement of the afterglow revolution, after proof of cosmological distances, and discovery that the relativistic outflow is a collimated jet, was the connection of gamma-ray bursts to

supernovae. The discovery of the galaxies that were host to gamma-ray bursts also brought suspicion that they were related to massive stars. The gamma-ray bursts were neither far out in the host galaxies, nor in the centers where active nuclei might lurk. Rather, they seemed to be in regions of active star formation. This provided circumstantial evidence that they were related to massive stars and hence, perhaps, to core-collapse supernovae. In the onrush of events that followed from the *BeppoSAX* discoveries, another surprise made the relation of gamma-ray bursts and supernovae explicit.

On April 25, 1998, *BeppoSAX* discovered a gamma-ray burst, GRB 980425, of otherwise ordinary properties in terms of its apparent brightness, energy, and timescale. *BeppoSAX* then swung to bring its fine-position-sensor X-ray detector into position and detected a couple of X-ray sources, one of which diminished in time and one of which seemed to be constant. A day later, optical astronomers caught up and found a strongly variable object. This object was not, however, the afterglow that one had quickly learned to expect. It was, rather, a supernova, one of rather strange properties. The supernova, SN 1998bw, was not exactly at the position of either of the two X-ray sources first reported by *BeppoSAX*. This raised some question about the association of SN 1998bw with GRB 980425. In the next few months, the *BeppoSAX* team recalibrated the positions of the X-ray sources they detected. The source that was at first observed to vary was determined to be much too far from SN 1998bw to be associated. The other source, at first thought to be constant, was shifted so that an association with SN 1998bw could not be ruled out. Then this source was discovered to be variable, if only slightly. This has left the issue of the association of SN 1998bw with the *BeppoSAX* X-ray sources somewhat befuddled. One must be wary of other sources of variable X-ray emission, such as active galactic nuclei, that could accidentally fall near the supernova, but an association of one of the X-ray sources with SN 1998bw cannot be ruled out.

A few days after the detection of SN 1998bw, Dale Frail of the National Radio Astronomy Observatory at Socorro, NM, Shri Kulkarni at Caltech, and their colleagues found a very bright radio source. This radio source was precisely at the position of SN 1998bw, so there was no question of their association. Analysis of the radio data showed that the radio source was brighter than could be easily explained without expansion of a shock wave at nearly the speed of light. Independent of the gamma-ray burst, SN 1998bw clearly produced a relativistic blast wave. All this evidence taken together suggests that

SN 1998bw and the gamma-ray burst GRB 980425 are one and the same thing. The likelihood of finding both GRB 980425 and SN 1998bw in the same part of the sky in the brief interval of time when they erupted is very low, so most astronomers think the connection must be real. In particular, even though gamma-ray astronomers tended at first to be leery of the association, supernova mavens embraced it with full passion.

Observations of SN 1998bw and its host galaxy showed that it was at a distance of about 40 million parsecs, or about 120 million light years. That is a great distance, but far less than, for instance, the 12 billion light years of GRB 971214. At 40 million parsecs, the total energy in the gamma-ray burst is deduced to be much less than that of the most powerful gamma-ray bursts, by a factor of about 1 million. On the other hand, at the same distance, SN 1998bw was exceptionally bright for a supernova. Both of these results are puzzles that have still not been fully assimilated in the ongoing attempt to understand gamma-ray bursts.

Although it is a step along an esthetically ugly path, one idea that emerged from this new event was that there were at least two kinds of gamma-ray bursts, one of very high energy seen at cosmological distances and one of lower energy seen relatively nearby. This is an uncomfortable hypothesis given that the gamma-ray properties of GRB 980425 were seemingly unexceptional. The similar nature of faraway energetic and nearby lower-energy gamma-ray bursts may arise because any physical events that can emit gamma-rays will have certain properties in common whether the total energy involved is high or low, but this remains to be shown. Another possibility that is actively discussed is that these bursts are all basically the same thing, but that the burst looks different, and dimmer, if you look at it from an angle rather than having it aimed right at you.

SN 1998bw brought its own set of questions. The early spectra seemed unlike any other supernovae we have discussed, Type Ia, Ib, Ic, or II. Closer study showed a similarity to Type Ic, but with especially high velocity causing an exceptionally large Dopper shift and ''broadening'' of the absorption features associated with atomic absorption. As it evolved, SN 1998bw looked more and more like a Type Ic with no evidence for hydrogen or helium. It certainly did not look like either a Type II or a Type Ia. With hindsight, there were a few other supernovae – SN 1997ef is a conspicuous example – that did bear some resemblance to SN 1998bw, and there have been a few more since.

The first models of the light curve and spectra assumed that SN 1998bw resulted from core collapse, and that enough radioactive nickel was produced to power the peak of the light curve. Because SN 1998bw was about as bright as a Type Ia (even though the spectrum is completely different), a comparable amount of radioactive nickel (Chapter 6, Section 6.6) is required, about 0.7 solar masses. Basic spherically symmetric models can produce this amount of nickel in a core-collapse explosion by shocking silicon layers, but they are extreme. Models that make this much nickel and that produce the observed light curve and spectra at some level of agreement (not perfect in the first models) require an exploding carbon/oxygen core of about 10 solar masses and an energy of expansion of the matter of more than ten times that normally associated with supernovae. These models suggest that SN 1998bw was a ''super'' Type Ic, and the term ''hypernova'' has been adopted in some circles. SN 1998bw was certainly exceptional in many ways. Other events labeled ''hypernovae'' have shown rather high velocities, but normal luminosity for a Type Ic, no relativistic outflow, no radio outburst, no gamma-ray burst. Exactly which events should bear the label ''hypernova'' is, at least, controversial.

Like Type Ic, SN 1998bw showed signs of asymmetry (Chapter 6), evidence that the flow of ejected matter departs rather strongly from spherical symmetry. This evidence was ignored in the first spherically symmetric ''hypernova'' models that require unprecedented amounts of energy to provide the supernova luminosity. Peter Höflich, Lifan Wang, and I considered models that are distorted by a sufficient amount to account for the asymmetries in Type Ic supernovae and in SN 1998bw itself. Preliminary models showed that, if the ejecta were in the shape of a fat pancake, they would be appreciably brighter if viewed from the top of the pancake compared to the edge, by about a factor of two. These models have the potential, at least, of accounting for the observed optical properties of SN 1998bw with ''normal'' amounts of energy and ejected nickel mass. Whether such models, or the ''hypernova'' models for that matter, can account for the gamma-ray properties remains to be seen.

The question of the connection of supernovae and gamma-ray bursts was further fueled by developments in the spring and summer of 1999. One gamma-ray burst from 1998 was later found by Shri Kulkarni, Josh Bloom, and their colleagues at Caltech to show evidence for a brightening about three weeks after the gamma-ray burst that interrupted the otherwise rather rapid (and hence from a jet?) decline

of the afterglow. This apparent new source of light was roughly consistent with the addition of the light from a ''SN 1998bw-like'' event that reached peak about three weeks after the gamma-ray burst, a reasonable time for a supernova to have attained maximum light output after its initiation. After this discovery, the original afterglow event, GRB 970228, was also reanalyzed by Dan Reichart, then a graduate student at the University of Chicago. Dan found evidence for a ''SN 1998bw-like'' brightening, and similar arguments were advanced for one or two more events. All this added to the growing circumstantial evidence that supernovae, most likely some variant of Type Ic, and gamma-ray bursts were connected.

Another strong piece of evidence in this direction was the occurrence of GRB 021004. This was the first gamma-ray burst that we successfully observed at McDonald Observatory with the Hobby–Eberly Telescope. Lots of other people got wonderful data on it as well. This burst showed rather direct evidence of material blown out from a massive star in a stellar wind prior to the explosion. This added to the growing conviction that gamma-ray bursts were associated with the death of massive stars.

At this point, essentially every major observatory on the planet was engaged in the supernova hunt. The proof came in March of 2003 with GRB 030329, discovered by *HETE 2*. This burst proved to be relatively nearby, only 3 billion light years away! Right next door compared to the 12 billion light years of GRB 971214. This was a statistically rare event, making this one discovery well worth all the effort that went into the disaster of *HETE 1* and the success of *HETE 2*, even if the latter had done nothing else. Everyone knew this was a good candidate from which to search for direct proof of the supernova connection. We certainly tried. We knew what to do: look after the gamma-ray burst for evidence of a rising contribution of supernova light and get a spectrum to prove what it was. Unfortunately our telescope was not quite sensitive enough for the task. Other observatories pinned it down, but there it was, just as expected. The early afterglow showed no evidence of a supernova, but about a week later, an extra contribution of light was seen. After the careful job was done of allowing for the still bright light of the afterglow itself, a spectrum was obtained and it was nearly identical to that of SN 1998bw, a *bona fide*, if somewhat strange, supernova. This was unambiguous proof that this gamma-ray burst arose in the explosion that created a supernova.

One has to be careful not to leap to the conclusion that every gamma-ray burst arises in a supernova, but that is clearly where all

the evidence is pointed, at least for certain classes of gamma-ray bursts. The gamma-ray burst and supernova communities have basically accepted this conclusion and are moving on to ask more detailed questions: what supernovae, why, and how?

11.5 THE POSSIBILITIES: BIRTH PANGS OF BLACK HOLES?

These years of mind-churning progress after the first *BeppoSAX* discovery have left a large range of issues concerning gamma-ray bursts that will take more work and ingenuity to resolve. Principal among these is the basic nature of the explosion. What sort of explosion is involved, and how is it related to ''normal'' supernovae? Other, closely related, issues are why the energy is collimated, how it gets out of the star without dragging so much star stuff that it cannot blast relativistically into space. How, exactly, is the blast converted to gamma rays? While some of the bursts show evidence for the circumstellar matter that is expected to be expelled in the wind from a massive star, others rather distinctly do not. How can that be, if gamma-ray bursts all come from massive stars? Is there, after all, more than one way to make a gamma-ray burst? Some of the *BeppoSAX* and *HETE 2* events showed optical afterglows, but others did not. Most of the afterglows decay so that the power fades inversely with time, but some decay more rapidly. In a real sense, the field is just beginning and will continue to explode with activity.

A plethora of models have been devised to address the gamma-ray burst energy issue head-on.

Some of these schemes involve colliding neutron stars at the end of a long gravitational in-spiral. That process has plenty of energy, enough for the most extreme events if the energy emerges in a jet. Another principal issue is turning the energy into gamma rays and a relativistic blast wave that is not so overloaded with protons that it cannot move rapidly enough to make the burst or the afterglow. One possibility that has been discussed is that the neutron stars do not collide directly but interact through their strong magnetic fields. That way, one can think about turning the pure magnetic energy into pure gamma-ray energy without getting the stuff of the neutron stars, those troublesome, slowing baryons, directly involved. The problem with that class of models is that neutron stars require a long time to spiral together under the grip of gravity waves, so they are expected to have drifted farther from the star-forming regions of host galaxies than gamma-ray bursts are observed to do. Such a model might still

account for some fraction of observed gamma-ray bursts (see Section 11.6 in this chapter).

Other models invoking neutron stars suggest that the powerful radiation from a newly born pulsar could result in a gamma-ray burst. These models have the possible advantage that they are the smallest step away from ''normal'' supernovae. In addition, as discussed in Chapter 6, we have found that normal supernovae that are most likely to involve neutron star (rather than black-hole) formation are asymmetric and might involve jets. Gamma-ray bursts seem to occur less frequently than core-collapse supernovae (but see the next section), so it must be the rare supernova that makes a burst. On the other hand, the highly magnetized magnetars (Chapter 8, Section 8.10) are more rare than ordinary pulsars. We do not know what the birth event of a magnetar is like; could that also be the rare explosion that produces a gamma-ray burst?

I have written several papers on this topic, analyzing the capability of a new-born neutron star to produce magnetic jets in normal supernovae, in extreme events like SN 1998bw, and even, perhaps, in gamma-ray bursts. In one paper, we envisaged a neutron star spinning like a pulsar with a simple dipole magnetic field, with the magnetic axis tilted with respect to the spin axis (Chapter 8, Section 8.2). Then we realized that when it is first born, the field is likely to be wrapped around the equator like a doughnut. In a paper with Dave Meier, a magnetic-jet expert from the Jet Propulsion Laboratory, and Jim Wilson, a pioneer of supernova-collapse calculations in general and magnetic collapse in particular, we analyzed how a torus of field might make a jet and explosion. We envisaged that there might be a first jet when the neutron star first forms that explodes the star; this would be a normal supernova. In some cases, however, we imagined that the subsequent rain of material crushes the neutron star to a black hole, and that launches a second, even faster jet that catches up to the first and creates the gamma-ray burst. A possible advantage of this picture is that both jets could be full of magnetic field which must be there to make gamma-ray bursts radiate as they do, but the origin of which is not well explained.

The most popular model to account for the production of gamma-ray bursts involves the collapse to form a black hole. This has also been termed the ''collapsar'' model, a word coined by Stan Woosley of the University of California at Santa Cruz, who has advocated such a model with great vigor. Strictly speaking, a stellar collapse could yield either a neutron star or black hole, but in its

popular usage, *collapsar* means the generic class of models based on black-hole formation.

Woosley and his colleagues envision collapse to form a spinning black hole. Subsequent infall forms an accretion disk of matter around that black hole. They assume that the accretion energy is channeled up the rotation axes by the natural axial nature of the rotating geometry or perhaps with the collimating aid of twisted magnetic fields. A jet of energy with plausibly sufficient energy and the capability of emerging relativistically into the surrounding space could be generated.

The appeal of this class of models is clear. Gamma-ray bursts are extreme events and black hole formation is an extreme event. We commonly see relativistic jets emerging from supermassive black holes in active galactic nuclei (Chapter 10, Section 10.9) and from miniquasars in some binary black-hole systems (Chapter 10, Section 10.8), so the parallels are compelling. In addition, detailed numerical models can account for various aspects of the problem, the formation of jet-like flow from the vortex around the black hole, the propagation of a jet out through the star with sufficiently large energy but small baryon load that it can emerge and accelerate to something like observed gamma-ray burst speeds.

Nevertheless, as in other contexts, invoking something as exotic as black holes requires a high standard of proof, and that proof is not yet forthcoming for gamma-ray bursts. The black hole explanation also brings some conundrums of its own. We do not know exactly what is the mass of stars that collapse to make black holes, but we suspect it is moderate, perhaps around 30 solar masses. Even allowing for the fact that we probably witness only one out of a hundred gamma-ray bursts because the others are aimed away from us, the rate of formation of gamma-ray bursts seems to be significantly less than the rate of death of 30-solar-mass stars. That would suggest that not every collapse which forms a black hole yields a gamma-ray burst. We need to understand why that is so. Black holes seem plausible because they can, in principle, provide a huge energy, but there is a puzzle of just the opposite sort. With collimation, we know that the typical energy in gamma-ray bursts is somewhat less than the typical expansion energy of a supernova, and is a factor of over one hundred less than the gravitational or rotational energy associated with formation of a black hole. How is it that such a small and yet well-defined fraction of the total reservoir of energy available is channeled into the gamma-ray burst?

There are also theoretical issues that remain to be resolved. There is a general perception that if a black hole launches a jet, that jet can both explode the star and produce the gamma-ray burst. This is not at all clear. For the jet to make a gamma-ray burst, it must be thin and fast to penetrate the star without slowing down too much. That means it cannot interact with the star very much, and that means it cannot explode the star. The analogy we invented in the paper by Wheeler, Meier, and Wilson referred to earlier is that this is like shooting a needle through a loaf of bread. The needle could penetrate without perturbing the loaf. How, then, does the star explode as a supernova? It could be that the ''standard'' processes of neutrino transport (Chapter 6, Section 6.4) do the trick, but that is far from proven, even for normal supernovae, and certainly in the case when a black hole, not a neutron star, forms. It also remains far from firmly established exactly how a new-born black hole produces a jet and under what circumstances that jet will have the right properties to be a gamma-ray burst. The role of magnetic fields in this process have scarcely been addressed.

People are thinking about these issues. There are a number of interesting papers discussing black-hole formation in a variety of contexts, from single stars or, even more interesting, from various binary systems. Some of these models-involving swallowing the black holes in common envelopes of normal stars or of helium stars-might be the progenitors of Type Ib or Type Ic supernovae. An advantage of the binary models is that they have some promise of spinning up the progenitor star and thus providing an especially rapidly spinning black hole, a seeming requirement for a successful gamma-ray burst model. This special requirement might also help to explain why not all black-hole formation yields a gamma-ray burst.

Other suggestions have problems as well. A key one for any model based on neutron stars, rather than black holes, is the danger that a jet emerging from near a neutron star would be far more contaminated with neutron-rich matter than observations allow.

All these pictures have a certain basic plausibility about them, given that we think our Universe is full of magnetic neutron stars and black holes of a range in mass from that of stars to that of galaxies. The devil is in the details. Having accounted for the energy, the first major requirement, can any of these models really account for gamma-ray bursts with the observed properties? All these models that are designed to give very high energy gamma-ray bursts at cosmological distances must also confront GRB 980425 and SN 1998bw. How is

it that a newly formed accreting black hole in the young Universe produces a gamma-ray burst with the same average observed properties as a relatively nearby, much less energetic, odd supernova?

I have written some papers exploring the question of whether or not gamma-ray bursts are related to the formation of neutron stars, in part just to keep this option on the table. If I were to bet, I would bet on some form of a black-hole model. To my mind, resolving this issue is the biggest problem remaining in gamma-ray burst research. Just what is the nature of the gamma-ray-burst machine, and how do we prove gamma-ray bursts involve black holes if, in fact, they do?

11.6 THE SHORT HARD BURSTS

As this gamma-ray burst story has unfolded, another aspect was revealed; BATSE showed evidence that there were two flavors of gamma-ray bursts. The majority were the type we have described so far and they have come to be known as ''long'' gamma-ray bursts, the type that typically last for tens of seconds. As the thousands of BATSE bursts accumulated, however, it became clear that there was another population of bursts, about a quarter of the total. These bursts lasted substantially less than a few seconds, frequently only a fraction of a second. The radiation from them also was, on average, of slightly higher energy, or ''harder'' in gamma-ray lingo, so they became known as ''short hard bursts.'' A stubborn puzzle of gamma-ray-burst research has been to understand this dichotomy in temporal behavior. Do the long and short bursts represent variations on a theme, or two distinct physical processes?

Some insight into this issue came from the behavior of the soft gamma-ray repeaters described earlier (Chapter 8, Section 8.10). While the majority of the energy in a soft gamma-ray-repeater outburst comes in relatively low energy or ''soft'' gamma-rays, the outburst that lit up the northern aurorae in August of 1998 was heralded by an initial intense, short-lived, energetic spike lasting a few tenths of a second. The source later showed a decaying, pulsing, flux of lower-energy radiation, as described in Chapter 10. The ''hard'' gamma-ray burst of that initial spike was indistinguishable from the short hard gamma-ray bursts. Because the soft gamma-ray repeaters are highly magnetic neutron stars or magnetars (Chapter 8, Section 8.10) in our Galaxy, this raised the question of whether or not all the short gamma-ray bursts could arise from neutron stars in our Galaxy. If this is so, their distribution should not be uniform on the sky because of

the Sun's offset position from the center of the Galaxy. The short hard bursts are, however, uniformly spread over the sky, so something else was going on. For technical reasons, *BeppoSAX* could not respond to these short bursts, so everything that has been learned about gamma-ray bursts and their afterglows in the *BeppoSAX* era pertained only to the ''long'' gamma-ray bursts. Until recently, even the distance to the short hard bursts remained a mystery, as it had for the long bursts for so many decades. We did not know whether they exploded in distant galaxies, or in the depth of intergalactic space, or somewhere else.

New insight into the nature of some of the short hard gamma-ray bursts came with the bright soft gamma-ray-repeater magnetar outburst that was detected on December 27, 2004 (Chapter 8, Section 8.10). This burst again began with a brief, intense, highly energetic spike that lasted only 0.2 seconds. As for the 1998 burst, that timescale put it in the range of the ''short'' gamma-ray bursts. The 2004 spike was, however, 100 times brighter than the initial spike of the 1998 burst. The teams of astronomers who analyzed the 2004 data, including my colleague Rob Duncan, deduced that such a burst could easily be observed to great distances, far beyond our Galaxy. They concluded that the BATSE sample of short hard bursts almost surely contained such magnetar bursts, perhaps half of all the short bursts BATSE detected. One still had to account for the other half.

The summer of 2005 brought a dramatic new chapter in this story. *Swift* found an X-ray afterglow of a short hard burst detected on May 9. An optical afterglow was not found, but the evidence pointed to the burst arising in an elliptical galaxy at modest distances by gamma-ray-burst standards, a few billion light years away. Elliptical galaxies are thought to have little star formation, so this association pointed to a significant difference compared to the long bursts that arise in short-lived massive stars and supernovae. Then the hardworking *HETE 2*, nearly overshadowed by the success of *Swift*, found another short hard burst on July 9. This burst had both X-ray and, even more importantly, optical afterglows. The host galaxy, again a few billion light years distant, was a modest-size galaxy with some star formation percolating along in it. Then *Swift* found two more; one on July 24 in another elliptical galaxy and one on August 13 in a very distant cluster of galaxies (with the specific host galaxy difficult to pinpoint). The sample is still small, but enough to start making some general deductions.

The short hard bursts are relatively nearby compared to the typical long bursts and produce a total energy output that is quite a bit

less, perhaps by a factor of ten. The emission does seem to be collimated, but somewhat less so than for the long bursts. In addition, the evidence suggests that the short hard bursts arise in an old population, even if they sometimes appear in galaxies with some star formation going on. Similar arguments apply to Type Ia supernovae that are thought to arise in an old population, even when they appear in a spiral galaxy where some of the stars are young. As remarked above, this evidence that the progenitor systems are old distinguishes them from the long bursts that are directly associated with young, massive stars. Even more critical, people looked very hard for supernova light in the optical afterglow of the July 9 *HETE 2* burst and found none. The limits are very tight. Any supernova-like optical display a couple of weeks after the burst must have been dimmer by at least a factor of 100 compared to ''normal'' Type Ic supernovae, and even more so compared to SN 1998bw.

The consensus is that the accumulating evidence is most consistent with a notion that has been pondered for the last few years. The idea is that the short hard bursts arise when two neutron stars, or perhaps a neutron star and a black hole, spiral together in a binary system under the influence of gravitational waves (Chapter 1, Section 1.10; Chapter 4, Section 4.4). Such a system would take a long time to coalesce, but the destruction of one or both neutron stars would produce a great deal of energy, plausibly in the gamma-ray portion of the spectrum and plausibly concentrated along the rotational axis of the orbiting pair. The great age expected for such systems is consistent with their appearance in elliptical galaxies (and one reason this model was rejected for the long bursts, after some initial interest), and with little matter around, no supernova light would be expected.

When the consensus arrives this quickly and with such force, my little contrarian itch needs scratching. I muse that an accretion-induced collapse of a bare white dwarf in a binary system could be old and would produce very little in the way of an optical display; there would be very little matter ejected and very little radioactive nickel-56 ejected to make it glow anyway. If the white dwarf collapsed to make a rapidly spinning, magnetized neutron star, one might get enough energy ejected up the rotation axis to make the relatively wimpy burst. As is widely discussed in the literature, the binary merger model is very likely (but not absolutely so) to make a substantial burst of gravity waves. The white-dwarf collapse picture would likely (but not absolutely so) generate very little in the way of gravity waves.

Future gravity-wave detection experiments might thus be able to distinguish between these two possibilities.

11.7 THE FUTURE

There has been immense progress in the afterglow era, establishing that gamma-ray bursts arise in explosions of Type Ic-like supernovae that produce highly collimated, relativistic jets in exceedingly distant galaxies. The short hard bursts also occur in distant galaxies, but appear among older stars and with no sign of an accompanying supernovae. There are also still many open questions. Among those scattered through this chapter pertaining to the long bursts are: what supernovae are associated with gamma ray bursts, and how often; what is the mechanism of explosion of the supernova; is a neutron star or a black hole involved; why is just a certain amount of energy emitted in the burst and how does that energy get out of the star; how are the gamma-ray burst itself and the subsequent afterglow produced; what is the effect of the burst on the environment of the galaxy in which it erupts? For the short hard bursts, are we observing coalescing neutron stars and if so, how do they produce collimated bursts of gamma rays?

One of the key open issues is whether or not there are explosions related to gamma-ray bursts, but with less energy, so that the gamma-ray bursts represent only the most easily observed eruptions due to their great power, that they are only the tip of the iceberg. One frontier in this regard is the study of what are known as X-ray flashes.

Over several decades, various X-ray satellites had witnessed brief flashes of X-ray light, lasting about a minute with no obvious origin. There was some speculation that they were related to the more energetic gamma-ray bursts, the origin of which was also unknown over most of this interval. Progress on this front came by combining *BeppoSAX* and BATSE observations that revealed that the X-ray flashes did have faint gamma-ray counterparts. *HETE 2* provided more evidence that linked the two phenomena; about one-third of the bursts discovered by *HETE 2* were X-ray flashes or X-ray-rich gamma-ray bursts, strongly suggesting a continuity of properties. There was some speculation that the X-ray flashes were identical to gamma-ray bursts but from exceedingly large distances, so that the cosmological red shift would make them appear dimmer and of lower energy. This notion was abused by the location of two X-ray flashes in star-forming galaxies ranging from perhaps six to eleven billion light years away;

this is very far, but typical of regular gamma-ray bursts and arising in the adolescent, but not the extreme infant Universe (see the next section).

Studies are now underway to better understand the nature of the X-ray flashes and how they relate to gamma-ray bursts. One possibility is that, for some reason, the X-ray flashes represent an explosion where the energy is shared with more matter, so the burst moves more slowly and generates less energetic photons. Another idea is that the X-ray flashes are, indeed, the same phenomenon as gamma-ray bursts, but seen from an angle to the main collimated flow, making them a sideshow to the main feature. In either case, the indications are that X-ray flashes are more common than gamma-ray bursts when allowance is made that they are dimmer and cannot be seen over as large a volume, on average, as gamma-ray bursts. Depending on the interpretation, some argue that essentially every Type Ic supernova must produce either a gamma-ray burst or an X-ray flash. There are countervailing arguments to this, but the discussion illustrates the range of issues yet to be fully studied and connected. The combination of *HETE 2* and *Swift* should produce a bounty of new X-ray flashes to study.

These are the conundrums that make astrophysics so exciting. Gamma-ray bursts will continue to provide all the stimulation an astrophysicist could want for some time to come. As better understanding of the gamma-ray bursts develops, so will a better understanding of the Universe on both stellar and cosmological scales. The gamma-ray bursts give us yet another means to look throughout the space and time of our visible Universe.

11.8 THE PAST IN OUR FUTURE: THE DARK AGES

Looking to the future brings yet another exciting possibility. After the epoch when the Universe was a million years old, the cosmic radiation streamed freely. The matter cooled and became dark. During the subsequent eons of expansion, the matter agglomerated into lumps that became galaxies. At some point, the gas in the lumps condensed and heated and started the first production of stars. The long interval between the release of the cosmic background radiation and the lighting up of the first stars has come to be called the ''Dark Ages.'' After a long period with no light, stars winked on and the Universe started to take the form we recognize around us now. The processes involved in forming the first stars and galaxies, the emergence from

the Dark Ages, is one of the frontiers of modern astronomy. It can be probed to some extent by the current generation of telescopes in the 8- to 10-meter class. The end of the Dark Ages will be the prime target of the *James Webb Space Telescope* currently under design by NASA, with plans to launch in 2013.

Some, maybe most, of those first stars to form will be massive. Some will evolve, collapse, and explode in just the way described in Chapter 6. When they do, their host galaxies will still be embryonic, small, and dim. There is a chance that, when astronomers peer back to the beginning of the end of the Dark Ages, they will see supernovae and gamma-ray bursts, the brightest beacons in the young Universe.

The first supernovae to arise should be from massive, short-lived stars. They should be predominantly some variety of Type II supernovae, although there could also be an admixture of Type Ib and Type Ic supernovae. The Type II supernovae might resemble SN 1987A by exploding as blue supergiants. As explained in Chapter 7, we do not fully understand why SN 1987A was a blue rather than a red supergiant when it exploded. Theoretical studies have shown, however, that when the amount of heavy elements in the atmosphere of an evolving massive star is low, the hydrogen envelope is likely to remain relatively compact so the star will look hot and blue, rather than expanding so that the star will look cool and red. In the very young Universe at the end of the Dark Ages, there will not have been much time to make heavy elements. My colleague Peter Höflich points out that whatever caused SN 1987A to be a blue supergiant, the paucity of heavy elements in the young Universe is likely to cause all the exploding stars to be blue supergiants, even if they retain their hydrogen envelopes against the ravages of winds and binary companions.

Another possibility, advocated by my colleague Volker Bromm, is that the first stars may be especially massive, perhaps up to hundreds of solar masses. The mechanism of explosion of these stars was studied in the late 1960s by Israeli astrophysicists, including my friend and colleague Zalman Barkat, with whom I shared a postdoc at Caltech long ago. Little use was found for the mechanism until now, but it is especially simple and elegant. When these massive stars produce a core of oxygen after helium burning, the core is hot enough to produce electron/positron pairs. Converting heat to mass in this way reduces the pressure and causes the star to collapse. Unlike an iron core, however, the oxygen core is very volatile; the oxygen ignites and explodes, blowing the star up completely, leaving no

compact remnant, but with a large production of radioactive nickel-56. These ''pair formation'' supernovae may be the first explosions to dispel the Dark Ages. At even greater mass, these stars might overcome the explosion of the oxygen core and collapse to produce black holes of hundreds or thousands of solar masses that could help to grow supermassive black holes (Chapter 10, Sections 10.9 and 10.10).

If the first supernovae at the end of the Dark Ages explode in blue supergiants, the resulting explosions, like SN 1987A, may be relatively dim and somewhat harder to see. If the first explosions are pair-formation supernovae, the task might be somewhat easier. As the Universe ages and more heavy elements collect in the interstellar gas from which new stars are born then, at some point, massive stars may begin to evolve into fully formed red supergiants before they die. They will then explode as what we consider to be ''normal'' Type II supernovae. With the full power of new telescopes to scan from the present epoch back to the end of the Dark Ages, we should be able to see that epoch when the normal Type II supernovae turn on.

Another exciting possibility that has attracted a lot of attention is the possibility to see gamma-ray bursts from this era. Because pair-formation supernovae explode completely, they will not produce gamma-ray bursts. If some of the stars in that very first epoch happen to have the more modest masses that evolve all the way to iron cores that collapse, then some of these stars should produce gamma-ray bursts. Since the gamma-ray bursts collimate their energy in jets, we will only see the ones pointed at us, but for those that are, what fireworks! These first gamma-ray bursts open up two exciting possibilities. One is to learn more about, and hence to better understand, the gamma-ray bursts themselves. Determining just how long after the end of the Dark Ages the first gamma-ray bursts began to erupt might give important clues to just which stellar collapses yield this phenomenon, and why. Another exciting possibility is simply to use the gamma-ray bursts (or supernovae, for that matter) as bright beacons to explore the early Universe. The notion is that the light from those distant explosions must traverse all the Universe between that distant, early time, and now. The radiation will be absorbed and affected in different ways as it travels, bringing with it a journal of its travels through that huge span of space and time during which the Universe made the transition from a uniformly dark place to one ablaze with stars and galaxies.

As one looks out in space and back in time, one runs out of both, since the Universe is only about 14 billion years old (Chapter 12). That

means than even the most distant objects are only about 14 billion light years away. The prediction is that gamma-ray bursts should be quite easy to see, even from that huge distance. In fact, the rather strong expectation is that in the nearly 3000 gamma-ray bursts recorded by BATSE, some must have been from this early era; we have just not yet figured out which ones. It was such a conviction that led people to propose that the X-ray flashes were gamma-ray bursts from this era of an infant Universe (Section 11.6). That particular idea was not correct, but that does not mean that some very distant gamma-ray bursts, the ones from the infant Universe at the end of the Dark Ages, do not await identification in the BATSE catalog. There is also a great expectation that *Swift*, and the global armament of follow-up that characterizes the afterglow era, will lead to the discovery of these very first bursts. Techniques have been developed to identify these especially distant and ancient bursts, and we await the first announcement with great anticipation.

This discussion has omitted Type Ia supernovae. That is because we think they have a ''fuse'' that must burn before they explode. As discussed in Chapter 6, we do not understand the binary evolution that leads to the explosion of a white dwarf as a Type Ia supernova. All the indications are, however, that considerable time must pass, a billion years or more in most cases, before these binary processes, perhaps the evolution of the smaller-mass companion, perhaps the decay of orbits through emission of gravitational radiation, lead to the explosion. That Type Ia supernovae have a long fuse compared to Type II means that when supernovae begin to explode at the end of the Dark Ages, they should all be due to the collapse of the cores of massive stars. There should be no thermonuclear explosions of white dwarfs and hence no Type Ia.

As the Universe ages and the binary evolution fuse burns, there will eventually be an epoch when the Type Ia supernovae begin to explode. Using the big new telescopes on the Earth and in space as time machines to probe these distant times, we should also be able to see this onset of Type Ia events. This would be a very exciting result because the time of the onset will give us critical new information on just what type of binary evolution constitutes the fuse. This, in turn, may finally teach us what binary evolution leads to Type Ia.

While we do not expect Type Ia supernovae to be the probe to tell us the cosmological story of the end of the Dark Ages, they have already been used to revolutionize cosmology in an entirely unexpected way. That is the story of the next chapter.

12

Supernovae and the Universe: probing the size, shape, and fate of the Universe with supernovae

12.1 OUR EXPANDING UNIVERSE

Distant galaxies, those so far away that, unlike the Magellanic Clouds, or our sister spiral Andromeda, we do not not sense their individual gravities, are moving away from us. Their speed is nearly proportional to their distance. One can get this effect by setting off a bomb. The faster fragments get further away in a given amount of time so, at a later instant, the faster fragments are further away with a distance that depends linearly on the speed. This, Einstein has taught us, is not how the Universe works. The bomb analogy requires there to be a preexisting space, independent of the matter in the ''bomb,'' into which the bomb explodes. Einstein has taught us, as we explored in Chapter 9, that space is a curving, dynamical entity that is shaped by the gravitating matter within it. Preexisting empty space with a bomb in the center makes no sense mathematically or conceptually in Einstein's Universe.

Rather, Einstein taught us that space itself can expand, carrying the essentially motionless galaxies apart. In this manner, all distant galaxies, those that do not share an immediate gravitational grip, move away from all others. There is no center of the explosion. The fact that we see all distant galaxies moving away from us is an effect created by the uniform expansion of space. With some thought, you can convince yourself that the apparent speed with which galaxies recede depends linearly on the distance, just as observed.

We expected this expansion to be slowing down. This is because the Universe is filled with matter that exerts gravity. For seventy years or so, the challenge to cosmology was to determine whether the expected gravitational deceleration was enough to halt the expansion, or too little, so the Universe would continue to expand, but at an ever

slower rate. One of the major glories of science is that with proper attention to Nature, preconceived notions as powerful as these can be overcome. It worked in this case!

12.2 THE SHAPE OF THE UNIVERSE

To use supernovae or any other technique to measure cosmological distances requires some perspective on what we are trying to accomplish and how we are doing the task. Recall from Chapter 9 the various two-dimensional analogs we have employed to picture curved space. The two-dimensional space around a gravitating object is funnel-like when viewed from the perspective of three dimensions. The two-dimensional analog of the Universe itself, at one moment of time, can be represented as the surface of a sphere, an infinite flat plane, or a saddle extending upward to infinity fore and aft and downward to infinity sideways, as shown in Figure 12.1. These two-dimensional analogs are the embedding diagrams for the Universe. They help picture curvature in three dimensions. These two-dimensional surfaces have no two-dimensional centers, no two-dimensional edges, and no two-dimensional outsides. Likewise, for the most basic conceptions of our real three-dimensional Universe, there is no three-dimensional center, no three-dimensional edge, and no three-dimensional outside.

We have stressed that looking down on a two-dimensional embedding diagram from a higher, three-dimensional perspective is cheating in a sense because there is no way we can look down on our three-dimensional curved space from an ''outside.'' That outside to our three-dimensional Universe, by analogy, would itself have to be a fourth spatial dimension. If there were an observer in that fourth spatial dimension, that observer could see the curvature of our Universe or that around the Earth or around a black hole in much the same way that we can see the curvature of the surface of a sphere. On a more direct and personal level, such an observer would also not be limited to viewing our surfaces, our skin, and our facial features as we do one another. An observer from a hypothetical fourth dimension would also be able simultaneously to see our volume, our guts, and our bones, much as we can see the interior of a circle inscribed on a sheet of paper. This is an amusing perspective, but it is not one of physics. Not until Chapter 14, at least.

Rather, the proper perspective is to recognize that a two-dimensional creature living in any of these curved two-dimensional

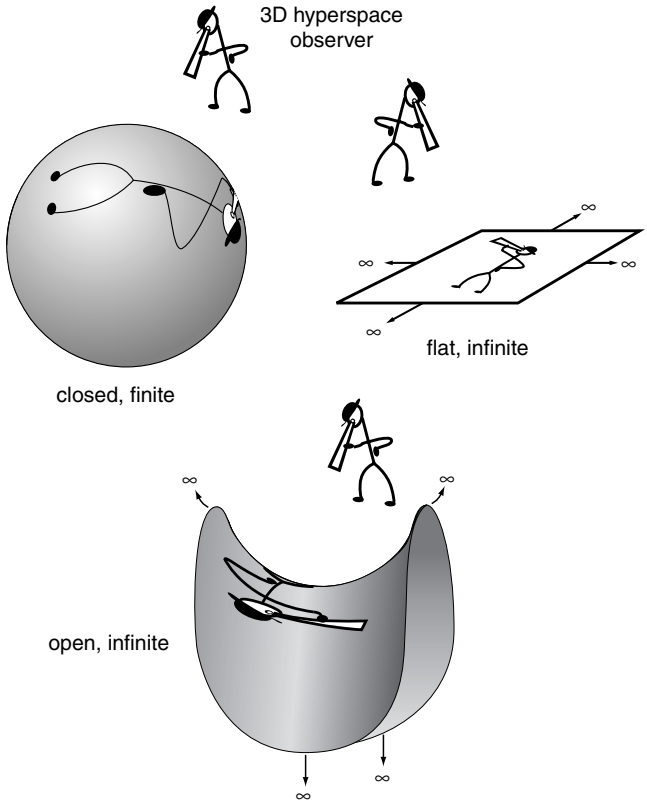


Figure 12.1 Einstein's theory tells us that the Universe must have one of three basic shapes. The two-dimensional analogs (embedding diagrams) for these cases are a spherical surface (a ''closed universe''), a flat plane extending to infinity in all directions (a flat universe), and a saddle shape that also extends to infinity in all directions (an open universe). Two-dimensional astronomers in two-dimensional universes cannot stand outside their universes to see the nature of the curvature the way a three-dimensional hyperspace observer can. Rather, they can do geometry in the context of their own space and determine the shape of their universe. Triangles in flat space will have their interior angles sum to 180 degrees, but the answer will be more than 180 degrees in the spherical universe and less than 180 degrees in the saddle-shaped case. As three-dimensional astronomers in our own three-dimensional Universe, we cannot stand outside of it in hyperspace, but we can do geometry to determine the nature of the Universe we occupy.

spaces of Figure 12.1 could determine that the space curves, and by how much, by doing geometry, by carefully measuring distances and angles. That is now our task! We are three-dimensional supernova observers trapped in our three-dimensional Universe. We must determine the curvature of our three-dimensional space without stepping outside of three dimensions, something we simply cannot do. Fortunately, we do not need to step outside. We just have to be careful with our geometry and our astrophysics.

12.3 THE AGE OF THE UNIVERSE

The Universe we see around us began in what we call the big bang. There are still mysteries surrounding how the Universe came to be. We will touch on some of them in Chapter 14. There is, however, no doubt that the visible Universe arose in a very dense, hot state, and expanded outward. Although the first instants are murky, ordinary particles, protons and electrons formed very quickly, and the Universe was pure hydrogen for a while. The light elements – helium, lithium – formed when this expansion was a few minutes old. When it was a million years old, the matter got sufficiently dilute that the radiation from its heat could stream freely. We see that radiation as the *cosmic background radiation* that comes at us from all directions. This cosmic radiation is red-shifted by the expansion that pulls everything in the Universe away from everything else. We understand this process very well. Further expansion of the Universe brought the agglomeration of matter into galaxies, stars, and planets in ways we are still striving to understand. Continued expansion pulls all the distant galaxies apart. Understanding the expansion of the Universe allows us to measure its age.

As emphasized in Section 12.1, it is important to realize that the big bang did not occur as an explosion in a preexisting space, like a bomb in outer space. Rather space itself expanded, carrying the matter with it. One popular analogy is the behavior of spots on the surface of an expanding balloon. The spots do not move with respect to the rubber surface as the balloon expands, but they become ever farther apart, as shown in Figure 12.2. A three-dimensional analogy is raisins in a rising loaf of bread. The raisins never drift in the dough, but again move ever farther apart until the loaf stops rising. The second analogy is limited and a little deceptive because the loaf of bread is finite. The three-dimensional loaf of bread is surrounded by ordinary three-dimensional space into which it expands, whereas the

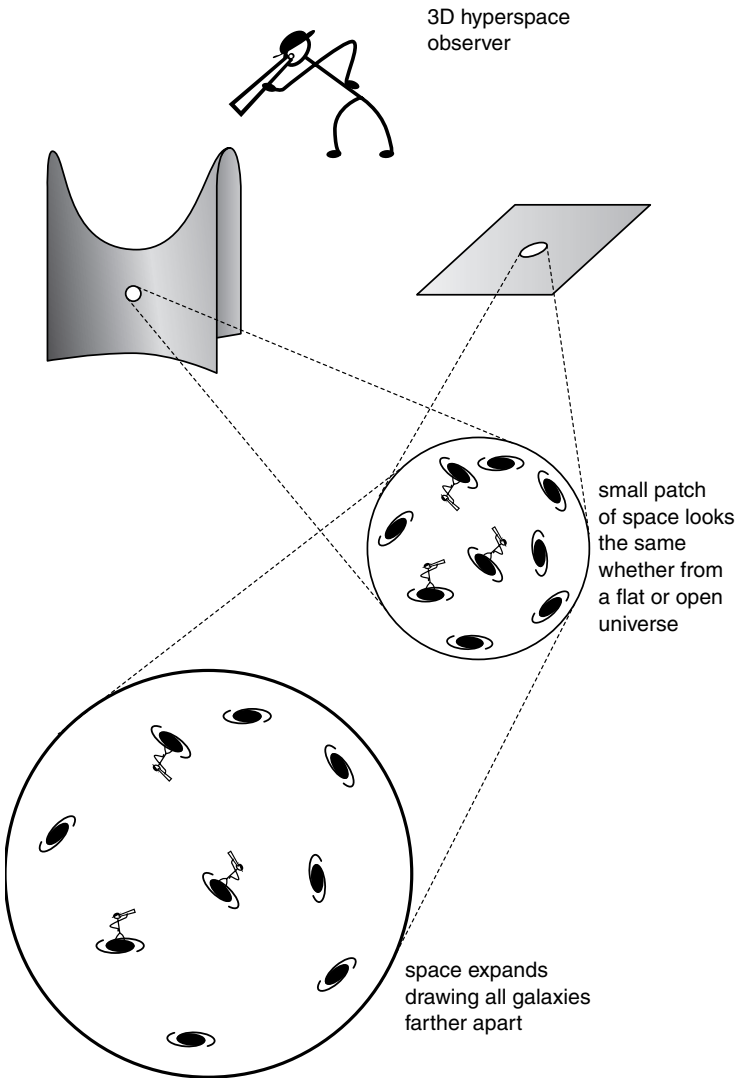


Figure 12.2 A small piece of any two-dimensional universe will appear flat. As the universe expands after its big bang, this piece of the universe will expand, drawing all the galaxies in it farther apart with time. A three-dimensional hyperspace observer could see this expansion, but two-dimensional astronomers resident in the two-dimensional universe could determine the expansion by registering the Doppler red shift as all distant galaxies move apart from all others. As three-dimensional astronomers in our own three-dimensional Universe we cannot stand outside, but we can measure Doppler shifts of distant galaxies and determine how fast the Universe is expanding.

space of the Universe is all-encompassing. The first analogy is limited because it is restricted to two dimensions, but it is more revealing in a way. One can see that the two-dimensional surface of the balloon has no two-dimensional outside, neither the outside as we understand it from our three-dimensional perspective nor what we regard as inside the balloon, which still requires going off into a third-dimensional ''hyperspace'' from the perspective of a two-dimensional creature inhabiting the two-dimensional surface. Likewise, the loaf of bread is perceived to have a center, whereas (ignoring the opening through which one blows) there is no two-dimensional center to the two-dimensional surface of a perfect sphere to which the balloon is an approximation. Unlike the loaf of bread, the balloon shows that if attention is restricted to the confines of the dimensions of the space, two for the surface of the balloon, three for our Universe as we perceive it, there is no center, there is no edge, and there is no outside. These are tricky and fascinating issues, and we will return to them in Chapter 14.

For our current purposes, it is sufficient to picture the expansion of the balloon and its dots or the bread and its raisins to understand how to measure the age of the Universe. The effect of the expansion of the Universe is still much the same as an explosion in preexisting space, even if the concepts are radically different. If you can measure how far away something is from you, say a distant supernova, and determine how fast it is traveling away from you, by measuring its Doppler shift to the red, then you can tell how long it has been traveling to get as far as it has. You get the same answer for every supernova and every galaxy. The faster they move away from us, the more distant they are, but they took the same time to get there, drawn by the expansion of the underlying space.

The parameter that is measured in this way is called the *Hubble constant*, after Edwin Hubble who pioneered this sort of measurement of distances and determined the nature of the Universal expansion. The Hubble constant tells you how fast something will be moving away from you at a given distance. Techniques for measuring the distances to Type Ia supernovae outlined in Section 12.5, and other techniques as well, say that velocity will be about 65 kilometers per second for every million parsecs in distance. The age is related to the inverse of the Hubble constant. Obtaining the age of the Universe from the Hubble constant involves another subtlety because it depends on the curvature of space and the acceleration of the Universe. Neglecting that subtlety for the moment, the corresponding age

of the expanding Universe is roughly just the inverse of the Hubble constant. If a supernova moving at 65 kilometers per second is 1 million parsecs away, it must have been moving away from us for about 10 billion to 15 billion years. If another supernova is moving away from us at 650 kilometers per second and is at 10 million parsecs, then the time for it to get there is just the same, 10 billion to 15 billion years. We get the same answer for every supernova, as we must because we are measuring the same age in every case, the age of the Universe.

The best current estimate is a remarkably precise 13.7 billion years, based on measurement of the cosmic background radiation (Section 12.5 in this chapter). The age estimated in this way does not depend on a detailed determination of the shape of the Universe. Whether our Universe is closed and finite in space and time, or open and infinite, its current age is about 14 billion years.

12.4 THE FATE OF THE UNIVERSE

The game is not over with the measurement of the Hubble constant. It is not enough to measure how old the Universe is. We want to know what will happen to it in the future. Since the days of Hubble, astronomers, particularly the subset known as cosmologists, have been engaged in a grand quest to determine the ''fundamental parameters of the Universe.'' This quest was shaped by Einstein's theory of gravity. The first attempts to apply Einstein's theory to the whole Universe showed that there were three parameters that would describe the whole shebang: the Hubble constant, the overall curvature of the Universe, and the rate at which the Universe is changing its speed of expansion due to the gravitational pull of the matter and energy within it. The issue of curvature is whether the Universe is the three-dimensional analog of the surface of a sphere, a flat plane, or a saddle, as shown in Figures 12.1 and 12.2. Einstein's theory showed that it had to be one of the three. Furthermore, with a key, but reasonable, simplifying assumption that the Universe had the same content, on average, everywhere, the theory showed that the fate is tied to the geometry. If the Universe were sphere-like, it would have a finite life and re-contract to a singularity; if it were flat, it would expand forever, just reaching zero expansion rate at the end of time; and if it were saddle-like, it would expand forever at a finite velocity. We will see later in this chapter and in Chapter 14 that these three parameters may not tell the whole story, but they make up a critical

part of it. Determining these parameters occupied cosmology for most of the twentieth century.

12.5 DARK MATTER

There are various ways of going about measuring the other two parameters in addition to the Hubble constant. The underlying theory requires the constraint of two specific quantities. One is the mass density of the gravitating matter in the Universe at the current epoch. In its simplest guise, this means determining the total mass of all kinds of stuff that has a finite mass and does not move at the speed of light. This mass includes stars, planets, and dust, but it also means any component of the mysterious *dark matter* that consists of particles, no matter how exotic. The photons of light that permeate the Universe also count. They have a mass-equivalent energy ($E = mc^2$), but the gravitational affect of this energy alone is small. The other quantity to be constrained (and ultimately measured) is the value of what is called the *vacuum energy density*. Recall that even a vacuum has an energy associated with it. This energy underlies the emission of Hawking radiation from black holes. The vacuum may have even more subtle properties that would only be manifested when its effects are determined on the scale of the whole Universe.

Dark matter is stuff that gravitates, but emits no detectable light. By detecting the gravitational effects of dark matter on the stars and gas that we can see, we have determined that there is about six times more of this stuff in the Universe than of what we think of as ordinary matter composed of protons, neutrons, and electrons; that is to say, ordinary matter like stars, planets, and people. Most of the mass of this ''ordinary'' matter is in protons and neutrons, the low-mass electrons contribute little to the total, so this component is known generally as the baryonic (Chapter 1) component of the Universe. Baryonic matter gravitates, but also, in proper circumstances, shines. That is how we find it. The dark matter gravitates, that is how we detect it. On the other hand, it must not have an electrical charge, or it would create electromagnetic radiation, light. Nor can it react by means of the strong nuclear force or it would behave far differently. The best guess is that it is composed of some particle, like a neutrino, only different, that reacts only to gravity and the weak nuclear force. There are ongoing experiments to try to detect a particle of dark matter, but there have been no unambiguous results.

One might wonder whether the dark matter could be black holes. The answer is no. The ratio of hydrogen to helium that emerged from the big bang depends on the amount of proton/neutron-like stuff, the amount of baryons. The observed ratio of hydrogen to helium says that there never was enough baryonic matter to account for all the dark matter, whether or not some of the baryons later fell into black holes. The dark matter is something different and something special, and it is the truly ''ordinary'' matter in the Universe; stuff like us is rare to the point of insignificance when it comes to determining the gravitational heft of the Universe. On the other hand, baryons, arranged into people, can think about the Universe, and the dark matter, undoubtedly, cannot.

The dark matter has played an amazing role in the Universe, given that we cannot see it. The *Cosmic Background Explorer* (*COBE*) satellite, launched in 1989, revealed that the cosmic background radiation left over from the big bang is of an exceedingly well-defined temperature, as expected. *COBE* also revealed faint irregularities in the temperature of the radiation from different parts of the sky. The *Wilkinson Microwave Anisotropy Probe*, or *WMAP*, launched in 2001, has provided the best measurement yet of those minute, but systematic fluctuations in the cosmic background radiation. These fluctuations were also expected and even inevitable, given our understanding of the big bang. The big bang grew out of a ''singularity.'' That singularity must have been subject to quantum fluctuations in its properties that are imposed on the expansion of the Universe and hence on the density and temperature of the matter in the Universe (Chapter 14, Section 14.2). Detection of these irregularities at the level of one part in one hundred thousand was another major vindication of the big-bang picture. The original explosion of the big bang left the same incredibly tiny quantum irregularities in the density of the dark matter, slight over-concentrations separated from pockets of ever so slight paucity.

As the Universe expanded, those density irregularities in the dark matter grew. When the Universe became transparent at the beginning of the Dark Ages (Chapter 11, Section 11.8) when it was only a million years old, these slight wrinkles in density deviated from the average by only one part in one hundred thousand. Yet those irregularities continued to grow and became large pockets of high and low density. Those rare protons, neutrons and electrons fell into the high-density pockets of dark matter. The protons, neutrons, and electrons, in turn, formed the stars and galaxies we see scattered

through the Universe. The whole structure of the Universe at which we can marvel now, and on which we depend for our existence, came from these initially tiny wrinkles in the dark matter that, in turn, trace back to the fluctuations of quantum uncertainty at the beginning. This is a truly amazing creation story, one backed by ever more detailed observational confirmation.

12.6 VACUUM ENERGY – EINSTEIN'S BLUNDER THAT WASN'T

There is also a story behind the vacuum energy. The vacuum energy is, in principle, related to the quantum properties of the vacuum, but something like it arises in Einstein's theory of gravity where it is called the *cosmological constant*. Astronomers who write the history of this subject tend to quote Einstein himself in this regard with great glee. Einstein called the cosmological constant ''the greatest blunder of my life.'' The historian's glee and Einstein's self-criticism are probably unfair. The cosmological constant emerges from the mathematics of Einstein in a perfectly natural way (it appears as a constant of integration, for those who know calculus). It is not a question of whether it exists in this mathematical sense. It certainly does. The issue is whether it is zero or not, and whatever its value, including zero, what the physics is that determines that value.

The reason Einstein regarded his treatment of the cosmological constant to be a ''blunder'' is that his first mathematical models for the Universe showed that the Universe would contract or expand. Einstein's intuition told him that the Universe could not possibly do such a radical thing. To render the solution static, Einstein went back to the equations and realized that he had implicitly set the value of the cosmological constant to zero. If he assigned it just the right nonzero value, then the cosmological constant could serve as an extra effect to balance the tendency of the Universe to expand or contract. Shortly afterward, Hubble proved that the Universe *is* expanding. It appeared to Einstein that the cosmological constant was unnecessary, a blunder.

Einstein may have blundered in guessing that the Universe was static, and hence in the value to which he set the cosmological constant, but he did not blunder in introducing the idea. In the long run, it is the latter that is more important, and another tribute to the power of Einstein's theory. The blunder was much less than it is often made out to be. We now see that even the issue of whether the cosmological constant might be exactly zero is not a trivial one, but one

that involves some of the deepest thinking about the Universe. More than that, there are hints that the cosmological constant is not zero, and that definitely raises profound issues of physics and cosmology.

12.7 Type Ia Supernovae as Calibrated Candles and Understood Candles

Apart from their intrinsic interest as star-destroying explosions, supernovae have other uses simply because they are so bright. Their great luminosity means that they are visible across the Universe. More specifically, supernovae are signposts that determine the distances to their host galaxies. Careful measurements of those distances allow astronomers to map out how fast the Universe is expanding and hence how old it is, the curvature of space, and clues to the fate of the Universe. The use of supernovae in this way has expanded extensively in the last decade and the results have been dramatic. Supernovae have provided clues that the Universe may expand forever, and that it is even now in the grip of powerful repulsive forces that accelerate its outward rush.

The use of supernovae to measure distances is based on a simple principle: things farther away look dimmer. Turned around, how dim a supernova appears to be is a measure of how far away it is. The basis for this intuitively reasonable notion is that, when light spreads out from a central source equally in all directions, the locus of the photons emitted at a given time defines a larger and larger surface. The light falling on a detector of a given area, a human eyeball or a telescope, then captures a smaller and smaller fraction of the total the farther away the detector is from the source. The fraction decreases just as the total area into which the radiation floods increases and that goes like the distance squared (the area is $4\pi D^2$, where D is the distance; this turns out to be a profound and important statement, as we will explore in Chapter 14). This means that the apparent brightness of a source of a given total luminosity decreases like the inverse of the square of the distance. In simple terms, the fainter a given kind of object appears, whether it is a porch light, a star, or a supernova, the farther away it must be. If you know how bright the object really is, then you can tell from how bright it apparently is how far away it must be. This gives us a powerful tool for measuring distances. The key is to figure out how bright a given object really is.

Recall that Type Ia supernovae are generally the brightest of all the different types (Chapter 6). This makes them especially good

signposts for measuring large distances. If we knew exactly how bright they were, the task of measuring distances would be rather easy. We would just see how bright a supernova looked in a given telescope and read off the distance. The immediate problem is to determine the intrinsic brightness of a given supernova.

For a long time, there was some reason to believe that Type Ia supernovae were all equally bright. That would have made the task of measuring their distances particularly easy. The jargon for this is that such identical supernovae would represent a *standard candle*. The idea is that, if you have a set of ''candles'' of identical, known brightness, they can serve as a ''standard'' with which to compare other sources of luminosity and to measure distances. In the last decade, we have determined that Type Ia supernovae are not exactly the same, but that the differences are systematic. That allows astronomers to make allowances for the differences between individual Type Ia supernovae.

In particular, astronomers have found that the Type Ia supernovae that are intrinsically brighter decline in brightness more slowly than those that are intrinsically dimmer. We believe that we even have a basic understanding of why this is true. Some variation in the exploding white dwarf causes variation in the amount of radioactive nickel-56 produced in the explosion. The extra energy from radioactive decay does not just make the supernova brighter, it also keeps the expanding matter opaque longer. The radiation takes longer to leak out, giving the slower decay. The trend that relates the brightness of the supernova to the rate of decline from peak light gives the means to determine the brightness of the supernova. One just needs to see how fast the supernova declines, and that tells you how bright it really is. Comparison with how bright it seems in the telescope then gives the distance.

There are two ways of doing this comparison. One uses only the empirical data from the supernova with no attempt at a theoretical understanding. This method requires some comparison with other astronomical objects for which the distances are established in some other way. This calibration sets the overall scale of just how bright a Type Ia supernova with a given rate of decline really is. This must be done for as many supernovae as possible for which the distance is already known (beginning with a dozen or so, with the sample growing steadily). Then the brightness–decline relationship gives the intrinsic brightness and hence the distance from a measurement of the decline rate alone. This technique uses Type Ia not as standard candles but as light sources for which the brightness of each

supernova can be calibrated compared with known sources, hence the phrase ''calibrated candles.''

The other technique to employ Type Ia supernovae to measure distances uses theoretical models of the explosions to determine how bright the supernova must be to produce a given light curve and spectrum. This technique thus attempts to employ ''understanding'' rather than ''calibration'' to provide the necessary information to turn the decline rate into a known intrinsic brightness. This technique thus uses Type Ia supernovae as ''understood candles.''

The first technique, using the Type Ia supernovae as calibrated candles, is only as good as the calibration and the implicit assumptions that underlie the empirical relation between peak brightness and the rate of decline. A key assumption is that the brightness–decline relation is unique. Two supernovae with identical decline rates are assumed to have the same intrinsic peak brightness. The second technique, using Type Ia as understood candles, is only as good as the rather complex underlying theory of the explosion and of the production of luminosity. This method can, in principle, allow for cases where, because of more subtle circumstances, other variables enter and two supernovae with the same decline rate do not have the same peak brightness. The two methods agree rather well. They both give the same age of the Universe (Section 12.3).

12.8 SUPERNOVAE AND COSMOLOGY

Using supernovae to determine the other fundamental parameters of the Universe has been a dream for decades. Many people have worked for a long time to bring it to pass. One of the pioneers, Stirling Colgate of the Los Alamos National Laboratory, estimated that to get the job done when he started working on an automated supernova search telescope in the early 1970s, he would have had to invent seven or eight brand new technologies. These included digital control of the telescope and its instrumentation, electronic detectors to replace photographic plates (Colgate called all this ''dig-as'' for digital astronomy; the tide of the digital revolution has fully enveloped astronomy by now, but the term never caught on), thin lightweight mirrors, time-sharing computers necessary for many people to work cooperatively on the complex computer code required to control the telescope and scan images, and cheap microwave links to allow remote control of the telescope from a distant site. The telephone company wanted $3 million for a microwave link from his telescope

to the headquarters in Socorro, New Mexico. Colgate had only $3000 for the job. He invented a simple method of error checking and installed the link with the funds and equipment he had.

In the 1990s, the technical capability, the development of critical techniques, and the willingness to devote a great deal of hard work came together to bring this dream to fruition, if not in quite the fully automated way Stirling Colgate envisioned. A key development has been the construction of large new telescopes and the special electronic detectors to record faint images over relatively large patches of the sky. Another was the launch, repair, and updating of the *Hubble Space Telescope*.

A team of astronomers at the Lawrence Berkeley Lab of the University of California, now headed by Saul Perlmutter, pioneered the breakthrough in technique. One of the inhibitions of research on supernovae is that their eruption is always a surprise. This means astronomers have to scramble to get data when an explosion occurs. Telescopes are often in the wrong configuration with the wrong instrumentation, the Moon is too bright to see the faint supernova light, or the weather is poor. The result is that we still do not get adequate information on most supernovae.

The LBL team realized that in certain circumstances they could discover supernovae ''on cue.'' They could then schedule procedures in advance to follow them up. These techniques work in precisely the context where one can use the resulting discoveries to do cosmology with supernovae. The trick is that if one looks out to very large distances, a given image obtained with a telescope spans a huge volume containing a huge number of galaxies. It is impossible to predict which of the many galaxies will produce a supernova, but if enough galaxies are in the image, one can be confident that some supernovae will erupt. It turns out that one does not even have to know which specific galaxies are there in advance. If one looks distant enough, there will always be plenty of galaxies and plenty of supernovae. The distances involved, billions of light years, are also just the distances astronomers needed to probe to learn about cosmology.

More particularly, the technique developed by the Berkeley team is to schedule time on a large telescope when the Moon is not up and the sky is dark. They obtain a first image of the sky. They then return and take another image of the same patch of sky two or three weeks later, after the Moon has passed through its bright phase and is no longer a problem. They compare the second image to the first and look for any new lights in the faint images. This is not trivial because

both the galaxies and the supernovae are very faint. Many person-decades have been invested in the computer codes that can automate this process and detect and eliminate flashes of man-made light, cosmic rays that strike the detector, asteroids, and other things that are just a nuisance for this project.

Nothing can be done about bum weather, but these procedures have brought the other factors under control. In addition to the LBL group, another group sprang up in competition, led by Brian Schmidt of Mt. Stromlo Observatory near Canberra and comprising astronomers in Chile, at Harvard, and elsewhere. The results were striking. The two groups of astronomers guaranteed the discovery of roughly a dozen very distant supernovae each time they returned to take the second image. Because they knew far in advance when they would take the second image, they could coordinate the prior scheduling of other telescopes. In this way, they were prepared to get critical spectral and photometric information as soon as they determine the precise location of the new discoveries. Rapid global communication, including the Internet, also played a key role here. Both teams also used the *Hubble Space Telescope* to examine closely the host galaxies after the supernovae have faded. This is a critical step because one must subtract off the light of the host galaxy to get a pure signal from the supernova. Determining the light of the galaxy alone can be done efficiently after the supernova has faded, but not when the supernova first goes off and the light is a complex admixture of supernova and galaxy emission. This technique requires patience. Several months must pass before the supernova has faded sufficiently, and many more months are required for careful calibration and analysis. Using these techniques, the number of supernovae discovered per year has shot up to around 100, most of them at distances that span a good fraction of the observable Universe.

Recall from Section 12.7 that for a given intrinsic luminosity, the apparent brightness of a supernova declines as the inverse of the distance squared. This result, like the ratio of the circumference to the radius of a circle and the sum of the interior angles of a triangle, depends on the curvature of the underlying space. The power of the method of using supernovae is that they can, in principle, give such precise measurements of the distance at such great distances that the effects of the curvature of the space can be gleaned; whether the Universe has a curvature that is the analog of a sphere, a flat plane, or a very big Pringle. The results of these efforts shocked the worlds of astronomy, cosmology, and physics.

12.9 ACCELERATION!

As mentioned earlier, the amount of gravitating mass of all kinds in the Universe affects the curvature of the Universe and tends to slow down the expansion because of the mutual self-gravity of all the mass energy. If the Universe is slowing down, then it was expanding more rapidly in the past. This means that, when we look at supernovae long, long ago and far, far away, with a given Doppler red shift, they will be a little closer and a little brighter than if the Universe had just been coasting at a constant speed, as shown in Figure 12.3. The Universe will also be younger than one would estimate from a given value of the Hubble constant and the assumption that the Universe had always expanded at the current rate.

That is all there is to it if the value of the vacuum energy density is zero. In the language of Einstein's cosmological constant, if the

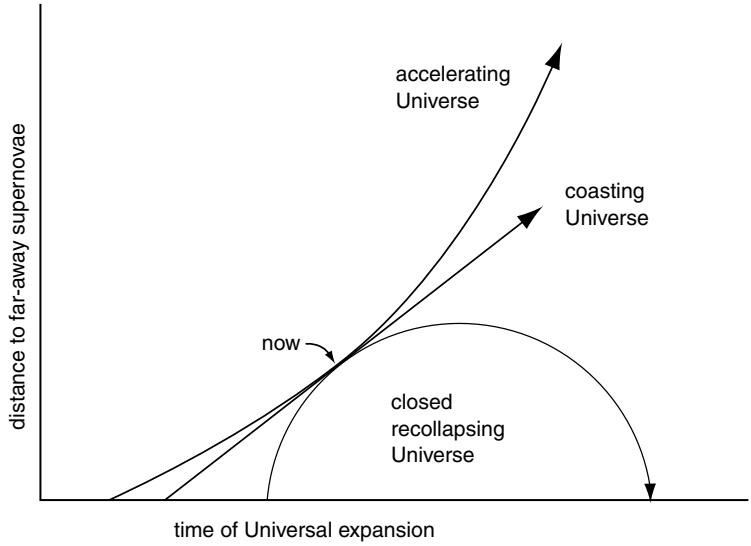


Figure 12.3 The size of the Universe as measured by the distance and Doppler shift of distant supernovae as a function of the age of the Universe. The three lines represent, schematically, the behavior of a closed Universe that is destined to recollapse, a flat Universe that will slowly coast to a halt in infinite time, and an accelerating Universe. The lines all have the same slope at the epoch marked "now." The slope of the lines at that point gives the Hubble constant. The beginning of the lines represent the origin of the big bang for each case. For a given slope of the lines now, the closed Universe gives the shortest time since the big bang, and the accelerating Universe gives the longest.

cosmological constant is not zero, then the effect depends on whether the cosmological constant is positive or negative. If it were negative, the energy of the vacuum would add to the gravity of the matter and slow the expansion even more. If the cosmological constant were positive, the vacuum energy has the effect of an antigravitating, repulsive force causing the Universe to fly apart ever faster as it ages. That sounds like a strange effect, but it is possible within the framework of Einstein's theory, and another measure of why the introduction of the cosmological constant was not a blunder but a very fascinating step. The mass density in the Universe must be positive, but the value of the cosmological constant could be positive or negative or zero and must be determined by observation or theory. If the cosmological constant were positive, it would act in the opposite way to the mass density. A positive cosmological constant would tend to make the Universe accelerate rather than decelerate, as shown in Figure 12.3. This means that a supernova at a given red shift will be a little farther away and a little dimmer than if the Universe had expanded at a constant rate. Likewise, the Universe would be a little older than one would estimate for a given Hubble constant and the assumption of a constant rate of expansion.

Because the effect of the positive mass density and of a positive vacuum energy density work in opposite directions to determine the dynamics of the Universe, the measurement of distances to supernovae tends to constrain the difference between the two effects. Using supernovae alone, the effects cannot be easily separated. Careful measurement of the apparent brightness and red shift of Type Ia supernovae of a given rate of decline and hence intrinsic brightness can, however, constrain the values of the mass density and the vacuum energy. From those constraints and a knowledge of the Hubble constant, the curvature of space and the rate of change of the speed of expansion of the Universe can also be estimated.

The measurement of distant supernovae gave two surprises. One was that there does not seem to be enough gravitating matter, mostly dark matter, to close the Universe. Other astronomical techniques give the same result. They all need to be further refined and considered, but astronomers have basically accepted the result. If this were all there were to it, the suggestion would be that the Universe did not contain enough stuff to close it, and hence it would expand forever.

The other result was even more surprising. Compared to the local sample of supernovae on which the calibration is done, and

compared to a Universe for which the vacuum energy is zero, the distant supernovae were a bit too dim. If this effect is caused purely by cosmological dynamics, then the implication is that the supernovae are a bit farther away for a given red shift. This effect, in turn, can only be explained if the Universe were not decelerating, nor even coasting, but *accelerating* its expansion! This was a striking and unexpected result. It was as if one tossed a ball in the air and rather than having it fall back into your hand, it raced every faster up into the sky! This expansion demands a finite and positive cosmological constant or an equivalent antigravitating effect of the vacuum energy density.

This result was so unexpected and dramatic, that there was an immediate frenzy to question the rather subtle results of the supernova work, not the least by the two teams among themselves and in the spirit of heated competition. The result has been a failure to impeach the result in any appreciable way. The distant supernovae are not materially different than nearby ones; there is no otherwise unexplained dust that could make them appear dimmer.

Even more important, other complementary, but completely independent, techniques have measured the same effect. The most significant is the careful measurement of the tiny fluctuations in the cosmic background radiation when the Universe became transparent at an age of a million years that were also imprinted in the dark matter (Section 12.5). Careful measurement of the fluctuations in the background radiation also constrain the matter density and the vacuum energy density. This technique is a critical complement to the research based on supernovae. The mass density and the cosmological constant tend to work in concert to make the fluctuations grow in amplitude as the Universe ages. The larger the mass density, the stronger the gravity and the faster the fluctuations will tend to grow. On the other hand, if there is a finite and positive vacuum energy density so that the Universe tends to accelerate, then the Universe will be a little older than it otherwise would be, other things being the same, and this gives the fluctuations more time to grow, again making them larger. The result is that the measurement of the fluctuations in the cosmic background will tend to measure the sum of the mass density and the effect of the vacuum energy density, whereas the supernova technique measures the difference between these quantities. Neither technique by itself gives the full picture. If, however, we have independent measures of both the sum and the difference of the mass density and the vacuum energy density, then, in an algebraic sense, we can solve for both unknowns. The incredible characterization

of the fluctuations of the temperature of the cosmic background radiation by *WMAP* has provided a critical source of complementary information. The precise pattern of radiation fluctuations on the sky gives a measurement of the age of the Universe, the amount of gravitating dark matter, and a measure of this antigravitating effect.

Combining supernovae, *WMAP*, and other results, has given rise to a new *concordance model* of the Universe; a Universe composed of about 1/3 dark energy, about 2/3 of this antigravitating influence, and a small smattering of stuff like us for garnish.

12.10 THE SHAPE OF THE UNIVERSE REVISITED

Although the dark matter gravitates and the dark energy antigravitates, they contribute in similar ways to determine the total energy density that in turn determines the curvature of the Universe. What we have learned is that there is enough dark matter and baryonic matter to give about 1/3 of that needed to render the Universe flat. Now we have learned that there is enough dark energy to give about 2/3 of that needed to render the Universe flat. The total $1/3 + 2/3 = 1$. Within current observational uncertainties, the best guess is that our Universe is flat, but accelerating!

This does not mean that our Universe is flat at absolutely each and every point. There are still real stars and black holes and galaxies that curve the space around them. This result means, rather, that when averaged over large volumes containing huge numbers of stars and galaxies, the average curvature is flat in three dimensions; the analogy of a flat plane, a space in which, on average, two initially parallel laser beams will always remain parallel and, if we could do the measurement, all very large triangles would always have their interior angles sum to 180 degrees.

Given the remaining uncertainties, we cannot rule out that the Universe is barely open or barely closed. There is an argument that it must be truly flat to extraordinary accuracy. This argument is based on the inflationary model of the Universe, that very early in its expansion, the Universe underwent a huge expansion in size, stretching all of space to a huge degree. The implication is that, whatever the shape might have been of the Universe before this, curved or flat, the final result would be essentially flat. This is equivalent to saying, for a two-dimensional surface, that if it were sufficiently large, we could not distinguish the curvature of a very large sphere or a very large saddle from a truly flat plane. The Earth

seems flat to casual observation because it is so large compared to the human scale. Imagine the surface of the Earth blown up to the size of the observable Universe. If we entertain that the Universe might be just the teensiest fit open or the teensiest fit closed, its fate is even more uncertain, as we shall see below.

12.11 DARK ENERGY

In a very deep way, we do not know what this antigravitating influence is that is causing the Universe to accelerate. It has been given the name *dark energy*, a term that has caught on broadly, but is just a mask to hide our ignorance of what is going on. What we do know is some things the dark energy is not. It cannot be composed of any ''normal'' particle like protons, neutrons, and electrons, nor even the yet unknown particles of dark matter, because those all gravitate. We also know that the dark energy cannot be accounted for by any currently known theory of physics. The dark energy was not just a surprise to astronomers and cosmologists; it represents a challenge to fundamental physics. That got the attention of physicists and, among other things, has profoundly changed the nature of supernova research. Supernovae are no longer just the plaything of astronomers. Physicists now think supernovae, at least Type Ia, are their experiment with Nature.

The current guesses are that the dark energy is some sort of force field that permeates the vacuum and pushes or antigravitates. It is perhaps useful to picture the force field that arises when you try to push two magnetic north poles together. There is no magnetic ''substance'' in the space between the poles (this experiment would work perfectly well in the vacuum of outer space), but the repulsive force is palpable. The dark energy is not a magnetic field, but this example serves to illustrate that there could be some repulsive field permeating empty space. An important aspect of the dark energy pictured in this way is that, since it is a property of empty space, it does not get diluted as the Universe expands; the expansion just yields more space, more volume, and hence more dark energy. The amount of dark energy per cubic centimeter of empty space could be the same as the Universe expands, whereas the gravitating dark matter would be diluted by the expansion and its gravitating effect would be ever less.

Given this perspective and the assumption that the vacuum energy density is roughly constant, the prediction is that in the young

Universe, the density of matter would dominate and the Universe would be decelerated by the gravity of that matter. The antigravitating effects of the dark energy would also be there, but too small in proportion to have much effect. As the Universe expands, however, the matter is diluted and its gravity becomes weaker. The dark energy remains undiluted, since it is a property of the empty space itself, and eventually there comes an epoch where the effect of the matter becomes less than that of the dark energy and the Universe begins to accelerate under that now dominating influence.

Remarkably, Adam Riess of the Space Telescope Science Institute and his collaborators have used the *Hubble Space Telescope* to measure just such an effect. Even more distant supernovae, observed when the Universe was even younger, show the effects of deceleration. The acceleration of the dark energy took over about 5 billion years ago, when the Universe was about 2/3 of its present age, coincidently about the time our Sun was born. Why the dark energy should be of the value that its effects would be revealed about ''now,'' in cosmological terms, is one of the mysteries associated with the dark energy.

By its mathematical appearance in Einstein's equations, the cosmological constant has a strictly imposed behavior. Inasmuch as the vacuum energy density is positive, the pressure associated with it must be negative and vary in exact proportion to the vacuum energy density. One can think of the negative pressure as the rough equivalent of a tension that pulls inward rather than a normal pressure that pushes outward. The latter gravitates; the former tends to antigravitate. The exact linearity between the pressure and the density if the dark energy is Einstein's cosmological constant gives a precise predicted behavior to the acceleration of the Universe. As far as we can tell from current observations, the Universe is behaving in exactly this way, as if the dark energy were exactly the same at all times and in all places in the Universe.

Even if it proves true that the Universe is behaving as if in the grip of exactly Einstein's cosmological constant, physicists will still want to know why the cosmological constant has the value it does in terms of fundamental quantum fields and forces. Physicists can estimate what the vacuum energy density should be, based on the ideas of the vacuum energy associated with particle creation and annihilation in the vacuum, as invoked to understand Hawking radiation (Chapter 10, Section 10.6). Doing so gives an answer that is wrong by a factor of 10^{120}. My colleague Steve Weinberg calls this ''the biggest

mistake ever made by physicists.'' Physicists faced with this dilemma had long speculated that on the cosmological scale there was some cancellation of the local vacuum energy by some other force field that yielded exactly zero when applied to the whole Universe. Now they are faced with the dilemma that there must be some cancellation that is nearly perfect, but not quite. That is, conceptually, a much more challenging problem, but the one Nature has apparently delivered.

The dark energy thus raises profound questions about what the nature of the vacuum must be that it contains a quantum property that acts as a repulsive, antigravitating force. In the inflationary model of the Universe (Section 12.10), when the Universe was first born, it had a vacuum energy that did act as a repulsive force, an antigravity, that caused a piece of the Universe to expand vastly and rapidly to form the Universe we see today. According to the theory, this energy of the vacuum should have decayed away to zero by now. If the vacuum still has some of this repulsive energy, new theories of the vacuum will have to be developed.

The suggested constancy of the dark energy, though consistent with Einstein's cosmological constant, is itself a deep challenge to physics. In the most general terms, forces in physics have the feature that they will vary in time and space. One early version of such a theory, based on some of the tenets of string theory that we will explore in Chapter 14, was called *quintessence* by Paul Steinhardt of Princeton and his collaborators. The name came from the ancient Greek notion of a ''fifth essence'' (after earth, air, fire, and water), but in this case, it represented the possible behavior of a quantum field theory of the vacuum energy that would manifestly be variable in space and time.

The next big push to understand the dark energy will be to attempt to determine if, despite current indications, it does vary in time and space. Whatever the case, dark energy is neither predicted nor described by current theories of physics. Understanding dark energy is one of the great challenges to modern physics, a challenge that emerged from simply wondering just how far away we might see Type Ia supernovae.

12.12 THE FATE OF THE UNIVERSE REVISITED

This discovery of dark energy has also upset the cosmological game plan to discover the fate of the Universe by measuring the three fundamental parameters of cosmology, as described in Section 12.4. It remains true that determining, directly or indirectly, the Hubble

constant, the matter density, and the vacuum energy density, one can determine the shape of the Universe – open, closed, or flat. With a vacuum energy density, however, that information alone may not reveal the fate of the Universe.

If the Universe has a low gravitating matter density and finite, positive antigravitating vacuum energy density, as current results suggest, so that the tendency to coast outward is even accelerated, then infinite expansion is certainly suggested. In principle, however, a positive cosmological constant could continue to push the Universe into infinite expansion, even if there were enough matter to close it, which there does not appear to be. If this were the fate of the Universe, the current ''best guess,'' the Universe is doomed to expand into a dark oblivion. Galaxies would get so far apart that inhabitants of one could not see another. Stars would die out. Black holes would eventually evaporate by Hawking radiation. Current theories suggest that baryons and leptons, and probably the dark matter, would all decay to photons. The Universe would finally be this accelerating void filled with dim, dilute flashes of light.

If the acceleration of the Universe were slightly stronger than seems the case today, if the antigravitating effect were slightly more sensitive to the vacuum energy density than strictly proportional, then the acceleration itself would accelerate. This might suggest that the Universe would reach its dark oblivion even faster, but the implications are even more dire. If the dark energy behaves in this way, the prediction of Robert Caldwell of Dartmouth and his colleagues is the Universe would be subjected to a *Big Rip*, in which the growing acceleration would overcome the grip of gravity, pulling galaxies apart, then overcome electromagnetic forces, pulling molecules and atoms apart (ouch!), then overcome the strong nuclear force pulling nuclei apart, and then, finally, pulling space–time itself apart. Most physicists consider this possibility so repugnant that they do not take it seriously.

On the other hand, given that the existence of a vacuum energy density raises issues of its origin that we clearly do not know how to answer, we cannot be sure that the cosmological constant is ''constant.'' If this vacuum energy should switch signs and the effective cosmological constant become negative, then, again in principle, the Universe could be doomed to recollapse in a *Big Crunch*, even though it did not contain enough gravitating matter to accomplish that feat on its own. These results have opened up new, if misty, vistas in both cosmology and physics; and this is before we peer into hyperspace.

13

Wormholes, and time machines: tunnels in space and time

13.1 THE MYSTERY OF TIME

''Time is the fire in which we all burn,'' says a character in a *Star Trek* movie. This quote captures the hold that time has on our imaginations. Time, especially the fascinating and philosophically thorny issue of time travel, has been a common topic of science fiction since the classic story of H. G. Wells. The ability to manipulate time remains beyond our grasp, but physicists have conducted a remarkable exploration of time in the last decade that once again brings us to the frontiers of physics.

Separation of time from space has been a part of physical thinking since at least the era of Galileo. The equations physicists use to describe Nature are symmetric in time. They do not differentiate time running forward from time running backward. A movie of dust particles floating in a sunbeam would look essentially the same run forward or backward. If the projectionist ran a regular film backward, you would notice immediately. Where does the difference, the ''arrow of time,'' arise? Why is it that we age from teenage to middle age, but not the other way around? Is that progression immutable?

New approaches to thinking about time came from new thinking about the connectedness of space, and all that came from the desire to make a film that could, among other things, explore issues of science and faith.

13.2 WORMHOLES

This particular attack on time travel arose from a work of science fiction. Carl Sagan envisaged a film that would invoke, among other inventive ideas, rapid travel though the Galaxy. The film stalled, and

Sagan turned to writing a novel first. The novel was a great success, and the film finally moved out of the perdition of production hell. The film, too, was a great success, but Sagan succumbed to a leukemia-related disease before it was released.

In the original draft of his novel, *Contact*, Sagan wrote of a mode of interstellar travel created by an ancient extraterrestrial civilization. He had in mind that his passageway was a black hole where you could fly into the event horizon and emerge – elsewhere. Sagan sent the draft of the book to Kip Thorne, a physicist at Caltech, and one of the world's experts on black holes. Thorne has written his personal version of this story in the book *Black Holes and Time Warps: Einstein's Outrageous Legacy*. Thorne realized that what Sagan proposed would not work. Thorne proposed a solution with both different physics and more imagination!

Einstein's equations for a black hole do describe a passage between two universes or between two parts of the same universe: a structure called an *Einstein–Rosen bridge*, or in more casual language, a *wormhole*. This is yet another phrase invented by the word-master physicist, John A. Wheeler. Black hole experts have known for decades that the apparent wormhole represents only a single moment in time in the two-Universe Schwarzschild solution for a nonrotating black hole described in Chapter 9 (Section 9.8.2). Just before or just after that instant, there is no passage, only the terrible maw of the singularity, waiting to destroy anything that passed into the event horizon. For an intrepid explorer who tried to race at anything less than the speed of light through the wormhole in the instant it opened, the wormhole would snap shut. The explorer would be trapped and pulled into the singularity. In principle, Sagan might have invoked a rotating Kerr black hole wherein there is the possibility of travel through the inner ''normal'' space where tidal forces are less than infinite if one avoids the singularity and thence out into another Universe as described in Chapter 9, Section 9.8.2. That passage might be slammed shut by the blue sheet of infalling star light. In any case, Thorne pursued a different route.

With further reflection, Thorne realized that there might be another approach. Suppose, he reasoned, you were dealing with a very advanced civilization that could engineer anything that was not absolutely forbidden by the laws of physics. Thorne devised a solution that was bizarre and unlikely, but could not be ruled out by the currently known laws of physics. His solution involved what he came to call *exotic matter*.

Ordinary matter has a finite energy and exerts a finite pressure, and creates a normal, pulling, gravitational field. One can envisage mathematically, however, matter that has a negative energy, that exerts a negative pressure, like the tension in a rubber band. For exotic matter, this tension is at such an extreme level that the tension energy is greater than the rest mass energy, $E = mc^2$, of the rubber band. Such material has the property one would label ''antigravity.'' Whereas ordinary matter pushes outward with pressure and pulls inward with gravity, exotic matter pulls inward with its tension and pushes outward with its gravity.

Remarkably, related stuff has become a prominent topic in cosmology, as described in Chapter 12. Cosmologists describe an inflationary stage occurring in the split seconds after the big bang, in which the Universe underwent a rapid expansion that led to its current size and smoothness. The condition that is hypothesized to cause inflation is some form of negative energy field that would have a negative pressure that pushed against normal gravity, resulting in rapid expansion. After a brief interval of hyperexpansion, this field is presumed to decay away, leaving what we regard today as the normal vacuum with its small but nonzero quantum vacuum energy density. Another version of these ideas arises in the context of the current apparently accelerating Universe presented in Chapter 12. If the Universe is accelerating its expansion, there must be something involved other than the gravitating matter in it, some quantum energy of the vacuum that antigravitates, the dark energy. Thorne did not attempt to make the nature of exotic matter explicit. In the most general sense, however, the exotic matter needed to create wormholes would share some of the repulsive properties of the inflationary energy and the dark energy.

Because it was not forbidden by physics, and might even be a part of physics, Thorne speculated that an advanced civilization could slather some of this exotic matter on a mortar board, pick up a trowel, and do something with it. Cleverly applied, the repulsive nature of the antigravity of the exotic cement could hold open an Einstein–Rosen bridge indefinitely! Thorne had discovered, conceptually at least, a way to traverse through hyperspace from one place in the Galaxy to a very distant one in a short time. The result would effectively be faster-than-light travel through a wormhole, just the mechanism that Sagan wanted to further his plot. Sagan adopted Thorne's basic idea and described such a wormhole in the book that went to press. The movie was finally released in the summer of 1997.

Having passed the basic idea on to Sagan, Thorne remained deeply intrigued. He continued to work on the idea with students and together they published a number of papers showing that a proper arrangement of exotic matter could lead to a stable, permanent wormhole.

It is tempting to ask what a wormhole would look like. A wormhole would not necessarily look black, like a black hole, even though the outer structure of their space–time geometries were similar. A black hole has an event horizon from within which nothing can escape. By design, however, you can both see and travel through a wormhole. In its simplest form, a wormhole might appear spherical from the outside, that is, all approaches from all directions would look the same. If you travel through one, you would head straight toward the center of the spherical space. Without changing the direction of your propagation, you would eventually find yourself traveling away from the center, to emerge in another place.

A wormhole is not literally a tunnel in the normal sense with walls you could touch, but from inside a spherical wormhole, the perspective would be tunnel-like. You would be able to see light coming in from the normal space at either end of the wormhole. The view sideways, however, would seem oddly constricted. The space–time of the interior of a wormhole is highly curved. Light heading off in any direction ''perpendicular'' to the radius through the center of the wormhole would travel straight in the local space but end up back where it started, like a line drawn around the surface of a sphere, only in three-dimensional space. If you faced sideways in a wormhole, you could, in principle, see the back of your head. In practice, the light might be distorted and your view very fuzzy. The effect might look like a halo of light around you that differentiated the ''sideways'' direction from that straight through the center of the wormhole. Figure 13.1 shows how it might look to you as you shined a flashlight on the interior of the wormhole.

A common misconception is to confuse the tunnel-like aspects of a wormhole with the funnel-like diagram that physicists use to make a two-dimensional representation, an embedding diagram, of the real three-dimensional space around a black hole or wormhole. In a two-dimensional embedding diagram, a circle in two-dimensional space is the analog of a sphere in three-dimensional space. The real curved space around a three-dimensional wormhole is represented in two dimensions by a stretched two-dimensional space that resembles a funnel, just as it was for a black hole, as discussed in Chapter 9. In

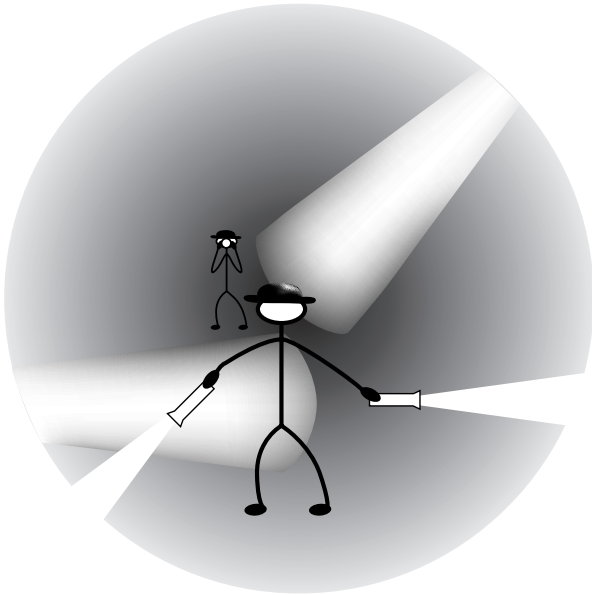


Figure 13.1 A flashlight beamed into a wormhole would shine out the other end, but one aimed sideways would illuminate the back of your head.

this two-dimensional analog, you cannot travel through what we perceive to be the mouth of the funnel. That is a third-dimensional hyperspace in the two-dimensional analog. You have to imagine crawling, spider-like, along the surface of the two-dimensional space to get the true meaning of the nature of that space and some feeling for the three-dimensional reality. A version of this two-dimensional analog of a wormhole is shown in Figure 13.2. The wormhole in Figure 13.2 connects two different parts of an open, saddle-shaped universe. One can also picture a wormhole cutting through a sphere in the two-dimensional analogy of a closed universe. It is more difficult to portray in an illustration, but wormholes can also provide such shortcuts in flat space. If they are properly designed, wormholes can, in principle, yield an arbitrarily short path between arbitrarily distant reaches of normal space in any sort of universe.

Some movies and TV programs have been based on these modern notions of wormholes, but there is still a tendency to confuse the actual tunnel-like nature with the two-dimensional funnel-like analog. In the first *Star Trek* movie, the *Enterprise* is captured in a wormhole when it jumps into warp drive too soon after leaving Earth.

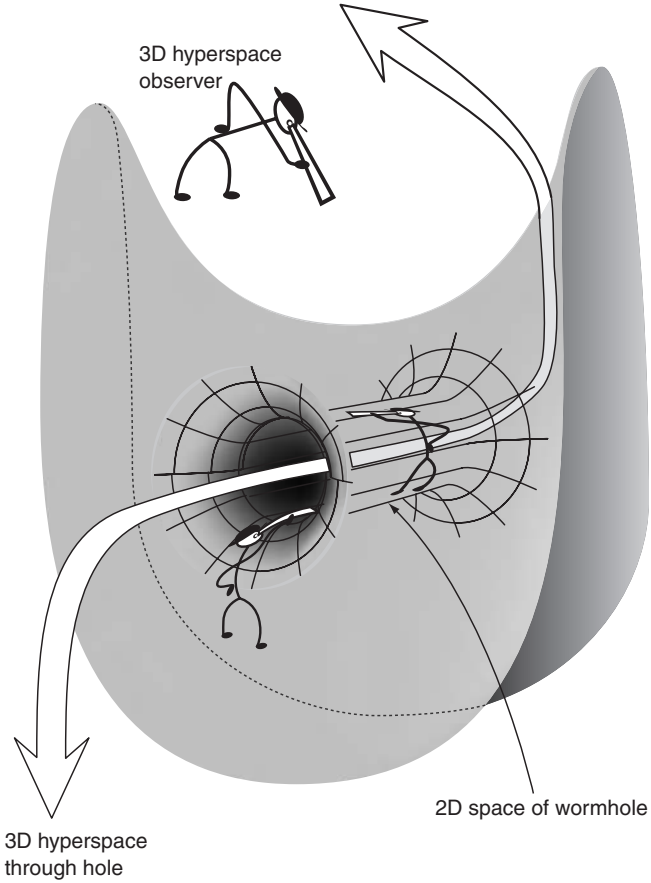


Figure 13.2 A two-dimensional wormhole giving a shortcut through an open saddle-shaped universe. In this representation, the three-dimensional space surrounding the universe and threading the wormhole is a hyperspace that two-dimensional residents of the universe could not perceive. A two-dimensional denizen of the two-dimensional universe could approach this wormhole from any direction in 360 degrees and pass through the wormhole along the two-dimensional surface to emerge on the other side of the universe. An astronomer near the ''mouth'' of the wormhole could see a colleague within the wormhole, and vice versa. The astronomer within the wormhole could travel ''straight'' on a path at right angles to the way in or out and end up back where he started.

You can see stars through the sides of the wormhole. That is definitely wrong. Light from stars could come in the end of the wormhole the *Enterprise* entered, or it could come in through the opposite end toward which the ship is headed. Inside the wormhole, however, light is trapped by the severe curvature of the space. There is no literal tunnel wall; hence, Kirk and his crew cannot look out ''sideways'' through it.

The TV series *Babylon 5* features a ''constructed'' wormhole, but its whirlpool-like nature is more reminiscent of the two-dimensional analogy than the proper manifestation in real space. In *Deep Space 9*, the wormhole can be approached from any direction and the tunnel-like interior is as close to ''reality'' as one can expect from graphic designers appealing to a TV audience. *Sliders* also does a pretty good job of capturing the spirit that the wormhole is basically spherical so the characters can enter and exit anywhere in three dimensions. The film *Stargate* and the TV program based on it show the wormhole portal to be a single flat, circular sheet. The characters enter and exit from only one side. That is Alice's looking glass, perhaps, but not well rooted in this particular bit of science.

The classic wormhole is that in the movie *2001: A Space Odyssey*. The fact that the monolith orbiting Jupiter is a wormhole is a bit obscure, but that is what it is. In that film, the exterior of the wormhole is three-dimensional, but it is a flattened rectangle. Matt Visser of Washington University of St. Louis designed a wormhole that looks much like that, with the exotic matter confined to struts along the boundaries of the rectangular body. In the movie version of *Contact*, the heroine is thrust into a wormhole by an alien-designed machine that opens the portal to the wormhole. The tunnel-like aspects are portrayed reasonably realistically, and there is an attempt to invoke the other amazing property of wormholes, the distortion of time.

13.3 TIME MACHINES

If exotic matter, antigravity, and superluminal travel were not enough, there is even more to the wormhole story, and time is its essence. As they worked on the nature of wormholes, Thorne and his coworkers realized to their amazement that wormholes must also function as time machines. In this phase, Thorne was joined by Igor Novikov, then of Moscow, now at the University of Copenhagen, and his colleagues. A key aspect of the next stage of their thinking is what has been called the ''twin paradox.''

This conundrum arises already in the context of Einstein's special relativity. Einstein's theory shows unequivocally that a pair of twins moving at some velocity with respect to one another will each measure the other to be aging more slowly. The twin paradox apparently arises when one of the twins rockets out into space and then returns while the other remains at home. The motion is relative, but the twins cannot each be younger than the other. Is one twin younger, and, if so, which one? The resolution to the paradox is that the one that traveled will be younger. That traveler must have experienced a force, an acceleration, upon turning around, and that makes all the difference. That is the answer when carefully analyzed, with special relativity accounting for the acceleration that the traveling twin felt and the stay-at-home did not.

Thorne realized that you could do this experiment, again conceptually at least, with the two ends of a wormhole. Grab one end (gravitationally), and rocket it out and back. It will be absolutely younger than the end that was not accelerated. Novikov realized that the same result will arise by putting one end of a wormhole in empty space and the other near a gravitating body. General relativity says that time will flow more slowly in the gravity well. The end of the wormhole deep in the gravity would be younger than the end in deep space.

In either of these arrangements, you have a time machine! You can walk into one end of the wormhole and emerge in an earlier era. If you walk to the first end of the wormhole though the exterior space, time passes, and you age normally. You could meet your younger self before you entered the hole! Because this is science, not fiction, there are limits. You cannot exit before the wormhole time machine was created, so you cannot travel arbitrarily far back in time.

Time travel, including that invited by wormhole time machines, leads to another classic paradox: the ''grandfather paradox.'' The idea is that a time traveler can go back in time and kill her grandfather before her mother, or she, was born, thus the paradox. Thorne thinks this is too paternalistic and invites the time traveler to kill her mother, giving rise to the ''matricide paradox.'' Novikov argues for leaving out the middleman. Kill your younger self in a time-contorted suicide. The result is the same. The time traveler could not have existed in the first place to commit any of the hypothesis-testing crimes.

All these examples invoke people and death to make them graphic, but people raise the issue of consciousness and free will and

those issues are messy for a physicist. Joe Polchinski, then of the University of Texas, now at the University of California at Santa Barbara, invented a simple mechanical paradox. Physicists often refer to ''pool ball'' physics, meaning the process of reducing a problem to something as visceral as pool balls bouncing off one another so that the physics – conservation of momentum, for instance – can be easily visualized. Polchinski adopted this metaphor to present the ''pool-ball crisis.'' In this thought experiment, a pool ball rolls into one end of a time machine. It comes out the other end in the past. It smacks its earlier incarnation, deflecting it so that it does not enter the wormhole. The paradox is the same in principle. How does the pool ball ''get there'' in the future if it never entered in the past? Polchinski argued that this simple setup showed that time machines could not exist and no kindly grandfathers or warm, loving mothers were threatened in the least.

The time-machine explorers did not buy it. The flaw in this argument, according to Novikov, is that the original pool ball is pictured as rolling unimpeded into the wormhole, and the collision is only considered when the ball emerges to collide with itself. That is not self-inconsistent. The original pool ball must be involved in the collision as it first rolls toward the opening of the wormhole. Physics must be self-consistent, Novikov insists, even in the presence of time travel. Novikov and his colleagues have carefully studied the pool-ball crisis and have shown that it cannot arise. They have looked at every conceivable interaction. Pool balls can miss, or they can strike a glancing blow, but they can never undergo a hard collision that leads to a paradox. Novikov's group even explored an exploding pool ball, one fragment of which manages to enter the wormhole, come back in time, and hit the exploding pool ball, causing it to blow up, rendering the whole experiment self-consistent. The notion that physics can incorporate time machines in this way is called, in some circles, the *Novikov consistency conjecture*.

Now we can reintroduce people. According to the consistency conjecture, any complex interpersonal interactions must work themselves out self-consistently so that there is no paradox. That is the resolution. This means, if taken literally, that if time machines exist, there can be no free will. You cannot will yourself to kill your younger self if you travel back in time. You can coexist, take yourself out for a beer, celebrate your birthday together, but somehow circumstances will dictate that you cannot behave in a way that will lead to a paradox in time. Novikov supports this point of view with

another argument: physics already restricts your free will every day. You may will yourself to fly or to walk through a concrete wall, but gravity and condensed-matter physics dictate that you cannot. Why, Novikov asks, is the consistency restriction placed on a time traveler any different?

What about the converse? If personal free will exists, does that mean time machines cannot? That question is unresolved. Physics cannot treat the issue of free will, but it may yet address the question of whether time machines can truly exist. The consistency conjecture does say that certain time-travel plots are allowed and others are not. In particular, the consistency conjecture would say that one cannot use time travel to change the future, the basic premise behind both the *Back to the Future* and the *Terminator* movies. Loops in time are allowed, but according to the consistency conjecture, the future is as fixed as the past and cannot be affected by an act of will or any other physical act.

Another way to resolve these issues is to say time somehow ''forks off'' at the moment of a paradox. The ''many worlds'' idea arose in another context as a way to understand some of the conundrums of the quantum theory, how a wave of probability can be turned into an experimentally measured certainty. In the context of time travel, the idea is that in one time-prong a time traveler lives on, even having killed her younger self. In this view, her younger self lives in the old time prong, but not in the current one. It is not clear that this resolves the origin of the memories of the time traveler of having been younger and having later wielded the knife.

Philosophical questions aside, the issues involved in time-machine research are right at the frontier of modern physics. We have known since the advent of quantum mechanics that the vacuum does not have zero energy. Having a specific energy, even zero, would violate the Heisenberg uncertainty principle. Rather, the vacuum is riven with fluctuations, particles of light, matter, and antimatter that constantly form and annihilate. The wormhole mouths, like the space near the event horizon of a black hole, will be endowed with these vacuum fluctuations. In the case of a black hole, these fluctuations lead to Hawking radiation and to the evaporation of the black hole (Chapter 9, Section 9.6). For a wormhole, the issue is, if anything, even deeper. The vacuum fluctuations can travel in normal space to the opposite mouth of the wormhole, zip inside, and emerge in the past just at the time they left. If that were to happen, there would be twice as much energy in vacuum fluctuations. The cycle might repeat

indefinitely and build up an infinite energy density, completely altering gravity and space and thus sealing off the wormhole or preventing it from having existed in the first place.

To properly address this issue, a full theory of quantum gravity is required. This theory must incorporate both violently curved space–time and the probabilistic nature of the quantum theory. Such a theory is the holy grail of modern physics. This theory is needed to understand the singularity of the big bang and that inside a black hole. There are great conceptual problems facing the development of such a theory of space–time that applies on scales where time and space themselves are uncertain in a quantum manner, where up and down and before and after lose their meaning. Only with the development of this ultimate theory of everything will we really know whether time machines are conceptually possible. Attempts to construct such a theory are the topic of the next chapter.

14

Beyond: the frontiers

Trispatiocentrism

''Egocentric.'' ''Enthnocentric.'' A variety of words in the English language describe the tendency of people to get locked into a limited perspective. ''Anthropocentric'' is a favorite word in some circles of astronomy. It describes the tendency of scientists, as well as *Star Trek* writers, to conjure up alien life forms that are fundamentally similar to us, not just physically, but emotionally and socially, with our motivations, drives, and dreams. The *anthropic principle* – that the Universe is as it is because we exist – is a related idea. In the never-ending battle to expand our perspectives, I write this to call attention to the existence of another limited, rarely questioned, viewpoint that affects us all: *trispatiocentrism*. Trispatiocentrism is the attitude that the ''normal'' three-dimensional space of our direct perceptions is all there is and all that matters.

This word arose in my substantial writing-component course at the University of Texas in Austin. We were exploring the nature of space and time with a particular emphasis on spaces of various dimensions. I wanted a word to connote the notion that our three-dimensional world view carries with it unrecognized restrictions. I came up with ''trispatiocentric'' and its obvious variations.

There is a serious scientific side to this. Some understanding of curved space is needed to picture how Einstein's theory of gravity works. To illustrate the basic ideas, gravitational physicists often have recourse to examples of curved two-dimensional spaces, the surfaces of spheres or of saddles or of doughnuts. In

these examples, our familiar three-dimensional space surrounds the surface so that we can easily envisage the curvature. The trick is to try to perceive what the corresponding curvature of our own three-dimensional space is like. The goal is to understand the arcanae of Einstein's theory: black holes, wormholes, time machines and the ramifications of string theory. In this context, it is quite natural for a logical, if naive, mind to ask: if the surface of a sphere curves in a three-dimensional space, then must our three-dimensional space curve in some four-dimensional space?

For the non-naive, these issues arise at the forefront of modern physics, the attempt to construct a ''theory of everything.'' This theory will allow us to understand the raging singularities predicted to be at the centers of black holes and from which the Universe was born. Singularities represent the place where our current concepts of space and time, indeed all of physics, break down. The most successful current attempts to develop a new understanding of space and time are based on ''string theory,'' where, to be self-consistent, the ''strings'' that constitute the fundamental elements of nature exist in a space of ten dimensions. Thus these developments have led physicists to ponder higher dimensions, perhaps ones so tightly packed we cannot perceive them directly. They speak in terms of surfaces or membranes in a space of *p*-dimensions and call them ''*p*-branes.'' Alas, I cannot resist pointing out that all this is not for pea brains like me. It is, however, the stuff that will push back the frontiers of knowledge and along the way help to resolve famous wagers made by Stephen Hawking concerning the nature of space and time.

In our course, we read the classic old tale *Flatland* by Edwin Abbott. Here we meet the Monarch of Line Land who, in blissful ignorance, suffers his monospatiocentrism. The hero of *Flatland* is a simple square who is ripped, to his ultimate chagrin, from his bispatiocentric world view by a visitor from a three-dimensional universe we would recognize.

Abbott, Einstein, and the work of string theorists would have us ponder a fundamental verity. We are gripped in a trispatiocentrism we rarely stop to recognize and even more rarely take the time to ponder. Why does our familiar space have three dimensions, no more, no less? Is the notion that this space is natural or even unique as archaic and limited as the notions that the Sun goes around the Earth or that the Solar System is in the center of the Universe? Is Heaven not ''up'' in a literal sense but in a higher

dimension we cannot perceive? If so, what of Hell? When Captains Kirk, Picard, or Janeway are transported to a different dimension, why is it always so boringly and trispatiocentrically of a familiar number of dimensions? We are trapped in this three-dimensional world of our direct perceptions and scarcely know it.

Is it possible that space can be prized open with ''exotic matter'' leading to wormholes that reconnect time and space? Are the ten dimensional spaces of string theory the first hint of the ''subspace'' of *Star Trek*? The work of physicists on the vanguard of knowledge provides the first glimpses of what may exist beyond or without.

The hero of *Flatland* was imprisoned for attempting to challenge the bispatiocentrism of his peers. My students seem to have the same dismal expectations for any departures from societal norms. The stories they wrote for class of other-dimensional worlds suggested that society is likely to find unwelcome any assault on cherished ''centrisms.'' With their stories as a guide, I should expect with this contribution to be summarily institutionalized, incarcerated, or executed. Nevertheless, the truth must be exposed.

Citizens of this three-dimensional Universe unite! You have nothing to lose but your branes!

14.1 QUANTUM GRAVITY

The search for quantum gravity, a theory that unites both the aspects of uncertainty from the quantum theory and the aspects of curved space from general relativity, a theory of everything, is the current frontier of physics. Black holes are at the center of the action. The current contender for this intellectual prize is what is called by physicists, *string theory*. The basic notion is that the fundamental entities of the Universe are not particles, dots of matter, but strings of energy, entities with one-dimensional extent.

That seems like a simple, maybe even unnecessary, generalization of our standard picture of elementary particles, electrons, neutrinos, protons, neutrons, and quarks. The doors that have been opened by this change in viewpoint are, however, wondrous.

For perspective, let us go back to the theory of Newton. Newton gave a rigorous mathematical framework in which to understand gravity and much else of basic physics, how things move under the

imposition of forces. Newton's law of gravity was based on the concept of a force between two objects. It was encapsulated in a simple formula that said that the force of gravity was proportional to the mass of two gravitating objects and inversely proportional to the square of the distance between them. This prescription was immensely successful. It is still used with great effect in most of astronomy to predict the motions of stellar objects from asteroids to the swirling of majestic galaxies. It is used to guide man-made satellites and rockets. We know now, however, that Newton's theory is wrong. It is wrong in concept and wrong in application.

A hint of the conceptual problem with Newton's theory comes by examining the law of gravity (see also Chapter 9, Section 9.1). Newton's version of this law tells of the dependence on the masses of the gravitating objects and the distance between them but is mute on the dependence on time. Newton knew that the speed of light was a speed limit, yet his theory demanded communication of information, the strength of gravity, at infinite speed. Another clue to problems with Newton's theory is that if you reduce the distance between two objects to zero the gravitational force between them is infinite. If one looks sufficiently closely at Newton, those errors exist. The ultimate test is comparison of theory with observation and experiment. Newton is exceedingly successful in many applications but fails in some. Newton's theory gives the wrong answer to carefully posed experimental situations.

Einstein's theory of gravity, general relativity, was based on an incredibly simple and elegant idea: that physics should behave the same, independent of the motion of the experimenter. The earlier version of this idea, Einstein's special theory of relativity, arose from the young Einstein asking another simple question: what would an electromagnetic wave look like if an observer moved along with it at the speed of light? To answer that question, to show that the observer could not move at the speed of light, Einstein had to show that the speed of light was the same, *independent of the motion of the observer*. This result, one of the deeply true aspects of physics, remains one of the most incredible of human insights. Einstein also proved with his special theory that the lengths and times measured by an observer depended on how the measured object was moving, not in an absolute sense, but moving with respect to the observer.

Einstein's general theory took another step and asked about observers not in uniform motion, the subject of special relativity, but observers in accelerated motion. He realized that an observer freely

falling in a gravitational field would measure physical effects and find them identical to an observer moving at uniform speed far from any gravitating object, but that an observer in an accelerating frame would feel exactly the same as one feeling the effects of gravity. This notion has been enshrined as Einstein's *equivalence principle*, that an acceleration gives the same effects as being at rest in a gravitational field. If you sat in a chair in a lecture hall that accelerated at a uniform rate, the floor would push on your feet and the seat would push on your rear end, exactly the same forces you feel sitting in your chair reading this book. The equivalence principle is elegantly simple to state. To put it into a self-consistent mathematical framework, Einstein found that he had to introduce the notions of curved space and a complex set of tensor equations to describe it. Our sense of the nature of space has never been the same.

Einstein's theory of gravity has passed every test put to it. It gets the right answers for the shift of Mercury's orbit and the deflection of light, and has passed numerous other tests to the limit of our current ability to devise those tests. This makes general relativity a better theory of gravity than Newton's. General relativity also becomes identical to Newton's theory, mathematically, and hence in its precise predictions, when gravity is weak, distances are large, and motion is small. It must do so in order to reproduce Newton's manifest success of predictability in those regimes. To accomplish this great success, Einstein had to abandon not just the mathematical structure adopted by Newton, but the fundamental concept behind gravity. Einstein abandoned the notion of a ''force'' of gravity, and replaced it with the notion of curved space and warped time. Space is curved, and that tells matter how to move, how to orbit, how to fall. Gravity is geometry, the geometry of curved space. The change in conception wrought by Einstein was deeply profound. General relativity is, however, wrong.

So far we only know that general relativity is wrong because of conceptual problems. We have not been able to devise a test sensitive enough to display the fact. The conceptual problem is in the prediction of the singularity. General relativity predicts that, right at the center of a black hole, a region of infinitesimal size, with infinite space–time curvature and infinite tidal forces, must form. Essentially identical conditions are predicted at the beginning of the Universe, a singularity from which all arose. Those predictions of infinity are the undoing of general relativity. To be specific, the prediction of singularities flatly violates the fundamental tenet of quantum theory, the

uncertainty principle (see Chapter 1, Section 1.2.4), which states that one cannot specify the position of anything exactly, including a ''singularity.'' As a predictive theory, general relativity is marvelous in the regimes where it works, just as Newton's theory was in its own regime. General relativity does everything that Newton's theory could do and more, including predictions of black holes and event horizons. Deep in its heart, however, general relativity contradicts quantum theory.

On the other side, quantum theory basically assumes that the underlying space in which particles are rendered uncertain is flat, or at least, not too curved. General relativity predicts conditions not as extreme as the singularity where its results should still be valid, but where the curvature of space is ''smaller'' than the size of a quantum-smeared particle. In this sense, the quantum theory breaks down at conditions where general relativity still rules. Each of these great theories of twentieth-century physics contradict one another at a fundamental level. We need a twenty-first-century theory to encompass and embrace both, but that also works where they fail.

A theory of everything must take its place in this hierarchy. It must incorporate everything that Newton accurately predicted. It must also incorporate everything that Einstein subsumed so elegantly. Then it must also answer the question: what is this amazing thing called a singularity? The theory must tell us what happens to space and time under conditions where quantum uncertainty dictates that the very notions of ''front,'' ''back,'' ''here,'' ''there,'' ''before,'' and ''after'' lose their meaning. There must be space without space as we know it and time without time as we know it. Is there any wonder that physicists since Einstein have labored against immense conceptual problems in attempting to cross this barrier?

14.2 WHEN THE SINGULARITY IS NOT A SINGULARITY

The singularity of Einstein's theory cannot exist. Something else must happen to space and time ''there.'' In the absence of the full development of quantum gravity, physicists are left to grope. When physicists grope, startling ideas emerge.

We know the scale on which Einstein's theory must break down, even if we do not fully understand what must replace it. This scale can be estimated from the simple idea of asking about the conditions where quantum uncertainty must be as important as the space–time curvature of gravity. The fundamental constants of

quantum gravity are the strength of gravity as measured by Newton's constant from the world of the large, the degree of quantum uncertainty as measured by Planck's constant from the world of the small, and Nature's speed limit, the speed of light from the world of the very fast. With values for these constants of Nature in some set of units, English or metric, it does not matter, one can estimate the scale where Einstein's theory, and ordinary quantum theory, fail. This scale, of length, time, density, is called the *Planck scale*. Newton's constant has units of length cubed, time squared, and the inverse of mass. Planck's constant has units of mass, length squared, and the inverse of time. The speed of light has units of length over time. There is only one way we can combine these three fundamental constants with their individual units to produce a quantity of only length, only one other way to produce a time, and only a single third way to produce a mass. This exercise is a simple one of sorting out units, but it has profound implications because the building blocks are the fundamental constants that tell us how space curves, the degree of quantum uncertainty, and how fast things can move. Their combination implicitly tells us where space gets so curved that a quantum wave cannot exist and simultaneously where quantum uncertainty is so large that speaking of a given curvature makes no sense. We learn the conditions where the two great theories of twentieth-century physics butt heads and contradict one another, the conditions that call for a new theory of physics.

The resulting value of the length, the Planck length, is about 10^{-33} centimeters. This is an incredibly small value, much smaller than the size of a proton, but it is not zero! This is roughly how large the singularity must be. At this level, space and time break down into something else, and Einstein's prediction of a singularity goes awry. The corresponding Planck time is about 10^{-43} seconds. This is again an incredibly short time, but not zero. Time as we know it probably does not exist at shorter intervals, so that asking what happened when the Universe was younger than 10^{-43} seconds or before the big bang may not make sense, at least not in the traditional way. The Planck mass is about 10^{-5} grams. This is a small number, but not incredibly small. It is vastly bigger than any elementary particle we know. One can also work out the Planck density, the Planck mass divided by the cube of the Planck length. The answer is about 10^{93} grams per cubic centimeter. This is a gigantic density, but it is not infinite. In some average way, this must be the density of a singularity, the density from which our Universe expanded in the big

bang, the density to which all is compressed in the centers of black holes.

One way to think about the singularity is as a bubbling sea of Planck masses, each a Planck length in extent winking in and out of existence for intervals of a Planck time. This quantum-bubbling mess has been called a *quantum foam*, another bit of etymological brilliance from John A. Wheeler. This term is a picturesque name intended to describe something we do not understand, yet to capture the flavor of the idea that it is not ordinary space and time. In the quantum foam, one could not speak of front and back because space itself would be so quantum-uncertain that such concepts are invalid. The same is true for the ideas of before and after, with time also a quantum froth.

Even in the absence of a full theory, if we picture the singularity not as a point of zero size and infinite density but a dollop of quantum foam, then other ideas begin to emerge. The Universe was not born from a point of infinite density but emerged as a bubble of ordinary space and time from this quantum foam. This bubble was highly energetic and expanded to become everything we see. As we discussed in Chapter 12, the expansion is pictured in the sense that all points of space move away from all other points of space, not an explosion of stuff into a preexisting three-dimensional space. Also, as three-dimensional physicists, we do not have to address the issue of what the three-dimensional Universe is expanding into, as much as that question seems to intrude.

That the Universe emerges from the quantum foam already gives some predictability to the nature of the Universe. There must have been quantum fluctuations in the density and temperature of the very young, hot big bang as it emerged from the quantum foam 10^{-33} centimeters across and 10^{-43} seconds old. These unavoidable fluctuations can be calculated from the quantum theory with some assumptions, and they later cause the tiny irregularities in temperature detected by *COBE* and *WMAP* that after billions of years grow to form all the structure we see – stars, galaxies, clusters of galaxies (Chapter 12, Section 12.5).

The notion of a quantum foam also plays a role in the thinking about wormholes (Chapter 13) and shows again that we cannot pursue the physics of wormholes without a theory of the quantum foam, a theory of the singularity, a quantum-gravity theory of everything. One way to picture the quantum foam is as quantum-connected fragments of space and time, connecting different places and different times willy-nilly in a probabilistic way. These connections, although dominated

by quantum uncertainty, are essentially tiny quantum wormholes. One can imagine making a wormhole by taking a little quantum loop of space and time and blowing it up to become a wormhole big enough to travel through.

Another way to imagine making a wormhole leads to similar issues of the quantum nature of space and time. If you start from ordinary space and want to make a black hole, you have to stretch and distort the space, but you do not have to rip or tear it (at least not until you get to that nasty singularity). That is not true for a wormhole. To make a wormhole, you have to tear and reconnect space. You have to change not just the curvature of space but its connectedness, its topology. If you think about it, a tea cup with a nice handle and a donut are the same basic thing in terms of how they are connected. They are both solid objects with one hole through them. You could make both from the same lump of clay by just molding a side of the donut shape to be the cup and shape the clay around the hole to be the handle. You would not have to tear the clay or reattach it at any point. You cannot, however, make a solid lump of clay into either a donut or tea cup without tearing a hole in the clay.

Think of how you could connect space on a large scale to make a wormhole. It helps to imagine this in two dimensions. Picture a balloon. Push two fingers inward from opposite sides until your fingers almost touch, separated only by the thin rubber of the balloon. You have almost made a wormhole. If the connection could be made there in the center of the balloon, there would be a way to travel on a shortcut through the center of the balloon, rather than taking the long way around on the surface. The balloon serves as a two-dimensional analog of our three-dimensional space, so all motion is confined to the rubber of the surface. Now think of what you need to do to make the connection between your fingers. You would have to cut the rubber and attach the ends of the two cones; but cutting the rubber is the analogy of cutting the very fabric of space. That would be the issue in our real three-dimensional space in order to make a three-dimensional wormhole. The cutting and reattaching of space would amount to, at least temporarily, introducing an end to space, a singularity, before the reattachment is made. To make a wormhole or a wormhole time machine in this way, we have to bring in the operation of introducing a tear in space–time, a tear in the quantum foam. We will not know whether such an operation even makes sense until we have a theory of quantum gravity that tells how space and time behave if such a rent is threatened. Once again, we cannot think

constructively about wormholes or time machines without a theory of quantum gravity to guide us.

If the Universe were born not from a singularity of infinite density, but from a spot of quantum foam, then the inverse is true. When a star collapses to make a black hole, the matter of the star does not disappear into a singularity of zero volume but is crushed into a froth of quantum foam of a Planck density. One of the most dramatic ideas to emerge in the last few years was to ask, if a black hole leads back to the quantum foam from which the Universe arose, why cannot the cycle repeat? This idea was first put forth by Andre Linde, a Russian physicist, now at Stanford University. Linde was striving for some new idea to present at a conference to which he had been invited. He was ill and contemplating skipping the meeting, when this notion came to him. He worked out the basic mathematical and physical picture and presented it at the meeting.

The idea is that the quantum foam that forms at the center of the black hole is identical to that from which the big bang, our whole Universe, arose. This means, Linde argued, that a new universe can arise from the quantum foam of the black hole. The dramatic implication is that the chain could be endless. A universe forms; it expands to form stars. Some of the stars collapse to make black holes. From the singularities of those black holes, new universes can be born elsewhere in hyperspace. Here, perhaps, is a way to answer the question of what came before and what comes after the big bang – endless universes forming endless black holes.

Like many grand ideas of physics, this one must be poked and pummeled and analyzed. How do you prove such a startling conjecture? We cannot travel to other universes to see how they work. We are stuck in this one but empowered with our imaginations and our mathematics and physics. Physicists are already at work generalizing the old cosmologies to see how these ideas could fit in. The easiest way to picture a bubble being blown in the quantum foam to become our Universe is to picture a literal bubble being blown. Such a bubble, basically a sphere, is a two-dimensional analog, an embedding diagram, for a closed three-dimensional universe. Such a universe would have a finite lifetime and would have to recollapse (neglecting the effects of dark energy). The results reported in Chapter 12 suggest that our Universe is not closed and ''spherical.'' It might be flat, but accelerating. Physicists and cosmologists are working now to develop models of inflating universes that are consistent with infinite expansion. Such universes, can, of course, make black holes as they

expand, and that is enough to raise Linde's conjecture of new universes being constantly created.

These ideas have been taken one more dramatic step by Lee Smolin, now at the Perimeter Institute for Theoretical Physics in Waterloo, Canada, in his book, *The Life of the Cosmos*. Smolin addresses the deepest issue that drives both physicists and theologians. Why are *we* here? What is it about our Universe that gave rise to life, to us. Smolin may not have the answer, but he has put the issues in an especially thought-provoking way by combining these ideas from physics with the basic ideas of biology, the power of natural selection. Smolin notes the amazing coincidences of numbers and physical conditions that are required to give rise to life as we know it. What if, Smolin wonders, each new universe had different numbers, for instance different values of the fundamental constants, Newton's constant of gravity, Planck's constant, the speed of light, and other physical constants of Nature. Most of those universes would fail. Some would not get out of the quantum foam or would quickly fall back. Others would expand so rapidly that stars did not have a chance to form, so there would be no black holes. In either case, those universes would be barren, unable to produce progeny, new universes with new properties. Smolin makes a natural-selection argument that after countless trials, the universes that survive would be those that maximize the production of black holes so that maximum progeny are ensured. Smolin argues that physicists may have to give up on a purely reductionist approach to science wherein the constants of Nature have set values that theory and experiment can reveal, and accept that our Universe has arisen from a process of trial and error, a result of probabilities, not certainty. To be fruitful, such a universe would have to expand about as fast as ours, make stars like ours, produce heavy elements like ours to control the heating and cooling of the interstellar gas to keep star formation going for billions of years. Such a universe, Smolin deduces, must have the properties of our Universe, and such a universe naturally gives rise to life to contemplate and make sense of it. Now that is a grand vision.

For all its inventiveness, Smolin's picture does not really address the fundamental issue. Given that there are infinite universes experimenting with all possible forms, how did it all arise in the first place? Was there a beginning to this process? Is there an end? James Gott of Princeton has put another wrinkle on the game by combining the self-reproduction of universes through black holes with the notions of time machines. If new universes emerge from the quantum

foam of a black hole singularity, can they emerge in the past? If that were possible, Gott conjectures, then the universe that emerges from a black hole could be the one that made the black hole from which it emerged, or a universe somewhere back in the chain of universes that Linde and Smolin contemplate. Recall from Chapter 13 that the Novikov consistency conjecture does not rule out time travel, it only demands self-consistency. Could it be that the Universe or a complex web of universes gave rise to itself in a closed but self-consistent time loop? Could it be that there is no ''beginning'' and no ''end'' but just an infinite closed loop? As Gott asks, could the Universe have created itself?

All these issues loom, but we cannot address them without a theory of quantum gravity. Fortunately, we have a candidate for that theory. Before forging into that area, a review of hyperspace notions is relevant.

14.3 HYPERSPACE PERSPECTIVES

To illustrate black holes and curved space, we have had recourse to embedding diagrams that reduce the fullness of the curved three-dimensional space to two so that we, as three-dimensional creatures, can view these warped spaces from our higher-dimensional perspective (Chapter 9, Section 9.5; Chapter 12, Section 12.2; Figure 13.2). From this perspective, it is clear to us that, even though there is no two-dimensional outside to the two-dimensional space, there is a very natural ''outside'' to the two-dimensional space, the very three-dimensional ''hyperspace'' that we occupy. This naturally leads one to wonder whether there is a ''real'' fourth spatial dimension that we, as three-dimensional creatures, cannot perceive, into which our three-dimensional Universe curves. This hyperspace would be where wormholes go when they go.

The issue of a fourth spatial dimension has been around for a long time, even predating Einstein. When Georg Reimann and Nikolai Ivanovich Lobachevsky laid the foundations for the mathematics of curved space in the mid nineteenth century, people already began to wonder to where might curved space curve. Notions of a four-dimensional hyperspace actually affected art and culture around the beginning of the twentieth century, as explored by my colleague, art historian Linda Henderson, in her book *The Fourth Dimension and Non-Euclidean Geometry in Modern Art*. People explored simple four-dimensional shapes like tesseracts, the four-dimensional extension of

a cube, and more complex shapes. Some founded religions and philosophies based on this hyperspace perspective.

It was in this context that Abbott's marvelous *Flatland* was written, misogyny and all. As Abbott described, an imagined two-dimensional creature could ''see'' (whether electromagnetic radiation could propagate in a two-dimensional space is another issue) the front of another denizen of two-dimensional space. From three-dimensions, however, we could see the front, back, sides, and *insides* of such a creature simultaneously. Likewise, when we greet a friend in our three-dimensional space, we perceive their smiling visage, but cannot simultaneously see their backsides, never mind the state of their heart and lungs. If there were a hypothetical four-dimensional creature who could look ''down'' on us as we look ''down'' on a two-dimensional creature sketched on a sheet of paper, that 4D creature could simultaneously perceive our front, back, sides, all our 2D surface, but also all of our 3D volume, all of our guts and plumbing, all with one glance.

A 3D creature passing through a 2D ''universe'' would first penetrate it at a point, then would ''fill'' a two-dimensional area, then would recede back to a point as the creature proceeded on into its own 3D ''hyperspace'' and no longer intercepted any part of the 2D ''universe.'' Likewise a 4D creature passing through our 3D space would first appear at a point, then expand to ''fill'' what we perceive as a 3D volume, but which would be a mere cross section to the 4D creature, then shrink back to a point and then vanish from our perspective as the creature proceeded on its 4D way.

These ideas floated through the salons of late nineteenth-century and early twentieth-century Paris. A case can be made that cubism arose in part out of an attempt to portray objects from different aspects and different times simultaneously (but not that Picasso influenced Einstein's thinking), in somewhat the manner that a hyperspace perspective invites. This cultural phenomenon of pondering a spatial four-dimensional hyperspace faded with Einstein and the powerful notion that the fourth dimension was time, but it has never quite vanished from the cultural landscape. The cross depicted in Salvadore Dali's famous *Crucifixion* is actually a representation of a 4D tesseract unfolded into 3D, each ''side'' of the tesseract itself being a 3D cube. The full title of the painting is *Crucifixion (Corpus Hypercubus)*. Even today, modern artists like the Brazilian Marcos Novak invent fantastic four-dimensional shapes and then represent them as they would be projected in our 3D space as they partially penetrated it.

Some of these ideas are even woven into Steve Martin's witty play, *Picasso at the Lapin Agile*.

Despite the intuitively natural sense that invokes this sort of higher dimension when one talks about curved space around black holes or the possibility that the entire Universe is the three-dimensional analog of the two-dimensional curved surface of a sphere, throughout most of the twentieth century, a true large four-dimensional hyperspace was not part of physics. Physicists can construct mathematical models of curved three-dimensional spaces and universes, even wormholes, completely within the confines of that three-dimensional space. There was no need, or means, to invoke any extensive higher dimension, no way to measure it, no way to do physics with it. Not until string theory, anyway.

14.4 STRING THEORY

Work on string theories is beginning to penetrate the barriers that separate Einstein's theory from the standard quantum theory and to bring a whole new perspective to hyperspace. The previous summary of the history of this area in Section gives some preparation for what is necessary. Whereas Einstein overthrew the concept of gravity as a force between two objects, the quantum gravity theory of everything is likely to bring with it entirely new ways to think about gravity and, indeed, about space and time. In the appropriate regime, one can still think of curved space as the origin of gravity, just as for weak gravity it is still useful to think of a force of gravity and to use Newton's theory in appropriate circumstances. One of the steps that energized string theory was the understanding that within the full mathematics of the theory, a subset described exactly Einstein's theory of general relativity. Just as Einstein's theory ''contains'' Newton's theory of gravity in the limit of weak gravity, string theory ''contains'' Einstein's theory.

String theory, however, holds a lot more. The underlying concepts of a theory of everything may require a shift in conceptual basis as profound as that from a force of gravity to gravity as curved space. The notion that the fundamental entities from which everything is constructed are strings is such a conceptual shift. Recent developments point strongly to the conclusion that, at a sufficiently small scale, physics will be very different from that which Newton, Einstein, or the founders of quantum theory envisaged.

To see how this idea has arisen, a sketch of string theory is necessary. An excellent introduction is given by Brian Green in his book *The Elegant Universe* and the PBS series of the same name. The roots of string theory go back to the 1960s when physicists were exploring the fundamental forces. In classic (nonstringy) quantum theory, the fundamental forces (Chapter 1, Section 1.2.1) have a very different cause than curved space or Newton's action at a distance. The strong and weak nuclear forces and the electromagnetic force arise from an exchange of particles between two interacting entities. This quantum exchange can yield either attractive or repulsive forces depending on circumstances. For the electromagnetic force, the exchange particles are photons, the fundamental entities of electromagnetic radiation. For the strong nuclear force, the exchanged particles are pi mesons and gluons. For the weak nuclear force, the particles are three special ones that can be charged either positively or negatively or not at all. In the 1960s, physicists realized that the equations that described the strong nuclear force by this sort of exchange also described entities that could stretch and wiggle, entities with the properties of dynamic strings of energy.

The basic notion is that particles, mathematical points, are too simple to contain the wonders of nature. True point particles have no inner structure, no richness. A string, on the other hand, by adding only one more dimension to the structure, can vibrate in many modes. You can't make music with four grains of sand, but with four violin strings you can have Mozart! In the view of string theory, different modes of vibrations of the string represent different particles, just as one string on a violin can give different notes depending on where the violinist's finger is placed.

Unlike violin strings, the strings that represent the fundamental entities in this theory do not exist only in our ordinary three-dimensional space. To make a mathematically self-consistent picture, one free of infinities and other inconsistencies, the space through which the strings thread must be of much higher dimension. The currently most viable versions of the theory have ten spatial dimensions plus one of time. Hyperspace, a notion that has floated through much of this book, is not a mere abstraction to string theory; hyperspace is absolutely intrinsic to the structure of string theory.

The nature of these multidimensional loops of energy is that they have a characteristic length or scale, roughly the distance ''along'' the loop. The exact size of this scale is not known; it is fantastically smaller than the size of an ordinary particle like a proton or

neutron, but somewhat bigger than the Planck scale by perhaps a factor of a thousand.

Concepts relating to black holes are woven throughout discussions of string theory. Here is an example. The way physicists have proceeded to explore ever more fundamental entities is to go to smaller scales: molecules to atoms to nuclei to protons to quarks. Experimentally, one probes these smaller scales by invoking ever higher energies. This is related to the fact that, in quantum theory where everything has a wave-like character, higher energy is associated with shorter wavelengths and hence, smaller length scales. Basically, one needs higher energy to probe smaller volumes and that is why physicists hunger for ever larger, more energetic ''atom smashers'' or particle accelerators in modern parlance. We know, though, that if one packs too much energy into a small volume, you make a black hole. We also know that black holes behave such that the more mass/energy you add to them the *bigger* they are, in terms of their event horizons, not smaller. The very nature of black holes thus suggests that there is a minimum size scale physicists can probe before they lose information inside event horizons. That length scale might be the string scale, or something related to it. The issue of information and black holes will come back again in a very profound way in Section . There is also an issue of how ''thick'' the strings are. By the same tenets of quantum uncertainty that limit the thickness of the ring singularity in a rotating black hole, strings cannot really be of zero thickness. Physicists assume for working purposes that they are of roughly a Planck size thick. One should not take the image of small rubber bands too literally; the strings are intrinsically quantum entities with all the wave-like uncertainty that entails.

With this string perspective, the ''singularities'' of Einstein are probably not of the Planck scale, but regions roughly the size of the string scale. Exactly what physics looks like, how space and time behave on the string scale, remains to be fully elucidated, but because they have finite length, strings smooth out physics on this string scale and remove the troublesome infinities that otherwise pop up in the mathematics.

Through much of its development, the higher dimensions invoked by string theorists were all ''compact.'' To picture a compact space, start again with a two-dimensional analog, a sheet of paper. As shown in Figure , roll the paper up into a tight roll. From a distance, the resulting object looks like a straight line, a string of length of perceptible extent, but no width. Imagine rolling the paper up lat-

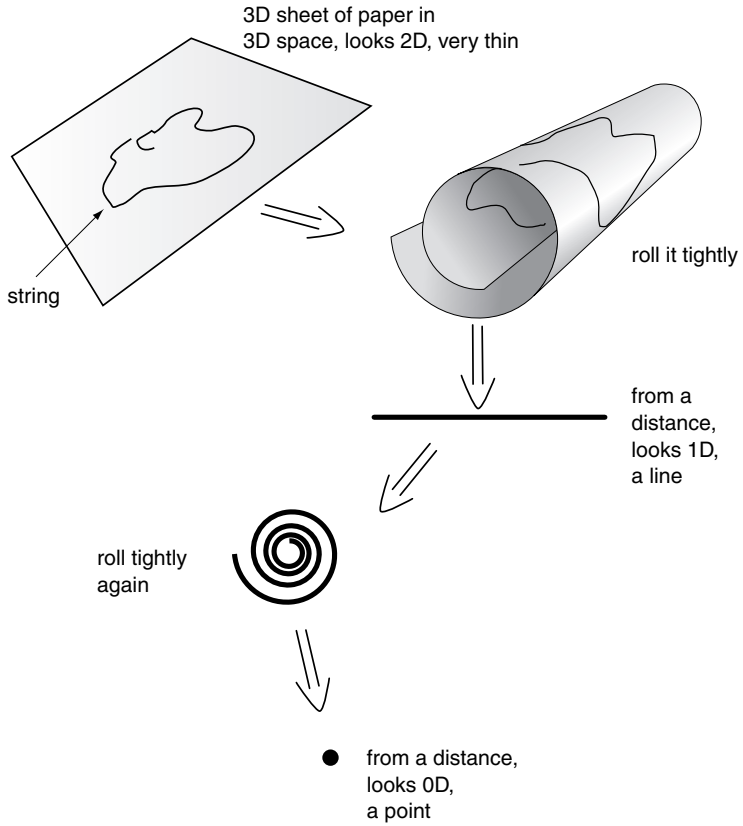


Figure 14.1 A schematic example of how a space could be compact and still contain a string capable of vibrating. A two-dimensional sheet containing a one-dimensional string can, in principle, be rolled up compactly so that it would appear to have only one dimension, length. The space is still two-dimensional and the one-dimensional string would still be there, just wound up in the compact space. If the space were rolled up again, it could, in principle, appear to be a point, a zero-dimensional space, yet it would still be two-dimensional and would still contain the string.

erally so you have a tiny ball. Now from a distance, the whole original sheet of paper resembles a point, a particle of no extent. A string in that original sheet of paper could still exist and vibrate away in that compact space that we could not directly perceive. We could, however, deduce that the higher dimensions exist because the nature of particles in our Universe demands it!

The last few years have seen some immense advances in string theory that have given great hope that it is the basis for the theory of everything. One step has been to prove that what looked like five or six different string theories are all versions of the same underlying theory, the full shape of which has not yet been elucidated. These connections were established by what physicists have called *duality*, a connection between the properties of the theories. In one version of the theory, a parameter could be small, and, as the parameter got large, the mathematics of the theory broke down. In another string theory, the dual to the first, there would be a parameter that was just the inverse of the first. In that second theory, as the first parameter got large, the inverse parameter got small, and the mathematics in that theory was well behaved. The middle ground is unknown, but this duality yields a signpost for how to link the disparate theories and show that they are deeply connected, that they are aspects of the same thing. This grand string theory that is taking shape is called *M theory, M* for matrix, or mystery, or an upside down W for Ed Witten of the Institute for Advanced Study, who developed it.

One of the concepts that has emerged from string theory is that there are not only strings threading the ten dimensions of the string theory hyperspace but also surfaces. These surfaces can be canted in hyperspace in just the same way that a sheet of paper can be oriented in all sorts of ways in our ordinary three-dimensional space. A more general word for a surface is a membrane, a term that also connotes a certain elasticity, a property that these surfaces have. These membranes can vibrate just as the strings can vibrate, and their modes of motion are also important to the behavior that emerges as ordinary physics in our ordinary space–time. To classify the membranes in spaces of various dimensions, they are referred to as p-branes, where p is a symbol denoting the dimension of the membrane; $p = 2$ for a two-dimensional surface, $p = 3$ for three-dimensions, $p = 9$ for nine dimensions. The surfaces must have at least one dimension less than the full dimensionality of the space they occupy. In a sense, strings themselves are 1-branes.

An important development of string theory in recent years has been the recognition of the critical nature of the interaction of strings with p-branes. The ends of the string can attach to the p-branes or snap off to form closed rings. Some of the seminal work on branes was done by Joe Polchinski at the University of California at Santa Barbara but many others are contributing to the fevered pace of development.

A striking feat that followed the development of the theory of *p*-branes and their interactions with strings has been the capacity to construct simple models of black holes. These black holes are not the creatures of the curved space–time of Einstein, but simpler versions in two dimensions constructed from the entities of *p*-branes and strings. Nevertheless, because string theory contains Einstein's theory, objects that exert gravitational pull and that have event horizons can be constructed. The difference is that string theorists can count the numbers of modes and vibrations of the strings within the black holes they have constructed and tell exactly what the temperature and entropy should be. They get precisely the same answer as Hawking did in predicting Hawking radiation (Chapter 9), even though the mathematics and, indeed, the conceptual framework they use, is completely different. This striking concordance is the sort of development that tells physicists that they are getting close to a universal truth and that string theory has deep lessons to reveal.

String theory has also brought new insight into another problem that arises from thinking about the nature of black holes. This is called the *information crisis*. Information, the bits and bytes of computers, is about as fundamental as you can get. The problem is that black holes seem to destroy information, and that bugs physicists. The idea was already there in our previous discussions of the nature of black holes in Chapter 9 and captured in John Wheeler's phrase ''black holes have no hair.'' You can throw stars, cars, people, and protons into a black hole, and all the information that described that ordinary stuff vanishes inside the event horizon. The only properties of a black hole that can be measured from the outside are its mass, spin, and electrical charge. Now Stephen Hawking enters the game. Black holes can evaporate, giving off Hawking radiation. Given enough time, the black hole will just disappear, leaving pure radiation with very little information content, essentially pure randomness. This process conserves energy, the energy equivalent of all the stuff that went down the black hole eventually emerges as the energy in the radiation. What happened to the information that defined that stars, the cars, the people, and the protons that went down the hole? Physicists have been debating this fundamental problem since the implications of Hawking's ideas of black hole evaporation were first assimilated.

One can sense a possible wrinkle in this argument. Hawking's theory was designed to work for ordinary-size black holes where the event horizon was well separated from the singularity at the center of the black hole. When a black hole evaporates down to the last of its

essence, one needs a theory that can simultaneously treat the event horizon and the singularity and that probably requires a quantum gravity, a theory of everything. In the absence of that theory, it is not clear that one can use Hawking's original theory to account for the final moments. String theory gives a different possibility. It suggests that the black hole cannot evaporate entirely, but that, as the process runs away, one is left with a string vibrating intensely somewhere in its eleven-dimensional space–time. In those vibrations could be the epitaph of all that entered the black hole, all that original information, the size of the stars, the bumper stickers on the cars, the personalities of the people, the number of protons. On the other hand, Hawking has proclaimed that the information might reside in the radiation emitted; that the radiation is not so simple as that of an object, of a single, well-defined temperature, and hence only one ''bit'' of information. This issue remains on the forefront.

Einstein wrote down a full and self-consistent set of equations to describe gravity (in the absence of quantum effects) in 1916. Those equations have yet to be fully solved. String theory is like that, only more so. The full mathematical structure of string theory is very complex, and only a few solutions have been wrested from it. Those solutions have been tremendously encouraging. Exactly what theory of space and time will emerge from string theory is thus not yet clear. One can see that, because string theory is a theory of quantum fields and forces, the fundamental concept of gravity will again be a force, but a quantum force, not that of Newton. Away from any singularity, this ''force'' of gravity will act just as in Einstein's theory. One will be able to speak in the language of curved space and time and dream of the construction of wormhole time machines.

On the microscopic scale, however, the new concepts of string theory will lead to different pictures, pictures that are only just now beginning to take hazy conceptual form. One can see that gravity will be represented by the familiar terms of Einstein's gravity plus ''something else'' that comes in ever more strongly as one approaches, intellectually at least, the string scale. At the string scale itself, Einstein's theory will be completely inapplicable, as Newton's theory is within the event horizon of a black hole. The point singularity of Einstein with infinite density and infinite tidal forces will not exist in this framework, but what will replace it is not entirely clear.

While string theory struggles to understand what physics is like at the string scale, the growing understanding of the properties of

strings and branes led to a revolution in our perspective of the Universe on the largest scales.

14.5 BRANE WORLDS

As outlined above, branes are surfaces that slice through multi-dimensional space. They must be of a dimension less than the full dimensionality of the space that contains them. In a 10-dimensional space, the largest dimension brane would be a 9-brane. The space ''surrounding'' a brane has come to be called the *bulk*. The bulk is effectively the hyperspace ''volume'' in which the brane is immersed. An example would again be a sheet of paper (or the two-dimensional surface of any ordinary object) in our normal three-dimensional space. In that case, the sheet of paper would represent the brane and the three-dimensional space above, below and around it would be the bulk in which it resides. From the notion of strings, branes, and bulk, came a new view of the hyperspace that may envelop our Universe.

Recall from Section the discussion of four-dimensional hyperspace, the space into which three-dimensional curved space might curve. Physicists did not merely ignore such a possibility. There was a very basic reason why physicists rejected the notion that such a hyperspace existed and why they insisted, in the development of string theory, that any higher dimensions must be tightly wrapped. The reasoning would have made sense to Newton.

In our common experience, the brightness of a light (the detected intensity of a distant supernova as discussed in Chapter 12), the electrical force due to a single electrical charge, or the effect of a star's gravity on an orbiting planet, all decrease like one over the distance squared. There is a very basic reason for that, and it is deeply connected with the dimensionality of our perceptions. For any of these three examples (for weak Newtonian gravity and light that shines equally in all directions, unlike gamma-ray bursts, Chapter 11), the effect of the light, the electrical charge, or the gravity spreads out through larger volumes of space as one gets more distant from the source. Specifically, the effect is spread over a larger and larger area at greater distance and that results in a dilution of the apparent brightness or electrical or gravitational ''force.'' The dilution factor is precisely the area through which the influence must flood at a given distance. If the area is bigger, most of the influence is ''wasted'' in other directions from the direction where the detection or measurement occurs. The area is just $4\pi D^2$ where D is the distance of the

observer or detector from the source of the light, or the electrical force, or the gravity. The effect at a distance is thus diluted by a factor of one over the area spanned at that distance and this, in turn, means one over the distance squared. The key point here is that the area goes like the distance squared only in a three-dimensional space where volumes scale as the size or distance cubed. The ''2'' that appears in the inverse distance squared law is exactly and precisely a factor of 1 less than the full dimensionality of the space, namely 3.

We can, in principle, extend this argument to hypothetical higher dimensions. Suppose we consider the possibility of a true, large, extended fourth spatial dimension, as some people did in the late nineteenth century. Setting aside for now the issue of how light or electrical force might penetrate that void, let's focus on gravity. Gravity is an entity of space. Gravity curves space. Gravity can send ripples though space. If there is an extended fourth spatial dimension, gravity ought to be able to go there. With a fourth dimension, however, ''volumes'' scale as length raised to the fourth power and ''areas'' scale like one power less, namely as length or size or distance raised to the third power; exactly and precisely a factor of 1 less than the full dimensionality of the space, namely 4.

If this were the case, then, physicists argued, the existence of an extended fourth spatial dimension would require that the strength of gravity would fall off like one over the distance *cubed*. Even Newton knew that was wrong! Planetary orbits would be completely bonkers and could not even exist if gravity worked that way. The best empirical attempts to measure the strength of gravity show that it does decrease like one over the distance squared.

The implication was, it was long thought, that if there were a fourth, or higher, spatial dimension, it must be tightly wrapped up. To the extent that gravity tried to ''go'' into this higher dimensional space, there would be very little ''volume'' or ''surface'' to dilute it, and so the inverse distance squared law would continue to work in the three-dimensional space of our perceptions, just as we observe it to do.

Various models of this wrapped-up space have been considered. One that seemed particularly amenable to the needs of physics and string theory was the six-dimensional Calabi–Yau space. The idea was that at each and every point in our three-dimensional space there were six other mutually perpendicular directions, each bending around in a tightly curved, complex, but systematic way to end up at exactly the same beginning point in three-dimensional space.

That perception that any higher-dimensional spaces must be tightly wrapped changed dramatically in 1999. Lisa Randall, now at Harvard, and Raman Sundrum, now at Johns Hopkins, realized that there was a technical flaw in this argument. The tacit assumption had been made that gravity must flood into a large fourth dimension with the same ease that it penetrates the three dimensions of our perceptions. Randall and Sundrum concluded that while that could be true, it was not necessarily true. Within a reasonable mathematical framework, there could be a large four-dimensional hyperspace and gravity would still go there only a little; there would be little effective ''area'' associated with this space, and gravity would still decrease very nearly as one over the distance squared. This idea opened the floodgates.

Within the framework that Randall and Sundrum revealed, our three-dimensional Universe would be a 3-brane immersed in this four-dimensional bulk. The bulk would represent a real, large (infinite) four-dimensional hyperspace in which our three-dimensional Universe is embedded. With this new vision, a number of deep issues of physics, quantum theory, gravity, and string theory fell into place.

In this picture, the ordinary forces – electromagnetism, nuclear forces – correspond to ''open'' strings that are not closed loops, but have open ends. These ends are not free to wiggle about, however; they must be anchored to a brane. In this case, the brane is the 3-brane of our Universe. This leads to an insight into why we cannot ''see'' higher dimensions. We ''see'' by receiving photons of electromagnetic radiation. In this view, photons are represented by certain vibrations of strings that themselves are locked onto the brane. The string cannot leave the 3-brane, the photons cannot leave the 3-brane, and so we cannot receive photons from, or send photons to, the bulk. It may also still very well be true that yet other higher dimensions are tightly wrapped up, so there is very little ''there'' to perceive even if photons could get there, which they cannot.

Even in this framework, gravity remains a different beast. The strings representing gravity, quanta of gravitational exchange ''particles'' called gravitons, are ''closed'' loops of strings. They are not attached to branes, and they can leave the brane to pervade the bulk. As for the pool-ball crisis of Chapter 13, an analogy is again the game of pool. Under normal circumstances the balls roll around on the table, confined to the two-dimensional flat plane. In this case, however, there is something that is *never* confined to the flat plane, and that is the *sound* of the pool balls as they click together. The sound

pervades the room, an intimate and intrinsic characteristic of the game. In our world, the electroweak force and the strong nuclear force (presumably all part of one grand unified force, Chapter 1) are represented by strings that cannot leave the brane, like the pool balls restricted to the green felt. Gravity carried by closed strings can leak out into the bulk as the sound of clicking pool balls can be heard throughout the bar. In the bar, the sound weakens as one over the distance squared, but, as Randall and Sundrum showed, gravity, while not completely restricted to our 3-brane, does not penetrate far into the bulk, so it also weakens very nearly like one over the distance squared even with the hypothesized immense bulk ''surrounding'' us.

Theoretical physicists and cosmologists are now on a rampage to explore all the implication of this amazing new intellectual vista. The models now flooding the literature are called *brane-worlds*. They are all built around the idea that our Universe is a 3-brane ''floating'' in this four-dimensional bulk. Virtually all the current models regard the other six dimensions of string theory's ten-dimensional space to be ''wrapped up,'' a Calibi–Yau space or some version of that. Whether having three ''normal,'' one large hyperspace bulk, and all the rest of the six higher dimensions wrapped up is merely the simplest extension of the Randall/Sundrum ideas or whether this configuration is somehow required by physics and mathematics is not completely clear. Virtually all the current work in this area assumes only one large extra dimension, although it is conceivable that there could be more than one of these large extra dimensions and correspondingly less wrapped-up dimensions.

With the notion that our Universe is a 3-brane in an immensely larger bulk, one is invited to consider other complete, even infinite, three-dimensional universes immersed in this bulk, but ''elsewhere'' in four-dimensional hyperspace. One early theory manifesting these ideas was the Ekpyrotic Theory (from the Greek *ekpyrosis*, or conflagration) developed by Paul Steinhardt of Princeton and his colleagues. In this theory, there would be two 3-branes floating in the bulk. These 3-branes could collide, with every three-dimensional point in one universe ''hitting'' a three-dimensional point in the other universe, as one can picture bringing two sheets of paper (2-branes) together in a room (the 3D bulk) so that each point of one sheet contacts a corresponding point on the other sheet. In the Ekpyrotic Theory, this collision would release immense energy and cause the two universes to spring apart in the bulk with an attendant expansion of their three-dimensional space. The result would be, from the

perspective within the 3-brane, an expansion from a very hot, dense state, a big bang. In this case, however, the big bang would not start from Einstein's singularity, but from this collision of 3-branes in a four-dimensional hyperspace bulk. This theory has not generally gained broad support, but it did suggest that the gravitational waves generated in the collision would be distinctly different than those in the standard big bang, so there is even some prospect for a test.

Is this bulk the place where our three-dimensional space curves when it curves? That link is invited, but is not necessitated in the current framework. Is the bulk the first hint of the hyperspace travel of *Star Trek* and *Star Wars* with only engineering details to be worked out? That also is extremely premature; but still physics, not science fiction, has given this peek behind the hyperspace curtain.

One of the lessons of science is that Nature follows the tenets of mathematics; sometimes there is no correspondence between an abstract aspect of mathematics and physical reality, but at other times pure mathematics has pointed the way to deep new understanding of Nature. String theory has been so rich and challenging, that it has opened new vistas for mathematical research as well as for physics. The hard and critical question for now is whether any of this is real or just mathematical fantasy. The key will be to put these ideas to observational or experimental test.

Physicists are straining to devise such tests. One question is whether gravity does, indeed, behave a little differently than one over the distance squared. Is it possible that gravity scales like $1/D^{2.001}$ rather than $1/D^2$? Such a difference might give a hint that some higher dimension or dimensions exists. Experiments are underway now to try to measure any minute departures from the inverse-distance-squared behavior of gravity. Another possibility, currently beyond the technical horizon, is the question of whether black holes might behave slightly differently than Einsteinian gravity right down near the event horizon. Perhaps someday that behavior could be measured with X-rays that emerge from the inner edges of accretion disks. People are exploring the idea that the dark energy of Chapter 12 is some manifestation of a ''nearby'' three-dimensional universe, another 3-brane, only a little distance from us in the four-dimensional bulk.

Black holes remain at the center of this quest. Black holes may behave differently in the presence of the bulk; in particular, small, primordial, black holes might extend into the bulk, changing their effective area and altering their Hawking temperature (Chapter 9,

Section 9.6). Recall that while radiating black holes will emit photons most easily (no rest mass to produce), they also can, in principle, emit any kind of particle, including antiprotons. Experiments to measure the abundance of antiprotons in cosmic rays have revealed evidence for a source of antiprotons other than normal cosmic ray interactions. Katsuhiko Sato and his colleagues in Japan have explored the notion that these excess antiprotons arise in primordial black holes and that the existence of the antiprotons hints at a large extra dimension. I would not take this to the bank, but this sort of work illustrates the range of exploration going into this topic today.

The take-away message is that hyperspace might be real. There will clearly be an immense amount of work on these topics in the near future. Stay tuned!

14.6 A HOLOGRAPHIC UNIVERSE?

Section referred to information that black holes do or do not have. That seems like an abstract and obscure topic, but thinking about it is at the frontier of modern physics. There are two key ideas that are familiar to anyone with a computer and a credit card. The information stored in a computer and whipped around the world on the Internet is digitized. It comes in patterns of bits, zeros and ones. The amount of information stored in a computer memory is then related to the number of bits that can be registered in its memory or on its hard drive. That amount of information is amazingly large in this day and age, and is destined to get larger, but it is finite. We have also learned to store information in holograms. The basic idea is to register information in the interference pattern of two lasers and to imprint that interference pattern, rather than a literal image, on a film surface. When another laser is shone upon that surface, a three-dimensional representation can be restored that seems to have depth and volume. The little ''hologram'' on your credit card is a basic version of this, giving at least some sense of three-dimensional depth, although you cannot walk around your credit card and see the image from all sides as you can a true reconstructed hologram. You can put these two ideas together and wonder whether there is a limit to the amount of bits one can store in a hologram, and hence the total information. If you follow that path, and recall that there is a smallest ''size'' to things, the Planck length, or perhaps the string length, then you find yourself contemplating deep issues of not just quantum gravity, but the nature of reality.

In 1993, Gerardus 't Hooft, who shared the 1999 Nobel Prize in physics for fundamental work on particle physics, proposed what he called the *holographic principle*. Leonard Susskind of Stanford and many other physicists have furthered the idea. The notion is that all the information about everything within a volume can be represented by a theory of the information on the surface of that volume and that each Planck area (the square of the Planck length; setting aside for the moment that the string length is larger than the Planck length) contains one ''bit'' of information. 't Hooft calls this ''Nature's bookkeeping system.''

The roots of this thinking go back to the nature of black holes. Black holes have a size, an event horizon, that increases with the mass. According to Hawking, they also have a temperature that decreases with the mass (Chapter 9, Section 9.6) and an *entropy* that increases with the mass. In a casual sense, entropy is a measure of the disorganization of a system. In the ''game'' of 52-card pickup, a deck of cards flung in the air to land scattered around a room is more disorganized than the original pack: after flinging, the cards have more entropy. Disorganization would seem to imply less information but, in fact, just the opposite is the case. If you flipped a coin 100 times and it came up heads every time, you would conclude the coin was rigged and could predict with essentially 100% accuracy that the 101st flip would produce a head. There would be no new information content in that 101st flip. A completely organized set of events, like all heads, or a string of all 1s, or a string of all 0s, has no entropy and no information content. An honest, random coin, would provide a new bit of information, whether you won or lost a bet on the outcome, for instance, with every flip. The randomness also represents a high entropy; each coin flip has one bit of entropy, one bit of information. According to *information theory*, entropy is a measure of information. Hawking also established that the entropy, and hence the information content, of a black hole increases with its mass in direct proportion to the area of the event horizon. It was the ability of string theory to provide an identical determination of the information content of a black hole that gave an impetus to string theory as a theory of gravity (Section).

Think, then, of a spherical volume, to keep things simple. A small-mass black hole with little entropy and hence little information can fit in that volume. There is a maximum mass, and hence size, and hence entropy and information, that will fit in that volume and that is when the event horizon of a black hole just fills the chosen volume.

For any smaller black hole, the information content is less. This means that the maximum entropy and information of a region is related not to its volume, as one might think, but to the area surrounding that volume. This suggests that the information about the volume is somehow related to the area surrounding that volume, not to the volume, per se. 't Hooft followed this line of logic to conjecture that the information about any volume, not just that containing a black hole, is related to the surface and that the surface, not what goes on within the volume, is the true reality. The little image on my credit card is really a flat surface with an imprinted interferogram. The idea that there is a little bird with some depth on my platinum card is an illusion. Could it be that all the information about the nature of the Universe is actually enscribed in some fashion on its surface and all that we perceive as three-dimensional reality is an ''illusion?'' These ideas currently have two manifestations, one in observational cosmology and one in the structure and meaning of string theory.

Craig Hogan of the University of Washington has considered some implications of holographic ideas in the context of the nature of the big bang. Hogan notes that the current theory of cosmology is that the Universe exploded from some hot dense state with matter/energy nearly uniform, but subject to wrinkles associated with the intrinsic quantum uncertainty of that early dense state. As the Universe expanded, those wrinkles were frozen in by the huge expansion of the inflation era; they remained the seeds of all the structure that ultimately formed in our visible Universe. Slightly overdense regions contracted under gravity to become denser and to attract surrounding matter, leaving irregularities in the temperature of the cosmic background radiation (Chapter 12, Section 5) and ultimately leading to the galaxies that litter deep *Hubble Space Telescope* images. Each patch of hotter or colder background radiation measured by the *WMAP* satellite (Chapter 12, Section 12.5) originated from a single quantum fluctuation. Hogan marvels that each such patch is at once the largest (in the current epoch) and the smallest (at the moment of the big bang) single entity we can image. Hogan notes that in ''classical'' quantum theory which assumes a continuous underlying space–time, there is no lower limit to the extent of the original perturbations, but there is in the context of holographic theory. Because no ''bit'' of universal information can be smaller than a Planck area, each quantum fluctuation contains a limited amount of information. An analogy, Hogan points out, is a digital photo that looks pixelated under high resolution. Perhaps, Hogan speculates, the space–time of the

Universe is fundamentally pixelated. Hogan estimates that the total amount of information that can be tiled on the surface that surrounded the causally connected volume of the inflating Universe was remarkably finite, only about 10 gigabits. You could store that amount of information on your personal computer! This is a quantum gravity notion; the information implied by standard quantum theory and standard gravity, Einstein's theory, considered separately would be tremendously greater, essentially infinite. From the holographic point of view, Hogan has estimated that the total number of bits in a given quantum fluctuation that grew to become a galaxy is less than a million, and that future maps of the temperature fluctuations of the cosmic background radiation might have the resolution to detect the fundamental pixelation of quantum gravity space–time. From such an observation might come fundamental understanding of how space and time form from conditions where space and time as we know them do not exist. That is a grand vision.

The other application of the holographic principle in physics operates in the new world of strings, branes, and the bulk. The key ideas were presented by Juan Maldecena, now of the Institute for Advanced Study, in the late 1990s. The ideas represented a conceptual breakthrough, yielding new insights into both quantum gravity and the standard model of particle physics. There is a mapping, an equivalence, of the theory of quantum gravity, string theory, in the bulk and the theory of ordinary physics on the brane. The two theories that sound so different can be mathematically identical. To make this work, the nature of the ''bulk'' must have four ordinary space dimensions plus time and be a so-called *anti-de Sitter space*, a space with an effective negative cosmological constant. Whether this mathematically defined space has anything to do with the implicit 4D hyperspace where wormholes go when they go is not clear. Anti-de Sitter space does not correspond to the space we live in, but it is mathematically more tractable. In certain mathematical circumstances, the boundary of this anti-de Sitter space is a flat space–time of three ordinary dimensions plus time; something like our observed Universe. Maldecena found that if one describes the physics on this boundary, our brane, in terms of certain classes of so-called supersymmetry theories of ordinary particles and forces, then the theory of gravity in the anti-de Sitter space bulk and the theory of physics on the surface brane are mathematically equivalent. In this rather subtle and sophisticated sense, the theory of physics on the brane, everything we know of physics in our 3D-plus-time Universe, is a

''hologram'' of the physics of gravity in the higher-dimensional bulk. We, everything we know, *are* the ''shadow.''

If this is the way physics works, all the physics on Earth, from atoms to you, could be contained on a surface around the Earth. All the physics in the Universe could be contained in the surface of the Universe, as if all the information that constitutes ''you'' could be enscribed in your shadow. In the context of M theory, branes and the bulk, we are the 3D shadow of the 4D bulk. How freaky is that?

There are also theories of the paranormal that label themselves as part of the ''holographic universe,'' so if you do a web search on this, use some discrimination.

14.7 CODA

This is heady stuff. It is amazing that these ideas have emerged, not from science fiction, but from hard-nosed physicists wrestling to make sense of the Universe of our observations. Examining these ideas for self-consistency will yield progress, and that enterprise will go forward with great energy. The real solution, or at least the one we can contemplate today, is to develop the theory of quantum gravity, the theory of everything. Today the best bet for that appears to be string theory, M theory. So one can ask, what does string theory say about the quantum foam? Quantum foam was just a name, a placeholder, until some physics came along. What exactly does string theory say about the conditions at the Planck scale? Does string theory allow new universes to be born from the conditions predicted by string theory for ''not time'' and ''not space'' at the center of a black hole constructed from strings?

Other, more speculative questions also arise. What are these higher dimensions that are forced on the string theorists by mathematical self-consistency? Do they simply dictate the properties of particles that appear in the three-dimensional Universe of our space–time, or can they be manipulated in some way? Does string theory allow wormholes and time machines? Does it prevent them?

While string theory remains the focus of intense effort, one can already glean hints that, as it stands today, it is not necessarily the theory of everything. As tantalizing and intellectually productive as it has been to study the vibrations of strings and branes in their higher dimensional spaces, one has to ask: whence those higher dimensional spaces; what of time? Einstein taught us to abandon preexisting space, to consider space as a dynamical entity. The space in which string and

branes vibrate is, however, just ''there'' and time is, mathematically, the same as we treat it in ''normal'' physics and in our everyday experience. As John A. Wheeler also said in yet another poetic summary, ''Time is what keeps everything from happening all at once.'' This is not fully satisfactory. A true theory of quantum gravity should have both space and time emerge from some aspect of the theory as emergent properties, not aspects that are assumed ad hoc. On a less fundamental but still sobering level physicists have been able to categorize string theories in the framework of M theory. They estimate that there may be 10^{500} different string theories constituting what Leonard Susskind has called a *string landscape*, in which only some might describe a universe we could know and love. That will take a while to sort out!

Papers exploring string theory, brane worlds, and the holographic principle are rampant. Some discuss the impact of these ideas on the ''real world.'' It is somewhat old fashioned, but my guess is that even with a theory of everything under discussion we are not about to see the end of physics.

Index