THE EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

ECMI

MATHEMATICS IN INDUSTRY 14

Janne Roos · Luis R. J. Costa Editors

# Scientific Computing in Electrical Engineering SCEE 2008

Springer

# MATHEMATICS IN INDUSTRY 14

THE EUROPEAN CONSORTIUM
FOR MATHEMATICS IN INDUSTRY

ECMI

*SUBSERIES*

Janne Roos
Luis R.J. Costa
*Editors*

# Scientific Computing in Electrical Engineering SCEE 2008

With 244 Figures, 83 in color and 48 Tables

Springer

*Editors*
Janne Roos
Luis R.J. Costa
Aalto University
School of Science and Technology
Faculty of Electronics, Communications and
Automation
Department of Radio Science and Engineering
P.O. Box 13000
FI-00076 AALTO
Finland
janne.roos@tkk.fi
luis.costa@tkk.fi

ISBN 978-3-642-12293-4 e-ISBN 978-3-642-12294-1
DOI 10.1007/978-3-642-12294-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010929626

*Cover design*: deblik, Berlin

Printed on acid-free paper

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

*To the memory of*
*Professor Angelo Marcello Anile*

# Preface

Scientific Computing in Electrical Engineering (SCEE) is an international conference series, which started as a national German meeting held in Darmstadt (1997) and Berlin (1998), both under the auspices of the Deutscher Mathematiker Verein. The first truly international SCEE conference was organized in 2000 in Warnemünde, Germany, by the University of Rostock. In 2002, the 4th SCEE conference took place in Eindhoven, The Netherlands, jointly organized by the Eindhoven University of Technology and Philips Research Laboratories Eindhoven. The 5th SCEE conference was held in 2004 in Capo D'Orlando, Italy, jointly organized by Universita di Catania and Consorzio Catania Ricerche. The venue of the 6th SCEE conference was Sinaia, Romania, in 2006, organized by the Politehnica University of Bucharest.

The 7th International Conference on Scientific Computing in Electrical Engineering (SCEE 2008) was held in Espoo, Finland, from September 28 to October 3, 2008. It was organized by the Helsinki University of Technology; Faculty of Electronics, Communications and Automation; Department of Radio Science and Engineering; Circuit Theory Group. (Details on the SCEE 2008 conference are at http://radio.tkk.fi/en/conferences/scee2008/).

The SCEE 2008 conference was sponsored by

- Nokia (http://research.nokia.com/),
- STMicroelectronics (http://www.st.com/),
- ABB (http://www.abb.com/),
- CST (http://www.cst.com/),
- AWR (http://web.awrcorp.com/),
- MunEDA (http://www.muneda.com/),
- MAGWEL (http://www.magwel.com/),
- Academy of Finland (http://www.aka.fi/),
- Helsinki University of Technology (http://www.tkk.fi/),
- City of Espoo (http://english.espoo.fi/),
- CoMSON Research Training Network (http://www.comson.org/),
- The Finnish Society of Electronics Engineers (http://www.eis.fi/).

The aim of the SCEE 2008 conference was to bring together scientists from academia and industry, primarily mathematicians and electrical engineers, with the goal of intensive discussions on modeling and numerical simulation of electronic circuits and of electromagnetic fields. The conference topics were the following:

1. Computational Electromagnetics
   - CAD/EDA tools and techniques
   - Modeling and parameter extraction
   - Discretization and solution methods (BEM, FEM, FDTD, FIT, FDS, PEEC, TLM, MoM, etc.)
   - Applications (antennas, microwaves, interconnects, on-chip passives, electrical machines, etc.)
2. Circuit Simulation
   - CAD/EDA tools and techniques
   - Modeling (passive, device, compact, behavioral, symbolic, etc.)
   - Simulation (DC, AC, transient, HB, envelope, noise, etc.)
   - Model-Order Reduction (MOR)
   - Applications (RF communication systems, power electronics, etc.)
3. Coupled Problems
   - Multi-physics modeling and simulation (electrical/thermal/mechanical, substrate coupling, etc.)
   - Co-simulation (EM–circuit, circuit–system, analog–digital, etc.)
   - Applications (interconnects, electromagnetic compatibility, bio-engineering, MEMS, etc.)
4. Mathematical and Computational Methods
   - Differential equations (PDEs and DAEs)
   - Solution methods for large linear systems
   - Multi-scale schemes
   - Parallel/grid computing
   - Optimization, space mapping, inverse problems, etc.

The SCEE 2008 Program Committee consisted of

- Prof. Gabriela Ciuprina (Politehnica University of Bucharest, Romania),
- Dr. Georg Denk (Qimonda, Germany),
- Prof. Michael Günther (University of Wuppertal, Germany),
- Dr. Jan ter Maten (NXP Semiconductors & TU Eindhoven, The Netherlands),
- Dr. Bastiaan Michielsen (ONERA, France),
- Prof. Ursula van Rienen (University of Rostock, Germany),
- Prof. Vittorio Romano (University of Catania, Italy),
- Dr. Janne Roos (Helsinki University of Technology, Finland),
- Prof. Wil Schilders (TU Eindhoven & NXP Semiconductors, The Netherlands),
- Prof. Thomas Weiland (TU Darmstadt & CST, Germany).

The Program Committee selected and invited, for each of the four main topics, (at least) one speaker from academia and one from industry. Thus, SCEE 2008 was honoured by the presence of the following 10 invited speakers:

- Dr. Sergey Yuferev (Nokia, Finland):
  "Challenges and Approaches in EMC/EMI Modeling of Wireless Devices"
- Dr. Emira Dautbegovic (Qimonda, Germany):
  "Wavelets in Circuit Simulation"
- Prof. Qi-Jun Zhang (Carleton University, Canada):
  "ANN/DNN-Based Behavioral Modeling of RF/Microwave Components and Circuits"
- Prof. Ansgar Jüngel (TU Wien, Austria):
  "Thermal Effects in Coupled Circuit–Device Simulations"
- Dr. Wim Schoenmaker (MAGWEL, Belgium):
  "Evaluation of the Electromagnetic Coupling Between Microelectronic Device Structures Using Computational Electrodynamics"
- Prof. Daniel Ioan (Politehnica University of Bucharest, Romania):
  "Parametric Reduced-Order Models for Passive Integrated Components Coupled with their EM Environment"
- Dr. Galina Benderskaya (CST, Germany):
  "Numerical Time Integration in Computational Electromagnetics"
- Dr. David Levadoux (ONERA, France):
  "New Trends in the Preconditioning of Integral Equations of Electromagnetism"
- Prof. Jan Hesthaven (Brown University, USA):
  "High-Order Discontinuous Galerkin Methods for Computational Electromagnetics and Uncertainty Quantification"
- Prof. Peter Benner (TU Chemnitz, Germany):
  "Advances in Balancing-Related Model Reduction for Circuit Simulation"

There were 95 participants from as many as 18 countries with altogether 84 abstracts accepted to be presented at the conference: 10 invited talks, 32 contributed talks, and 42 posters.

The 65 papers appearing in this SCEE 2008 post-conference book were selected from the 84 abstracts presented at the conference through a second review round. Each paper was carefully reviewed by two experts in this second round as well. The overall review process was coordinated by the editors of this book and supervised by the Program Committee.

The selected 65 papers are organized under the four main topics of the SCEE 2008 conference and an additional one on model-order reduction. Adding the fifth category seemed appropriate due to the abundance of papers on this topic. Thus, this book is divided into five parts: Computational Electromagnetics (14 papers), Circuit Simulation (15), Coupled Problems (10), Mathematical and Computational Methods (12), and Model-Order Reduction (14).

Thanks to the suggestion of the Editorial Board of the Springer series "Mathematics in Industry", the readability of this book has been improved by writing a short introduction on the scientific rationale of the entire book and writing an umbrella-like introduction for each of the five parts. An attempt has also been made to arrange the papers within each part to create a relatively smooth transition from one paper to the next.

This SCEE 2008 post-conference book is dedicated to the memory of Professor Angelo Marcello Anile, who was an active member of the SCEE community. His close friends and colleagues Dr. G. Alì, Prof. M. Günther, and Prof. V. Romano remember him in the in memoriam that follows this preface.

On behalf of the SCEE 2008 Local Organizing Committee, we would like to thank all the authors for submitting high-quality papers, the reviewers for their hard and exacting work within a tight schedule, the members of the Program Committee for their efforts and time, and those of the Scientific Advisory Committee for their support. Also, we are grateful to the financial and material support received from our sponsors.

Espoo,
December 2009

*Janne Roos*
*Luis R.J. Costa*

# Angelo Marcello Anile In Memoriam

**3.1.1948–16.11.2007**

We wish to express our deep regret that SCEE has to go on without its Program Committee member over many years, Professor Angelo Marcello Anile from Catania, Sicily, Italy. Marcello passed away on Friday, 16th November, 2007. With the passing away of Marcello, we lost a colleague respected throughout the field of applied mathematics both in academia and industry, an inspiring scientist, and a dear friend. Without his enthusiasm, his devotion to interdisciplinary research on a European scale, his mediative attitude and extensive knowledge, international activities such as the biannual SCEE conferences would never have been the success they are. We will always treasure the fond memories of Marcello the scientist, colleague, and dear friend.

## Marcello — the scientist

Angelo Marcello Anile graduated from the Scuola Normale of Pisa, Italy, in 1971 and got his Ph.D. at the Oxford University under the supervision of Prof. Dennis Sciama, one of the best known astrophysicists in those years. Marcello had a brilliant academic career, becoming full professor of mathematical physics in 1980 and one of the most internationally representative researchers of the field. He started his research activities studying problems arising from general relativity and, in particular, relativistic fluids with applications to astrophysics and cosmology. A comprehensive review of his work on relativistic waves is given in the monograph "Relativistic fluids and magnetofluids", published by the Cambridge University Press (1989). Later, his interests turned towards applied mathematics and he started collaborations with companies, in particular STMicroelectronics (the former SGS-Thomson), becoming a leader in the field of industrial mathematics. His interest was in trying to solve problems arising from real industrial cases. He focused his attention on mathematical models for charge transport in semiconductors and nanoscale devices. In the last years before his death, Marcello also worked on fuzzy logic, neural networks, and optimization.

He was one of the founders of the SCEE series of conferences and a member of either the SCEE Program Committee or the Scientific Advisory Committee over the years. He organized the SCEE 2004 conference and the Summer School in 2005, both at Capo D'Orlando, Sicily. He was also a member of the ECMI council and one of the organizers of ECMI 2000, the 11th ECMI conference, at Palermo, Italy. He was one of the inspiring fathers of the EU-FP6 Research Training Network on Coupled Multiscale Simulation and Optimization in Nanoelectronics (CoMSON), linking his home university Catania with four other European universities and three leading European semiconductor companies. Before his untimely death, he founded in Catania a center for applied mathematics jointly with the Fraunhofer Institute of Kaiserslautern and the University of Florence. At the University of Catania Marcello created a research group with several collaborators from and partnerships with other Italian and foreign universities. Many of his Ph.D. students now have permanent university positions. Marcello's idea of research embraced interdisciplinary activity where mathematics plays a fundamental role allowing modeling and simulation of challenging technological problems.

## Marcello — a colleague and a friend

Marcello was not only a man of science, but had a polyhedric passion for culture. He was attracted by all kinds of arts: painting, music, literature. He studied piano and knew very well the English literature, with a special predilection for James Joyce and his 'stream of consciousness'. During his travels around the world at conferences and meetings for scientific projects he investigated the social and economic features of the visited places and tried to acquaint himself with the local customs.

The book “The travelling mathematician”, a brief biography and collection of images and words published by the Angelo Marcello Anile Association, is proof of his deep interest in culture.

All the authors of this note are deeply indebted to Marcello for his influence at the most critical moment of their career as researchers, that is during their years of youth and formation.

One of the authors (G.A.) still recalls vividly the memory of the first lecture on rational mechanics given by Marcello during his second year at university. That lecture made it clear to the young student that he was in front of a master. And, later on, he would decide to follow that master rather than devote himself to those topics which were his initial motivation for starting a university course in mathematics.

Another of the authors (V.R.) remembers with a special emotion his participation in his first conferences during his doctoral studies under Marcello’s supervision, particularly the international conference held in La Scuola Internazionale Superiore di Studi Avanzati (SISSA), in Trieste, Italy, dedicated to the 65th birthday of Prof. Dennis Sciama. The foremost experts in the field of general relativity were present on that occasion. He still remembers the stimulating atmosphere, the surreal presence of Steven Hawking, and the depth of the talk delivered by Roger Penrose and by the other speakers. This episode well represents the sense of wonder that accompanied most young researchers close to Marcello. He was, for them, a sort of window to the world as well as a naturally authoritative person.

At the same time, he was friendly with everybody; he displayed a deep familiarity, irrespective of roles or titles. One of the authors (G.A.) remembers the day after his admission to the Ph.D. course in applied mathematics when encountering Marcello on the stairway of the department of mathematics. He greeted him, “Buongiorno Professore,” and his reply was, “You are a Ph.D. student now, you can call me Marcello.”

This benevolence and friendliness was directed not only to his closest collaborators, but it was a general trait that characterized Marcello. One of the authors (M.G.) remembers well that it was Marcello who introduced him as a young Ph.D. student into the European scale of research and industrial mathematics. He also loves remembering Marcello’s hospitality and pride when he showed him the beauty of Sicily and its long and rich history.

With the loss of Marcello we have lost not only a master, but also a colleague and a friend.

May 2009

*Giuseppe Alì*[1]
*Michael Günther*[2]
*Vittorio Romano*[3]

[1] Università della Calabria, Italy.

[2] Bergische Universität Wuppertal, Germany.

[3] Università di Catania, Italy.

# Contents

**Part II Circuit Simulation**

**Part III Coupled Problems**

**Part IV Mathematical and Computational Methods**

## Part V Model-Order Reduction

# List of Contributors

**Wolfgang Ackermann**
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, Schlossgartenstraße 8, 64289 Darmstadt, Germany
ackermann@temf.tu-darmstadt.de

**Giuseppe Alì**
Dipartimento di Matematica, Università della Calabria, via Ponte P. Bucci 30/B, 87036 Arcavacata di Rende (CS), Italy
giuseppe.ali@unical.it
g.ali@mat.unical.it

**Björn Andersson**
FCC–Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Göteborg, Sweden
bjorn.andersson@fcc.chalmers.se

**Athanasios C. Antoulas**
Department of Electrical and Computer Engineering, MS-380, William Marsh Rice University, P.O. Box 1892, Houston, TX 77251-1892, USA
aca@rice.edu

**Ashish Awasthi**
University of Applied Science of Upper Austria, Hagenberg, Austria
aawasthi@fh-hagenberg.at

**Christian Rüdiger Bahls**
Universität Rostock, A.-Einstein-Str. 2, 18051 Rostock, Germany
christian.bahls@uni-rostock.de

**Bastian Bandlow**
FG Theoretische Elektrotechnik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany
bandlow@tet.upb.de

**Andreas Bartel**
Bergische Universität Wuppertal, Gaußstrasse 20, 42119 Wuppertal, Germany
bartel@math.uni-wuppertal.de

**S. Baumanns**
Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Cologne, Germany
sbaumann@math.uni-koeln.de

**F. van Belzen**
Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
f.v.belzen@tue.nl

**Galina Benderskaya**
CST - Computer Simulation Technology AG, Bad Nauheimer Straße 19, 64289 Darmstadt, Germany
galina.benderskaya@cst.com

**Peter Benner**
Mathematik in Industrie und Technik, Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany,
benner@mathematik.tu-chemnitz.de

**M.C. van Beurden**
Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands
m.c.v.beurden@tue.nl

**Andreas Blaszczyk**
ABB Corporate Research, 5405 Baden, Switzerland
andreas.blaszczyk@ch.abb.com

**Hans Georg Brachtendorf**
University of Applied Science of Upper Austria, Hagenberg, Austria
brachtd@fh-hagenberg.at

**Markus Brunk**
Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway
markus.brunk@math.ntnu.no

**Angelika Bunse-Gerstner**
University of Bremen, Bremen, Germany

**C. Chauviere**
Laboratoire de Mathématiques, Université Blaise Pascal, 63177 Aubière, France
cedric.chauviere@math.univ-bpclermont.fr

**Carlos Christoffersen**
Department of Electrical Engineering, Lakehead University, 955 Oliver Road, Thunder Bay, ON, Canada P7B 5E1
c.christoffersen@ieee.org

**Gabriela Ciuprina**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
gabriela@lmn.pub.ro

**Marissa Condon**
RF Modeling and Simulation Group, Research Institute for Networks and Communications Engineering (RINCE), School of Electronic Engineering, Dublin City University, Dublin 9, Ireland
marissa.condon@dcu.ie

**Luis R.J. Costa**
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland
luis.costa@tkk.fi

**Jeroen Croon**
NXP-TSMC Research Center, High Tech Campus 37, PostBox WY4-01, 5656 AE Eindhoven, The Netherlands
jeroen.croon@nxp.com

**Massimiliano Culpo**
Bergische Universität Wuppertal, Gaußstrasse 20, 42119 Wuppertal, Germany
culpo@math.uni-wuppertal.de

**Hasan Dağ**
Information Technologies Department, Kadir Has University, Istanbul, Turkey
hasan.dag@khas.edu.tr

**Emira Dautbegovic**
Qimonda, 81726 Munich, Germany
emira.dautbegovic@qimonda.com

**Georg Denk**
Qimonda AG Munich, Am Campeon 1-12, 85579 Munich, Germany
georg.denk@qimonda.com

**Patrick Dewilde**
Circuits and Systems, EEMCS, TUDelft, Delft, The Netherlands
p.dewilde@ewi.tudelft.nl

**Tom Dhaene**
Department of Information Technology (INTEC), Ghent University - IBBT, Gaston Crommenlaan 8 Bus 201, 9050 Ghent, Belgium
tom.dhaene@ugent.be

**Michal Dobrzynski**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
d_michal@lmn.pub.ro

**C.R. Drago**
Dipartimento di Matematica e Informatica, Università di Catania, Viale Andrea Doria 6, 95125 Catania, Italy
drago@dmi.unict.it

**Fredrik Edelvik**
FCC–Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Göteborg, Sweden
fredrik.edelvik@fcc.chalmers.se

**Carlo de Falco**
Dublin City University, Glasnevin, Dublin 9, Ireland
carlo.defalco@dcu.ie

**Uwe Feldmann**
HiSIM Research Center, Hiroshima-University, 1-3-1 Kagamiyama, Higashi-Hiroshima 739-0046, Japan
uwe.feldmann@online.de

**Lihong Feng**
Mathematics in Industry and Technology, Faculty of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany
lihong.feng@mathematik.tu-chemnitz.de

**Herbert De Gersem**
KU Leuven, Leuven, Belgium
herbert.degersem@kuleuven-kortrijk.be

**Sebastián Gim**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
seb@lmn.pub.ro

**Erion Gjonaj**
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, 64289 Darmstadt, Germany
gjonaj@temf.tu-darmstadt.de

**Dirk Gorissen**
Department of Information Technology (INTEC), Ghent University - IBBT, Gaston Crommenlaan 8 Bus 201, 9050 Ghent, Belgium
dirk.gorissen@ugent.be

**M.M. Gourary**
IPPM RAS, 3 Sovetskaya, Moscow, Russia
gourary@ippm.ru

**Georgi G. Grahovski**
School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland
Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, 72 Tsarigradsko chaussée, 1784 Sofia, Bulgaria
grah@eeng.dcu.ie

**Michael Günther**
BU Wuppertal, Wuppertal, Germany
guenther@math.uni-wuppertal.de

**D. Harutyunyan**
Technische Universiteit Eindhoven, Den Dolech 2, 5612 AZ Eindhoven, The Netherlands
NXP Semiconductors, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands
d.harutyunyan@tue.nl

**Magnus Herberthson**
Department of Sensor Informatics, Swedish Defence Research Agency, Box 1165, SE-581 11 Linköping, Sweden
Department of Mathematics, Linköpings Universitet, 581 83 Linköping, Sweden
magnus.herberthson@foi.se
maher@mai.liu.se

**J.S. Hesthaven**
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
jan.hesthaven@brown.edu

**Mikko Honkala**
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland
mikko.a.honkala@tkk.fi

**Hsuan-Ming Huang**
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu 300, Taiwan

**Chih-Hong Hwang**
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu 300, Taiwan

**Daniel Ioan**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
lmn@lmn.pub.ro

**Roxana Ionutiu**
Department of Electrical and Computer Engineering, MS-380, William Marsh Rice University, P.O. Box 1892, Houston, TX 77251-1892, USA
School of Engineering and Science, Jacobs University Bremen, 28725 Bremen, Germany
roxana.ionutiu@rice.edu
r.ionutiu@jacobs-university.de

**Satoru Iwata**
Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan
iwata@kurims.kyoto-u.ac.jp

**Stefan Jakobsson**
FCC–Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Göteborg, Sweden
stefan.jakobsson@fcc.chalmers.se

**Ansgar Jüngel**
Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria
juengel@anum.tuwien.ac.at

**Takahiro Kajiwara**
HiSIM Research Center, Hiroshima-University, 1-3-1 Kagamiyama, Higashi-Hiroshima 739-0046, Japan
ancient-future@hiroshima-u.ac.jp

**Sebastian Kula**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
sebastia@lmn.pub.ro

**D.J.P. Lahaye**
Delft Institute of Applied Mathematics (DIAM), Technical University of Delft, Mekelweg 4, Delft, The Netherlands
d.j.p.lahaye@tudelft.nl

**Siegmar Lampe**
University of Bremen, Bremen, Germany

**Barbara Lang**
University of Bremen, Bremen, Germany

**Thomas Lau**
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, 64289 Darmstadt, Germany
lau@temf.tu-darmstadt.de

**Sanda Lefteriu**
Rice University, MS-366, 6100 Main Street, Houston, TX 77005, USA
Sanda.Lefteriu@rice.edu

**Peter Lenaers**
Mathematics for Industry, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
plenaers@gmail.com

**David P. Levadoux**
ONERA, Chemin de la Hunière, 91761 Palaiseau, France
Mathematics Laboratory, University Paris XI, 91405 Orsay, France
david.levadoux@onera.fr

**Yiming Li**
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu 300, Taiwan
ymli@faculty.nctu.edu.tw

**Giovanni Mascali**
Dipartimento di Matematica, Università della Calabria, and INFN-Gruppo c. Cosenza, via P. Bucci 30/B, 87036 Cosenza, Italy
g.mascali@unical.it

**E. Jan ter Maten**
NXP Semiconductors, Eindhoven, The Netherlands
jan.ter.maten@nxp.com

**Wolfgang Mathis**
Institut für Theoretische Elektrotechnik, Leibniz Universität Hannover, Appelstr. 9A, 30167 Hannover, Germany
mathis@tet.uni-hannover.de

**Peter Meuris**
MAGWEL NV, Martelarenplein 13, 3000 Leuven, Belgium
peter.meuris@magwel.com

**B.L. Michielsen**
ONERA, 2, av E. Belin, 31055 Toulouse Cedex, France
bastiaan.michielsen@onera.fr

**Pekka Miettinen**
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland
pekka.miettinen@tkk.fi

**Diana Mihalache**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
mihaladi@lmn.pub.ro

**Florence Millot**
CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse, France

**Mitiko Miura-Mattausch**
HiSIM Research Center, Hiroshima-University, 1-3-1 Kagamiyama, Higashi-Hiroshima 739-0046, Japan
mmm@hiroshima-u.ac.jp

**Masataka Miyake**
HiSIM Research Center, Hiroshima-University,
1-3-1 Kagamiyama, Higashi-Hiroshima
739-0046, Japan
miyake-m053223@hiroshima-u.ac.jp

**Kasra Mohaghegh**
Bergische Universität Wuppertal, Wuppertal,
Germany
mohaghegh@math.uni-wuppertal.de

**B.J. Mulvaney**
Freescale Semiconductor Inc., 7700 W. Parmer
Lane, Austin, TX, USA
brian.mulvaney@freescale.com

**Claus-Dieter Munz**
Institut für Aerodynamik und Gasdynamik,
Universität Stuttgart, Stuttgart, Germany
munz@iag.uni-stuttgart.de

**Carsten Neff**
NEC Laboratories Europe, Rathausallee 10,
53757 St. Augustin, Germany
neff@it.neclab.eu

**Marko Neitola**
University of Oulu, Oulu, Finland
marko.neitola@ee.oulu.fi

**Keijo Nikoskinen**
Department of Radio Science and Engineering,
Faculty of Electronics, Communications
and Automation, Helsinki University of
Technology, P.O. Box 3000, FI-02015 TKK,
Finland
keijo.nikoskinen@tkk.fi

**Clemens Pechstein**
Institute of Computational Mathematics,
Johannes Kepler University, Altenberger
Str. 69, 4040 Linz, Austria
clemens.pechstein@numa.uni-linz.ac.at,

**Sébastien Pernet**
CERFACS, 42 Avenue Gaspard Coriolis,
31057 Toulouse, France
sebastien.pernet@cerfacs.fr

**Walter Pflanzl**
austriamicrosystems AG, Schloss Premstätten,
8141 Unterpremstätten, Austria
walter.pflanzl@austriamicrosystems.com

**Jagoda Plata**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of
Bucharest, Spl. Independenţei 313, 060042
Bucharest, Romania
plata@lmn.pub.ro

**Jan Pomplun**
Zuse Institute Berlin, Takustrasse 7, 14195
Berlin, Germany
pomplun@zib.de

**Gisela Pöplau**
Universität Rostock, A.-Einstein-Str. 2, 18051
Rostock, Germany,

**Thomas Preisner**
Institut für Theoretische Elektrotechnik,
Leibniz Universität Hannover, Appelstr. 9A,
30167 Hannover, Germany
preisner@tet.uni-hannover.de

**Roland Pulch**
Lehrstuhl für Angewandte Mathematik und
Numerische Mathematik (Chair of Applied
Mathematics and Numerical Analysis,
Department of Mathematics and Sciences),
Bergische Universität Wuppertal, Gaußstr. 20,
42119 Wuppertal, Germany
pulch@math.uni-wuppertal.de

**Martin Quandt**
Institut für Aerodynamik und Gasdynamik,
Universität Stuttgart, Stuttgart, Germany
quandt@iag.uni-stuttgart.de

**Timo Rahkonen**
Department of Electrical and Information
Engineering and Infotech Oulu, University of
Oulu, Oulu, Finland
timo.rahkonen@oulu.fi

**Timo Reis**
Institut für Mathematik, MA 4-5, TU Berlin,
Straße des 17. Juni 136, 10623 Berlin, Germany
reis@math.tu-berlin.de

**Rob Remis**
EM-lab, EWI, TUDelft, Delft, The Netherlands
r.f.remis@ewi.tudelft.nl

**Ursula van Rienen**
Universität Rostock, A.-Einstein-Str. 2, 18051
Rostock, Germany

**Vittorio Romano**
Dipartimento di Matematica e Informatica (Department of Mathematics and Computer Science), Università di Catania, viale A. Doria 6, 95125 Catania, Italy
romano@dmi.unict.it

**Werner Römisch**
Inst. of Mathematics, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
romisch@math.hu-berlin.de

**Joost Rommes**
NXP Semiconductors, Corporate I&T/DTF, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands
joost.rommes@nxp.com

**Janne Roos**
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland
janne.roos@tkk.fi

**Salvatore La Rosa**
Dipartimento di Matematica e Informatica, Università di Catania, viale A. Doria 6, 95125 Catania, Italy
larosa@dmi.unict.it

**S.G. Rusakov**
IPPM RAS, 3 Sovetskaya, Moscow, Russia
rusakov@ippm.ru

**Robert Scheichl**
Dept. of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK
masrs@bath.ac.uk

**W.H.A. Schilders**
Technische Universiteit Eindhoven, Den Dolech 2, 5612 AZ Eindhoven, The Netherlands
NXP Semiconductors, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands
wil.schilders@nxp.com

**Frank Schmidt**
Zuse Institute Berlin, Takustrasse 7, 14195 Berlin, Germany
frank.schmidt@zib.de

**André Schneider**
Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany
andre.schneider@mathematik.tu-chemnitz.de

**Rudolf Schneider**
Forschungszentrum Karlsruhe, Institut für Hochleistungsimpuls und Mikrowellentechnik, Karlsruhe, Germany
rudolf.schneider@ihm.fzk.de

**Wim Schoenmaker**
Magwel NV, Martelarenplein 13, 3000 Leuven, Belgium
wim.schoenmaker@magwel.com

**Sebastian Schöps**
BU Wuppertal, Wuppertal, Germany
schoeps@math.uni-wuppertal.de

**Rolf Schuhmann**
FG Theoretische Elektrotechnik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany
schuhmann@tet.upb.de

**C. Scordia**
University of Wuppertal, Wuppertal, Germany
scordia@dmi.unict.it

**Ehrenfried Seebacher**
austriamicrosystems, 8141 Schloss Premstaetten, Austria
ehrenfried.seebacher@austriamicrosystems.com

**M. Selva Soto**
Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Cologne, Germany
mselva@math.uni-koeln.de

**N. Serap Şengör**
Electrical and Electronics Engineering Faculty, Istanbul Technical University, Maslak, Istanbul, Turkey
sengorn@itu.edu.tr

**Zhifeng Sheng**
Circuits and Systems, EEMCS, TUDelft, Delft, The Netherlands
z.sheng@ewi.tudelft.nl

**Thorsten Sickenberger**
Dept. of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, UK
t.sickenberger@hw.ac.uk

**Murat Şimsek**
Electrical and Electronics Engineering Faculty, Istanbul Technical University, Maslak, Istanbul, Turkey
simsekmu@itu.edu.tr

**Alexandra Stefanescu**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
alexar@lmn.pub.ro

**Alexander Steinmair**
austriamicrosystems AG, Schloss Premstätten, 8141 Unterpremstätten, Austria
alexander.steinmair@austriamicrosystems.com

**Oliver Sterz**
CST - Computer Simulation Technology AG, Bad Nauheimer Straße 19, 64289 Darmstadt, Germany
oliver.sterz@cst.com

**Michael Striebel**
Technische Universität Chemnitz, Chemnitz, Germany
michael.striebel@mathematik.tu-chemnitz.de

**Tatjana Stykel**
Institut für Mathematik, MA 4-5, TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
stykel@math.tu-berlin.de

**O.O. Sy**
Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands
o.o.sy@tue.nl

**Mizuyo Takamatsu**
Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan
mizuyo_takamatsu@mist.i.u-tokyo.ac.jp

**A.G. Tijhuis**
Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands
a.g.tijhuis@tue.nl

**Caren Tischendorf**
Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Cologne, Germany
tischendorf@math.uni-koeln.de

**Luciano De Tommasi**
Ghent University - IBBT, Department of Information Technology (INTEC), Gaston Crommenlaan 8 Bus 201, 9050 Ghent, Belgium
luciano.detommasi@ua.ac.be

**M.V. Ugryumova**
Eindhoven University of Technology, Den Dolech 2 Postbus 513, 5600 MB Eindhoven, The Netherlands
m.v.ugryumova@tue.nl

**S.L. Ulyanov**
IPPM RAS, 3 Sovetskaya, Moscow, Russia

**J.A.H.M. Vaessen**
Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands
j.a.h.m.vaessen@tue.nl

**Martti Valtonen**
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland
martti.valtonen@tkk.fi

**Alexander Vasenev**
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
vasenev@lmn.pub.ro

**Arie Verhoeven**
VORtech Computing, Delft, The Netherlands
arie.verhoeven@na-net.ornl.gov

**Steffen Voigtmann**
Qimonda AG Munich, Am Campeon 1-12, 85579 Munich, Germany
steffen.voigtmann@qimonda.com

**T. Warburton**
Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA
timwar@rice.edu

**S. Weiland**
Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
s.weiland@tue.nl

**Thomas Weiland**
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, Schlossgartenstraße 8, 64289 Darmstadt, Germany
thomas.weiland@temf.tu-darmstadt.de

**L. Wilcox**
Institute for Computational Engineering and Sciences (ICES), University of Texas at Austin, Austin, TX 78712, USA
lucasw@ices.utexas.edu

**Renate Winkler**
Dept. of Mathematics, Bergische Universität Wuppertal, 42199 Wuppertal, Germany
winkler@math.uni-wuppertal.de

**Tao Xu**
RF Modeling and Simulation Group, Research Institute for Networks and Communications Engineering (RINCE), School of Electronic Engineering, Dublin City University, Dublin 9, Ireland
taoxu@eeng.dcu.ie

**E. Fatih Yetkin**
Informatics Institute, Computational Science and Engineering Program, Istanbul Technical University, Istanbul, Turkey
fatih@be.itu.edu.tr

**Sergey Yuferev**
Nokia Corp., P.O. Box 1000, FI-33721 Tampere, Finland
sergey.yuferev@nokia.com

**Lei Zhang**
Department of Electronics, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K2C 1N5
leizhang@doe.carleton.ca

**Q.J. Zhang**
Department of Electronics, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K2C 1N5
qjz@doe.carleton.ca

**M.M. Zharov**
IPPM RAS, 3 Sovetskaya, Moscow, Russia

# Introduction

Janne Roos

This post-conference book contains 65 accepted full papers of the 7th International Conference on Scientific Computing in Electrical Engineering (SCEE 2008).

*Scientific computing* is a field of study concerned with the following kind of steps: 1) formulate a mathematical model of the object/phenomenon being studied, 2) develop a numerical algorithm for analyzing/solving the mathematical model, 3) implement the algorithm in the form of a computer program, 4) simulate the model by running the computer program, and 5) analyze the results. In real life, as also in the papers of this book, the main focus at a time may be on one or more steps. Also, these steps are seldom carried out from scratch; for example, one may use (parts of) an existing computer program rather than implementing a new one. Finally, note that scientific computing, which has considerably contributed to our current (material) well-being, is used in various scientific disciplines like, e.g., biology, economy, social sciences, and engineering.

*Electrical engineering*, which in this book also includes electronic engineering, is a field of engineering that deals with the study and application of electricity, electromagnetism, and electronics. Nowadays, electrical engineering covers a range of subtopics including, e.g., antennas, electrical machines, control systems, circuit design/modeling/simulation, signal processing, and telecommunications.

*Scientific computing in electrical engineering* could include, based on the above discussion, a very large spectrum of different topics. Only a part of this wide spectrum was covered at the SCEE 2008 conference, reflecting the current interests of the “SCEE community”. So, this book contains the following five parts: I. Computational Electromagnetics, II. Circuit Simulation, III. Coupled Problems, IV. Mathematical and Computational Methods, and V. Model-Order Reduction. Each of these five parts consists of an introduction followed by the actual papers.

Janne Roos
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland, e-mail: janne.roos@tkk.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_1,

# Part I
# Computational Electromagnetics

# Introduction to Part I

Gabriela Ciuprina

Simulations are indispensable to industry nowadays. They allow one to perform virtual experiments which are faster and cheaper than the physical ones. Simulation environments are created or improved on the basis of numerical methods applied to solve specific problems. Electrical and electronics engineering need at least the computation of the electromagnetic (EM) field. For instance, in electronics, the increase of the operating frequency makes that effects specific to the EM field, and neglected until now, be relevant. Consequently, on the one hand, the development of a new approach based on the EM field computation is needed, since the old techniques, based mainly on circuits, do not correspond to the necessities any more. On the other hand, solely the numerical approach of obtaining the field cannot be effective, due to the enormous computational resources needed. Thus, in spite of the commercial advertisements, software tools' functionality is low at high frequencies reaching 60 GHz. Even if the general software EM field packages solve many of the designers' problems, they need appropriate solutions that are not offered yet by the present available software commercial tools.

Hence, research related to *computational electromagnetics* still has to be done and the 14 papers in this part address some of the current problems related to electromagnetic compatibility (EMC), the use of various discretization methods such as boundary-element method (BEM), the finite-element method (FEM) with its discontinuous Galerkin variant, the finite-difference time-domain (FDTD) method, the finite integration technique (FIT), the surface-integrated field equation (SIFE), and the computation of eigenvalues. Special cases are envisaged, such as the consideration of dispersive materials or high-contrast materials, magnetic-force microscopy, treatment of multistage problems, computation of phase-space coordinates of plasma particles, and even the modelling of the relativistic movement of particles. Both time-domain and frequency-domain cases can be found in this collection of papers. For validation, some papers use comparisons with well-known

Gabriela Ciuprina
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independentei 313, 060042 Bucharest, Romania, e-mail: gabriela@lmn.pub.ro

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_2,

software tools (like CST and APLAC), or use simple test problems with known analytical solutions, or, even better, they compare the numerical results with the results obtained from physical experiments. A brief description of these papers follows.

The invited paper by Yuferev focuses on the numerical modelling of high frequency EMC problems. The author speaks from his position in a well-known company that manufactures complex mobile phones that incorporate other products such as cameras, sensors, computers, etc. For such devices, EMC regulations have to be strictly observed. The paper addresses two tasks. The first task is to overcome the computational difficulties caused by multi-scale features. The solution proposed consists of an iterative procedure, each iteration involving two steps, the first of which ignores the device and computes the near-field radiation of the printed circuit board (PCB) and the second which analyzes the field inside the device without the PCB, the latter being replaced by a certain field distribution at the interface. The procedure is demonstrated by using the boundary integral equation method. The second task addressed in the paper refers to obtaining absolute values of the magnetic field emissions radiated from PCBs. This is carried out by calibrating the parameters of the EM source in the numerical models for EM computations by using measured data. Two practical applications of both techniques are presented. One of them is the investigation of a typical radiation from a shield located on the PCB. Another example refers to the computation of the magnetic field around a voltage-controlled oscillator, well-known in EMC design as a source of radiation emission requiring to be reduced.

When measurement data, such as scattering parameters, are available, low complexity macromodels of EM devices, such as chips, packages, or boards, can be obtained by vector fitting (VF) or by a new method proposed by Lefteriu and Antoulas. The authors use a Loewner matrix pencil constructed in the context of tangential interpolation. It uses no heuristics, but only available data, makes no assumption of the underlying system, and allows to identify a system if enough measurements are provided. The approach is especially suited for devices with a large number of ports and it proved to be successful for problems for which the VF method is not. Two tests are shown. The first one is a theoretical system for which the proposed method was able to identify the original system whereas VF was not. The second starts from real measurements performed using a vector network analyzer for which the proposed method was more efficient than VF, considering the order of the system obtained and the required computational time.

Stefanescu et al. also discuss the modelling required by integrated circuits. In this case, interconnects modelled as transmission lines are considered, the contribution of the paper being the inclusion of the parameterization which is useful when variability with respect to the geometric parameters has to be modeled. The parametric models are based on the computation of first-order sensitivities of line parameters. Three multiparametric models are proposed, called additive, rational, and multiparameteric. For the one-parameter case, the proposed method avoids the evaluation of higher-order sensitivities. The multiparametric model is based on the assumption that the quantity of interest can be expressed with separated variables. It can be a better choice than the use of traditional models based on first-order Taylor series truncation. The case study used is a microstrip transmission line for which measurements are also available.

A procedure to obtain models for 3-D passive integrated components that take into consideration the variability is proposed in the paper by Ciuprina et al. The paper uses the electromagnetic circuit element (EMCE) formulation proposed previously by the authors. The contribution of the paper is that it describes how parameterization can be taken into consideration for the EMCE formulation. A specific section is dedicated to the generation of the parametric semi-state space system when FIT is used. Two ways of interconnecting the models obtained, one based on the system matrices and the other on transfer matrices, are described. The approach is validated for a two-coupled-inductors configuration for which measurements are available. The advantage of this approach is that it bears an inherent parallelism, the sub-models can be treated independently both from the point of view of the variability and EM-field formulation.

The paper by Vasenev et al. describes a graphical-based tool for the extraction of magnetic reluctances between on-chip current loops. They are useful to build magnetic circuits that can be connected to the magnetic terminals of devices that can be modelled with the EMCE formulation to model inductive coupling between components or between components and the environment. The approach is a comprehensive multi-scale modelling solution using domain decomposition, hierarchical substrate structuring, and compact parametric models to model passive integrated structures and functional blocks and the electric and magnetic parasitic interactions between them.

Costa et al. look at a way to implement stable nonlinear lumped elements (LE), like a diode, for use in an FDTD-based EM simulation. The embedded linear or nonlinear LE FDTD models may span multiple cells of the 3-D FDTD grid. This is useful to correctly model sources and loads within complex electronic subsystems. The technique does not increase the complexity of the desired nonlinear model and has a high operational stability, as demonstrated for a diode working far beyond normal operational voltages. The simulation results are validated with the circuit simulator APLAC.

The paper by Bandlow and Schuhmann presents a formulation based on FIT to handle EM eigenvalue problems from structures containing frequency-dispersive materials. This is useful when the operating frequency corresponds to the infrared spectrum case for which noble metals do not act like perfect conductors. The problem obtained by using the Drude dispersion model is solved with the Jacobi–Davidson method. As an example, a unit-cell structure from the literature is used, the geometric modelling being carried out with the commercial tool CST Microwave Studio.

The paper by Blaszczyk proposes a new BEM-like approach applied for the calculation of the electric field in arrangements with extreme differences in material properties. The method is called “region oriented” because the space is divided into three types of regions, each region including a homogeneous, linear, and isotropic material, the field being calculated by means of layers of charge located on the region boundaries. The method is tested for a simplified geometry of a surge-arrester, its convergence being faster than the traditional BEM.

Sheng et al. describe in their paper how SIFE is implemented to solve 3-D time-domain EM problems on substrates in which high-contrast materials occur. To satisfy the partial-continuity conditions on the material interfaces, a discretization scheme is built that meets the continuity requirements across interfaces exactly, using a tetrahedron mesh combined with a consistent linear interpolation of electric and magnetic field strengths. The advantages of the method are its high flexibility and accuracy for a given discretization level. This is obtained at the cost of high computational complexity. The numerical validation is done for a problem with an analytical solution.

The paper by Pomplun and Schmidt uses the reduced basis (RB) method applied to EM field computation with FEM for the simulation of light scattering from geometrically parameterized phase shift masks. The RB method allows the splitting of the solution process into two parts, an expensive offline part in which the model is solved rigorously several times for different values of the geometrical parameters and a fast online part in which a reduced problem, obtained after projection onto the RB, is solved.

The discontinuous Galerkin FEM has recently become popular as a method for the numerical solution of partial differential equations. It is used by Bahls et al. to solve Poisson's equation on unstructured grids. For this, the Nudg++ library is used to solve a problem with analytical solution.

In the approach proposed by Preisner and Mathis, a theoretical model of magnetic force microscopy was developed to verify and improve the results of laboratory measurements. A scanning process is simulated and different force-calculation methods, based on the Maxwell stress tensor, the virtual work principle, and the local interaction forces, are implemented. The results obtained by the various methods are compared with each other in order to obtain the total magnetic force acting on a cantilever as well as local magnetic force densities.

The paper by Quandt et al. proposes a numerical method to compute phase-space coordinates of charged particles driven by the Lorentz force. The new relativistic particle-push method developed is based on a truncated Taylor series expansion up to the desired order of convergence. Both non-relativistic and relativistic test cases are in good agreement with available analytical solutions.

Finally, a statistical characterization of random EM interactions affected by resonances is presented by Sy et al. It is based on the analysis of the variance and the kurtosis to evaluate the intensity of the resonances. The analyses of these two quantities are complementary. The variance is useful in a dimensioning process as it measures the physical variation of the quantity (a voltage), whereas the kurtosis is valuable in a protection stage to foretell extreme values of the response parameter, which could damage the device under study. As an example, a randomly varying thin wire modelled by a Pocklington integral equation is used.

# Challenges and Approaches in EMC Modeling of Wireless Consumer Devices

Sergey Yuferev*

*Invited speaker at the SCEE 2008 conference

**Abstract** This paper focuses on the following key tasks in numerical modeling of high frequency EMC (electromagnetic compatibility) problems: overcoming computational difficulties caused by multi-scale features and obtaining absolute (as opposed to normalized) values of the magnetic field emissions radiated from a printed circuit board. The first task is solved by using an iterative procedure combining codes for 2.5D and 3D EM field computations. The second task is considered using the technique to approximate ("tune") parameters of the EM source in numerical models for 3D EM computations using measured data. Examples of applications of both techniques are included.

## 1 Introduction

With decreasing design cycles in modern electronic industry, simulations play a more and more important role as an alternative to the traditional way of making prototypes and measurements. High frequency wireless consumer devices such as mobile phones are becoming increasingly complex since they are actually combinations of other products such as cameras, sensors, radio, computers, etc. At the same time, they are getting smaller. So complying with EMC regulations is now a challenging task. At least some of the potential EMC problems can be predicted well before physical prototypes are built by the application of numerical analysis at an early design stage. Significant efforts have been recently made by the IEEE EMC Society to develop standards and recommended practices of the use of computational packages for the simulation of real EMC problems [1].

Sergey Yuferev
Nokia Corp., P.O. Box 1000, FI-33721 Tampere, Finland, e-mail: sergey.yuferev@nokia.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_3,

In spite of such obvious benefits, CAD simulation packages have not become yet everyday tools of EMC designers, unlike antenna, thermal or signal integrity designers of wireless devices. Even now, EMC problems of wireless devices are frequently considered as a “black magic area” where rigorous numerical analysis is impossible due to the high complexity of the problems. In the present paper some key challenges in the numerical analysis of EMC problems of wireless devices are discussed.

## 2 Tasks for Numerical Analysis of HF EMC Problems

EMC regulations are imposed as maximum acceptable values (limits) of the electromagnetic field at certain distances from the product. Limiting values of the EM field parameters are also specified inside the handset around modules containing strong EM radiators (emission interoperability limits) and modules sensitive to excessive levels of the EM field (immunity interoperability limits). By calculating the electromagnetic field distribution inside and nearby outside the module/device, it is very attractive to check compliance with interoperability limits and other EMC regulations well before prototypes are built. Obviously these limits are expressed in terms of absolute numbers and do not depend on conditions of either measurements or numerical modeling. Thus EMC simulation technology should also provide a numerical solution in the form of absolute numbers to be compared with EMC standards and interoperability limits.

A typical emission problem is optimization of the shield of the module containing RF IC (EMI source). The highest magnetic field around the module should not exceed the interoperability limit defined as equal to $x$ dBm at $y$ mm from the module. Rigorous consideration requires field–circuit co-simulations using SPICE-like models of components. With rapid progress in the development of commercial codes, such simulation workflows may become routine procedures in the near future, but today they are still considered as advanced “state of the art”. One of the reasons is that models of active components are frequently unavailable, which makes the circuit part of the workflow meaningless. In these cases source parameters are defined using reference data known a priori (for example, 2 mW and $50 Omega$) and simulations are performed using code for three-dimensional electromagnetic field computations only. This approach cannot take into account real electromagnetic behavior of the source (IC) and, therefore, cannot provide accurate computation of the magnitude of the field. Of course, if the excitation in the numerical model is assigned correctly (for example, to the PCB net carrying the highest current), numerical results will indicate accurately enough the areas of highest and lowest concentration of the electromagnetic field. However it does not answer the following question frequently posed by EMC designers: “how *high* are simulated high fields and how *low* are low fields?”. In other words, inaccurate setup of the parameters of the EM source in numerical model does not allow comparing computed fields with interoperability limits.

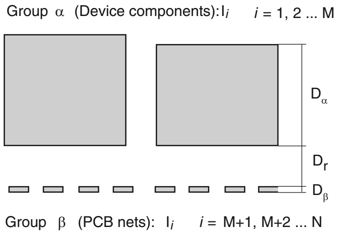

Group α (Device components): $I_i$ $i$ = 1, 2 ... M
$D_\alpha$
$D_r$
$D_\beta$
Group β (PCB nets): $I_i$ $i$ = M+1, M+2 ... N

**Fig. 1:** Two-dimensional problem containing "multi-scale" feature

Another challenge is related with the fact that the characteristic size of the module or device is between one and ten centimetres, whereas the characteristic size of the PCB details is much less than one millimetre. The difference between the characteristic dimensions is two orders of magnitude that leads to a very serious computational problem, namely to provide the necessary resolution in the simulation of the PCB details, the average size of the cell in the computational mesh used by the software for three-dimensional electromagnetic simulations should be hundreds or even tens of microns. Thus discretization of the computational space including the device by cells of this size leads to a mesh that no one modern computer will be able to simulate such a device "as is" in reasonable time. This problem having the name of "multi-scale" feature is well known in computational electromagnetics [2].

## 3 Simulation Approaches

Computational challenges related with "multi-scale" features may be overcome using so-called multi-stage modeling: the problem is divided into sub-problems, each of them is analyzed by special software, and the results are combined [3]. In the case of wireless devices like mobile phones, the basic idea of this approach is to separate the numerical solution of Maxwell's equations in the domain occupied by the printed circuit board from rest of the device. The realization of this idea can be done by iterations; each of them consists of the following two steps:

a The electromagnetic behavior of the printed circuit board is analyzed separately from the device (the device is ignored) and the near field radiation of the board is calculated. Then the electromagnetic field distribution around the board is transferred to the next iteration to be used as input data (equivalent source or boundary condition).
b The three-dimensional problem of the electromagnetic field distribution is solved everywhere inside the device except the domain occupied by the printed circuit board: the PCB domain is eliminated from the numerical procedure and replaced by the field distribution at the interface (obtained at the previous iteration).

Let us demonstrate the realization of this procedure for the boundary integral equation method (BIEM) in a simplified 2-D approximation of long parallel conductors. Consider a system of $N$ conductors classified into two groups denoted as $\alpha$ (device) and $\beta$ (PCB), and numbered from 1 to $M$ and from $M+1$ to $N$, respectively (Fig. 1). The material for all conductors is assumed to be copper. Let the characteristic scales of the conductors of the two groups be $D_\alpha$ and $D_\beta$, respectively, and the characteristic distance $\mathrm{d}_r$ between them be the following:

$$D_\alpha \approx 10^{-2} \div 10^{-3}\,\mathrm{m};\ D_\beta \approx 10^{-4} \div 10^{-5}\,\mathrm{m};\ D_r \approx 10^{-3}\,\mathrm{m} \tag{1}$$

Let an external source produce time harmonic currents $I_i$, $i = M+1, M+2, \dots N$, flowing in the conductors of the group $\beta$. The operating frequency $f$ of the source is assumed to be 3 GHz that corresponds to the wavelength $\lambda$ equal to 10 cm in free space and the skin depth $\delta$ approximately equal to 1 $\mu$m in copper. Comparison of $\lambda$ and $\delta$ with the scales in (1) enables us to apply further considerations from the quasi-static approximation (displacement current is neglected) and to apply the surface impedance boundary conditions (SIBCs) to eliminate the conducting region from the numerical procedure.

In the 2-D case the magnetic vector potential has only one component and can be considered as scalar. Since its distribution in the dielectric space separating conductors is governed by the Laplace equation, the use of BIEM employing SIBCs yields the following integral equation formulation [4, 5]:

$$A^{\mathrm{s}} - \sum_{i=1}^{N} \oint_{L_i} GK\mathrm{d}l = cF[K] + \sum_{i=1}^{N} \oint_{L_i} F[K]\frac{\partial G}{\partial \mathbf{n}}\mathrm{d}l \tag{2}$$

$$\oint_{L_i} K\mathrm{d}l = \begin{cases} I_{i,\alpha},\ 1,2,\dots,M \\ I_{i,\beta},\ M+1,\dots,N \end{cases}; \qquad G(\mathbf{r},\mathbf{r}') = -(2\pi)^{-1}\ln(|\mathbf{r}-\mathbf{r}'|) \tag{3}$$

where $A^{\mathrm{s}}$ is the source component of the magnetic vector potential, $K$ is the surface current density and $F[K]$ is the known surface impedance operator [6]. Without loss of generality in further derivations we will assume $F[K] = 0$ (perfect electrical conductor limit). We emphasize that the integration in (2)-(3) should be performed over contours of cross sections of all conductors that cause the "multi-scale" problem since $D_\beta \ll D_\alpha$. It can be avoided by transformation of (2)-(3) using iterative an procedure in such a way that the left hand side of the equations contain the integrals along conductors of only one group ($\beta$ or $\alpha$). The first two steps of the procedure are the following:

*Step 1a*

Only group $\beta$ (PCB) is considered and group $\alpha$ (rest of device) is ignored as is shown in Fig. 2.

$$A^{\mathrm{s}}_{1\mathrm{a}} - \sum_{i=M+1}^{N} \oint_{L_i} G_{\beta\beta}K_{1a}\mathrm{d}l = 0; \qquad G_{\beta\beta} = -(2\pi)^{-1}\ln|\mathbf{r}_\beta - \mathbf{r}'_\beta| \tag{4}$$

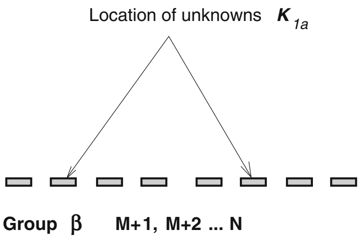


**Fig. 2:** Iterative procedure: step 1a

$$\oint_{L_i} K_{1\mathrm{a}} \mathrm{d}l = \mathrm{i}_{i,\beta}, \ \ i = M+1, M+2, \ldots, N \tag{5}$$

*Step 1b* (Fig. 3)

The distribution of unknowns $K_{1\mathrm{b}}$ over the conductors of group $\alpha$ is sought treating the distribution of $K_{1\mathrm{a}}$ over group $\beta$ as known (obtained in the previous step).

$$A^{\mathrm{s}}_{1\mathrm{b}} - \sum_{i=1}^{M} \oint_{L_i} G_{\alpha\alpha} K_{1b} \mathrm{d}l = \sum_{i=M+1}^{N} \oint_{L_i} G_{\alpha\beta} K_{1\mathrm{a}} \mathrm{d}l \tag{6}$$

$$\oint_{L_i} K_{1\mathrm{b}} \mathrm{d}l = \mathrm{i}_{i,\alpha}, \ \ i = 1, \ldots, M G_{\alpha\beta} = -(2\pi)^{-1} \ln|\mathbf{r}_\alpha - \mathbf{r}'_\beta| \tag{7}$$

This approach is frequently referred as the 2.5D-3D technique: simulation tools

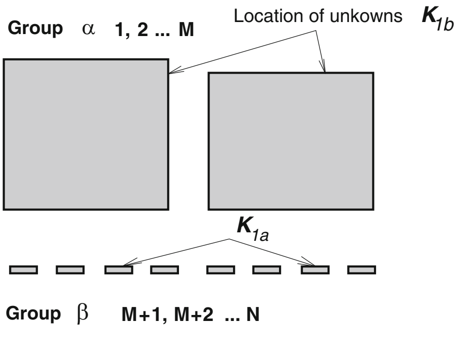


**Fig. 3:** Iterative procedure: step 1b

based on the so-called 2.5D approximation are used for PCB analysis (stage a), whereas full wave 3D EM simulators are applied at stage b. The main advantage of the described procedure is that computer grids for the PCB and device are used separately at different stages so that the multi-scale problem is resolved. However, a practical realization of the approach requires a high level of compatibility between the codes applied at steps a and b (files produced by one code should be used as the input data to the other).

It can be shown that the iterations can be interrupted after the first two steps (described above) if the following condition is met:

$$\left| \frac{D_\beta}{D_r \ln D_r} \right| \ll 1 \tag{8}$$

When detailed specifications and models needed for circuit-field co-simulations are not available, one of the most popular approaches to setup the excitation in commercial tools for 3D EM field computation is the definition of the characteristics of the current flowing through a conductor using the so-called discrete or lumped port (whose physical representation is similar to the Hertzian dipole). Usually the power ($P$) and resistance ($R$) of the port should be defined (default values are 1 W and 50 $\Omega$, respectively), but how can we know the actual values in a given case? The answer becomes clear if we have already measured the near field distribution over IC: parameters of the source in the numerical model are tuned (calibrated) in order to reach agreement between measured and computed fields. This task is an example of *inverse problems* that are widely known in nondestructive testing, geophysics, medical imaging (such as computed axial tomography), remote sensing, etc. Therefore, near field measurements can be applied at two stages of EMC analysis: for improvement of the numerical model and the final verification of compliance with interoperability limits. In the first stage the electromagnetic field is measured over the IC (if possible), whereas in the second stage measurements should be done over whole module. In both cases we want to obtain the absolute values of the field magnitudes that are related to the output data of measurements via a correction or calibration factor [7].

In many practical cases the problem is linear and advanced commercial codes allow $P$ and $R$ to vary at the post-processing stage so the procedure consists of the following steps [8]:

1. Model the real IC by a set of ports assigned to the nets carrying the highest current. This requires a priori knowledge about the general EM behavior of the module/device.
2. Run simulations with default values of the port parameters.
3. Compare the computed and measured distributions of the EM field at a specified height over the PCB and "tune" the parameters of the port to reach agreement between the measured data and the numerical results.

It is natural to ask, "will the port with parameters tuned for one distance from the IC to the observation point describe correctly the EM field at other distances?" The answer is: as long as this distance remains much larger than the characteristic size of the IC. In other words, until the Hertzian dipole approximation can be applied. Therefore, it is enough to tune the port parameters just once for a certain frequency and then apply them for the EM field computations over a wide range of distances.

In practice, however, there are restrictions related to the validity of the near field measurements.

## 4 Examples

Probably, the most typical EMC problem is undesirable radiation ("leakage") from the shield. Factors at the root of this phenomenon are not only mechanical defects of the shield construction, but also the structure of the layout under the shield. Some modern packages for 3-D EM field computations enable to import selected traces from software for the PCB design and include them into the computational model. However, the PCB ground also should be taken into consideration to provide a return path for the current flowing in the nets. Since the ground usually includes thousands traces, it is practically impossible to import it to the software for 3-D EM analysis. Thus this problem has been selected to demonstrate the application of the 2.5D-3D approach using commercial CAD packages.

Figure 4 shows the mechanical model of the shield located on the PCB. All apertures in the shield are so small that the field cannot really "leak" through them in the range of 0.5-5 GHz. So the only possible source of radiation may be the current flowing in the layout that is shown in Fig. 5. In the first step of the 2.5D-3D approach, the current distribution is calculated and the results are shown in Fig. 6 for two frequencies: 1 and 3 GHz. This distribution has been exported to the software used in the next step and used as a source in the 3-D EM field computations including the shield. Numerical results—the electric field distributions at 1 and 3 GHz—are shown in Fig. 7. It is easy to see that at 1 GHz shielding effectiveness may be poor.

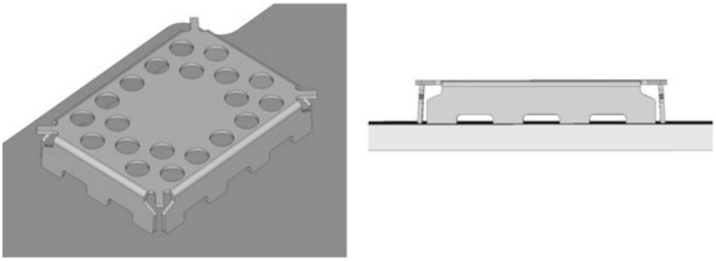


**Fig. 4:** CAD model of the shield and PCB

In practice IC components are supplied by third-party vendors, and frequently it is much easier to get results of near field measurements than the detailed specifications and models needed for circuit-field co-simulations. In such cases measured data can be used for the numerical modeling of EM behavior of the module/device and verification of compliance with interoperability limits. As an example, consider

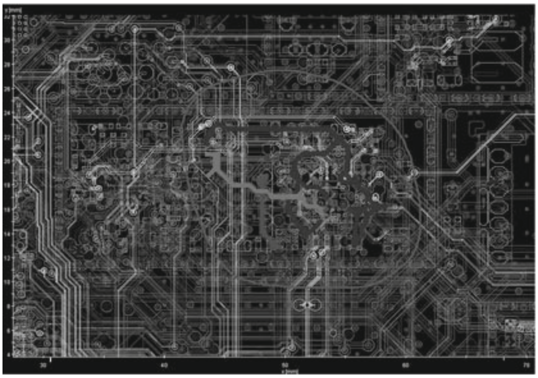


**Fig. 5:** Part of the layout under the shield

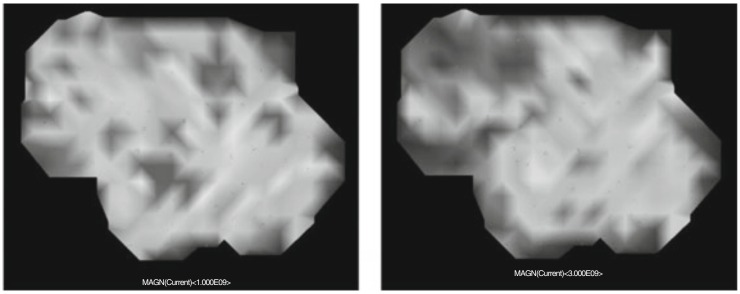


**Fig. 6:** Distribution of the current flowing in the PCB nets under the shield at 1 GHz (*left*) and 3 GHz (*right*)

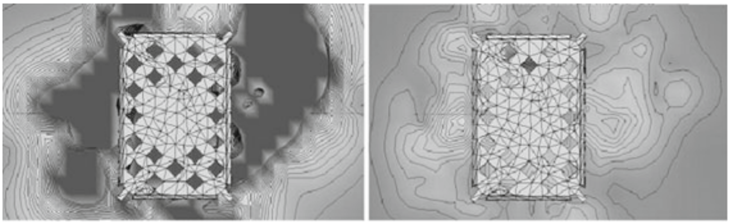


**Fig. 7:** Distribution of the electric field around the shield at 1 GHz (*left*) and 3 GHz (*right*)

the computation of the magnetic field around a voltage-controlled oscillator (VCO), well-known in EMC design as a source of radiation emission that needs to be reduced by shielding.

Our first aim was the measurement of the magnetic field over the module at the fundamental frequency 3.975 GHz and detection of the highest values of emission (maximum amplitudes of the magnetic field components). Measurements have been performed in the XY-plane (parallel to the PCB) for the opened VCO shield using a commercial EMC scanner and magnetic probes. Figure 8 presents the distributions of components of the magnetic field measured at 3.2 mm from the PCB (1 mm from top of the shield; height of the shield is 2.2 mm). The measured output voltage has been converted to the magnetic field following methodology described in [7–10].

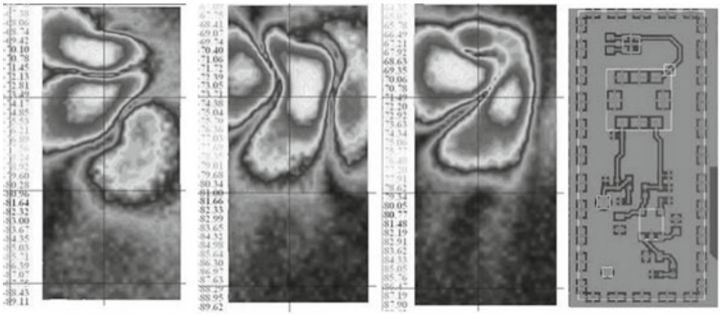


**Fig. 8:** X-, Y- and Z- components of the magnetic field measured over the VCO block

The next step in the methodology is the development of a simplified numerical model for 3-D EM field computations. From the VCO specification it is known that the highest current flows in the "RF output" net. The simplest model of the VCO block consists of this net with an assigned port, the PCB represented as a solid metal brick, and the walls of the shield (Fig. 9). All parts are assumed to be copper. Calculations are first performed with default values of the port parameters $P$ and $R$ that are then tuned using measured data.

Numerical results before and after tuning are shown together with measured data in Table 1. It is easy to see that setting up the port power and resistance equal to 0.012 W and 50 $\Omega$, respectively, provides good agreement between measured and computed components of the magnetic field at 3.2 mm from the PCB. These parameters can then be used in 3-D EM simulations of all problems where this VCO is the excitation source.

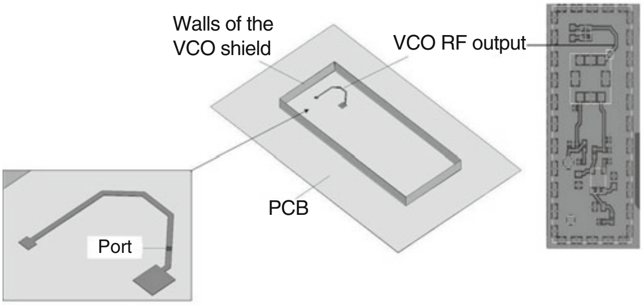


**Fig. 9:** Simplest model of the VCO block without cover of the shield

**Table 1:** Magnetic field over the VCO (3.2 mm from PCB, 3.975 GHz): measured, computed with default parameters of the port, and computed with tuned parameters of the port

| | $(H_x^{\mathrm{VCO}})^{\mathrm{max}}$<br>A/m | $(H_y^{\mathrm{VCO}})^{\mathrm{max}}$<br>A/m | $(H_z^{\mathrm{VCO}})^{\mathrm{max}}$<br>A/m |
|---|---|---|---|
| Computed with default parameters (1 W, 50 $\Omega$) | 0.68 | 0.47 | 0.75 |
| Computed with "tuned" parameters (0.012 W, 50 $\Omega$) | 0.074 | 0.052 | 0.082 |
| Measured | 0.071 | 0.063 | 0.082 |

Although the port in our numerical model is assigned to one particular net (RF output), the parameters of the port are tuned using the highest values of emission of the whole VCO block. Therefore, the VCO output power, related to the current flowing in the RF output net, should not be mixed with the tuned power of the port in the numerical model. As is seen in Fig. 8, the maximum of the EM field distribution does not occur not over the RF output net, so the VCO output power, equal to 2 mW according to specification, is significantly less than 0.012 W.

Figure 10 shows measured and computed distributions of the maximum amplitudes of the components of the magnetic field with increase in the distance from the VCO. The radiation in the VCO occurs not only from RF output net, as is assumed in our simplest numerical model. This is the most probable reason of some disagreements between the measured and computed curves in Fig. 10.

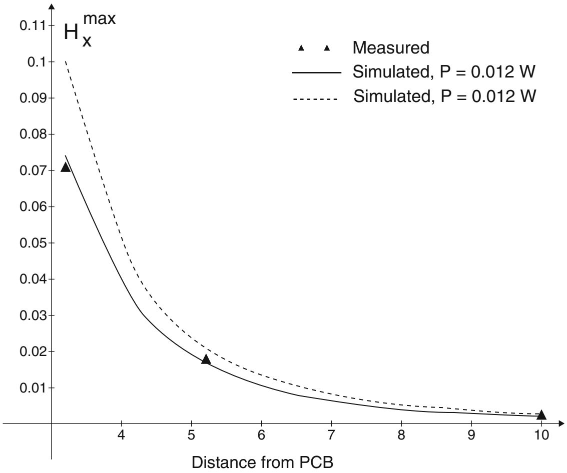


**Fig. 10:** Maximum amplitude of the x-component of the magnetic field as a function of the distance from PCB

# 5 Conclusion

In the present paper two sets of methodologies are considered:

1. A methodology to combine codes for 2.5D and 3D EM field computations around a device or module containing printed circuit. The distribution of the current flowing in the PCB nets is calculated by one code, exported and used as input data in the other, solving a 3-D problem in the surrounding domain (for example, the interior of a device containing the board). The main advantage of the proposed approach is that the cells for modeling the PCB and surrounding domain are not combined in the same mesh. This enables the simulation of real industrial EMC problems of wireless devices that cannot be analyzed using 3D simulators only due to very high computational expenses.
2. A methodology to calibrate (tune) EM sources in numerical models using measured data by solving the inverse problem. This approach enables the simulation of the absolute (as opposed to normalized) values of the EM field emission without detailed knowledge of the source properties. For this purpose, EMC scanner output data (voltages measured in dBm) are converted into absolute values of magnetic fields independent of the measurement conditions and expressed in A/m. This is done using a reference PCB that is simple enough to enable accurate numerical modeling. The measurement results and the modeling of the reference PCB are used to obtain the calibration factor that is then applied to the IC measured data.

# References

1. Drozd, A.L.: Progress on the development of standards and recommended practices for CEM computer modeling and code validation. In: Proceedings of 2003 IEEE International Symposium on Electromagnetic Compatibility, 8-22 Aug., 313–316 (2003)
2. Yilmaz, A., Choi, M.L., Jin, J., Michielsson, E., Cangellaris, A.C.: Multi-scale hybrid electromagnetic modeling and transient simulation of multi-layered printed circuit boards. In: Proceedings of Progress in Electromagnetic Research Symposium (PIERS), 28-31 Mar. (2004)
3. Archambeault, B., Ramahi, O.M., Brench, C.: EMI/EMC Computational Modeling Handbook. Kluwer Academic Publishers (1998)
4. Yuferev, S., Yufereva, J.: A new boundary element formulation of coupled electromagnetic and thermal skin effect problems for non-harmonic regimes of current passage. IEEE Trans. Magn. **32**(3), 1038–1041 (1996)
5. Di Rienzo, L., Yuferev, S., Ida, N.: Computation of the impedance matrix of multiconductor transmission lines using high order surface impedance boundary conditions. IEEE Trans. Electromagn. Compat. **50**(4), 974–984 (2008)
6. Barmada, S., Di Rienzo, L., Ida, N., Yuferev, S.: Time domain surface impedance concept for low frequency electromagnetic problems. II. Application to transient skin and proximity effect problems in cylindrical conductors. IEE Proc. Science, Measurement and Technology **152**(5), 207–216 (2005)
7. Measurement of conducted emissions - magnetic probe method. Intern. Standard IEC 61967-6, Part 6 (2002)

8. Yuferev, S., Saunamäki, E.: Practical techniques for measurements and computations of near-field absolute values. IEEE EMC Society Newsletters Winter **220** (2008). Accepted for publication
9. Gao, Y., Lauer, A., Ren, Q., Wolff, I.: Calibration of electric coaxial near-field probes and apprlications. IEEE Trans. Microwave Theory Techn. **46**(11), 1694–1703 (1998)
10. Slattery, K., Neal, J., Cui, W.: Near-field measurements of VLSI devices. IEEE Trans. Electromagn. Compat. **41**, 374–384 (1999)

# A New Adaptive Approach to Modeling Measured Multi-Port Scattering Parameters

Sanda Lefteriu and Athanasios C. Antoulas

**Abstract** This paper addresses the problem of building a low complexity macromodel of an electromagnetic device based on measurements of its scattering parameters. For devices with a large number of ports, currently available techniques are very expensive. The approach we propose is based on a system-theoretic tool, the Loewner matrix pencil constructed in the context of tangential interpolation. Several implementations are possible. They are fast, accurate and robust; they construct models of low order and are especially designed for devices with a large number of terminals. Moreover, they allow to identify the underlying system, rather than merely fitting the measurements. We compare our algorithms to industry standard vector fitting method on two examples. This paper is a summary of [1].

## 1 Introduction and Motivation

To model electromagnetic effects of complex structures such as chips, packages or boards, one of the following techniques is used. Discretizing Maxwell's equations leads to a high-order representation for which model reduction techniques need to be applied to reduce the dimension to a manageable size [2]. Alternatively, the frequency response (impedance, admittance or scattering parameters) of the device is measured over the desired frequency band. Using the data, a macromodel of low complexity which is consistent with the measurements is constructed. The problem of building a system which approximates given measurements is known as rational interpolation and has been studied thoroughly (see [3] for a survey).

Several techniques have been developed in the electronics community. Most algorithms are based on least-squares approximations, for example [4], but, due to ill-conditioning, their application is restricted to narrow frequency bands and small orders of the model. The algorithm in [5] uses Nevanlinna-Pick interpolation for bounded-real interpolation of S-parameters in which mirror images of the original points are used as additional constraints. Frequency domain subspace identification [6] fails often or requires large computational time, based on experiments in [7]. Nevertheless, vector fitting [8] is the current industry standard.

Sanda Lefteriu, Athanasios C. Antoulas
Rice University, MS-366, 6100 Main Street, Houston, TX 77005, USA, e-mail: sanda.lefteriu@rice.edu, aca@rice.edu. Part of this work was in the framework of the project "MRTN-CT-2006-035559 A Computer Aided Engineering Approach to Smart Structures Design" while S. Lefteriu was with Fraunhofer ITWM, Kaiserslautern, Germany.

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_4,

Algorithms developed in this paper employ a common framework: tangential interpolation and the Loewner matrix pencil [9]. They are fast, accurate and robust; they construct low-order models and are especially designed for devices with many ports. Moreover, they identify the underlying system, rather than merely fit the measurements. Using a black-box approach allows us to model systems with no knowledge of their internal logic [10]. Moreover, we construct the models exclusively from the available measured data by arranging them in an appropriate way.

This paper is organized as follows. Sect. 2 states the problem as a rational interpolation problem, while Sect. 3 shows how to apply the concept of tangential interpolation to S-parameter modeling. Afterwards we describe two different implementation approaches in Sects. 4 and 5. Sect. 6 presents numerical examples which validate the proposed procedures. Finally, Sect. 7 concludes this paper.

## 2 Problem Statement

S-parameter modeling is formulated as a rational interpolation problem as follows. A linear time invariant system models the data set containing $k$ measurements of the scattering coefficients of a device with $p$ ports

$$\left( f_i, \mathbf{S}^{(i)} := \begin{bmatrix} S_{11,i} & \dots & S_{1p,i} \\ \vdots & \vdots & \vdots \\ S_{p1,i} & \dots & S_{pp,i} \end{bmatrix} \right), \; i = 1, \cdots, k$$

if the associated transfer function of the system evaluated at $j \cdot 2\pi f_i$ is close to $\mathbf{S}^{(i)}$:

$$\mathbf{H}(\mathrm{j} \cdot 2\pi f_i) \approx \mathbf{S}^{(i)}, \; i = 1, \dots, k. \tag{1}$$

**Definition 1.** We define *the error matrix* at a specific frequency as:

$$\mathbf{H}(\mathrm{j} \cdot 2\pi f_i) - \mathbf{S}^{(i)} = \mathrm{Err}(f_i), \; i = 1, \dots, k. \tag{2}$$

Clearly, if the norms of all $k$ error matrices are small, the model is accurate.

Let us start from the simple case of model construction from scalar data: $(s_i, \phi_i)$, $i = 1, \dots, P,\; s_i \neq s_j,\; i \neq j$, where $s_i, \phi_i \in \mathbb{C}$ and $s_i$ is not necessarily on the imaginary axis, as it was the case in (1). The rational interpolation problem is equivalent to finding $H(s) = \frac{n(s)}{d(s)}$, with $n, d$ coprime polynomials, such that $H(s_i) = \phi_i$, $i = 1, \dots, P$. There always exists a solution (e.g. Lagrange interpolating polynomial). The main tool in our approach is the Loewner matrix, denoted as $\mathbb{L}$, which is constructed by partitioning the data in disjoint sets: $(\lambda_i, w_i)$, $i = 1, \dots, k$ and $(\mu_j, v_j)$, $j = 1, \dots, h$, where $h, k \approx \left[\frac{P}{2}\right]$ such that $k + h = P$, using the following formula:

$$\mathbb{L} = \begin{bmatrix} \frac{v_1 - w_1}{\mu_1 - \lambda_1} & \dots & \frac{v_1 - w_k}{\mu_1 - \lambda_k} \\ \vdots & \ddots & \vdots \\ \frac{v_h - w_1}{\mu_h - \lambda_1} & \dots & \frac{v_h - w_k}{\mu_h - \lambda_k} \end{bmatrix} \in \mathbb{C}^{h \times k} \tag{3}$$

Several reasons indicate that this is a good tool to use. The rank of the Loewner matrices built using all the possible partitions encodes the degree of the minimal interpolant of the data. Moreover, the Loewner matrix has system theoretic interpretation in terms of the generalized controllability and observability matrices. In particular, when data is obtained by sampling the transfer function $\mathbf{H}(s)$ with minimal state space representation $[\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$ (i.e., $\mathbf{H}(s) = \mathbf{C}\,(s\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}+\mathbf{D}$), then

$$\mathbb{L} = -\underbrace{\begin{bmatrix} \mathbf{C}\,(\lambda_1\mathbf{E}-\mathbf{A})^{-1} \\ \vdots \\ \mathbf{C}\,(\lambda_h\mathbf{E}-\mathbf{A})^{-1} \end{bmatrix}}_{\mathscr{O}} \mathbf{E} \underbrace{\left[\,(\mu_1\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B} \,\ldots\, (\mu_k\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}\,\right]}_{\mathscr{R}} \tag{4}$$

Last, for data consisting of a single point with multiplicity, $(s_0; \phi_0, \phi_1, \ldots, \phi_{P-1})$, i.e. the value of a function at $s_0$ and that of a number of its derivatives are provided, the Loewner matrix has Hankel structure. Thus the Loewner matrix generalizes the Hankel matrix when interpolation at finite points is considered.

## 3 Tangential Interpolation

We will now define the tangential interpolation problem as a rational interpolation problem in the general framework. We are given right interpolation data of the form

$$\{(\lambda_i, \mathbf{r}_i, \mathbf{w}_i) \mid \lambda_i \in \mathbb{C}, \mathbf{r}_i \in \mathbb{C}^{m\times 1}, \mathbf{w}_i \in \mathbb{C}^{p\times 1},\ i = 1, \cdots, k\},\ \text{or}$$
$$\Lambda = \text{diag}\,[\lambda_1,\ \cdots,\ \lambda_k],\ \mathbf{R} = [\mathbf{r}_1,\ \cdots,\ \mathbf{r}_k],\ \mathbf{W} = [\mathbf{w}_1,\ \cdots,\ \mathbf{w}_k]$$

and left interpolation data as

$$\{(\mu_j, \ell_j, \mathbf{v}_j) \mid \mu_j \in \mathbb{C}, \ell_j \in \mathbb{C}^{1\times p}, \mathbf{v}_j \in \mathbb{C}^{1\times m},\ j = 1, \cdots, h,\},\ \text{or}$$
$$M = \text{diag}\,[\mu_1,\ \cdots,\ \mu_h],\ \mathbf{L}^* = \left[\ell_1^*,\ \cdots,\ \ell_h^*\right],\ \mathbf{V}^* = \left[\mathbf{v}_1^*,\ \cdots,\ \mathbf{v}_h^*\right]$$

and the goal is to find $\mathbf{H}(s)$ such that $\mathbf{H}(\lambda_i)\mathbf{r}_i = \mathbf{w}_i$ and $\ell_j\mathbf{H}(\mu_j) = \mathbf{v}_j$. In particular, a minimal realization $[\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$, with $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{n\times m}, \mathbf{C} \in \mathbb{R}^{p\times n}, \mathbf{D} \in \mathbb{R}^{p\times m}$ is desired. A solution to this problem was proposed in [9] in terms of the Loewner matrix and the newly introduced shifted Loewner matrix, constructed as

$$\mathbb{L}_{j,i} = \frac{\mathbf{v}_j \cdot \mathbf{r}_i - \ell_j \cdot \mathbf{w}_i}{\mu_j - \lambda_i},\ \sigma\mathbb{L}_{j,i} = \frac{\mu_j\mathbf{v}_j \cdot \mathbf{r}_i - \ell_j \cdot \mathbf{w}_i\lambda_i}{\mu_j - \lambda_i},\ i = 1, \ldots, k, j = 1, \ldots, h. \tag{5}$$

The following lemma gives a formula for a minimal realization in terms of the Loewner and shifted Loewner matrices together with the data matrices [9].

**Lemma 1.** *If $k = h$, the matrix pencil $(\sigma\mathbb{L}, \mathbb{L})$ is regular and $\mu_j, \lambda_i \notin \lambda\,(\sigma\mathbb{L}, \mathbb{L})$, then*

$$\mathbf{E} = -\mathbb{L},\ \mathbf{A} = -\sigma\mathbb{L},\ \mathbf{B} = \mathbf{V},\ \mathbf{C} = \mathbf{W}\ \textit{and}\ \mathbf{D} = \mathbf{0} \tag{6}$$

*is a minimal realization of an interpolant (i.e., the transfer function $\mathbf{H}(s) = \mathbf{W}(\sigma\mathbb{L} - s\mathbb{L})^{-1}\mathbf{V}$ satisfies the left and right constraints: $\ell_j\mathbf{H}(\mu_j) = \mathbf{v}_j$ and $\mathbf{H}(\lambda_i)\mathbf{r}_i = \mathbf{w}_i$).*

For modeling measured S-parameters, the right tangential data can be chosen as

$$\left(\lambda_i = \mathrm{j}\omega_i, \mathbf{r}_i, \mathbf{w}_i = \mathbf{S}^{(i)}\mathbf{r}_i\right),\ i = 1, \cdots, k, \tag{7}$$

with $\omega_i = 2\pi f_i \in \mathbb{R}$, $\mathbf{r}_i = \mathbf{e}_m \in \mathbb{R}^{p\times 1}$ ($m$-th unit vector), where $m = p$ for $i = p \cdot c_1$ and $m = 1, \cdots, p-1$ for $i = p \cdot c_1 + m$, for some $c_1 \in \mathbb{Z}$, and

$$\left(\mu_i = -\mathrm{j}\omega_i, \ell_i, \mathbf{v}_i = \ell_i\overline{\mathbf{S}}^{(i)}\right),\ i = 1, \cdots, k \tag{8}$$

as left tangential data, with $\ell_i = \mathbf{r}_i^T \in \mathbb{R}^{1\times p}$. Using the tangential data, the Loewner and shifted Loewner matrices are built as in Eq. (5).

**Remark**. The fact that the $p^2$ entries of the S-parameters can be collapsed into a vector of dimension $p$ makes our method suitable for devices with many ports.

## 4 SVD Approach

The first idea is to use all measurements to build the Loewner matrix pencil. However, when too many samples are available, the pencil is singular, so the realization in Lemma 1 is not minimal. The singular part can be projected out via an SVD:

$$x\mathbb{L} - \sigma\mathbb{L} = \mathbf{Y}\Sigma\mathbf{X},\ x \in \{\mathrm{j}\omega_i, -\mathrm{j}\omega_i\},\ i = 1, \cdots, k \tag{9}$$

where $\operatorname{rank}(x\mathbb{L} - \sigma\mathbb{L}) =: n$ (the *dimension of the regular part*; it is precisely the order of the underlying system, for noise-free measurements), $\mathbf{Y} \in \mathbb{C}^{k\times n}$ and $\mathbf{X} \in \mathbb{C}^{n\times k}$. Using the singular vectors as projectors, the realization is given as $\mathbf{E} = -\mathbf{Y}^*\mathbb{L}\mathbf{X}$, $\mathbf{A} = -\mathbf{Y}^*\sigma\mathbb{L}\mathbf{X}$, $\mathbf{B} = \mathbf{Y}^*\mathbf{V}$, $\mathbf{C} = \mathbf{W}\mathbf{X}$, $\mathbf{D} = \mathbf{0}$. Nevertheless, real-world measurements are noisy, so the zero singular values of the pencil are corrupted by noise (they are larger than the noise by a factor proportional to the square root of the number of samples [11]). Thus, one can identify the order of the system based on the drop in the singular values of the Loewner pencil.

Clearly, this approach is expensive for data sets with a large number of samples $k$, as the complexity of the SVD of $x\mathbb{L} - \sigma\mathbb{L}$ is $O(k^3)$. This is overcome by the next approach which identifies immediately when the pencil becomes singular.

## 5 Adaptive Approach

We choose a certain number of samples from the available ones adaptively to construct the desired model. We start with a low order system of order $p$ from $p$ measurements selected from the $k$ available (the $p$ indices are linearly distributed between 1 and $k$) using unit vectors as sampling directions, building the $\Lambda$, $M$, $\mathbf{R}$, $\mathbf{L}$, $\mathbf{W}$, $\mathbf{V}$, $\mathbb{L}$ and $\sigma\mathbb{L}$ matrices as in Sect. 3 and setting $\mathbf{E}$ as $-\mathbb{L}$, $\mathbf{A}$ as $-\sigma\mathbb{L}$, $\mathbf{B}$ as $\mathbf{V}$, $\mathbf{C}$ as $\mathbf{W}$ and $\mathbf{D}$ as $\mathbf{0}$. For each sample, we compute the $p$ singular values of the error matrices defined in (2). We update our model by adding $p$ new measurements where the largest errors occur, with sampling directions taken as the singular vectors associated to the largest singular values of those particular error matrices. In the next step, a system of order $2\cdot p$ is constructed, the singular values of the error matrices are computed again and new se ts of $p$ measurements are selected according to the same criterion. Lastly, we improve the accuracy by adding more measurements and projecting to the desired order $n$ using the singular vectors obtained from the SVD of $\mathrm{j}\omega_1\mathbb{L} - \sigma\mathbb{L}$. If $n$ is a multiple of $p$, we add $p$ more measurements, otherwise we add $\mathrm{mod}(n, p)$. Note that the accuracy of the current model is directly available as the largest singular value of all error matrices, so we stop the procedure as soon as the pencil becomes singular (for noise-free measurements) or the accuracy is below the threshold given by the noise level scaled by the square root of the sample size (for noisy measurements). Adding blocks of $p$ samples makes our method better suited for a large number of ports due to the fact that the complexity scales with $O(kp^3)$, as opposed to column-wise vector fitting (VF), which scales with $O(kp^4)$.
**Remark on stability and passivity.** Given an appropriate accuracy of fit or desired

order of the macromodel, the resulting system will be stable and passive, provided that measurements come from a stable and passive device. Due to measurements errors, the data may not passive (i.e. the maximum of the largest singular value is larger than one), so the resulting macromodel may not be passive. To make the model passive, one can divide the **B** or **C** state-space matrix by that maximum. Another a posteriori passivation enforcement is described in [12].

## 6 Results

This section analyzes a theoretical example and an example obtained from measurements. We compare our methods to state-of-the-art *vector fitting*, in terms of accuracy of the macromodels and required CPU time. The accuracy was assessed using two error measures: the normalized $\mathscr{H}_\infty$-norm of the error system,

$$\mathscr{H}_\infty \text{ error } = \frac{\max_{i=1\ldots k} \sigma_1 \left(\mathbf{H}(\mathrm{j}\omega_i) - \mathbf{S}^{(i)}\right)}{\max_{i=1\ldots k} \sigma_1 \left(\mathbf{S}^{(i)}\right)},$$

which evaluates the maximum deviation in the singular values and the normalized $\mathscr{H}_2$-norm of the error system,

$$\mathscr{H}_2 \text{ error } = \frac{\sum_{i=1}^{k} \left\|\mathbf{H}(\mathrm{j}\omega_i) - \mathbf{S}^{(i)}\right\|_F^2}{\sum_{i=1}^{k} \left\|\mathbf{S}^{(i)}\right\|_F^2}$$

where $\|\cdot\|_F^2$ stands for the Frobenius-norm, which evaluates the error in the magnitude of all entries, proving to be a good estimate of the overall performance. For devices with many ports, computing the error in each entry of the S-parameters is unfeasible, as for $p = 50$ ports, the errors in all $50^2 = 2500$ entries would have to be assessed. Thus, these error measures give a good indication of the model's quality.

As a visual tool, it is common to plot each entry of the measured data against the corresponding entry in the transfer function of the model. For large number of ports, this is impractical, so in this paper, we compare the singular values of the measured S-parameters to the singular values of the transfer function of the model evaluated at each frequency (sigma plot). If the singular values of the model are close to those of the data, the fit will be of good quality for all entries.

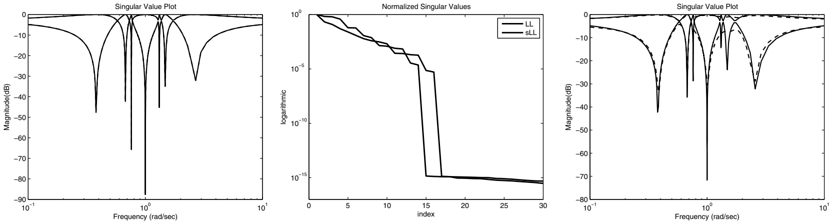
All experiments employed column-wise vector fitting with the same options:

- the starting poles are real and stable, linearly distributed in the frequency band
- the starting poles of each column are obtained by fitting the column sum with $N_1 = 5$ iterations which are used to fit the column itself with $N_2 = 5$ iterations
- no asymptotic terms were required, unless otherwise specified
- the fast implementation of relaxed VF [8, 13, 14] was used.

The tests were performed on a Pentium Dual-Core at 2.2GHz with 3GB RAM.

### *6.1 Noise-Free System with 2 Ports, 14 Poles and Non-zero* **D** *Matrix*

We consider a theoretical system of order 14 with $p = 2$ ports and a non-zero **D**-term. We compare the algorithms when trying to recover the original system from 608 noise-free measurements between $10^{-1}$ rad/sec and $10^{1}$ rad/sec.



**(a)** Original system **(b)** Singular value drop **(c)** $n = 14$ model with VF

**Fig. 1:** Original system, singular value drop of the Loewner matrix pencil and model built with VF

Figure 1a shows the sigma plot of the original system.Figure 1b shows the normalized singular values of the Loewner and shifted Loewner matrices (only the first 30 of all singular values are shown as the rest are zero). We notice that the Loewner matrix has rank 14 while the shifted Loewner matrix has rank 16, so by generating models of order 16 which have a singular $\mathbf{E}$ matrix and an invertible $\mathbf{A}$ matrix, we obtain the 14 poles of the original system and 2 infinite eigenvalues. Vector fitting was given $N = 7$ starting poles and was required to produce a $\mathbf{D}$ matrix.

**Table 1:** Results for $k = 608$ noise-free measurements of an order 14 system with $p = 2$ ports

| Algorithm | CPU time (s) | $\mathscr{H}_\infty$ error | $\mathscr{H}_2$ error |
|---|---|---|---|
| **Vector Fitting** | **0.78** | **1.0956e+000** | **4.7563e-002** |
| SVD Approach | 5.85 | 4.8050e-011 | 6.1323e-023 |
| **Adaptive Approach** | **0.39** | **1.7736e-010** | **7.7177e-022** |

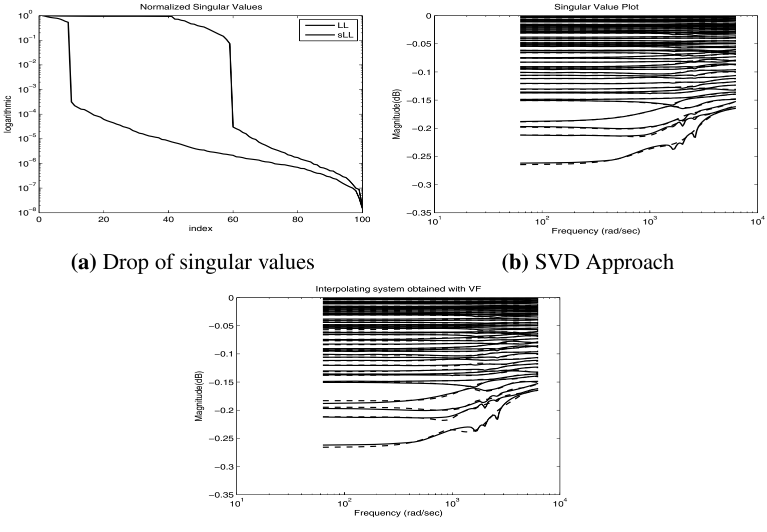
Table 1 presents the CPU time and the errors for the resulting models. We conclude that the proposed algorithms were able to identify the original system, while VF did not (Fig. 1c). If VF is given $N = 14$ starting poles, the resulting errors are similar to ours. However, the realization will have order $n = 28$ and each pole will have multiplicity 2. This requires an additional compacting step [15].

### *6.2 Example Obtained from Measurements*

Measurements were performed using a vector network analyzer (VNA). The data set was provided by CST and contains $k = 100$ frequency samples between 10MHz and 1GHz from a device with $p = 50$ ports. To avoid numerical instabilities, all frequencies were scaled by $10^{-6}$. Figure 2a shows the same behaviour as in Fig. 1b. The singular values of the Loewner matrix and of the shifted Loewner matrix drop several orders of magnitude between the 9th and 10th, and the 59th and 60th, respectively. We conclude that there is a non-zero $\mathbf{D}$ matrix and the underlying system is of order 9. Thus, we generate models of order $n = 59$ with $\mathbf{D} = 0$. VF was given $N = 5$ starting poles for each column and was required to produce a $\mathbf{D}$ matrix, so the order of the resulting VF model was $n = 5 \cdot 50 = 250$.

Table 2 presents a summary of the results. The model obtained with the SVD approach (order $n = 59$ and $\mathbf{D} = \mathbf{0}$) is shown in Fig. 2b, while that obtained with VF (order $n = 250$ and $\mathbf{D} \neq \mathbf{0}$) is presented in Fig. 2c (the x-axis has a dB scale and the frequencies are scaled by $10^{-6}$). Note that to obtain comparable errors to our models, VF needs to built a model of order $n = 250$.

Table 3 shows that all our models were stable and very close to being passive.



(**a**) Drop of singular values (**b**) SVD Approach

(**c**) VF ($n = 250$)

**Fig. 2:** Singular value drop of the Loewner matrix pencil, sigma plots for different models

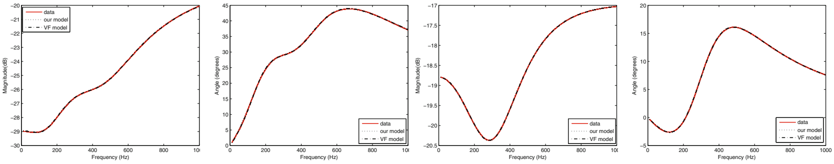
**Table 2:** Results for a data set with $k = 100$ samples obtained from a device with $p = 50$ ports

| Algorithm | CPU time (s) | $\mathscr{H}_\infty$ error | $\mathscr{H}_2$ error |
|---|---|---|---|
| **SVD Approach** | **0.07** | **6.0440e-003** | **1.9506e-007** |
| Adaptive Approach | 1.09 | 7.0829e-003 | 6.9701e-007 |
| Vector Fitting ($n = 250$) | 8.29 | 4.6797e-003 | 3.7886e-007 |

**Table 3:** Stability and passivity results for models of a device with $p = 50$ ports

| Algorithm | Stable | $\mathscr{H}_\infty$-norm |
|---|---|---|
| SVD | Yes | 1.0002 |
| Adaptive | Yes | 1.0003 |
| VF ($n = 250$) | Yes | 1.9869 |

Figure 3 compares the measured $S_{1,1}$ and $S_{10,20}$ entries to the model obtained with the SVD approach ($n = 59$ with $\mathbf{D} = \mathbf{0}$) and with VF ($n = 250$ with $\mathbf{D} \neq \mathbf{0}$). Thus, the sigma plots in Fig. 2b and 2c indeed predict that the models are good for all entries of the S-parameters.



(**a**) Magnitude of $S_{1,1}$ (**b**) Angle of $S_{1,1}$ (**c**) Magnitude of $S_{10,20}$ (**d**) Angle of $S_{10,20}$

**Fig. 3:** Modeling entries of the measured S-parameters obtained from a device with $p = 50$ ports

## 7 Conclusion and Future Work

This paper proposes accurate and efficient algorithms for modeling measured multiport scattering parameters. They are based on the system theoretic concept of the

Loewner matrix pencil constructed in the framework of tangential interpolation. The approach we are presenting is especially suited for devices with a large number of ports. It uses no heuristics, but only the available data, makes no assumptions of the underlying system and last, but not least, allows to identify the system if enough measurements are provided. We compared the performance of the new algorithms to the state-of-the-art vector fitting method and concluded that our approaches are faster and yield better models with dimensions smaller than the ones produced by VF. More numerical examples can be found in [16], but the quality of the results shown by the examples we included is similar. The extension to the case where derivatives are also provided as measurements is currently under investigation. Moreover, we are interested in generalizing this approach to time-domain data.

## References

1. Lefteriu, S., Antoulas, A.C.: A new approach to modeling measured multi-port scattering parameters. IEEE Trans. on CAD Submitted
2. Ioan, D., Ciuprina, G.: Reduced order electromagnetic models of on-chip passive components and interconnects, workbench and test structures for inntegrated passive components. In: W. Schilders, H. van der Vorst (eds.) Model Order Reduction: Theory, Research Aspects and Applications. Springer, Heidelberg (2008)
3. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia (2005)
4. Elzinga, M., Virga, K.L., Prince, J.L.: Improved global rational approximation macromodeling algorithm for networks characterized by frequency-sampled data. IEEE Trans. Microw. Theory Tech. **48**(9), 1461–1468 (2000)
5. Coelho, C., Silveira, L., Phillips, J.: Passive constrained rational approximation algorithm using Nevanlinna-Pick interpolation. In: DATE '02: Proceedings of the conference on Design, automation and test in Europe, p. 923. IEEE Computer Society, Washington, DC, USA (2002)
6. Overschee, P.V., Moor, B.D.: Continuous-time frequency domain subspace system identification. Signal Processing **52**(2), 179–194 (1996)
7. Ebert, F., Stykel, T.: Rational interpolation, minimal realization and model reduction. Tech. Rep. 371-2007, DFG Research Center MATHEON (2007)
8. Gustavsen, B., Semlyen, A.: Rational approximation of frequency domain responses by vector fitting. IEEE Trans. Power Del. **14**, 1052–1061 (1999)
9. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. Linear Algebra and Its Applications **405**, 634–662 (2007)
10. Deschrijver, D.: Broadband macromodeling of linear systems by vector fitting. Ph.D. thesis, Universiteit Antwerpen (2007)
11. Stewart, G.W.: Perturbation theory for the singular value decomposition. Tech. Rep. CS-TR-2539, Dept. of Computer Science and Inst. for Advanced Computer Studies, Univ. of Maryland (1990). URL citeseer.ist.psu.edu/stewart90perturbation.html
12. Grivet-Talocia, S.: Passivity enforcement via perturbation of Hamiltonian matrices. IEEE Trans. Circuits Syst. **51**, 1755–1769 (2004)
13. Gustavsen, B.: Improving the pole relocation properties of vector fitting. IEEE Trans. Power Del. **21**(3), 1587–1592 (2006)
14. Deschrijver, D., Mrozowski, M., Dhaene, T., De Zutter, D.: Macromodeling of multiport systems using a fast implementation of the vector fitting method. IEEE Microwave and Wireless Components Letters **18**(6), 383–385 (2008)
15. Gustavsen, B., Semlyen, A.: A robust approach for system identification in the frequency domain. IEEE Trans. Power Del. **19**, 1167–1173 (2004)
16. Lefteriu, S.: New approaches to modeling multi-port scattering parameters. Master's thesis, Rice University, Houston, TX (2008)

# Parametric Models of Transmission Lines Based on First Order Sensitivities

Alexandra Stefanescu, Daniel Ioan, and Gabriela Ciuprina

**Abstract** Further downscaling of the integrated circuits pushes the limits of lithographic technologies and certain variability effects previously considered negligible now should be taken into account. This paper proposes an efficient approach that addresses the problem of interconnect process variations. New models for line parameters parameterized with respect to the geometric transversal dimensions, subject to small or large variations are proposed. The parametric models are solely based on the computation of first order sensitivities. In the multiparametric case the use of multiplicative models can be a better choice than the use of traditional models based on first order Taylor Series truncation.

## 1 Introduction

Continuous improvements in today's fabrication processes determine smaller chip sizes and smaller device geometries. The impact of interconnect performances has become important as millions of closely spaced interconnections in one or more levels connect various components on the integrated circuit [1]. Process induced variations induce changes in the properties of metallic interconnect between devices, pushing the limits of lithographic technologies. Parasitic capacitances, resistances and inductances of the interconnections have become major factors in the evolution of very high speed IC technology. This paper focuses on the variability of the numerical extracted models for long interconnects modeled as transmission lines with respect to geometric parameters. The authors investigate promising alternatives beside the classic models of first-order truncations of Taylor expansions. The self - imposed restriction is to use in the extracted model exclusively the values of first-order sensitivities and not those of superior orders. The advantage of this approach

Alexandra Stefanescu, Daniel Ioan, Gabriela Ciuprina
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania, e-mail: alexar@lmn.pub.ro, lmn@lmn.pub.ro, gabriela@lmn.pub.ro

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_5,

is obvious. This represents one of the goals of the research carried out within the European project FP6/IST/Chameleon [2].

This paper is structured as follows: first the basic approach used is discussed, second, the approach is validated in the case of a microstrip line having one or multiple variable parameters. Next, results on technology variability are shown and conclusions are drawn at the end.

## 2 Parametric Models Based on First Order Sensitivities

First order sensitivities are essential for the analysis of the parameter variability [3, 4]. Parametric models are often obtained by truncating the Taylor series expansion for the quantity of interest. This requires the computation of the derivatives of the device characteristics with respect to the design parameters [5]. Let us assume that $y(p_1, p_2, \cdots, p_n) = y(\mathbf{p})$ is the device characteristic which depends on the design parameters $\mathbf{p} = [p_1, p_2, \cdots, p_n]$. The quantity $y$ may be, for instance the real or the imaginary part of the device admittance at a given frequency. In our case this quantity is any of the p.u.l. parameters. The parameter variability is thus completely described by the real function, $y$, defined over the design space $S$, a subset of $\mathbb{R}^n$. The nominal design parameters correspond to the particular choice $\mathbf{p}_0 = [p_{01}\ p_{02}\ \cdots\ p_{0n}]$.

### 2.1 Additive Model (A)

If $y$ is smooth enough then its truncated Taylor Series expansion is the best polynomial approximation in the vicinity of the expansion point $\mathbf{p}_0$. For one parameter $(n = 1)$, the additive model is the first order truncation of the Taylor series:

$$\hat{y}(p) = y(p_0) + \frac{\partial y}{\partial p}(p_0)(p - p_0). \tag{1}$$

If we denote by $y(p_0) = y_0$ the nominal value of the output function, by $\frac{\partial y}{\partial p}(y_0)\frac{p_0}{y_0} = S_p^y$ the relative first order sensitivity and by $(p - p_0)/p_0 = \delta p$ the relative variation of the parameter $p$, then the variability model based on (1) defines an *affine* [6] or *additive* model (A):

$$\hat{y}(p) = y_0(1 + S_p^y \delta p). \tag{2}$$

According to the Taylor Series theory the neglected terms can be expressed function of the second order derivative in an intermediate point, $\xi$:

$$y(p) = y(p_0) + \frac{\partial y}{\partial p}(p_0)(p - p_0) + \frac{\partial^2 y}{\partial^2 p}(\xi)(p - p_0). \tag{3}$$

It follows that the relative variation of the output quantity $\delta y = (y(p) - y_0)/y_0$ can be expressed as

$$\delta y = S_p^y \delta p + \varepsilon, \tag{4}$$

where the approximation error $\varepsilon$ depends on the second order derivative of the output quantity:

$$\varepsilon = \frac{p_0^2}{2y_0}\frac{\partial^2 y}{\partial^2 p}(\xi)(\delta y)^2. \tag{5}$$

Thus, to ensure a relative validity range of the first order approximation of the output quantity less a given threshold $t_1$, the absolute variation of the parameter must be less than

$$V_d = \sqrt{\frac{2y_0 t_1}{D_2}}, \tag{6}$$

where $D_2$ is an upper limit of the second order derivative of the output quantity $y$ with respect to parameter $p$.

The validity range of the first approximation can be increased in some cases if the Taylor Series expansion is used for the "reversed" quantity $1/y(p)$. In this case, to obtain the same validity range of the first order approximation for the reversed output quantity, the variation of the parameter has to be less than

$$V_r = \sqrt{\frac{2t_1}{y_0 D_2'}}, \tag{7}$$

where $D_2'$ is an upper limit of the second order derivative of the reversed output quantity.

For the multiparametric case, one gets:

$$y(\mathbf{p}) = y(\mathbf{p}_0) + \nabla y(\mathbf{p}_0)\cdot(\mathbf{p}-\mathbf{p}_0) = y_0 + \sum_{k=1}^{n}\frac{\partial y}{\partial p_k}(\mathbf{p}_0)(p_k - p_{0k}). \tag{8}$$

Similar with one parameter case, the relative sensitivities w.r.t. each parameter are denoted by $\frac{\partial y}{\partial p_k}(\mathbf{p}_0)\frac{p_{0k}}{y_0} = S_{p_k}^{y}$ and the relative variations of the parameters by $\delta p_k = (p_k - p_{0k})/p_{0k}$, the additive model (A) for $n$ parameters being given by:

$$\hat{y}(\mathbf{p}) = y_0(1 + \sum_{k=1}^{n} S_{p_k}^{y}\delta p_k). \tag{9}$$

Thus, each new independent parameter taken into account adds a new term to the sum [7]. The additive model is simply a normalized standard version of a linearly truncated Taylor expansion. Instead of using this truncated expansion may be numerically favorable to expand some transformation $F(y)$ of $y$ instead. Two particular choices for $F$ have practical importance: identity and inversion as it will be indicated below. The originality of the algorithm for parametric model extraction proposed by authors is the automation of the choice of transformation $F$, based on the numerical estimation for the validity ranges (6), (7).

## 2.2 Rational Model (R)

The rational model is the additive model for the reverse quantity $1/y$. It is obtained from the first order truncation of the Taylor Series expansion for the function $1/y$. For $n = 1$, if we denote by $r(p) = \frac{1}{y(p)}$, it follows that:

$$\hat{r}(p) = r(p_0) + \frac{\partial r}{\partial p}(p_0)(p - p_0). \tag{10}$$

We define the relative first order sensitivity of the reverse circuit function: $\frac{\partial r}{\partial p}(p_0) \frac{p_0}{r(p_0)} = S_p^r = S_p^{1/y}$. Consequently, we obtain the rational model for $n = 1$:

$$y(p) = \frac{y_0}{1 + S_p^{1/y}\delta p}. \tag{11}$$

It can be easily shown that the reverse relative sensitivity is $S_p^{\frac{1}{y}} = -S_p^y$. For the multiple parameter case, the rational model is:

$$\hat{y}(\mathbf{p}) = \frac{y_0}{1 + \sum_{k=1}^{n} S_{p_k}^{1/y}\delta p_k}. \tag{12}$$

## 2.3 Multi-parametric Model (M)

Let us assume that in the multiparametric case the quantity of interest can be written as a product of functions with separated variables:

$$y(\mathbf{p}) = y_1(p_1)y_2(p_2)\cdots y_n(p_n). \tag{13}$$

Each component function, $y_k$ depends only on a single parameter, $p_k$ and for each one we can use either an additive or a rational model:

$$\hat{y}(\mathbf{p}) = \frac{y_0(1 + \sum_{k=1}^{m} S_{p_k}^{y}\delta p_k)}{1 + \sum_{k=m+1}^{n} S_{p_k}^{1/y}\delta p_k}. \tag{14}$$

The tensor product representation (13) seems to be a very particular case, however it fits perfectly the variation of RLC parameters w.r.t. geometric parameters extracted from uniform electric or magnetic field. The factorization and the choice of $m$ are dictated by physics of the problem itself, however the modeling algorithm we propose is a numerical approach based on the expressions (6) and (7) for the validity ranges. For instance, in the case of two variable parameters, $p_1, p_2$ four versions of model M are possible:

- $M_{AA}$- additive models for both parameters

$$\hat{y}(\mathbf{p}) = y_0(1 + S_{p_1}^{y})(1 + S_{p_2}^{y}); \tag{15}$$

- $M_{RR}$ - rational models for both parameters

$$\hat{y}(\mathbf{p}) = y_0 \frac{1}{(1 + S_{p_1}^{1/y})(1 + S_{p_2}^{1/y})}; \tag{16}$$

- $M_{AR}$ - additive model for the first parameter and rational model for the second one

$$\hat{y}(\mathbf{p}) = y_0 \frac{(1 + S_{p_1}^{y})}{(1 + S_{p_2}^{1/y})}; \tag{17}$$

- $M_{RA}$ - rational model for the first parameter and additive model for the second one

$$\hat{y}(\mathbf{p}) = y_0 \frac{(1+S_{p_2}^{y})}{(1+S_{p_1}^{1/y})}. \tag{18}$$

Together with the two "classical" A and R models, there are six possible parametric models for the two parameter case.

# 3 Case Study

In order to validate our approach and to evaluate different parametric models, several experiments have been performed on a test structure that consists of a microstrip (MS) transmission line having one Aluminum conductor embedded in a SIO$_2$ layer. The line has a rectangular cross section, parameterized by several parameters (Fig. 1). The return path is the grounded surface placed at $y = 0$. The nominal values used are: $x_{max} = 20\mu$m, $h_2 = 10\mu$m, $h_3 = 5\mu$m, $h_0 = 1\mu$m, $p_1 = 1\mu$m, $p_2 = 0.67\mu$m, $p_3 = 3\mu$m, $\sigma_{Si} = 10000$ MS/m, $\sigma_{Al} = 3.3$MS/m, $\varepsilon_{r-SiO_2} = 3.9$. In order to comply with designer's requirements, the model should include the field propagation along the line, taking into consideration the distributed parameters and the high frequency effects.

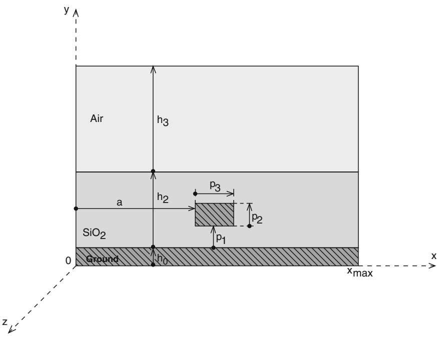


**Fig. 1:** Stripline parameterized structure

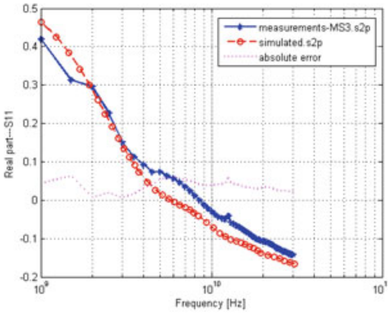


**Fig. 2:** Frequency characteristic Re(S$_{11}$): numerical model vs measurements

## 3.1 Validation of the Nominal Model

Before considering the parametric model, the results obtained for the nominal values of p.u.l. parameters were validated by deriving from them the scattering parameters (**S**) and compare the results with the measurements provided within the European project FP5/Codestar (www.imec.be/codestar). For the nominal case, by using dFIT + dELOB [7], at low frequencies, the following values are obtained:

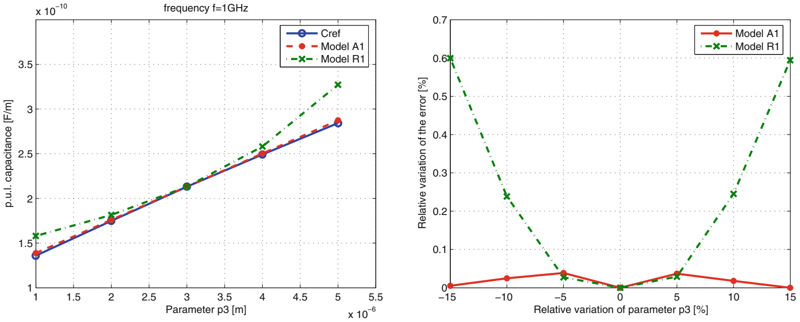


**Fig. 3:** *Left*: Reconstruction of the p.u.l. C from Taylor Series first order expansion; *Right*: Relative error w.r.t. the relative variation of parameter $p_3$

$$R = 18.11k\Omega/m, L = 322nH/m, C = 213pF/m \tag{19}$$

Actually p.u.l. resistance and inductance are frequency dependent, and they can be computed with the method described in [7]. The frequency response of the entire line having the length $d$ was computed using Transmission Line equations [6]. The comparison between the simulations and the measurements is shown in Fig. 2 and validates the nominal model described before. The sensitivities of p.u.l. parameters are computed using the CHAMY software [2], by direct differentiation method applied to the state space equations [5]. They can also be computed by Adjoint Field Technique (AFT) [8,9].

## 3.2 Parametric Models

In this section, the accuracy of the A, R and M models for the line capacitance is investigated.

### One Parameter Case

The first sets of tests considered only one parameter that varies, namely the width of the line, $p_3$. The nominal value chosen was $p_3 = 3\mu$m and samples in the interval $[1,5]\mu$m were considered. The reference result was obtained by doing "exact" simulations for the samples. These were compared with the approximate values obtained from models A and R (Fig. 3). As expected intuitively, the dependence w.r.t. $p_3$ is almost linear and the A model is better than the R model. Considering the relative variation of the parameters less than 15% (which is the typical limit for the technological variations nowadays) the relative variation of the output parameter is obtained (Fig. 3, right). The errors of both affine and rational first order models for p.u.l. parameters are given in Table 1. Model A based on the first order Taylor series approximation has a maximal error for technologic variations 1.78% for p.u.l. resistance when $p_3$ is variable, while model R has an approximation error of only 0.6% for the same range of the technological variations for p.u.l. capacitance when

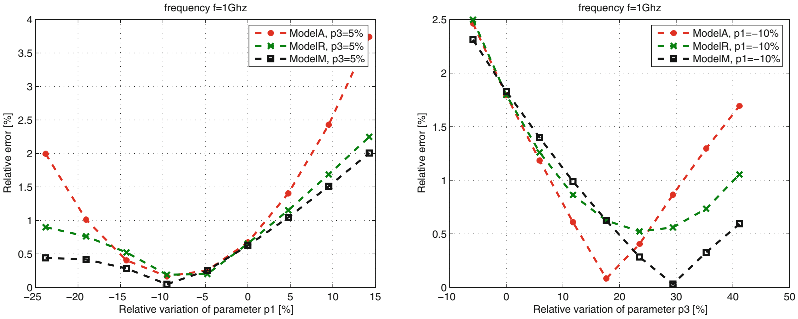


**Fig. 4:** *Left*: Relative error w.r.t. the relative variation of parameter $p_1$, for a variation of $p_3$ of 5%; *Right*: Relative error w.r.t. the relative variation of parameter $p_3$, for a variation of $p_1$ of 10%

$p_3$ is variable. Using (6) and (7) can be easily identified which is the best model in any case.

**Table 1:** Maximal errors [%] of p.u.l. parameters for technology variation of ± 15%

| Parameter | Quantity | Affine (A1) | Rational (R1) |
|---|---|---|---|
| $p_1$ | $L$ | 0.11 | 0.15 |
| | $C$ | 0.65 | 0.25 |
| $p_3$ | $R$ | 1.78 | 0.22 |
| | $L$ | 0.34 | 0.04 |
| | $C$ | 0.035 | 0.6 |

### Multiple Parametric Case

Let us consider now two parameters that vary simultaneously: $p_1$ and $p_3$. For reference, a set of samples in $[0.8, 1.2]\mu$m × $[2, 4]\mu$m were considered. The p.u.l. capacitance was approximated using the additive, rational and multiplicative models described above. In this case, model M is computed using an additive model for $p_3$ and a rational one for $p_1$, which is the best choice. Fig. 4a compares the relative variation of the errors w.r.t. a relative variation of parameter $p_1$ for a variation of $p_3$ of 5%. Model M provides lower errors (maximum error is 2%) than models A (3.7%) and R (2.2%). Fig. 4b illustrates that in the range from 20% to 40% model M is the best one if we look at the variation w.r.t. $p_3$ for a variation of $p_1$ of 10%. Thus, by using the appropriate multiplicative models in the modeling of the technological variability, the necessity of higher order approximations may be eliminated.

## 4 Conclusions

This paper analyzes variability models for TL structures considering the dependency of p.u.l. parameters w.r.t. geometric parameters, at a given frequency. A detailed study of the line sensitivity was made by using numeric techniques. For one parameter case, the proposed methods avoid the evaluation of higher order sensitivities, maintaining the accuracy by introducing rational models. The multi-parametric case has been analyzed, in addition, a multiplicative parametric model (M) has been proposed. This is based on the assumption that the quantity of interest can be expressed with separated variables, for which A and/or R models are used. Model M is sometimes better than A and R models obtained from Taylor Series expansion. Its specific terms (products of first order sensitivities) can thus approximate higher order, cross-terms of Taylor Series. In order to automatically select the best first order model for a multiparametric problem, the validity ranges of direct and reversed quantities have to be evaluated. Once we establish the best model (A or R) for each parameter, the M model will be easily computed by multiplication of individual submodels. Our numerical experiments with the proposed algorithm in all particular structures we investigated prove that the technological variability (e.g. ±20% variation of geometric parameters, which is typical for the technology node of 65 nm) can be modeled with acceptable accuracy (relative errors under 5%) using only first order parametric models for line parameters.

**Acknowledgements** This study was partially supported by the European Commission in the framework of the CoMSON RTN project, grant number MRTN-2005-019417.

## References

1. Goel, A.K.: High-Speed VLSI Interconnections. Wiley Series in Microwave and Optical Engineering (2007)
2. CHAMELEON–RF site: www.chameleon-rf.org
3. Kinzelbach, H.: Statistical variations of interconnect parasitics: Extraction and circuit simulation. In: Proceedings of the 10th IEEE Workshop on Signal Propagation on Interconnects, pp. 33–36. Budapest, 09–12 May (2006)
4. Labun, A.: Rapid Method to Account for Process Variation in Full-Chip Capacitance Extraction. IEEE Transactions Computer-Aided Design, **23**(6), 941–951 (2004)
5. Ciuprina, G., Ioan, D., Niculae, D., Villena, J.F., Silveira, L.M.: Parametric models based on sensitivity analysis for passive components. Intelligent Computer Techniques in Applied Electromagnetics, in bookseries *Studies in Computational Intelligence*, vol. 119, pp. 231–239, Springer 2008.
6. Stefanescu, A., Ciuprina, G., Ioan, D.: Models for Variability of Transmission Line Structures. In: Proceedings of the 12th IEEE Workshop on Signal Propagation on Interconnects, Avignon, 12–15 May (2008)
7. Ioan, D., Ciuprina, G., Kula, S.: Reduced Order Models for HF Interconnect over Lossy Semiconductor Substrate. In: Proceedings of the 11th IEEE Workshop on Signal Propagation on Interconnects, pp. 233–236. Ruta di Camogli, 13–16 May (2007)
8. Ioan, D., Ciuprina, G., Schilders, W.: Parametric Models Based on the Adjoint Field Technique for RF Passive Integrated Components. IEEE Transactions of Magnetics, **44**(6), 1658–1661 (2008)
9. Bi, Y., van der Meijs, N.P.,Ioan, D.: Capacitance Sensitivity Calculation for Interconnects by Adjoint Field Technique. In: Proceedings of the 12th IEEE Workshop on Signal Propagation on Interconnects, Avignon, 12–15 May (2008)

# Domain Partitioning Based Parametric Models for Passive On-Chip Components

Gabriela Ciuprina, Daniel Ioan, Diana Mihalache, and Ehrenfried Seebacher

**Abstract** This paper shows how to obtain models for passive integrated components that take into consideration the variability inherent to their design. To achieve this, the computational domain is split into sub-domains in which the electromagnetic circuit element (EMCE) formulation is used. The variability is described by using first order Taylor Series representation for the semi-state space matrices. The novelty of the paper is that it describes how the EMCE based parametric models can be obtained. The parametric sub-models can be interconnected afterwards to obtain a global parametric model that can be simulated or reduced. The advantage of this approach is that it bears an inherent parallelism. The sub-models can be treated independently both from the point of view of the variability, and from the point of view of electromagnetic field formulation. Both aspects are illustrated with a simple test case as well as a real benchmark designed and characterized at austriamicrosystems.

## 1 Introduction

The design of the next-generation of integrated circuits is challenged by an increased number of difficulties since electromagnetic (EM) field effects at high frequencies are too relevant to be neglected. In this respect, one of the issues of the European research project CHAMELEON-RF was to develop methodologies and tools able to simulate RF blocks up to 60 GHz by taking the electromagnetic coupling and design variability into account (www.chameleon-rf.org).

Gabriela Ciuprina, Daniel Ioan, Diana Mihalache
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania, e-mail: gabriela@lmn.pub.ro, lmn@lmn.pub.ro, mihaladi@lmn.pub.ro

Ehrenfried Seebacher
austriamicrosystems, 8141 Schloss Premstaetten, Austria, e-mail: ehrenfried.seebacher@austriamicrosystems.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_6,

In this framework, the concept of magnetic terminals (“hooks” or “connectors”) was used for the first time, to describe the interaction of on-chip components with their environment [1]. These magnetic hooks are special boundary conditions that allow the extension of the electric circuit element (ECE) [2] to the electromagnetic circuit element (EMCE) [1]. Such an EMCE allows the connection to an external magnetic circuit, and thus inductive coupling effects between the device and its environment can be considered.

This paper shows how domain partitioning (DP) can be further exploited by taking into account the variability. The EMCE based model sensitivities are computed by using first order Taylor Series (TS) expansion with respect to the parameters. The parametric sub-models can be interconnected afterwards to obtain a global parametric model that can be simulated or reduced. A standard for parametric representation of systems has been defined and the interested reader can find details in [3].

## 2 Parametric Full Wave Discretized Models for the EMCE

The EM field effects at high frequencies are quantified by Maxwell equations in Full-Wave (FW) regime. The most appropriate formulation for passive devices with distributed parameters, compatible both with external electric and magnetic circuits is the Electro-Magnetic Circuit Element. This represents a first level of approximation, i.e an EM field problem correctly formulated, with appropriate boundary and initial conditions. The next level of approximation is obtained by applying a numerical method to obtain a discretized model of the EMCE formulation. In this respect we used Finite Integration Technique (FIT), as it is described in [4]. Thus, a parametric time-domain model can be obtained:

$$\mathbf{C}(\mathbf{p})\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}+\mathbf{G}(\mathbf{p})\mathbf{x}=\mathbf{B}\mathbf{u}, \qquad \mathbf{y}=\mathbf{L}\mathbf{x}, \tag{1}$$

where $\mathbf{x} = [\mathbf{u}_m^T, \mathbf{u}_e^T, \mathbf{y}^T]^T$ is the state space vector, consisting of electric voltages $\mathbf{u}_e$ defined on the electric grid used by FIT, magnetic voltages $\mathbf{u}_m$ defined on the magnetic grid and output quantities $\mathbf{y}$. Equations can be written such that only two semi-state space matrices ($\mathbf{C}$ and $\mathbf{G}$) are affected by the parameters $\mathbf{p}$. The input quantities $\mathbf{u}$ and the output quantities $\mathbf{y}$ are solely related to the terminals. Each terminal introduces exactly one input and one output quantity. For instance, if an electric terminal is voltage excited, its voltage is a component of the input vector and the current flowing through it (entering in the domain) is an output quantity. Similarly, if a magnetic terminal is excited in magnetic voltage, this one will be an input quantity and the magnetic flux entering in the domain through this terminal will be an output quantity. Thus, the number of inputs is always equal to the number of outputs. Also, the output quantities are considered as degrees of freedom, so as to be computed simultaneously with the other unknowns. Therefore, the matrix $\mathbf{L} = \mathbf{B}^{\mathbf{T}}$ is merely a selection matrix, as is explained in [4].

For instance, the structure of the matrices in the case of voltage excitation (both for electric and magnetic terminals) is the following:

$$\mathbf{C}=\begin{bmatrix} \mathbf{G}_{\mathrm{m}}(\mathbf{p}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{C}_{\mathrm{i}}(\mathbf{p})\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathrm{Sl}}(\mathbf{p}) & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathrm{TE}}(\mathbf{p}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathrm{TM}}(\mathbf{p}) & \mathbf{0} \end{bmatrix} \quad \mathbf{G}=\begin{bmatrix} \mathbf{0} & \mathbf{B}_1\ \mathbf{B}_2 & \mathbf{0} \\ \mathbf{B}_1^T & -\mathbf{G}_{\mathrm{i}}(\mathbf{p})\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\ \mathbf{B}_{\mathrm{Sl}} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathrm{Sl}}(\mathbf{p}) & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathrm{TE}}(\mathbf{p}) & -\mathbf{S}_{\mathrm{E}} \\ \mathbf{G}_{\mathrm{m-TM}}(\mathbf{p}) & \mathbf{0} & -\mathbf{S}_{\mathrm{M}} \\ \mathbf{0} & \mathbf{P}_{\mathrm{E}} & \mathbf{0} \\ \mathbf{P}_{\mathrm{M}} & \mathbf{G}_{\mathrm{TM}}(\mathbf{p}) & \mathbf{S}_{\mathrm{M}}(\mathbf{R}_{\mathrm{m-TM}}(\mathbf{p})+\mathbf{R}_{\mathrm{m-gnd}}(\mathbf{p})) \end{bmatrix} \tag{2}$$

There are eight sets of rows, corresponding to the eight sets of equations. The first group of equations is obtained by writing Faraday's law for inner elementary electric loops. $\mathbf{G}_{\mathrm{m}}$ is a diagonal matrix holding the magnetic conductances that pass through the electric loops. $\left[\mathbf{B}_1\ \mathbf{B}_2\right]$ has only 0, 1, $-1$ entries, describing the incidence of inner branches and branches on the boundary to electric faces. The second group corresponds to Ampere's law for elementary magnetic loops. $\mathbf{C}_{\mathrm{i}}$ and $\mathbf{G}_{\mathrm{i}}$ are diagonal matrices, holding the capacitances and electric conductances of the inner branches. The third group represents Faraday's law for electric loops on the boundary. $\mathbf{B}_{\mathrm{Sl}}$ has only 0, 1, $-1$ entries, describing the incidence of electric branches included in the boundary to the electric boundary faces. The fourth row is obtained from the current conservation law for all nodes on the boundary excepting nodes on the electric terminals. $\mathbf{G}_{\mathrm{Sl}}$ and $\mathbf{C}_{\mathrm{Sl}}$ hold electric conductances and capacitances directly connected to boundary. The fifth set of equations represents current conservation for electric terminals. $\mathbf{G}_{\mathrm{TE}}$ and $\mathbf{C}_{\mathrm{TE}}$ hold electric conductances and capacitances that are directly connected to electric terminals. $\mathbf{S}_{\mathrm{E}}$ is the connection matrix between electric branches and terminals path. The sixth set means flux conservation for magnetic terminals. Entries of $\mathbf{G}_{\mathrm{m-TM}}$ are magnetic conductances that are connected to magnetic terminals. $\mathbf{S}_{\mathrm{M}}$ is the connection matrix between magnetic branches and terminals path. The seventh row is obtained from expressing the voltages of electric terminals as sums of voltages along open paths from terminals to ground, $\mathbf{P}_{\mathrm{E}}$ being a topological matrix that holds the paths that connect electric terminals to ground. The last group of equations represent magnetic voltages of magnetic terminals. These are function of the capacitances and conductances that pass through the surface defined by the magnetic reluctances, magnetic paths on the magnetic grid and the actual path on the boundary. The reluctances ($\mathbf{R}_{\mathrm{m-TM}}$) correspond to magnetic voltages that have to be taken into consideration since the magnetic grid is strictly inside the boundary. $\mathbf{G}_{\mathrm{TM}}$ and $\mathbf{C}_{\mathrm{TM}}$ hold magnetic conductances and capacitances that are directly connected to magnetic terminals. $\mathbf{P}_{\mathrm{M}}$ is a topological matrix that holds the paths that connect magnetic terminals to ground.

Thus, the top left square block of $\mathbf{C}$ is diagonal and the top left square bloc of $\mathbf{G}$ is symmetric. The size of this symmetric bloc corresponds to the useful magnetic branches and to the useful inner electric branches. Its size is dominant over the size of the matrix, therefore, solving or reduction strategies that take into consideration this particular structure are useful.

The simplest way to analyse the parameter variability is to compute first order sensitivities. These are derivatives of the device characteristics with respect to the design parameters $\mathbf{p} = (p_1 \ldots p_n)$. Considering as design parameters the geometrical variables and material constants, then only the Hodge matrices $\mathbf{G}_*$, $\mathbf{C}_*$ and $\mathbf{R}_{\mathrm{m}}$ may be influenced by these design parameters. Thus, the possible variations of parameters do not affect all entries in the semi-state space matrices. The affected blocks are marked with $(\mathbf{p})$ in (2).

The model parameterization and the extraction of the state space matrices sensitivities can be easily included in the FIT discretization scheme for the EMCE, since the assembling of sensitivities is similar to the assembling of matrices, the only difference is that only the affected cells add contribution to the sensitivity matrices $\partial \mathbf{C}/\partial p_k$, $\partial \mathbf{G}/\partial p_k$. That is why, the first order sensitivities are assembled simultaneously with the matrices, by direct differentiation.

## 3 Interconnecting the Models

The models described as parameterized linear time invariant system (1) can be coupled with models of other devices by means of terminals (Fig. 1). We will illustrate this for two models that are interconnected. We will assume that all terminals are voltage excited and the two systems are described by

$$\mathbf{C}_1 \frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}t} + \mathbf{G}_1\mathbf{x}_1 = \mathbf{B}_1\mathbf{u}_1, \qquad \mathbf{y}_1 = \mathbf{L}_1\mathbf{x}_1, \tag{3}$$

$$\mathbf{C}_2 \frac{\mathrm{d}\mathbf{x}_2}{\mathrm{d}t} + \mathbf{G}_2\mathbf{x}_2 = \mathbf{B}_2\mathbf{u}_2, \qquad \mathbf{y}_2 = \mathbf{L}_2\mathbf{x}_2. \tag{4}$$

According to the number of electric/magnetic terminals that are interconnected, the input vector of each system is partitioned as follows $\mathbf{u}_1 = \left[\mathbf{u}_{01}^T \; \mathbf{u}_{c1}^T\right]^T, \mathbf{u}_2 = \left[\mathbf{u}_{02}^T \; \mathbf{u}_{c2}^T\right]^T$, where $\mathbf{u}_{01}$ and $\mathbf{u}_{02}$ are the voltages of terminals of the first and, respectively, the second sub-model that will not be coupled, whereas the inputs $\mathbf{u}_{c1}$ and $\mathbf{u}_{c2}$ will be coupled. Therefore, the vectors $\mathbf{u}_{c1}$ and $\mathbf{u}_{c2}$ must have the same size, whereas the number of external input of the interconnected system will be given by the sum of sizes of the vectors $\mathbf{u}_{01}$ and $\mathbf{u}_{02}$. Since the output quantities are placed in the state space vector on the last positions, it is useful to partition as well the state space vectors in $\mathbf{x}_1 = \left[\mathbf{x}_{01}^T \; \mathbf{y}_{01}^T \; \mathbf{y}_{c1}^T\right], \mathbf{x}_2 = \left[\mathbf{x}_{02}^T \; \mathbf{y}_{02}^T \; \mathbf{y}_{c2}^T\right]$, where $\mathbf{y}_{01}$ and $\mathbf{y}_{02}$ are the vector of currents flowing through the terminals that will not be coupled, corresponding to the first, and respectively to the second sub-system, and $\mathbf{y}_{c1}$ and $\mathbf{y}_{c2}$ are the currents flowing through the terminals that will be interconnected.

The partitioning of input vector and state space vector conduces to the partitioning of state space matrices in $\mathbf{C}_1 = \left[\mathbf{C}_{11} \; \mathbf{C}_{12} \; \mathbf{C}_{13}\right], \mathbf{G}_1 = \left[\mathbf{G}_{11} \; \mathbf{G}_{12} \; \mathbf{G}_{13}\right], \mathbf{B}_1 = \left[\mathbf{B}_{11} \; \mathbf{B}_{12}\right], \mathbf{C}_2 = \left[\mathbf{C}_{21} \; \mathbf{C}_{22} \; \mathbf{C}_{23}\right], \mathbf{G}_2 = \left[\mathbf{G}_{21} \; \mathbf{G}_{22} \; \mathbf{G}_{23}\right], \mathbf{B}_2 = \left[\mathbf{B}_{21} \; \mathbf{B}_{22}\right]$.

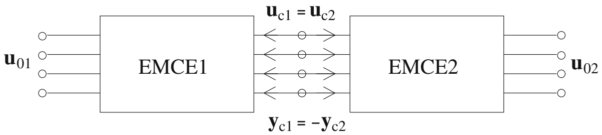

$\mathbf{u}_{c1} = \mathbf{u}_{c2}$
$\mathbf{u}_{01}$
EMCE1
EMCE2
$\mathbf{u}_{02}$
$\mathbf{y}_{c1} = -\mathbf{y}_{c2}$

**Fig. 1:** Interconnection of the partitioned sub-domains

By imposing the coupling conditions for coupled terminals voltages ($\mathbf{u}_c$) and currents ($\mathbf{y}_c$)

$$\mathbf{u}_{c1} = \mathbf{u}_{c1} \overset{\text{not}}{=} \mathbf{u}_c, \qquad \mathbf{y}_{c1} = -\mathbf{y}_{c2} \overset{\text{not}}{=} \mathbf{y}_c, \tag{5}$$

and assuming that in the global model the state space are ordered as follows $\mathbf{x} = \begin{bmatrix} \mathbf{x}_{01}^T \ \mathbf{x}_{02}^T \ \mathbf{u}_c^T \ \mathbf{y}_c^T \ \mathbf{y}_{01}^T \ \mathbf{y}_{02}^T \end{bmatrix}^T$, it can easily be proved that the semi-state space model of the interconnected system is given by

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{C}_{13} & \mathbf{C}_{12} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{21} & \mathbf{0} & -\mathbf{C}_{23} & \mathbf{0} & \mathbf{C}_{22} \end{bmatrix}, \tag{6}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{0} & -\mathbf{B}_{12} & \mathbf{G}_{13} & \mathbf{G}_{12} & \mathbf{0} \\ \mathbf{0} & -\mathbf{G}_{21} & \mathbf{B}_{22} & -\mathbf{G}_{23} & \mathbf{0} & \mathbf{G}_{22} \end{bmatrix}, \tag{7}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{21} \end{bmatrix}. \tag{8}$$

The same procedure applies for the sensitivities of the global model. Thus, the sensitivities of the global model are easily obtained by sticking blocks of the sensitivities of the sub-models:

$$\frac{\partial \mathbf{C}}{\partial p_k} = \begin{bmatrix} \frac{\partial \mathbf{C}_{11}}{\partial p_k} & \mathbf{0} & \mathbf{0} & \frac{\partial \mathbf{C}_{13}}{\partial p_k} & \frac{\partial \mathbf{C}_{12}}{\partial p_k} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \mathbf{C}_{21}}{\partial p_k} & \mathbf{0} & -\frac{\partial \mathbf{C}_{23}}{\partial p_k} & \mathbf{0} & \frac{\partial \mathbf{C}_{22}}{\partial p_k} \end{bmatrix}, \tag{9}$$

$$\frac{\partial \mathbf{G}}{\partial p_k} = \begin{bmatrix} \frac{\partial \mathbf{G}_{11}}{\partial p_k} & \mathbf{0} & \mathbf{0} & \frac{\partial \mathbf{G}_{13}}{\partial p_k} & \frac{\partial \mathbf{G}_{12}}{\partial p_k} & \mathbf{0} \\ \mathbf{0} & -\frac{\partial \mathbf{G}_{21}}{\partial p_k} & \mathbf{0} & -\frac{\partial \mathbf{G}_{23}}{\partial p_k} & \mathbf{0} & \frac{\partial \mathbf{G}_{22}}{\partial p_k} \end{bmatrix}. \tag{10}$$

From the parameterization point of view, the partitioning in sub-domains has the advantage that it may isolate the sub-domain in which the effect of a variation is sensed (for example, in the computations above, all sensitivities of the second systems can be zero) and ease the task of simulation or reduction. Moreover, different formulations (e.g. full-wave, magneto-static, etc.) can be used in each part (Fig. 2).

In order to inspect the validity range of the first order Taylor Series expansion for the system representation, the output quantity computed as

$$\mathbf{y}_{\mathrm{TS}} = \mathbf{L}\left(\mathrm{j}\omega\mathbf{C}_{\mathrm{TS}} + \mathbf{G}_{\mathrm{TS}}\right)^{-1}\mathbf{Bu}, \tag{11}$$

where first order TS expansions are used for $\mathbf{C}$ and $\mathbf{G}$:

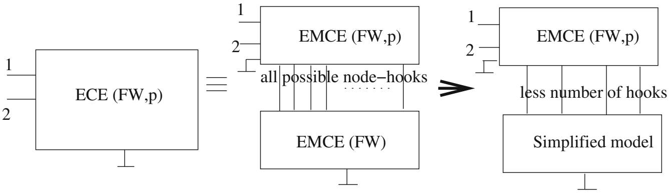


**Fig. 2:** Domain partitioning of complex models: parameters affect only some submodels whereas simplified formulations can be used for other submodels

$$\mathbf{C}_{\mathrm{TS}} = \mathbf{C}_{\mathrm{nom}} + \sum_{k=1}^{n} \frac{\partial \mathbf{C}}{\partial p_k} \left(p_k - p_{\mathrm{nom}_k}\right), \quad \mathbf{G}_{\mathrm{TS}} = \mathbf{G}_{\mathrm{nom}} + \sum_{k=1}^{n} \frac{\partial \mathbf{G}}{\partial p_k} \left(p_k - p_{\mathrm{nom}_k}\right) \quad (12)$$

with $p_k$ the varying design parameter and $p_{\mathrm{nom}_k}$, the nominal value of the parameter, is compared to the simulation results obtained with nominal model extracted for every parameter value. Results are given in the next section.

## 4 Numerical Results

The benchmark presented to exemplify the procedure consists of two coupled inductors above a Silicon substrate, as shown in Fig. 3. Its domain was decomposed in three parts, the top part (air) and the bottom part (Si), modeled by a quasistatic field, while the middle part, which contains the coils is modeled with full wave field. Results shown in Fig. 4 are a validation for the EMCE concept used in correlation with the partitioning of the domain. In order to validate the modeling of sensitivities and to investigate the validity of the TS expansion, a much simpler test was considered, consisting of two U-shape conductors, placed above a silicon substrate. Each conductor has one terminal voltage excited and one terminal connected to ground. The first parameter that is investigated is the distance between two conductors $d$ (Fig. 5). Since $d$ does not affect the shape of the coils, the domain partitioning used is in three parts: left, middle and right. By combining together only the left and the right parts the configuration corresponding to the minimum value $d_{\mathrm{min}}$ is obtained. Therefore, the parameter that varies is $d^* = d - d_{\mathrm{min}}$, and it is associated solely to the middle domain. Figures 6 and 7 show the impact of this parameter on the admittance $Y_{12}$ at 3 GHz. The reconstruction using the TS for system matrices is accurate enough (less than 5 %) even if the relative variation of the parameter is $\pm 100\%$. However, this behaviour is not the same at 40 GHz (Fig. 8), when an accurate representation is obtained only for a relative variation of $\pm 20\%$.

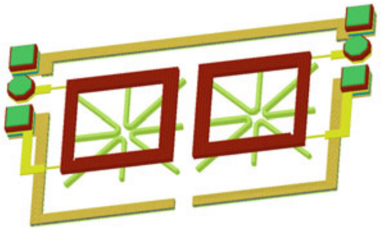


**Fig. 3:** Real benchmark: two coupled inductors (CHRF 202)

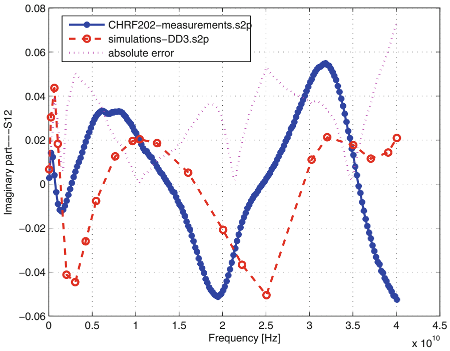


**Fig. 4:** Comparison between results of measurement and simulation. Scattering parameter of CHRF 202

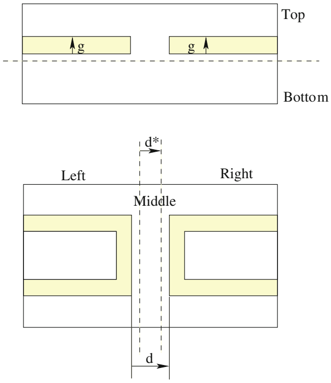


**Fig. 5** Simple test: two U shaped conductors

**Table 1** Size of models (no.of DoFs; no.of terminals) obtained by using DP and matrices coupling

| Model | Domain | Size |
|---|---|---|
| CHRF 202 | Top (MS) | 17138; 14 |
| | Middle (FW) | 81453; 26 |
| | Bottom (MS) | 15427; 14 |
| | Complete | 104102; 2 |
| U-coupled a) | Top (FW) | 5466; 129 |
| | Bottom (FW) | 5351; 127 |
| | Complete | 10817; 2 |
| U-coupled b) | Top (FW) | 5394; 57 |
| | Bottom (MS) | 950; 55 |
| | Complete | 6344; 2 |

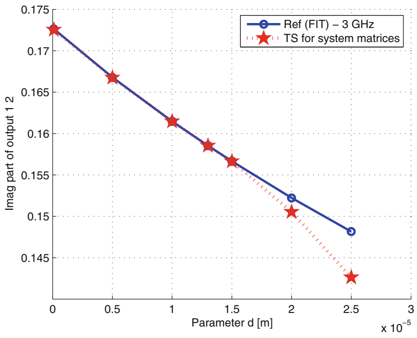


**Fig. 6:** Parameter $d$ impact on $Y_{12}$ at 3 GHz

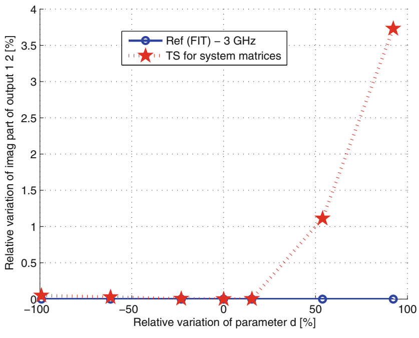


**Fig. 7:** Relative impact of distance $d$

A second test used a technological parameter, namely the thickness $g$ of the corresponding metal layer. Figure 9 show that, in this case, the reconstruction using TS for system matrices is very accurate (less than 1 %) even at 40 GHz.

Different decomposition schemes were tested, dividing the domain in a bottom part, which encloses the substrate, and a top part, enclosing the metal layers and upper air. Different regimes were used for different domains, reducing the overall model complexity with equivalent accuracy, as presented in Table 1.

## 5 Conclusions

The approach we propose for the extraction of parametric models of passive on-chip components is based on domain partitioning and use of the EMCE formulation. Its main advantages are the reduction of computational complexity for the model

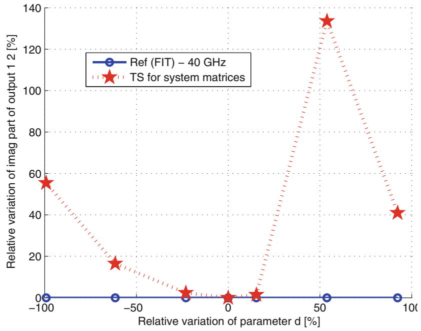


**Fig. 8:** Relative impact of distance *d* at 40 GHz

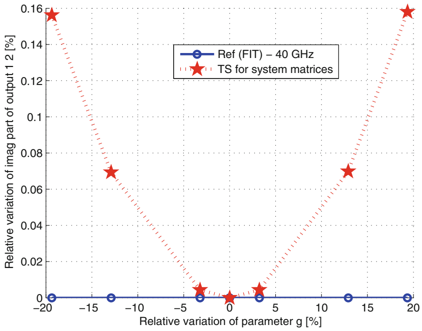


**Fig. 9:** Relative impact of thickness *g* at 40 GHz

extraction process and the possibility of using different, independent grids or formulations in several sub-domains, locally refined and adapted to the local modeled structure. In this manner the main drawback of numerical methods based on the rectangular, uniform grids is eliminated. The global modeling effort is thus reduced, replaced by the independent model extraction for each sub-domain.

The computation of sensitivities of the state-space matrices for the EMCE formulation is straightforward when using FIT as discretization method. For our tests, the approximation based on Taylor Series for the system matrices was accurate enough (less than 1 %) if the variations were technological (less than 20 %), or if the variations were due to the design but the operation frequency is low. If accurate enough, such parametric models can be further submitted to parameterized model order reduction procedures.

**Acknowledgements** Financial support by UEFISCSU (RO) via contract IDEI 609/16.01.2009, allowing the continuation of the research initiated in FP6/Chameleon-RF, is acknowledged.

## References

1. Ioan, D., Schilders, W., Ciuprina, G., Meijs, N., Schoenmaker, W.: Models for integrated components coupled with their EM environment. COMPEL Journal **27**(4), 820–829 (2008)
2. Ioan, D., Ciuprina, G.: Reduced order models of on-chip passive components and interconnects, workbench and test structures. in (W.H.A. Schilders, H.A. van der Vorst, J. Rommes, Eds). Model Order Reduction: Theory, Research Aspects and Applications, Springer series on Mathematics in Industry, Springer-Verlag **13**, Ch. 20 (2008)
3. Silveira, L., et al.: Report on parametric model order reduction techniques (d1.3a). Tech. rep., FP6/Chameleon-RF (2007). Available at `http://www.chameleon-rf.org/`
4. Ciuprina, G., Ioan, D., Mihalache, D.: Magnetic hooks in the finite integration technique: a way towards domain decomposition. In: Proceedings of the IEEE CEFC. Athens, Greece (2008)

# A Novel Graphical Based Tool for Extraction of Magnetic Reluctances Between On-Chip Current Loops

Alexander Vasenev, Sebastián Gim, Alexandra Stefanescu, Sebastian Kula, and Diana Mihalache

**Abstract** Continued device scaling into the nanometer region has given rise to new effects that previously had negligible impact but now present greater challenges and unprecedented complexity to designing successful mixed-signal silicon. This paper presents a novel graphical tool for semi automatic extraction of magnetic reluctances between on-chip current loops. The novel graphical tool seamlessly integrates within the workflow of the CHAMELEON-RF software prototype developed.

## 1 Introduction

One of the major challenges in the nano electronic design industry (EDA) is the management of design complexity. As integrated circuits are being scaled into the nanometer scale, more and more functionality can be integrated on-die [1]. However, the greater integration leads to a number of design challenges, amongst them, the mutual coupling between interconnects, sub-circuits and functional blocks. All these need to be effectively managed by the EDA software in an efficient but intuitive manner to ensure a successful design. A coherent framework and comprehensive workflow model is needed.

There are several reasons to find an optimal procedure to support communication between the various data formats within a framework's workflow model and automating the tasks. The two most pressing reasons are reduction of human effort and elimination of user errors. Furthermore for the effective use of simulation software it is necessary to find a proper way for correct representation and transformation of initial data from schematic parameters and layout to simulation.

A. Vasenev, S. Gim, A. Stefanescu, S. Kula, D. Mihalache
Numerical Methods Lab., Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania, e-mail: vasenev@lmn.pub.ro, seb@lmn.pub.ro, alexar@lmn.pub.ro, sebastia@lmn.pub.ro, mihaladi@lmn.pub.ro

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_7,

In this paper, we present a novel graphical tool and a proposed workflow model for the extraction of magnetic reluctances between on-chip current loops that integrates seamlessly with the Chamy electromagnetic simulator [2]. Based on the principles identified previously, code for a novel graphical tool called Mag-Tris (Magnetic Tetris) was developed in LayoutEditor and Matlab. This novel graphical tool that is specially designed for Chamy software is based on Matlab and can effectively work with Matlab's standard files .mat and ASCII files as an import format. A few rules should be observed to solve this task. First, the solution should comply with requirements of Chamy. For example, it must support any Manhattan geometry structure and other principles, defined by the software. Secondly, it must be based on Graphical User Interface (GUI). A designer as the user can easily make modification, see and examine changes. Finally, implementation should be done by using non-commercial general public license software.

The remaining sections of this paper are organized as follows. Section 2 describes the key features of Mag-Tris and presents the proposed workflow model and seamless integration with Chamy software. This approach is then applied to extracting the magnetic reluctances between fundamental current loops in a 24 GHz LNA. The results from this extraction process are then presented in Section 3. The paper is finally concluded in Section 4.

## 2 Workflow for Multi-scale Compact Modeling

Contemporary mixed signal design workflows involve ad-hoc iterations of incremental design improvements based upon tweaking a baseline circuit using a gamut of discrete software tools, some analytic methods and plenty of heuristics. It is an iterative cycle of SPICE circuit simulation, silicon fabrication, measurement and parameter tweaking to meet design performance targets. Many fabrication spins and prior experience is often needed to isolate coupled effects before tweaking SPICE parameters in order to compensate for these undesirable parasitic effects. The Mag-Tris graphical tool which integrates seamlessly together with Chamy presents the designer with an intuitive integrated development environment (IDE) workflow for EDA of electromagnetically coupled circuits.

In this new workflow, the designer creates a particular RF block and simulates the theoretical performance using existing SPICE simulators. However, instead of fabricating the layout and subsequently measuring the performance many months later, the designer first uses Chamy to simulate and extract interconnect parasitic effects directly from the layout [3, 4]. The parameters of the interconnect parasitic when included into the SPICE netlist will more accurately model RF block performance than a topological electrical netlist alone. In the second step, coupling within the circuit is modeled by creating a dual magnetic circuit to the original electric circuit. Voltage sources representing induced voltages and current sources of the magnetic field are placed into the original electric and also the dual magnetic netlist or schematic (Fig. 1a). The sources are actually linear controlled sources that model mutual interaction between fundamental current loops of the RF block because in integrated circuits magnetic interaction between conductive loops essentially occurs between the holes.

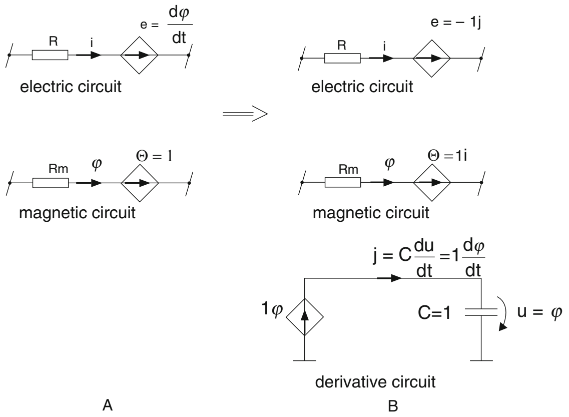


**Fig. 1:** Elimination of time derivative flux controlled source by means of a third derivative circuit. **a** Original circuit (*left*), **b** equivalent circuit (*right*). *i* electric current, *j* magnetic flux

The dual magnetic circuit can be thought of as having magnetic flux as the ‘current’ which flows within it. Magnetic reluctances or “resistances” impede the flow of this “current”. The magnetic flux originates because of a magnetic motive force or magnetic voltage source due to actual current flow in the electrical circuit within the fundamental loops. A time derivative of flux controlled voltage source however is not a native element in SPICE. To get around this limitation, an equivalent Spice sub-circuit was defined. It contains a third derivative circuit as shown in Fig. 1b. This technique is known as the magneto-electric equivalent circuit (MEEC).

Next Mag-Tris, a novel graphical interface developed and implemented using Layout Editor and Matlab and integrates seamlessly with Chamy is used to calculate the magnetic reluctances between fundamental current loops. The dataflow of Mag-Tris is shown in Fig. 2. The designer needs to visually identify and mark the fundamental current loops in an open source layout file viwer Layout Editor (Fig. 4). Mag-Tris macros extract the relevant polygons to files before passing the formatted data to Chamy in order to calculate the required magnetic reluctances. Mag-Tris also post processes Chamy output to return the desired magnetic reluctances.

The completed MEEC circuit can now be simulated using existing SPICE simulators. Together with the layout interconnect parasitic extracted earlier, the MEEC circuit allows more realistic simulation compared with conventional electric SPICE simulation alone.

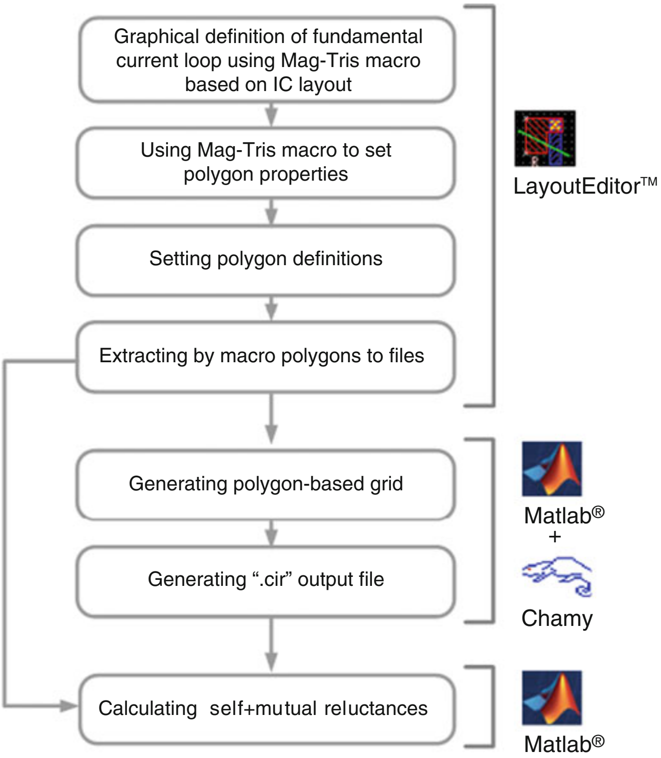

Graphical definition of fundamental current loop using Mag-Tris macro based on IC layout
Using Mag-Tris macro to set polygon properties
Setting polygon definitions
Extracting by macro polygons to files
Generating polygon-based grid
Generating ".cir" output file
Calculating self+mutual reluctances
LayoutEditor™
Matlab®
+
Chamy
Matlab®

**Fig. 2:** Dataflow of Mag-Tris

## 3 Calculation of Magnetic Reluctances

The fluxes and magnetic voltages of the magnetic terminals are related by means of the linear relation:

$$\varphi = \mathrm{Pv}_m \tag{1}$$

where P is is the nodal magnetic permeance matrix of magnetic terminals. Matrix P is symmetric, diagonal dominant and positive defined with positive diagonal entries and negative off-diagonal terms. By expressing voltage between two terminals as potential differences, the branch reluctances in the complete polygonal topology are obtained.

$$\begin{aligned} &\phi_{kj} = \textstyle\sum_{j,k=1}^{n} p_{jk}v_k = \sum_{j,k=1}^{n} G_{mjk}(v_j - v_k) \\ &G_{mkj} = -p_{kj} > 0, \qquad G_{mk0} = \textstyle\sum_{i=1}^{n} p_{ki} > 0 \\ &R_{mkj} = 1/G_{mkj}, \qquad R_{mk0} = 1/G_{mk0} \end{aligned} \tag{2}$$

The test case consists of a 24GHz Car Radar LNA (Fig. 3:LNA schematic). This benchmark is a 50Ω single ended input and output impedance 24 GHz LNA for car radar applications designed and fabricated using the NXP QUBiC4X process.

This benchmark also demonstrates that the QUBiC4X 0.25 $\mu$m SiGe:C BiCMOS process is adequate for emerging microwave applications between 10-30 GHz. The LNA core circuit essentially consists of a common emitter first stage that is cascade loaded which is then followed by the emitter follower output. The 50$\Omega$ input impedance matching is achieved by means of a low Q matching network ($C_m$ $L_m$) at the input stage. This is then followed by an inductively degenerated matched common emitter first stage to achieve good power gain over broadband frequencies. Current sources (I1 I3) in the final test structure are implemented through a transistor current mirror network. In addition to this LNA core circuit, several other practical details such as a MIM decoupling capacitor network, voltage divider network, a 60 $\mu$m long GSG transmission line at the input feed and bondpads are implemented in the final layout (Fig. 4). According to the modeling procedure described in the previous section, using just topological data from the LNA core schematic, a simple tree co-tree graph is created to identify a system of independent fundamental loops. (Fig. 3: Tree co-tree graph). The tree is colored blue and the co-tree is colored red. The topological nodes are blue squares. The red circles indicate voltage sources representing the induced voltage and are placed in the co-tree branches. There are four fundamental loops (1-4), carrying independent currents, which are sources of the parasitic magnetic field. This implies that there are 4 self reluctances and 6 mutual reluctances as part of a fully connected magnetic network. In each fundamental loop there is a source for magnetic field attached (Fig. 3: green circle). The self reluctances of each source have been omitted for simplicity. After that, in order to model the parasitic induced voltages, a dual magnetic circuit was coupled to the original electrical circuit by means of controlled sources. The demo Chameleon

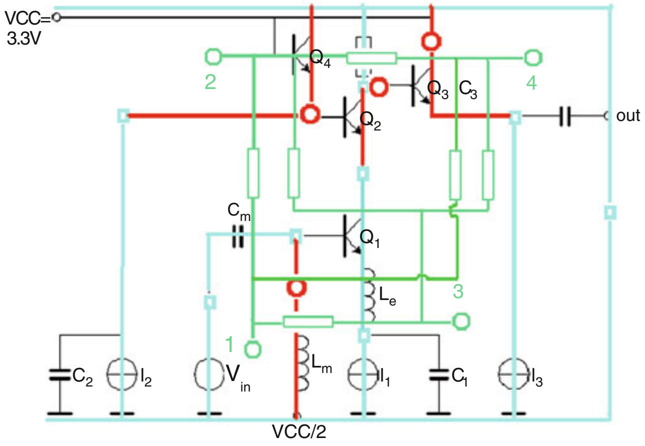


**Fig. 3:** LNA core schematic with tree (*blue*)  co-tree (*red*) with induced sources and magnetic circuit (*green*) overlayed

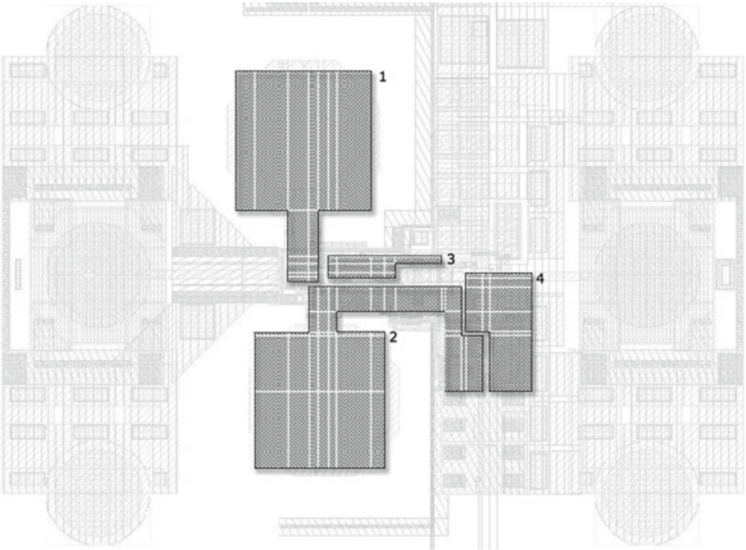


**Fig. 4:** Identified fundamental current loops in a 24 GHz LNA

software prototype [2] and Mag-Tris was then used to demarcate the fundamental loops on the LNA gds layout file and extract the self and mutual reluctances using Finite Integral Technique (FIT) [5,6] between the hooks of Manhattan shapes (union of rectangles) (Fig. 4). The values are summarized in the Table 1.

**Table 1:** Extracted self and mutual reluctances

| Fundamental loop | Reluctance [1/H] |
|---|---|
| 11 | 2.40630e+010 |
| 22 | 3.14921e+011 |
| 33 | 4.30807e+010 |
| 44 | 2.29307e+011 |
| 12 | 2.53936e+011 |
| 13 | 3.83304e+010 |
| 14 | 2.00067e+011 |
| 23 | 3.97307e+011 |
| 24 | 1.97317e+012 |
| 34 | 2.76450e+011 |

With the reluctances obtained, the entire MEEC of the LNA circuit can then be simulated in SPICE. As can be seen from Fig. 5, the MEEC simulation is a closer match to measurement than conventional electrical only uncoupled simulation.

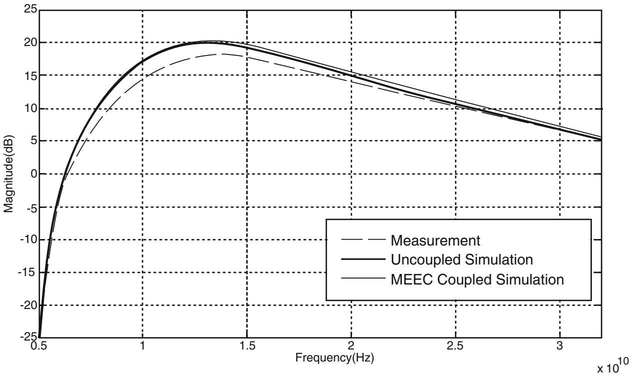


**Fig. 5:** LNA measurement of S21 (*thin green*) with electrical uncoupled simulation (*red*) and MEEC coupled simulation (*cyan*) overlayed

## 4 Conclusion

Based on the results obtained from using the Chameleon RF prototype software on the above benchmarks, our approach based on the original concepts of MEEC to model coupled problems in VLSI design using electromagnetic circuit elements (EMCE) and electromagnetic hooks is feasible and promising. The approach is a comprehensive multi-scale modeling solution using domain decomposition, hierarchical substrate structuring and compact parametric models to model passive integrated structures and functional blocks and the electric and magnetic parasitic interactions between them.

The concept of a simply connected EMCE facilitates the multi-scale modeling of coupled electromagnetic effects within a cohesive framework. These electromagnetic couplings between functional blocks are described by means of electromagnetic hooks which are special boundary conditions associated to the electromagnetic field problem [3]. The electromagnetic hooks enable the concise decomposition (partitioning) and description of the conductive, capacitive and/or inductive coupling between a component and the environment. These parasitic couplings between components are coupled in reality through the air and/or the silicon substrate [7, 8].

## Acknowledgement

The authors would like to gratefully acknowledge support and insightful discussion with Prof. Daniel Ioan, Asc.Prof. Gabriela Ciuprina. The authors would also like to acknowledge the financial support offered by the European Commission under the Marie Curie Fellowship (FP6) programme and also the Romanian CEEX/nEDA program. The views stated herein reflect only the author's views and the European Commission is not liable for any use that may be made of the information contained.

## References

1. E. Mollick, Establishing Moore's Law. IEEE Annals of the History of Computing, **28**, 62–75, (2006)
2. Niehof, J., Janssen, H., Schilders, W.: Comprehensive High-Accuracy Modeling of Electromagnetic Effects in Complete Nanoscale RF blocks: CHAMELEON RF. In: Proceedings of IEEE Workshop on Signal Propagation on Interconnects (2006)
3. Niehof, J., Janssen, H.H.J.M., Schilders, W.H.A., Schoenmaker, W., Ioan, D., Ciuprina, G., Pflanzl, W.: Domain Decomposition via Electromagnetic Hooks for the Modeling of Complete RF blocks. In: Proceedings of Signal Propagation on Interconnects (2008)
4. Ioan, D., Schilders, W.H.A., Ciuprina, G., van der Meijs, N., Schoenmaker, W.: Models for Integrated Components Coupled with their EM Environment. ISEF 2007 – XIII International Symposium on Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering, Prague (2007)
5. Ioan, D., Ciuprina, G., Mihalache, D.: Reduced Order Electromagnetic Models on on-chip passives based on dual Finite Integrals Technique. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 161–166. Springer, Berlin Heidelberg New York (2007)
6. Weiland, T., Clemens, M.: Discrete electromagnetism with the finite integration technique. Progress in Electromagnetics Research (PIER), **32**, 65–87, (2001)
7. Pflanzl, W.C., Seebacher, E.:Investigation of Substrate Noise Coupling and Isolation Characteristics for a 0.35 $\mu$m HV CMOS Technology. Mixed Design of Integrated Circuits and Systems, 2007. MIXDES'07. 14th International Conference, pp. 429–432 (2007)
8. Bronckers, S., Vandersteen, G., De Locht, L., van der Plas, G., Rolain, Y.: Study of the different coupling mechanisms between a 4 GHz PPA and a 5-7 GHz LC-VCO. RFIC Symposium, 15-17 June (2008)

# A Robust Technique for Modelling Nonlinear Lumped Elements Spanning Multiple Cells in FDTD

Luis R.J. Costa, Keijo Nikoskinen, and Martti Valtonen

**Abstract** A robust technique for modelling linear and nonlinear lumped elements spanning multiple cells in an FDTD-based electromagnetic field simulator is presented. The nonlinear models require iteration as part of the model. The technique is applied to produce a highly stable LE-FDTD diode model that works well far beyond normal operational voltage ranges. Simulation results are in good agreement with those obtained with the circuit simulator APLAC and those in the literature [1].

## 1 Introduction

Simulating complex electronic systems typically requires the use of electromagnetic field simulators to simulate parts of the whole. Often lumped elements are required to be embedded within the field simulation to correctly model sources and loads within the simulated subsystem. Embedding lumped elements in a 3D finite-difference time-domain (FDTD) field simulation was first accomplished in Ref. [1] for the passive elements, the resistive voltage source, and the diode, each spanning one cell, and the transistor spanning two cells. The lumped element–FDTD (LE-FDTD) method, as this technique is called, requires an iteration routine to solve the transcendental equations resulting in the diode and transistor models due their exponential current–voltage relation. The one-celled resistive voltage source LE-FDTD model was extended in Ref. [2] to span multiple cells by solving a system of linear equations to simultaneously update the electric field in the region occupied by the source. Modelling the linear passive elements with this approach is straightforward. The authors presented a model for the resistive voltage source that does away with the need for a linear equation solver [3], and the approach is readily applied to the

Luis R.J. Costa, Keijo Nikoskinen, Martti Valtonen
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland, e-mail: luis.costa@tkk.fi, keijo.nikoskinen@tkk.fi, martti.valtonen@tkk.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_8,

linear passive elements as well. The method, suitably extended, turns out to be an effective technique for modelling nonlinear lumped elements, like the diode, spanning multiple cells without having to solve a system of equations, thus making the implementation simpler. The strongly nonlinear diode requires iteration to obtain a stable model, which turns out to be stable over a very wide voltage range.

## 2 Modelling Technique

In the LE-FDTD method, in order to model an embedded lumped element, the current density term in Ampère's circuital law is split into the conduction current density $\mathbf{J}_\mathrm{c} = \sigma\mathbf{E}$ of the lossy medium and the current density $\mathbf{J}_\mathrm{l}$ due to the lumped element. The standard FDTD algorithm is then applied to evaluate the $\mathbf{E}$ field at time $n+1$ and the $\mathbf{H}$ field at $n+1/2$ [4]:

$$\oint_C \mathbf{H}^{n+\frac{1}{2}} \cdot \mathrm{d}\mathbf{l} = \int_S \left( \varepsilon \frac{\partial \mathbf{E}^{n+\frac{1}{2}}}{\partial t} + \sigma \mathbf{E}^{n+\frac{1}{2}} + \mathbf{J}_\mathrm{l}^{n+\frac{1}{2}} \right) \cdot \mathrm{d}\mathbf{s}, \tag{1}$$

where $\mathbf{E}^{n+\frac{1}{2}}$ is computed semi-implicitly as the average $(\mathbf{E}^{n+1} + \mathbf{E}^{n})/2$.

Assuming, without loss of generality, a $z$-directed lumped element current in the following, (1) is discretised in the standard manner and some short-hand notation is introduced at the same time to derive the update equation for $E_z$, the $z$ component of $\mathbf{E}$. The equations for the $x$ and $y$ directed currents are similarly derived.

So, denoting the discretised left-hand side of (1), which is the current at position $i,j,k$ due to the circulating magnetic field, as

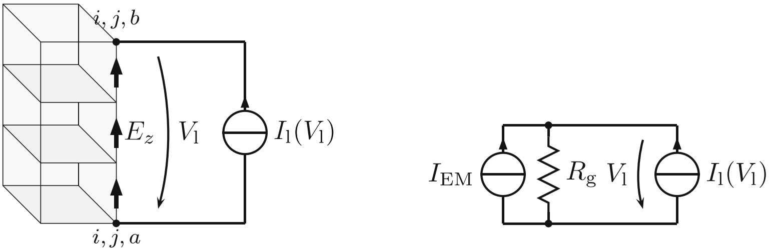
$$\mathscr{I}_z\big|_{i,j,k}^{n} = \left(H_y\big|_{i+\frac{1}{2},j,k+\frac{1}{2}}^{n+\frac{1}{2}} - H_y\big|_{i-\frac{1}{2},j,k+\frac{1}{2}}^{n+\frac{1}{2}}\right)\Delta_y - \left(H_x\big|_{i,j+\frac{1}{2},k+\frac{1}{2}}^{n+\frac{1}{2}} - H_x\big|_{i,j-\frac{1}{2},k+\frac{1}{2}}^{n+\frac{1}{2}}\right)\Delta_x, \tag{2}$$

the update equation for the $E_z$ field at a cell containing the lumped element is

$$E_z\big|_{i,j,k+\frac{1}{2}}^{n+1} = \left( \frac{1 - \frac{\sigma\Delta_t}{2\varepsilon}}{1 + \frac{\sigma\Delta_t}{2\varepsilon}} \right) E_z\big|_{i,j,k+\frac{1}{2}}^{n} + \left( \frac{\frac{\Delta_t}{\varepsilon}}{1 + \frac{\sigma\Delta_t}{2\varepsilon}} \right) \times \left\{ \frac{\mathscr{I}_z\big|_{i,j,k}^{n}}{\Delta_x\Delta_y} - \frac{I_\mathrm{l}(V_\mathrm{l}^{n+\frac{1}{2}})}{\Delta_x\Delta_y} \right\}. \tag{3}$$

The position indices of $\varepsilon$ and $\sigma$, which are the same as that of $E_z$, are omitted for notational compactness and must be added by the reader. The current $I_\mathrm{l}$ through the lumped element is a function of the voltage $V_\mathrm{l}$ across itself. Compared to the standard update equations elsewhere, the update equation (3) of the cells with the embedded lumped element contains the additional current density term $I_\mathrm{l}(V_\mathrm{l})/(\Delta_x\Delta_y)$ accounting for the lumped element current flowing through these cells across which the lumped element is connected.

The next step is to find a generic expression for $I_\mathrm{l}(V_\mathrm{l}^{n+\frac{1}{2}})$ to be plugged into (3). Applying the semi-implicit approximation, the voltage of the lumped element connected across multiple cells, as shown in Fig. 1a, is



(a) (b)

**Fig. 1: a** Lumped element connected across three cells, and **b** the circuit equivalent of (8)

$$V_{\mathrm{l}}^{n+\frac{1}{2}} = -\sum_{m=a}^{b} \frac{E_z|_{i,j,m}^{n+1} + E_z|_{i,j,m}^{n})}{2} \Delta_z \tag{4}$$

$$= -\sum_{m=a}^{b} \left\{ \frac{2}{1+\frac{\sigma\Delta_t}{2\varepsilon}} E_z|_{i,j,m}^{n} + \frac{\frac{\Delta_t}{\varepsilon}\frac{\mathscr{I}_z|_{i,j,m}^{n}}{\Delta_x\Delta_y}}{1+\frac{\sigma\Delta_t}{2\varepsilon}} \right\} \frac{\Delta_z}{2} + I_{\mathrm{l}}(V_{\mathrm{l}}^{n+\frac{1}{2}}) \sum_{m=a}^{b} \frac{\Delta_z}{2\Delta_x\Delta_y} \frac{\Delta_t}{\varepsilon\left(1+\frac{\sigma\Delta_t}{2\varepsilon}\right)}. \tag{5}$$

Above, (5) results from (4) by substituting $E_z|^{n+1}$ from (3). Denote

$$R_{\mathrm{g}} = \sum_{m=a}^{b} \frac{\Delta_z}{2\Delta_x\Delta_y} \frac{\Delta_t}{\varepsilon\left(1+\frac{\sigma\Delta_t}{2\varepsilon}\right)}, \tag{6}$$

which can be interpreted as the grid resistance seen by the lumped element. Note here that the exact role of the summation index $m$ is not evident in (6) because the position indices associated with $\varepsilon$ and $\sigma$ have been omitted for notational compactness. Next, divide both sides of (5) by $R_{\mathrm{g}}$, and denote

$$I_{\mathrm{EM}} = -\sum_{m=a}^{b} \left\{ \frac{2}{1+\frac{\sigma\Delta_t}{2\varepsilon}} E_z|_{i,j,m}^{n} + \frac{\frac{\Delta_t}{\varepsilon}\frac{\mathscr{I}_z|_{i,j,m}^{n}}{\Delta_x\Delta_y}}{1+\frac{\sigma\Delta_t}{2\varepsilon}} \right\} \frac{\Delta_z}{2} \Big/ R_{\mathrm{g}}, \tag{7}$$

which can be interpreted as the current due to the FDTD grid flowing through the cells containing the embedded lumped element, to arrive at

$$\frac{V_{\mathrm{l}}^{n+\frac{1}{2}}}{R_{\mathrm{g}}} = I_{\mathrm{EM}} + I_{\mathrm{l}}(V_{\mathrm{l}}^{n+\frac{1}{2}}). \tag{8}$$

This equation is Kirchhoff's current law for the circuit in Fig. 1b which also provides the generic expression sought.

Thus, the interface necessary for LE-FDTD modelling is formed by (3) and (8). For a time-invariant medium, the gridresistance $R_{\mathrm{g}}$, being a constant, can be precomputed before starting the time stepping. The following two sections discuss the model implementation details for the linear and nonlinear elements.

## 2.1 Modelling Linear Elements

For a linear element, the current source $I_\mathrm{l}$ in Fig. 1b is in fact an impedance that must have a suitable time-domain current–voltage description for it to be realisable as an LE-FDTD model. The element's current–voltage dependence may contain an additional constant term representing a source within. The implementation of the linear element entails solving from the circuit in Fig. 1b either $I_\mathrm{l}$ directly or $V_\mathrm{l}$, in which case the voltage is put into (8) to get $I_\mathrm{l}$. The current $I_\mathrm{l}$ is finally substituted into (3) to get the desired LE-FDTD model, i.e. the update equation. This technique was used in [3] to implement an LE-FDTD resistive voltage source.

## 2.2 Modelling Nonlinear Elements

For a nonlinear element, however, Newton-Raphson iteration, for example, must be used to solve $V_\mathrm{l}$ from the circuit in Fig. 1b, since otherwise the simulation can be rendered unstable; the stability is very sensitive to the model parameters used. Thus, the general LE-FDTD nonlinear iteration model required is derived as follows:

1. Iterate $V_\mathrm{l}^{n+\frac{1}{2}}$ using

$$V_\mathrm{l}^{n+\frac{1}{2},l+1} = V_\mathrm{l}^{n+\frac{1}{2},l} - \frac{V_\mathrm{l}^{n+\frac{1}{2},l}/R_\mathrm{g} + I_\mathrm{l}(V_\mathrm{l}^{n+\frac{1}{2},l}) - I_\mathrm{EM}}{1/R_\mathrm{g} + I_\mathrm{l}'(V_\mathrm{l}^{n+\frac{1}{2},l})}, \tag{9}$$

   where $I_\mathrm{l}'$ is the derivative with respect to $V_\mathrm{l}$ and $l$ is the iteration index. A good initial guess to start the iteration at each new time point is the previous value $V_\mathrm{l}^{n-\frac{1}{2}}$, i.e. $V_\mathrm{l}^{n+\frac{1}{2},0} = V_\mathrm{l}^{n-\frac{1}{2}}$. (The iteration converges rapidly for the diode model discussed in Sect. 3.)
2. Compute current $I_\mathrm{l}(V_\mathrm{l}^{n+\frac{1}{2}})$ from the known function, or equivalently from (8). The former is typically heavier to compute making the latter the preferred choice.
3. Use the computed $I_\mathrm{l}(V_\mathrm{l}^{n+\frac{1}{2}})$ to update $E_z$ in the cells spanned by the nonlinear element using (3), having the corresponding position indices $i, j, m$, where $m$ spans the $k$ index from $a$ to $b$.

The advantage of this technique is that $I_\mathrm{l}$ needs to be computed only once at each time point and then used in the update equation of the stack of cells containing the modelled lumped element; no system of nonlinear equations needs to be solved.

Also, the presence of the resistance $R_\mathrm{g}$ in (9) improves the stability of the iteration model.

An alternative approach to model the nonlinear element is to express the element current as a function of $E_z|_{i,j,m}$ ($m$ ranging from $a$ to $b$) using (4), plug this $I_\mathrm{l}(E_z)$ into (3) for all the relevant cells (ranging from $i,j,a$ to $i,j,b$), and then simultaneously solve all the $E_z$s at time $n+1$ using multivariable Newton-Raphson iteration. The disadvantage of this approach is that more computational effort is required to obtain the updated $E_Z$ values and the implementation is more complicated.

## 3 Model Example: Diode

The above technique for implementing a nonlinear lumped element is exemplified using a diode model. The implemented LE-FDTD diode is simulated and the simulation results are compared with those obtained using the circuit models of a circuit simulator.

A simple diode current–voltage relation is given by

$$I_\mathrm{d} = I_\mathrm{s}\left(\mathrm{e}^{\frac{q}{kT}V_\mathrm{d}} - 1\right), \tag{10}$$

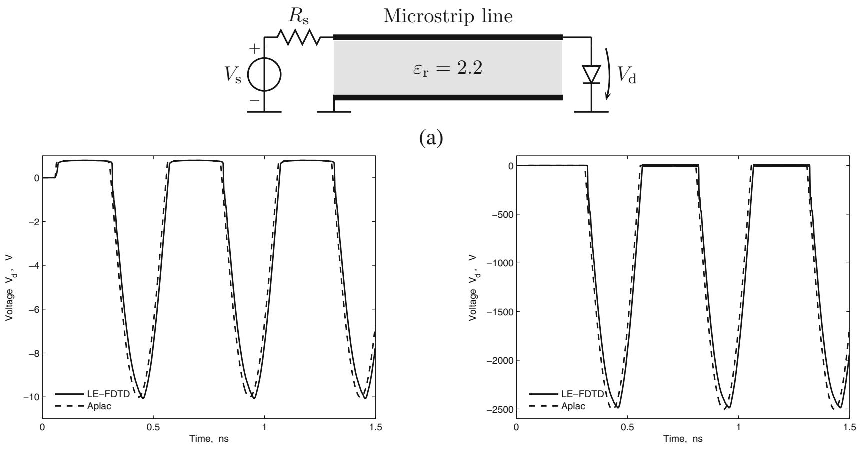
where $I_\mathrm{s}$ is the saturation current, $q$ is the electron charge, $k$ is Boltzmann's constant, and $T$ is the temperature in Kelvin. The diode voltage $V_\mathrm{d}$ ($= V_\mathrm{l}$) iteration equation is derived from (9):

$$V_\mathrm{d}^{n+\frac{1}{2},l+1} = V_\mathrm{d}^{n+\frac{1}{2},l} - \frac{V_\mathrm{d}^{n+\frac{1}{2},l}/R_\mathrm{g} + I_\mathrm{s}\left(\mathrm{e}^{\frac{q}{kT}V_\mathrm{d}^{n+\frac{1}{2},l}} - 1\right) - I_\mathrm{EM}}{1/R_\mathrm{g} + \frac{q}{kT}I_\mathrm{s}\mathrm{e}^{\frac{q}{kT}V_\mathrm{d}^{n+\frac{1}{2},l}}}, \quad V_\mathrm{d}^{n+\frac{1}{2},0} = V_\mathrm{d}^{n-\frac{1}{2}}. \tag{11}$$

The iterated $V_\mathrm{d}^{n+\frac{1}{2}}$ is then plugged into (10) to determine the diode current $I_\mathrm{d}$ ($= -I_\mathrm{l}$). Alternatively, $I_\mathrm{d}$ is found using (8), which is the preferred option. Then $I_\mathrm{d}$ is substituted into (3) to update $E_z$ in the cells spanned by the diode from $i,j,a$ to $i,j,b$.

To test this model, the microstrip circuit in Fig. 2a was simulated with a self-written C-language FDTD field solver. The microstrip of width $6\Delta_x$, length $30\Delta_y$, and substrate height $3\Delta_z$ has a $50\,\Omega$ characteristic impedance. The FDTD space of size $64 \times 50 \times 12$ cells, with $\Delta_x = 0.4064\,\mathrm{mm}$, $\Delta_y = 0.4233\,\mathrm{mm}$ and $\Delta_z = 0.265\,\mathrm{mm}$, was truncated with a first-order Mur boundary condition. The LE-FDTD resistive voltage source was implemented as described in Ref. [3] with source resistance $R_\mathrm{s} = 50\,\Omega$ and frequency of the sinusoid excitation 2 GHz. The diode parameters used were $I_\mathrm{s} = 10^{-14}\,\mathrm{A}$ and $T = 300\,\mathrm{K}$. The simulation was run for 2286 time steps using different values for the sinusoidal excitation amplitude.

Figures 2b and 2c show the diode voltage for the source amplitudes 10 V and 2.5 kV, respectively. The LE-FDTD diode model displays stable behaviour even for the 2.5 kV excitation. To verify these results, the simulation results are compared



(a)

(b)

(c)

**Fig. 2:** **a** Microstrip test circuit used to verify the diode model. **b** Diode voltage when the source amplitude is 10 V and **c** 2.5 kV. The simulation results are compared with those got with APLAC

with those obtained using the circuit models in the circuit simulator APLAC [5] for the test circuit and are seen to be in good agreement. At 200 MHz, the 10 V source amplitude yields the simulation result obtained in Ref. [1], whose one-cell diode model is reported to be stable using an excitation amplitude of 15 V. Extending the model in Ref. [1] to span multiple cells would require the solution of a nonlinear system of equations, making the implementation complicated. The Newton-Raphson iteration of the diode voltage of the proposed LE-FDTD diode converged rapidly, typically in two to four iterations.

At 2.6 kV the LE-FDTD simulation became unstable, but the instability originated in the Newton-Raphson iteration, which was implemented without any time-step control. A more sophisticated time-step control will extend the working range of this diode model further. In RF applications, the useful operating voltage range for the diode is typically up to about 15 V. The high voltages were used in these simulations only to test the stability of the diode model.

At frequencies above about 3 GHz the circuit models for microstrip structures, as also for other planar transmission line structures, become increasingly inaccurate, thus making field simulations imperative. This justifies the development of robust LE-FDTD models to be used in such planar transmission line and waveguide structure field simulations.

## 4 Conclusion

A technique for modelling lumped elements spanning multiple cells in an FDTD field simulator is presented. The technique does not increase the complexity of desired nonlinear element models and yet displays great operational stability. The technique is used to build a novel diode model that is stable far beyond useful operational voltages. The diode model produces simulation results that are in agreement with those produced by the circuit models in the circuit simulator APLAC and those in the literature.

## References

1. Picket-May, M., Taflove, A., Baron, J.: FD-TD modeling of digital signal propagation in 3-D circuits with passive and active loads. IEEE Trans. Microwave Theory and Tech. **MTT-42**(8), 1514–1523, (1994)
2. Xu, J., Zhao, A.P., Räisänen, A.V.: A stable algorithm for modeling lumped circuit source across multiple FDTD cells. IEEE Microwave and Guided Wave Letters **7**(9), 308–310 (1997)
3. Costa, L.J., Nikoskinen, K., Valtonen, M.: Models for the LE-FDTD resistive voltage source spanning multiple cells. In: Proceedings of the 2007 European Conference on Circuit Theory and Design, ECCTD 2007, August 26-30, pp. 671–674 (2007)
4. Taflove, A., Hagness, S.C.: Computational Electrodynamics: the Finite-Difference Time Domain Method, 2nd edition. Artech House, Boston (2000)
5. APLAC — Circuit Simulation and Design Tool, AWR–APLAC Corporation, Finland (2008) http://www.awrcorp.com/

# Computation of Eigenmodes in Periodic Structures with Dispersive Materials

Bastian Bandlow and Rolf Schuhmann

**Abstract** In the infrared spectrum noble metals do not act like perfect conductors, but have to be described by dispersive material models. Eigenvalue problems including such frequency-dependent material properties occur for instance, when the dispersion relations of periodic structures such as photonic crystals and metamaterials are analyzed by electromagnetic field simulations of the corresponding unit cells. We show that the commonly used Drude dispersion model leads to a polynomial eigenvalue problem which can be solved by a modified Jacobi-Davidson method.

## 1 Introduction

Periodic electromagnetic structures with lattice constants $\Delta$ (i.e. length of spatial periodicity) comparable to the wavelength of an incident wave are favorably described by their dispersion relation. The electromagnetic eigenmodes of such periodic structures may be cast into an electromagnetic band structure, which gives information about the propagating modes. Note that we use *electromagnetic band structure* synonym to the expression *dispersion relation*. A sample dispersion relation is shown in Fig. 1 (left). On the abscissa the phase shift $\varphi_z$ over one spatial period can be interpreted as a scaled macroscopic wave number $\varphi_z = k_z\Delta$. On the ordinate there is the angular frequency of several modes, which can be of forward or backward type (depending on the slope of the curve, which can be interpreted as group velocity). Moreover, the phenomenon of Bragg reflection leads to stop bands between the propagating modes.

A convenient way to obtain the dispersion relation by electromagnetic field simulation is a series of several eigenmode computations of one spatial unit cell for

Bastian Bandlow, Rolf Schuhmann
FG Theoretische Elektrotechnik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany, e-mail: bandlow@tet.upb.de, schuhmann@tet.upb.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_9,

different (fixed) phase shifts $\varphi_z$. These phase shifts are introduced into the model by imposing periodic boundaries in propagation direction, see Fig. 1b, relating the fields on both interfaces to each other by the factor $e^{\mathrm{i}\varphi_z}$.

After discretizing the structure with the finite integration technique (FIT, [1]), the resulting algebraic eigenvalue problem has the form

$$\mathbf{A}_{cc,\varphi}\widehat{\mathbf{e}} = \omega^2 \mathbf{M}_\varepsilon \widehat{\mathbf{e}}. \tag{1}$$

Here, $\mathbf{A}_{cc,\varphi}$ is the curl-curl system matrix derived from the discrete form of Maxwell's equations. It includes the double curl operation, the permeability distribution of the structure, and also the periodic boundary conditions for one specific phase shift $\varphi_z$. The diagonal matrix $\mathbf{M}_\varepsilon$ is the generalized permittivity operator. The searched eigenvalue is the squared angular frequency $\omega^2$, and the field distribution is obtained by the eigenvector $\widehat{\mathbf{e}}$. Based on a three-dimensional grid model, the FIT discretization leads to large sparse matrices with the dimension $N_e \times N_e$, with $N_e$ the number of grid edges.

An example for a metamaterial unit cell is shown in Fig. 1 (right). It consists of two rectangular structures of silver and a dielectric spacing in between [2]. This structure is supposed to operate at optical wavelengths around 1.4 microns. At these wavelengths, noble metals like gold and silver show a dispersion of their dielectric constant. Since now the entries in the permittivity matrix $\mathbf{M}_\varepsilon$ depend on the searched eigenfrequency $\omega$, we obtain a non-linear eigenvalue formulation.

A first approach is to evaluate the material properties at a specific frequency $\omega_0$ and to perform a fixed point iteration over several frequencies $\omega_i$, which may converge to the desired eigenfrequency (and eigenmode) of interest. For many materials, however, the dependency $\varepsilon(\omega)$ can be approximated by explicitly known *rational* functions, e.g. in case of the commonly used Drude [3] and Lorentz-models, and the problem can be reformulated leading to a polynomial eigenvalue problem (PEP). This gives rise to better solution approaches than a fixed point iteration.

The rest of the paper is organized as follows: In section 2, we show how to formulate the eigenvalue problem including dispersive materials. In section 3 we discuss some strategies to solve this problem, and in section 4 we give a numerical example. Further ideas and suggestions are given in section 5.

## 2 Formulation

To simplify the presentation, we start with a homogeneous medium and the continuous rather than the discrete notation. From Maxwell's equations in frequency domain we derive the eigenmode formulation for the electrical field strength $\mathbf{E}$

$$\operatorname{curl} \frac{1}{\mu} \operatorname{curl} \mathbf{E} = \omega^2 \varepsilon \mathbf{E}. \tag{2}$$

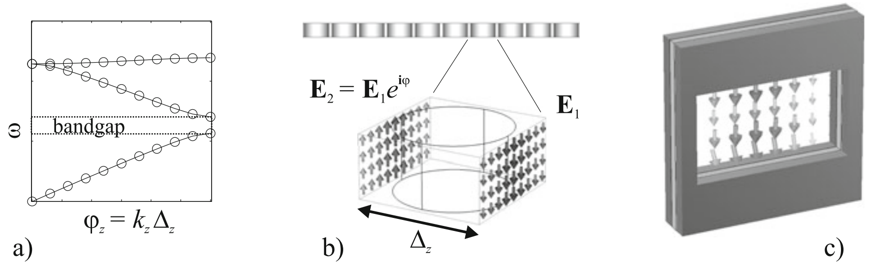


**Fig. 1:** *Left*: Example of a typical dispersion relation with forward and backward modes and a photonic band gap. *Center*: Unit Cell from an one-dimensional lattice and the relation of the fields at the interface planes. *Right*: Sample unit cell structure taken from [2] consisting of two layers of Drude-dispersive silver and a dielectric spacing in between

The dependency on frequency of the permittivity $\varepsilon(\omega)$ shall be given by a general 2nd order model

$$\varepsilon(\omega) = \varepsilon_0 \left( \varepsilon_\infty + \frac{\beta_0 + \mathrm{i}\omega\beta_1}{\alpha_0 + \mathrm{i}\omega\alpha_1 - \omega^2} \right), \tag{3}$$

with the real-valued parameters $\varepsilon_\infty$, $\alpha_0$, $\alpha_1$, $\beta_0$ and $\beta_1$ implying an $e^{\mathrm{i}\omega t}$ time dependency. Inserting equation (3) into equation (2) leads to a complex, non-hermitian, polynomial eigenvalue problem (PEP) in $\omega$, which can be notated by

$$(\omega^4 A_4 + \omega^3 A_3 + \omega^2 A_2 + \omega A_1 + A_0)\mathbf{E} = 0. \tag{4}$$

The coefficients $A_i$ are given by

$$A_4 = -\varepsilon_0\varepsilon_\infty, \quad A_3 = \mathrm{i}\varepsilon_0(\alpha_1\varepsilon_\infty + \beta_1), \quad A_2 = \varepsilon_0(\varepsilon_\infty\alpha_0 + \beta_0) + A_{cc}$$

$$A_1 = -\mathrm{i}\alpha_1 A_{cc}, \quad A_0 = -\alpha_0 A_{cc}, \quad A_{cc} = \mathrm{curl}\,\frac{1}{\mu}\mathrm{curl}.$$

Using the Finite Integration Technique (FIT), this representation of the PEP can be transformed into a discrete formulation in a straight forward manner

$$\Psi(\omega)\widehat{\mathbf{e}} = (\omega^4\mathbf{A}_4 + \omega^3\mathbf{A}_3 + \omega^2\mathbf{A}_2 + \omega\mathbf{A}_1 + \mathbf{A}_0)\widehat{\mathbf{e}} = 0, \tag{5}$$

where $\mathbf{A}_i$ are the coefficient matrices corresponding to the expressions above. Of course, this FIT model also supports arbitrary inhomogeneous material distributions, and it turns out that the usual facet-weighted averaging procedure at interfaces does not need any special treatment.

The periodic boundary condition assures that the tangential components of the electrical grid voltage of two opposed boundary interfaces are related by a predefined factor $e^{\mathrm{i}\varphi}$. The dependency of the corresponding components is inserted into the curl-curl-operator. The overall number of degrees of freedom is then reduced by the components at the dependent boundary plane.

# 3 Solving the Polynomial Eigenvalue Problem

There are various methods to solve the PEP from the previous section, and we will discuss three variants.

## *3.1 Fixed-Point Iteration*

The PEP can be formulated as a fixed-point iteration $\omega_{i+1} = \Phi(\omega_i)$, where the operator $\Phi$ contains a standard linear eigenvalue problem. The dispersive material properties in $\mathbf{M}_\varepsilon(\omega)$ are evaluated at a certain frequency $\omega_i$ and then we solve the (linearized) eigenvalue problem from equation (5). This leads to the iterative scheme

$$\mathbf{A}_{cc,\varphi}\widehat{\mathbf{e}} = \omega_{i+1}^2 \mathbf{M}_\varepsilon(\omega_i)\widehat{\mathbf{e}}. \tag{6}$$

This scheme works well (even if not very fast) in many cases. However, to prove its convergence for general cases — e.g. using Banach's fixed-point theorem — is quite challenging due to the complex nature of the eigenvalue problem involved.

## *3.2 Linearization*

A more direct way to solve the PEP uses the so-called companion matrix of the PEP. The PEP (5) of order 4 can be transformed into a generalized linear eigenvalue problem of the form $\mathbf{Ax} = \lambda \mathbf{Bx}$ of higher dimension, where

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{I} & 0 & 0 \\ 0 & 0 & \mathbf{I} & 0 \\ 0 & 0 & 0 & \mathbf{I} \\ -\mathbf{A}_0 & -\mathbf{A}_1 & -\mathbf{A}_2 & -\mathbf{A}_3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{I} & & & \\ & \mathbf{I} & & \\ & & \mathbf{I} & \\ & & & \mathbf{A}_4 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \widehat{\mathbf{e}} \\ \lambda\widehat{\mathbf{e}} \\ \lambda^2\widehat{\mathbf{e}} \\ \lambda^3\widehat{\mathbf{e}} \end{bmatrix}. \tag{7}$$

For the square coefficient matrices of (5) of dimension $n$, this approach leads to $3 \times n$ additional eigenvalues, which are not necessarily all solutions of the original PEP and therefore have to be dropped. Since this approach leads to large and typically ill-conditioned matrices, it is only feasible for low-dimensional problems.

## *3.3 Jacobi–Davidson Algorithm for PEPs*

The PEP in the form (5) can be solved using a Jacobi-Davidson method (JD) [4, 5]. This method is intended to find one or more interior eigenvalues of the spectrum near a given estimation $\tau$. In the JD method, the PEP is projected on a low-dimensional orthogonal subspace $\mathbf{V}$, leading to a low-dimensional PEP with coefficient matrices $\mathbf{M}_i = \mathbf{V}^*\mathbf{A}_i\mathbf{V}$. This PEP is solved, e.g. by the companion matrix approach from section 3.2. The low-dimensional eigenvector $\mathbf{s}$ is expanded again to generate the current approximation of the searched eigenvector $\mathbf{u} = \mathbf{V}\mathbf{s}$. From this approximation a residual is calculated, and the iteration is stopped if it is small enough. If this is not the case, a so-called correction equation is formulated and solved, which produces an additional vector to be added to the subspace. This procedure is repeated until convergence. The computationally most expensive task inside the Jacobi-Davidson iteration is the solution of the correction equation, which reads

$$\left(\mathbf{I} - \frac{\mathbf{p}\mathbf{u}^*}{\mathbf{u}^*\mathbf{p}}\right)\Psi(\theta)(\mathbf{I} - \mathbf{u}\mathbf{u}^*)\mathbf{t} = -\mathbf{r} \tag{8}$$

Here, $\Psi$ is the polynomial from equation (5) evaluated at the last, best estimation $\theta$ (namely the Ritz value), $\mathbf{t}$ the correction vector, $\mathbf{r}$ the residual, $\mathbf{u} = \mathbf{V}\mathbf{s}$ the best actual approximation of the searched eigenvector, and $\mathbf{p} = \Psi'(\theta)$. In the beginning of the JD process the user-defined estimation $\tau$ may be more accurate than an extracted Ritz value. Therefore we set $\mathbf{p} = \Psi'(\tau)$ until the residual is below a predefined tolerance, in order to get a better correction vector $\mathbf{t}$. (This was proposed in [6].)

For the solution of this equation we use a preconditioned bicgstab(l) method [7], where the preconditioner is an LU decomposition of the polynomial evaluated at the estimation $\tau$. Therefore this LU decomposition has to be established only once per JD run. In [8] it has been proposed to replace equation (8) by

$$\mathbf{t} = -\Psi(\theta)^{-1}\mathbf{r} + \varepsilon_O\Psi(\theta)^{-1}\mathbf{p} \quad \text{where} \quad \varepsilon_O = \frac{\mathbf{u}^*\Psi(\theta)^{-1}\mathbf{r}}{\mathbf{u}^*\Psi(\theta)^{-1}\mathbf{p}}. \tag{9}$$

[8] also suggests to replace $\Psi(\theta)^{-1}$ by an appropriate preconditioner, but here we use the original matrix.

### 3.3.1 Scaling and Balancing

Since we look for eigenfrequencies $\omega$ which are in the magnitude of $10^{14}$ we may easily run into numerical troubles. In order to improve the situation we use two scaling and balancing steps.

In order to transform the eigenvalues to values near one, we apply the transformation proposed in [9] for the low-dimensional coefficient matrices $\mathbf{M}_i' = \alpha_{opt}^i\mathbf{M}_i = \alpha_{opt}^i\mathbf{V}^*\mathbf{A}_i\mathbf{V}$. The scaling factor is calculated by $\alpha_{opt} = (||\mathbf{M}_0||_2 \,/\, ||\mathbf{M}_i||_2)^{\frac{1}{i}}$. Of course it is also possible to calculate the 2-norms of the unprojected coefficient matrices $\mathbf{A}_i$ of (5) and to perform the scaling with these larger matrices. Since the calculation of a 2-norm is a rather expensive task, it is computationally more efficient to do that on the lower dimensional projected matrices, especially if we only need a very small number of JD iterations.

Additionally we follow an idea from [9], proposing a balancing transformation of the coefficient matrices. The norms of the coefficient matrices are balanced by $\widetilde{\mathbf{M}}_\mathbf{i} = \mathbf{D}_1\mathbf{M}_i'\mathbf{D}_2$, keeping the eigenvalues unchanged. The balancing matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal and contain only powers of two. Since a multiplication or division by two is realized through a bit-shift, no further numerical errors occur and the overall condition of the eigenvalues near the user-defined estimation $\tau$ is improved. The back-transformed low-dimensional eigenvector is obtained by $\mathbf{s} = \mathbf{D}_2\widetilde{\mathbf{s}}$.

### 3.3.2 Validation

A simple way to validate the results from the JD method is to go back to the original eigenvalue problem in equation (5) and to execute one single step of the fixed-point iteration described in section 3.1. In all numerical tests we have obtained accuracies in the range of numerical noise (if the JD process has fully converged).

## 4 Numerical Example

As an example we take the unit cell structure from [2] shown in Fig. 1. It consists of two layers of silver, which are 45 nm thick and a dielectric spacing of $\varepsilon_r = 1.9044$ and 30 nm thickness in between. The square edge length is 600 nm and the lattice constant is 200 nm. The rectangular aperture is 316 nm by 100 nm and the permeability is chosen to be $\mu = \mu_0$. The coefficients of the dispersive model (3) for silver are $\varepsilon_\infty = 5$, $\alpha_0 = 0$, $\alpha_1 = 3.22e13$, $\beta_0 = 1.96e32$ and $\beta_1 = 0$ [3]. Therefore, the general second order dispersive model from equation (3) results in the Drude-dispersive model. Hence, the resulting PEP is of order three. The transversal boundary conditions are also periodic with the transversal phase shifts $\varphi_{x,y} = 0$, leading to a field distribution at the input and output planes, which is similar to a plane wave. For the computation of the dispersion diagram, the periodic boundary phase shift in propagation direction is sampled 13 times. For each step the complex non-hermitian polynomial eigenvalue problem has to be solved. For the geometric modeling we use the commercial tool CST MICROWAVE STUDIO [10]. The discretization with a maximum mesh step of 19.75 nm leads to a 33x32x15 mesh with 41664 complex degrees of freedom (dof). The PEP is solved by the algorithm from section 3.3, and the dimension of the subspace $\mathbf{V}$ is kept between 6 and 13. Convergence is reached when the norm of the residual is less than 1e-9.

### *4.1 Results*

The computed dispersion diagram (featuring only the real part of the eigenvalue $\omega$) is shown in Fig. 2. It clearly shows an backward mode with $\frac{\partial\omega}{\partial k} < 0$, as it is expected for this kind of metamaterial unit cell [11]. As a measure of the losses inside the structure we also compute the $Q$-factor, which relates the energy stored in the structure to the energy dissipated per oscillation. It is calculated by $Q = 0.5\Re\{\omega\}/\Im\{\omega\}$ and has a value around 100 (Fig. 2). The algorithm from section 3.3 has been implemented in MATLAB and the average computation time is around 7 min per each of the 13 samples. Note that the major part of these 7 min is spent on the LU decomposition, which needs a large amount of memory in our implementation. In spite of this disadvantage the LU decomposition of the estimated target value yields a good preconditioner, and the bicgstab solver for equation (9) converges within a few iterations.

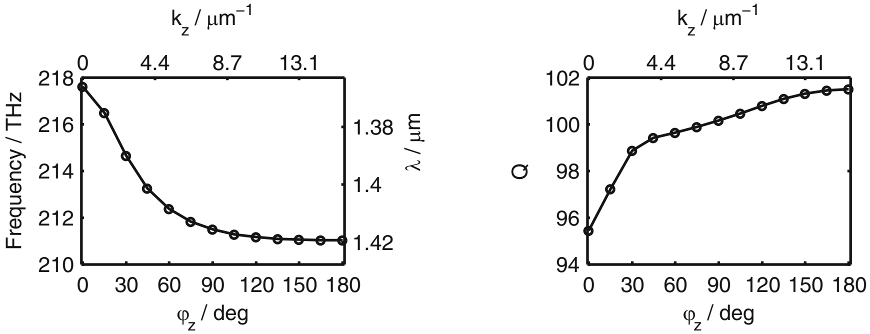


**Fig. 2:** *Left*: Dispersion relation $\Re\{\omega_i\}$, *Right*: Q-factors $Q_i = 0.5\Re\{\omega_i\}/\Im\{\omega_i\}$

For the first sample, the initial subspace of the JD run was chosen randomly. For the consecutive JD runs the eigenvector of the last JD run was used and it turned out that this choice has a dramatic influence on the convergence behavior. In Fig. 3 the number of JD iterations of all 13 runs are shown. The computation of the first sample ($\varphi_z = 0$) took 17 JD iterations, the phase shift $\varphi_z = 15°$ only 5 JD iterations, and each of the following samples took less than 5 JD iterations to converge.

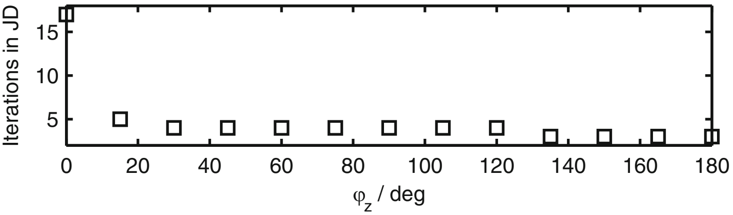


**Fig. 3:** The number of JD iterations of the 13 consecutive eigenvalue simulations suggests that the convergence behavior is improved by reusing the previous eigenvector as initial subspace for the following JD run

## 5 Outlook

An extension of the JD algorithm in [4] has recently been published in [6]. The harmonic and refined Rayleigh-Ritz approach was generalized for PEPs. Especially for start spaces of poor quality an improved convergence is expected.

The application to lossy waveguide ports, which contain a dispersive filling, is straight forward in the FIT formulation. Especially for so-called photonic crystal fibers including noble metals [12] a mode calculation is possible with our approach.

## 6 Conclusion

In this paper we have presented a formulation to handle electromagnetic eigenvalue problems from structures containing frequency dispersive materials. Our formulation is based on the finite integration technique (FIT) and leads to a polynomial eigenvalue problem (PEP). This PEP is solved using a Jacobi-Davidson method from [4, 5]. A numerical example of a fishnet-type metamaterial unit cell has been presented, where a Drude model for the permittivity of silver has to be used. The simulation converges within reasonable CPU-time and produces dispersion curves with negative group velocity as expected.

## References

1. Weiland, T.: Time Domain Electromagnetic Field Computation with Finite Difference Methods. International Journal of Numerical Modelling **9**(4), 295–319 (1996)
2. Dolling, G., Enkrich, C., Wegener, M., Soukoulis, C.M., Linden, S.: Simultaneous Negative Phase and Group Velocity of Light in a Metamaterial. Science **312**(5775), 892–894 (2006)
3. Johnson, P.B., Christy, R.W.: Optical Constants of the Noble Metals. Phys. Rev. B **6**(12), 4370–4379 (1972)
4. Sleijpen, G.L.G., Booten, A.G.L., Fokkema, D.R., der Vorst, H.A.V.: Jacobi-Davidson type methods for generalized and polynomial eigenproblems. BIT **36**(3), 595–633 (1996)
5. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.): Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM, Philadelphia (2000)
6. Hochstenbach, M.E., Sleijpen, G.L.G.: Harmonic and refined Rayleigh-Ritz for the polynomial eigenvalue problem. Numerical Linear Algebra with Applications **15**(1), 35–54 (2008)
7. Sleijpen, G.L.G., Fokkema, D.R.: Bi-cgstab(l) for linear equations involving unsymmetric matrices with complex spectrum. Elec. Trans. Numer. Anal. **1**, 11–32 (1993)
8. Hwang, T.M., Lin, W.W., Liu, J.L., Wang, W.: Jacobi-Davidson methods for cubic eigenvalue problems. Numerical Linear Algebra with Applications **12**(7), 605–624 (2005)
9. Betcke, T.: Optimal Scaling of Generalized and Polynomial Eigenvalue Problems. Tech. Rep. 2007.120, Manchester Institute for Mathematical Sciences, The University of Manchester, UK (2007)
10. Computer Simulation Technology AG (CST): CST Studio Suite 2008. Http://www.cst.com
11. Smith, D.R., Vier, D.C., Kroll, N., Schultz, S.: Direct calculation of permeability and permittivity for a left-handed metamaterial. Applied Physics Letters **77**(14), 2246–2248 (2000)
12. Schmidt, M.A., Sempere, L.N.P., Tyagi, H.K., Poulton, C.G., Russell, P.S.J.: Waveguiding and plasmon resonances in two-dimensional photonic lattices of gold and silver nanowires. Physical Review B (Condensed Matter and Materials Physics) **77**(3), 033417 (2008)

# Region-Oriented BEM Formulation for Numerical Computations of Electric Fields

Andreas Blaszczyk

**Abstract** The paper presents a concept of the region-oriented 3D formulation for the boundary element method (BEM) applied to computation of electric fields. Differences between the region-oriented and the traditional BEM are explained. Numerical tests performed for simple arrangements with high permittivity components show that the new approach leads to better accuracy than the traditional BEM.

## 1 Introduction

The calculation of electric fields based on the boundary element method became during last 10 years very popular in design of high voltage devices [1, 2]. The fundamental advantage of the BEM approach is related to the fact that only surface mesh on boundaries of solid parts is needed. The modeling of so called "air box" as well as the generation of solid mesh is not required, which significantly simplifies the discretization of complex models.

The traditional BEM approach [3,4] provides a good accuracy for dielectric problems with typical values of the relative permittivity (in the range between 1 and 10). For values of material properties far beyond the typical range significant numerical errors or inconsistent results may occur. An improvement of this behavior is of great interest from the view point of industrial applications.

In this paper the region-oriented formulation concept has been introduced to improve the numerical performance of the BEM approach.

This type of formulation has been used for the charge simulation method in the 1990s [5]. It has been successfully applied to calculate models with large permittivity values. This paper presents results of a numerical experiment in which the integration technique of the traditional BEM has been applied to the region-oriented formulation.

Andreas Blaszczyk
ABB Corporate Research, 5405 Baden, Switzerland, e-mail: andreas.blaszczyk@ch.abb.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_10,

## 2 Concept and Formulation

The basic concept is shown in Fig. 1. The model space has been divided into 3 regions. Each of the regions includes a homogeneous, linear and isotropic material. The field and potential in each region is calculated based on a single layer of charge located on the region boundaries. Consequently, on boundaries between two regions two single layers of charges are defined while along triple junctions (like point T in Fig. 1a) three different single charge layers meet together. Typically, the model definition includes one unbounded, open region as well as a few of non-computational regions that are excluded from field computations, see region 0 in Fig. 1. On the boundary of non-computational regions (electrodes) a potential value is prescribed.

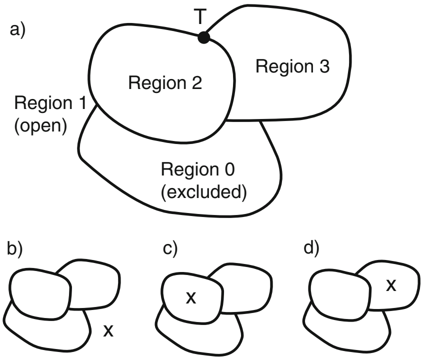


**Fig. 1** Basic concept of the region-oriented formulation: **a** geometrical configuration **b**, **c** and **d** charge assigned to regions 1, 2 and 3 respectively. *Comment:* For each of regions 1, 2 and 3 a separate, unique layer of charge is defined. An observer inside a region (denoted by *crosses*) can only see the charge layer of his region. All other layers are not used for field computation in his region

The formulation is based on boundary elements created on region boundaries. We formulate equations for collocation points located in corner nodes of the elements (in this implementation we consider triangles). For each corner node we assign one unknown value of surface charge density per region connected to this node. Additionally, an unknown value of reference potential is assigned to each region (except of the open region). The following types of equations are formulated:

- prescribed potential $\Phi_i$ at point $i$ on electrode boundary in region $K$:

$$\sum_{j\in K} p_{ij}\sigma_{jK} + \Phi_{rK} = \Phi_i \tag{1}$$

- potential continuity on interfaces between regions – this equation is formulated for each region $L$ that meets with region $K$ at point $i$:

$$\sum_{j\in K} p_{ij}\sigma_{jK} + \Phi_{rK} = \sum_{j\in L} p_{ij}\sigma_{jL} + \Phi_{rL} \tag{2}$$

- electric displacement (flux) continuity at the boundary point $i$ – one equation for all regions that meet at point $i$ (it applies the Gauss law to point $i$)

$$\sum_{K\in i} S_{iK}\varepsilon_{rK} \sum_{j\in K} \mathbf{n_{iK}} \cdot \nabla \mathbf{p_{ij}} \sigma_{jK} = 0 \tag{3}$$

- charge compensation (Gauss law) for closed regions – one equation per region (This equation is not formulated for open regions. In case of open regions the reference potential used in (1) and (2) is equal to zero.):

$$\sum_{i \in K} S_{iK} \sum_{j \in K} \mathbf{n_{iK}} \cdot \nabla \mathbf{p_{ij}} \sigma_{jK} = 0 \tag{4}$$

where $\varepsilon_{rK}$ is relative electric permittivity of region $K$ while $S_{iK}$ and $\mathbf{n_{iK}}$ are surface area and normal vector at collocation point $i$ assigned to region $K$. When calculating normals and surface area for junction points only the elements facing the region of interest are included. The normal vectors are pointing always to the region for which they are calculated (independent of the element orientation) At discontinuity points like sharp edges or triple junctions the normal vector contributions from all elements at corner node $i$ are averaged.

The characteristic feature of the region-oriented BEM is the fact that the equations are formulated at corner nodes of elements (in contrast to the region-oriented charge simulation where the equations are formulated for the center of surface elements). This has a significant impact on handling the triple junctions. For example, for the point T shown in Fig. 1 two potential continuity equations (2) and one flux continuity equation (3) are formulated. The latter one includes flux components of all 3 regions that meet at point T. The formulation presented here does not include any special handling of triple junctions on electrodes. For example, for each of triple points on electrode in Fig. 1) only 2 equations of type (1) are formulated.

The unknowns in (1)–(4) are surface charge densities $\sigma_{jK}$ at the charge locations $j$ and reference potentials $\Phi_{rK}$ (both assigned to region $K$). A solution of the equation system is got iteratively using GMRES without preconditioning. The potential $p_{ij}$ and field coefficients $\nabla \mathbf{p_{ij}}$ are calculated between the collocation point $i$ and the charge location $j$. The next section explains how these coefficients are calculated.

## 3 Traditional Versus Region-Oriented Approach

The traditional BEM introduces an equivalent single layer of surface charge distributed in vacuum. The formulation is based on the Fredholm integral equations with Green's function kernels [3]. The potential in collocation points is calculated according to the Fredholm integral equation of the first order:

$$\Phi_i = \frac{1}{4\pi\varepsilon_0} \sum_j \int_{S_j} \frac{1}{r_{ij}} \sigma_j \, \mathrm{d}S = \sum_j p_{ij} \sigma_j \tag{5}$$

where $\varepsilon_0$ is the permittivity of vacuum, $r_{ij}$ is the distance between collocation point $i$ and integration point $j$, $\sigma_j$ and $S_j$ are surface charge density and surface area at the charge location $j$. The potential coefficients $p_{ij}$ in equation (1) and (2) are computed exactly in the same way as above. The only difference is related to the summation, which is performed in equations (1) and (2) for the region of interest while in the equation (5) all charges are considered.

For a collocation point $i$ on a dielectric interface between the regions with relative permittivity $\varepsilon_{rK}$ and $\varepsilon_{rL}$ the continuity of electric displacement is formulated as

$$\varepsilon_{rK} E_{niK} = \varepsilon_{rL} E_{niL} \tag{6}$$

In the traditional BEM the normal components of the electric field $E_{niK}$ and $E_{niL}$ on both sides of the single charge layer are calculated as a superposition of 2 components. The first, $E_{ni}^{-}$, includes the contribution of all charges except of a very small flat area around point $i$ while the second (the singular jump term) corresponds to the contribution of the local surface charge density $\sigma_i$ on both sides of this area:

$$E_{niK} = E_{ni}^{-} + \frac{\sigma_i}{2\varepsilon_0} \tag{7}$$

$$E_{niL} = E_{ni}^{-} - \frac{\sigma_i}{2\varepsilon_0} \tag{8}$$

After applying (7) and (8) to (6) the Fredholm integral equation of the second order can be obtained:

$$\sigma_i = 2\varepsilon_0 \frac{\varepsilon_{rK} - \varepsilon_{rL}}{\varepsilon_{rK} + \varepsilon_{rL}} E_{ni}^{-} = 2\varepsilon_0 \lambda_{KL} E_{ni}^{-} \tag{9}$$

$$E_{ni}^{-} = \frac{1}{4\pi\varepsilon_0} \sum_j \int_{S_j} \frac{\mathbf{n_{iK}} \cdot \mathbf{r_{ij}}}{r_{ij}^3} \sigma_j \, \mathrm{d}S = \sum_j \mathbf{n_{iK}} \cdot \nabla \mathbf{p}_{\mathbf{ij}}^{-} \sigma_j \tag{10}$$

According to (9) and (10) the matrix line for collocation point $i$ is determined by the array of coefficients $\mathbf{n_{iK}} \cdot \nabla \mathbf{p}_{\mathbf{ij}}^{-}$ where from the diagonal element ($j = i$) a correction factor representing material properties $1/(2\varepsilon_0 \lambda_{KL})$ is subtracted. The approach based on separation of the singular component enables a better numerical treatment of the strongly singular kernel in (10). In this way the numerical integration, in particular handling of singularities, is already incorporated in the traditional BEM formulation (in contrast to the region-oriented).

In the region-oriented approach we reuse the integration technique of the traditional BEM. We apply the same field coefficients $\nabla \mathbf{p}_{\mathbf{ij}}^{-}$ defined in (10). In order to obtain the $\nabla \mathbf{p_{ij}}$ used in (3) and (4) we have to add a correction factor to the coefficients representing the singular integration point according to (7):

$$\mathbf{n_{iK}} \cdot \nabla \mathbf{p_{ii}} = \mathbf{n_{iK}} \cdot \nabla \mathbf{p}_{\mathbf{ii}}^{-} + \frac{1}{2\varepsilon_0} \tag{11}$$

An observer inside of a region calculates field in his region based on a single charge layer that he can see from his location shown in Fig. 1bcd. In contrast to the traditional BEM he disregards everything on the other side of the charge layer. Consequently, the equation (8) is not used in the region oriented approach.

For the numerical tests presented in this paper the existing industrial code [1] has been applied for the calculation of $p_{ij}$ and $\nabla \mathbf{p}_{\mathbf{ij}}^{-}$. The same numbers representing $p_{ij}$ and $\nabla \mathbf{p}_{\mathbf{ij}}^{-}$ have been obtained for the traditional and the region-oriented BEM formulations. The only difference is related to the way how these numbers are arranged in the final equation system. A detailed explanation of the integration technique is beyond the scope of this paper. However, it is important to mention that the integration

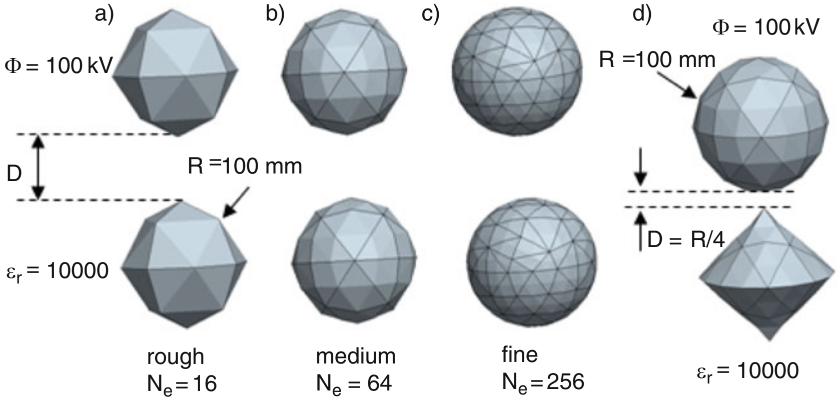


**Fig. 2:** Numerical test arrangement consisting of 2 spheres for: **a** rough **b** medium and **c** fine discretization level. **d** Sphere-cone arrangement (only medium discretization level is shown)

is based on the parabolic shape function for triangles as well as the linear approximation of surface charge density between the corner nodes. For the singular and near singular integration the Gaussian integration is used while the far field computation is based on multi-pole expansion.

## 4 Numerical Tests

For the numerical test an arrangement shown in Fig. 2 has been selected. It consists of one spherical electrode at fixed potential of 100 kV with the radius of R = 100 mm and another dielectric body with a very high permittivity value of $\varepsilon_r$ =10000 placed in a certain distance from the electrode. We consider 3 following geometrical variants of the dielectric body:

- a sphere with the same radius as the electrode located at a distance D = R/4.
- a similar sphere as above but located at a distance D = R, see Fig. 2abc.
- a symmetrical conical body with the same radius of base and height as the radius of the electrode located at a distance D = R/4, see Fig. 2d.

For each of the geometrical variants 3 different discretization cases have been defined: rough, medium and fine with 16, 64 and 256 elements respectively, see Fig. 2abc. The goal of computations is to obtain the potential of the dielectric body, which due to high permittivity value behaves like floating electrode. The results of computations based on traditional and region-oriented BEM formulations as well as the 2D results (used as reference) are shown in Tables 1 and 2. The 2D results have been obtained by a region-oriented charge simulation solver [5] for axially symmetric case (as well as the results shown in section 5).

The comparison shows that the traditional BEM formulation has significant problems to achieve correct results. Even with the increased discretization density the solution diverges from the the accurate one [1]. Furthermore, it is difficult to obtain a constant value of potential on the high permittivity body, in particular for the conical shape that includes discontinuity points (sharp corners and edges).

For the region-oriented formulation the result converges much better to the correct value. Only in case of rough discretization and small distance between bodies the potential error approaches the value of 15 %. The accuracy of all other region-oriented BEM results is satisfactory.

*Discussion of reference potential:* According to the kernel of (5) the potential at a large distance from charge approaches zero: Implicitly we get always zero potential at infinity. On the other hand the high permittivity provides a kind of *short-circuit to infinity* [2]. Consequently, the reference potential calculated for a detached dielectric body with a high permittivity value is equal to the real potential of such a body. It results from capacitive coupling to electrodes. In general case (for an arbitrary value of permittivity) the value of the reference potential depends on discretization. The reference potential provides an additional mechanism to compensate numerical discretization errors based on explicitly formulated Gauss law (4).

**Table 1:** Potential calculated for the high permittivity sphere from Fig. 2abc (in kV)

| | $D = R = 100\,\text{mm}$ | | | $D = R/4 = 25\,\text{mm}$ | | |
|---|---|---|---|---|---|---|
| | Rough | Medium | Fine | Rough | Medium | Fine |
| Traditional BEM | 15.8[a] | 34.3[a] | 58.7[a] | 195.4[a] | 43.0[a] | 85.6[a] |
| Region-Oriented BEM | 33.7 | 33.7 | 33.8 | 55.2 | 48.4 | 48.3 |
| 2D result | | 33.9 | | | 48.7 | |

[a] Average value – with a difference between top and bottom of the sphere up to ±10 %

**Table 2:** Potential calculated for the high permittivity conical body from Fig. 2d (in kV)

| | Rough | Medium | Fine |
|---|---|---|---|
| Traditional BEM | 28.0 / 58.6[a] | 37.5 / 68.8[a] | 41.1 / 66.3[a] |
| Region-Oriented BEM | 52.3 | 46.3 | 45.5 |
| 2D result | | 45.7 | |

[a] Potential of the bottom corner / potential of the top corner

[1] The reason of the inconsistent behavior has not been investigated in scope of this work. A possible reason may be related to numerical integration errors for irregular mesh generated by the CAD system Pro/Engineer. The convergence behavior becomes consistent when the permittivity of the dielectric sphere is in the range between 1 to 10.

[2] If we skipped the formulation of equation (4) and did not use the reference potential in (2) for the dielectric sphere in Fig. 2 the potential calculated for this sphere would be close to zero.

## 5 Example

Figure 3a shows an example of a surge arrester. It consists of 3 sections of zinc oxide cylinders supported by porcelain tubes. The goal of designers is calculation of the potential distribution along the arrester axis for the purely capacitive case. In order to keep the properties of zinc oxide in linear range a uniform potential gradient is required for normal operation.

In order to focus on essential numerical aspects the arrangement has been simplified to one long zinc oxide cylinder without floating armatures and porcelain supports, see Fig. 3b. The main numerical problem is related to the fact that the tangential field strength along the cylindrical surface of the zinc oxide is much larger than the normal component. On the other hand the equations for traditional BEM formulation are based on the normal component only.

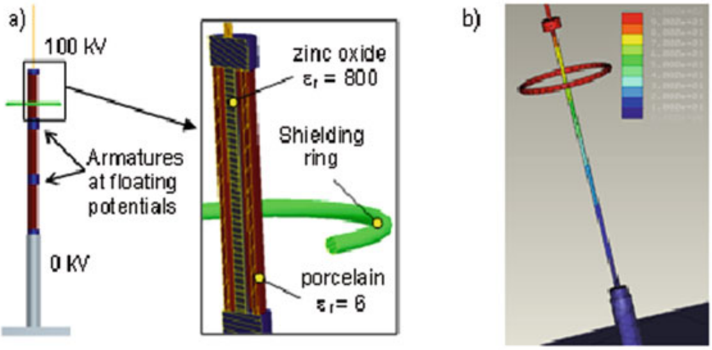

a)
100 kV
Armatures at floating potentials
0 kV
zinc oxide $\varepsilon_r = 800$
Shielding ring
porcelain $\varepsilon_r = 6$
b)

**Fig. 3: a** Example of surge arrester used by IEC 60099-4 as benchmark model. **b** Simplified model used for comparison between the traditional and the region-oriented BEM

The results in Fig. 4 show that the region-oriented BEM based on the potential continuity equations (2) as well as the explicit formulation of Gauss law for the high permittivity region (4) converges much better to the correct solution than the traditional BEM formulation.

## 6 Conclusion

The region-oriented BEM formulation is suitable for the calculation of electric fields in arrangements with extreme differences in material properties. The numerical tests and example of a simple power device show a better accuracy of this approach than the traditional BEM formulation.

The results presented in this paper should be regarded as a feasibility study. Effort is still needed to make the region-oriented BEM efficient for industrial applications. The following steps are proposed:

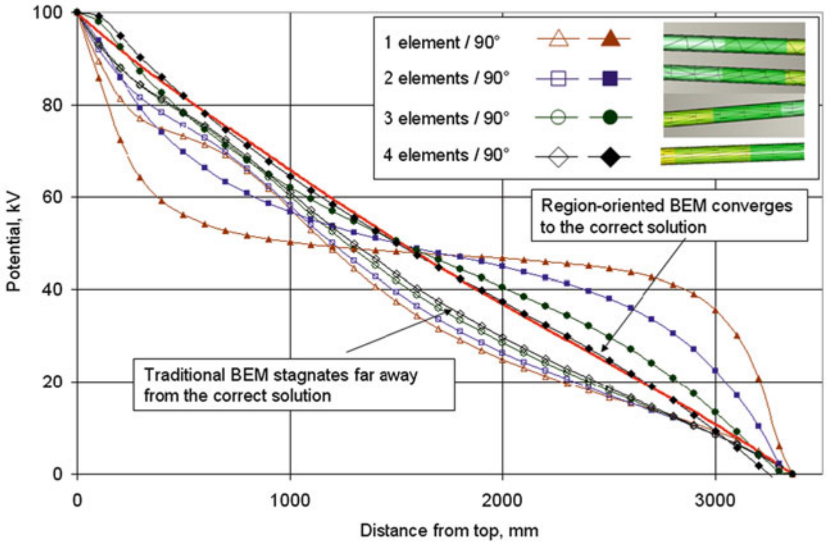


**Fig. 4:** Potential distribution along the axis of the surge arrester for different discretization density. The *hollow markers* denote traditional BEM solutions while the *filled markers* correspond to the region-oriented ones. The correct (2D) solution is represented by the *thick line without markers*

- Adjust integration technique. The currently used integration algorithms are tuned for traditional BEM. The region-oriented formulation is more flexible with regard to singularities: e.g. the charge need not to be located on the surface of elements.
- Improve stability of triple junctions.
- Preconditioning of GMRES solver: in contrast to the traditional BEM the region-oriented formulation shows poor GMRES convergence behavior.
- Parallelization based on distributed memory (MPI).

## References

1. De Kock, N., Mendik, M., Andjelic, Z., Blaszczyk, A..: Application of 3D boundary element method in the design of EHV GIS components. IEEE Magazine on Electrical Insulation **14**(3), 17–22 (1998)
2. Blaszczyk, A., Ketterer, H., Pedersen, A.: Computational electromagnetism in transformer and switchgear design: Current trends. SCEE 2000. *Lecture Notes on Computational Science and Engineering*, vol. 18, Springer Verlag, ISBN 3-540-42173-3, pp.55–62 (2001)
3. Tozoni, O.B., Mayergoyz, I.D.: Calculation of three dimensional electromagnetic fields. Published by Technika, Kiev (1974) (in Russian)
4. Andjelic, Z., Krstajic, B., Milojkovic, S., Blaszczyk, A., Steinbigler, H., Wohlmuth, M.: Integral methods for the calculation of electric fields. Scientific Series of the International Bureau Research Center Juelich, ISBN 3-89336-084-0 (1992)
5. Blaszczyk, A., Steinbigler, H.: Region-oriented charge simulation. IEEE Trans. on Magnetics **30**(5), 2924–2927 (1994)

# Surface Integrated Field Equations Method to Solve 3D Electromagnetic Problems

Zhifeng Sheng, Patrick Dewilde, and Rob Remis

**Abstract** This paper describes how the Surface Integrated Field Equations method (SIFE) is implemented to solve 3D Electromagnetic (EM) problems on substrates in which high contrast materials occur. It gives an account of the promising results that are obtained with it when compared to traditional approaches. Advantages of the method are the highly improved flexibility and accuracy for a given discretization level, at the cost of higher computational complexity.

## 1 Introduction

In our previous work, we have used the Surface Integrated Equations(SIFE) method for solving 2D electromagnetic problems[1], in which domains are present that exhibit highly contrasting material properties (electric and/or magnetic) with each other. In this paper, we develop the method further to solve 3D electromagnetic problems. Limitation of space prohibits us from giving a detailed description of the SIFE method and its spatial and temporal discretization schemes. For more details, we refer the readers to our previous papers [1, 2], and a full paper documenting the underlying theory will be published soon.

In strongly heterogeneous media such as modern chips, the constitutive parameters can jump by large amounts upon crossing the material interfaces. On a global scale, the EM field components are not differentiable and Maxwell's equations in differential form cannot be used, one has to resort to the original integral form of the EM field relations as the basis for the computational method.

Zhifeng Sheng, Patrick Dewilde
Circuits and Systems, EEMCS, TUDelft, Delft, The Netherlands, e-mail: z.sheng@ewi.tudelft.nl, p.dewilde@ewi.tudelft.nl

Rob Remis
EM-lab, EWI, TUDelft, Delft, The Netherlands, e-mail: r.f.remis@ewi.tudelft.nl

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_11,

Let $\mathscr{D}$ be the domain of interest with boundary $\partial\mathscr{D}$, $S$ be any (sufficiently smooth and small) surface ($S \in \mathscr{D}$) with boundary $\partial S$. For any $S$, Maxwell's equations in integrated form are:

$$-\oint_{\partial S} \mathbf{H} \cdot \mathrm{d}\mathbf{l} + \partial_t \int_S \mathbf{D} \cdot \mathrm{d}\mathbf{A} = -\int_S \mathbf{J}^{\mathrm{tot}} \cdot \mathrm{d}\mathbf{A}, \tag{1}$$

$$\oint_{\partial S} \mathbf{E} \cdot \mathrm{d}\mathbf{l} + \partial_t \int_S \mathbf{B} \cdot \mathrm{d}\mathbf{A} = 0, \tag{2}$$

where **E** is the electric field strength, **H** the magnetic field strength, **D** the electric flux density, and **B** the magnetic flux density. Moreover, $\mathbf{J}^{\mathrm{tot}} = \mathbf{J} + \mathbf{J}^{\mathrm{ext}}$, where **J** is the induced (field dependent) electric-current density, and $\mathbf{J}^{\mathrm{ext}}$ is the external electric-current densities. In addition to Maxwell's equations, the compatibility equations have to be satisfied as well. In integrated form, these equations are

$$\oint_{S'} (\partial_t \mathbf{D} + \mathbf{J}^{\mathrm{tot}}) \cdot \mathrm{d}\mathbf{A} = 0 \text{ and } \partial_t \oint_{S'} \mathbf{B} \cdot \mathrm{d}\mathbf{A} = 0,$$

where this time $S'$ is a smooth and closed surface. We also have to describe the type of matter that we are dealing with. The constitutive relations provide us with such a description and for the materials that we consider these relations are:

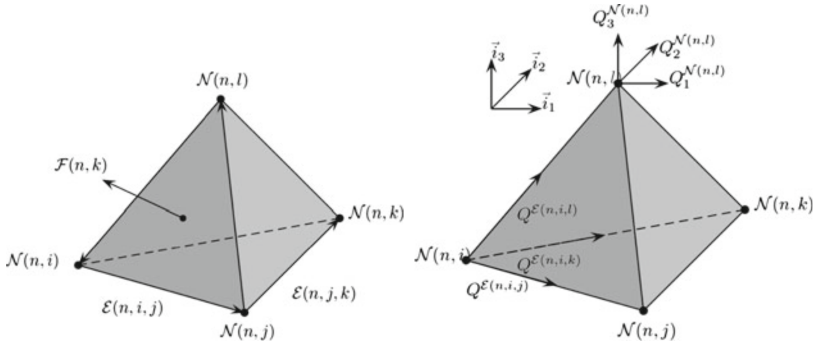
$$\mathbf{J} = \sigma \mathbf{E}, \quad \mathbf{D} = \varepsilon \mathbf{E}, \quad \text{and} \quad \mathbf{B} = \mu \mathbf{H}, \tag{3}$$

where $\sigma$ is the conductivity, $\varepsilon$ the permittivity, and $\mu$ the permeability. These three material parameters are all position dependent and are piecewise continuous in general. We are mostly concerned with media for which the medium parameters are piecewise constant, however, and at source-free interfaces where the parameters exhibit a jump, the tangential components of the electric and magnetic field strength have to be continuous, while the normal components of the electric and magnetic field strength are discontinuous because of the contrast.

## 2 Discretization Scheme

To satisfy the partial continuity conditions on material interfaces, we construct a so called "consistent linear discretization scheme" [3–5] that meets the continuity requirements across interfaces exactly, using a tetrahedron mesh combined with a consistent linear interpolation of electric and magnetic field strengths.

In this section we briefly present our discretization scheme. In all the experiments that we shall present, we use a nonuniform tetrahedron mesh generated by netgen [6] or msh [7]. For good results it is necessary that the tetrahedrons are "well formed", i.e. that they are not too skewed or too flat in one or more directions so that a vector decomposition along the edges yields a numerically accurate representation. We assume that the material parameters in each tetrahedron are constant (actually taking average values). This is consistent with the fact that a piecewise



**(a)** Tetrahedron **(b)** Hybrid element

**Fig. 1: a** Tetrahedron $\mathscr{T}_n$ and some geometrical quantities defined on it. $(i,j,k,l)$ is an even permutation of $(0,1,2,3)$. **b** The coefficients of the linear, hybrid expansion functions on $\mathscr{T}_n$

constant approximation for material parameters is sufficient to ensure continuity of the solutions. A more refined approximation such as continuous linear splines is certainly possible.

### *2.1 Geometrical Quantities*

Before introducing the linear expansion functions, we define a few geometrical quantities as shown in Fig. 1a.

- The coordinate vector is $\mathbf{x} = x_1\mathbf{i}_1 + x_2\mathbf{i}_2 + x_3\mathbf{i}_3$.
- A node with global node index $n$ is denoted as $\mathscr{N}_n$, and $\mathbf{x}_n$ is its coordinate vector.
- A tetrahedron with global tetrahedron index $n$ is denoted as $\mathscr{T}_n$.
- The four nodes delimiting $\mathscr{T}_n$ are locally denoted as $\mathscr{N}(n,i), i = \{0,1,2,3\}$.
- For every node with local label $\mathscr{N}(n,i)$, a unique global node index $m$ can be found, and: $\mathscr{N}(n,i) = \mathscr{N}_m$, $\mathbf{x}(n,i)$ denotes its coordinate vector.
- $\mathbf{x}_n^b = \frac{1}{4}\sum_{i=\{0,1,2,3\}}\mathbf{x}(n,i)$ is the coordinate vector of the barycentre of $\mathscr{T}_n$.
- Let $\varepsilon$ be an arbitrary small, positive real number, $\bar{\mathbf{x}}(n,i)$=$\mathbf{x}(n,i) + \varepsilon\left[\mathbf{x}_n^b - \mathbf{x}(n,i)\right]$.
- $\mathscr{E}(n,i,j); j \neq i$ denotes the edge pointing from $\mathscr{N}(n,i)$ to $\mathscr{N}(n,j)$.
- Let $\mathbf{e}(n,i,j)$ be the vectorial length of $\mathscr{E}(n,i,j)$: $\mathbf{e}(n,i,j) = \mathbf{x}(n,j) - \mathbf{x}(n,i)$.
- Let $\mathscr{F}(n,k)$ be the facet of $\mathscr{T}_n$, which is not delimited by $\mathscr{N}(n,k)$.
- Let $\mathbf{A}(n,k)$ be the vectorial area of $\mathscr{F}(n,k)$,e.g $\mathbf{A}(n,0) = \frac{1}{2}\left[\mathbf{e}(n,1,2) \times \mathbf{e}(n,2,3)\right]$.
- Let $V(n)$ be the volume of $\mathscr{T}_n$:$V(n) = \frac{1}{3}\left[\mathbf{x}(n,1) - \mathbf{x}(n,0)\right] \cdot \mathbf{A}(n,0)$.
- Let $\phi_i(\mathscr{T}_n, \mathbf{x})$ be the local linear scalar interpolation function,

$$\phi_i(\mathscr{T}_n, \mathbf{x}) = 1/4 - (\mathbf{x} - \mathbf{x}_n^b) \cdot \frac{\mathbf{A}(n,i)}{3V(n)}, \forall \mathbf{x} \in \mathscr{T}_n,$$

## 2.2 Spatial Discretization Scheme

Let $\mathbf{Q}(\mathbf{x})$ be a vectorial function of space representing electric field strength or magnetic field strength at a time instance $t$; that is: $\mathbf{Q}(\mathbf{x})$ represents $\mathbf{E}(\mathbf{x},t)$ or $\mathbf{H}(\mathbf{x},t)$. Its tangential component is continuous across the interface while its normal component is discontinuous. We can represent the value of $\mathbf{Q}(\mathbf{x})$ on the nodes delimiting a tetrahedron $\mathscr{T}_n$ with the well defined components. The value of $\mathbf{Q}(\mathbf{x})$ inside $\mathscr{T}_n$ can be interpolated with linear, hybrid expansion functions which are built upon *continuity nodes* for nodes inside domains with uniform material parameters and *discontinuity nodes* on the interfaces.

### 2.2.1 Continuity Node

Let $\mathscr{N}(n,i)$ be a node where all components of $\mathbf{Q}(\mathbf{x})$ on this node are continuous, $\mathbb{N}_{\mathbf{Q}}^{C}$ be the set of nodes in the mesh where $\mathbf{Q}(\mathbf{x})$ is totally continuous. The value of $\mathbf{Q}(\mathbf{x})$ on node $\mathscr{N}(n,i)$ can be represented as:

$$\mathbf{Q}^{\mathscr{N}(n,i)} = \sum_{j=1,2,3} Q_j^{\mathscr{N}(n,i)}\mathbf{i}_j, \forall \mathscr{N}(n,i) \in \mathbb{N}_{\mathbf{Q}}^{C} \tag{4}$$

where $\mathbf{Q}^{\mathscr{N}(n,i)}$ denotes the value of $\mathbf{Q}(\mathbf{x})$ on the node $\mathscr{N}(n,i)$.

### 2.2.2 Discontinuity Node

Let $\mathscr{N}(n,i)$ be a node on the interface of material discontinuity, $\mathbb{N}_{\mathbf{Q}}^{D}$ be the set of nodes in the mesh where $\mathbf{Q}(\mathbf{x})$ is discontinuous in its normal component. The normal component of $\mathbf{Q}(\mathbf{x})$ is not well defined for the nodes on the interfaces. However, if we offset the node into the tetrahedron $\mathscr{T}_n$, then the value of $\mathbf{Q}(\mathbf{x})$ on that node can be represented as:

$$\mathbf{Q}^{\mathscr{N}(n,i)} = \sum_{j=\{0,1,2,3\},j\neq i} -Q^{\mathscr{E}(n,i,j)}\frac{|\mathbf{e}(n,i,j)|}{3V(n)}\mathbf{A}(n,j), \forall \mathscr{N}(n,i) \in \mathbb{N}_{\mathbf{Q}}^{D}. \tag{5}$$

$Q^{\mathscr{E}(n,i,j)}$ is the projection of $\mathbf{Q}(\mathbf{x})$ on the node $\mathscr{N}(n,i)$ to the direction $\mathbf{e}(n,i,j)$.

### 2.2.3 Linear, Hybrid Expansion

Let $[\mathbf{Q}](\mathscr{T}_n,\mathbf{x})$ be the local linear approximation of $\mathbf{Q}(\mathbf{x})$ in the tetrahedron $\mathscr{T}_n$:

$$[\mathbf{Q}](\mathscr{T}_n,\mathbf{x}) = \sum_{i\in\{0,1,2,3\}} \phi_i(\mathscr{T}_n,\mathbf{x})\mathbf{Q}^{\mathscr{N}(n,i)}, \forall \mathbf{x} \in \mathscr{T}_n \tag{6}$$

in which $\mathbf{Q}^{\mathcal{N}(n,i)}$ is defined by Eq.4 or Eq.5, $Q_j^{\mathcal{N}(n,i)}$ and $Q^{\mathcal{E}(n,i,j)}$ are the linear, hybrid expansion coefficients as shown in Fig. 1b. A list of properties of the linear, hybrid interpolation functions follows:

- The linear, hybrid expansion functions are consistently linear functions [1], they permit a completely linear expansion of the partially continuous vectorial function $\mathbf{Q}(\mathbf{x})$ inside each tetrahedron. the approximation errors of the linear, hybrid expansion functions are of order $O(h^2)$.
- Assuming the *discontinuity nodes* are assigned in the right place, the linear, hybrid expansions functions ensure continuity in tangential components across material interfaces, and ensure total continuity in homogeneous sub-domains .
- With the linear, hybrid expansion functions, it is easy to apply the boundary conditions that prescribe tangential components.

We refer to the elements with linear, hybrid expansion functions as "hybrid elements", and refer to the elements with only *continuity nodes* as "nodal elements".

#### 2.2.4 Field Strength Discretization

With all these properties above, the linear, hybrid expansion functions are a very good choice for interpolating electric field strength and magnetic field strength. Note that an interface can be with electric and/or magnetic contrast. The set of *discontinuity nodes* for magnetic field strength $\mathbb{N}_{\mathbf{H}}^{D}$ does not have to be the same as that for the electric field strength $\mathbb{N}_{\mathbf{E}}^{D}$. With the graphic user interface we implemented, it is very easy to assign the *discontinuity nodes*.

### *2.3 Temporal Discretization Scheme*

To implement a time stepping scheme for the spatially discretized Maxwell's equations, let $t_0$ be the initial time and $\Delta t > 0$ be the time step size, we introduce the time instance $t_n = n\Delta t + t_0$ and integrate Maxwell's equations from $t = t_{n-1}$ to $t = t_n$. All integrals that can not be computed analytically are approximated using the trapezoidal rule, which is known to be unconditionally stable in time. The approximation of the trapezoidal rule is of order $O(\Delta t^2)$, which is verified in Section 4.

## 3 The Surface Integrated Field Equations Method

In this section, we briefly present the Surface Integrated Field Equations method for solving electromagnetic problems in the time domain. We apply the Ampere's equation (Eq.1) and constitutive relations (Eq. 3) on every facet of every element product a time interval, i.e. $\mathcal{F}(n,i) \times [t_{m-1}, t_m]\,; \Delta t = t_m - t_{m-1}$, and use the linear spatial and temporal approximation presented in Sec. 2.2 and Sec. 2.3 to discretize the electromagnetic field strengths, we obtain discretized Ampère equations:

$$\frac{\Delta t}{4}\left[\mathbf{e}(n,l,k)\cdot\mathbf{H}^{\mathcal{N}(n,j)}(t_m)+\mathbf{e}(n,j,l)\cdot\mathbf{H}^{\mathcal{N}(n,k)}(t_m)+\mathbf{e}(n,k,j)\cdot\mathbf{H}^{\mathcal{N}(n,l)}(t_m)\right]$$
$$-\sum_{h=j,k,l}\left[\frac{\Delta t}{6}\sigma(\overline{\mathbf{x}}(n,h))+\frac{1}{3}\varepsilon(\overline{\mathbf{x}}(n,h))\right]\mathbf{A}(n,i)\cdot\mathbf{E}^{\mathcal{N}(n,h)}(t_m)=-$$
$$\frac{\Delta t}{4}\left[\mathbf{e}(n,l,k)\cdot\mathbf{H}^{\mathcal{N}(n,j)}(t_{m-1})+\mathbf{e}(n,j,l)\cdot\mathbf{H}^{\mathcal{N}(n,k)}(t_{m-1})+\mathbf{e}(n,k,j)\cdot\mathbf{H}^{\mathcal{N}(n,l)}(t_{m-1})\right]$$
$$+\sum_{h=j,k,l}\left[\frac{\Delta t}{6}\sigma(\overline{\mathbf{x}}(n,h))-\frac{1}{3}\varepsilon(\overline{\mathbf{x}}(n,h))\right]\mathbf{A}(n,i)\cdot\mathbf{E}^{\mathcal{N}(n,h)}(t_{m-1})$$
$$+\sum_{h=j,k,l}\frac{1}{3}\int_{t=t_{m-1}}^{t_m}\left[\mathbf{A}(n,i)\cdot\mathbf{J}^{\text{ext}}(\overline{\mathbf{x}}(n,h),t)\right]\mathrm{d}t$$

for any $\mathcal{T}_n$, and $\{i,j,k,l\}$ an even permutation of $\{0,1,2,3\}$. With similar procedure applied on the surface-time integrated Faraday's equation, we obtain the discretized Faraday's equation. Note that discontinuous field quantities are not well defined for the nodes on the interfaces. Therefore, we take values pertaining to a vanishing offset towards the barycentre of $\mathcal{T}_n$, where all field quantities are well defined.

With appropriate boundary conditions, we shall have an over-determined system of linear discrete surface-time integrated field equations. Such a system may have no solution at all. The best thing we can do is to find an approximate solution which minimizes a quadratic functional. With the weighted least-squares method[8], we can easily construct normal equations, which we then solve iteratively. After solving the system for the coefficients, we get the approximated electromagnetic field strength in the domain of computation (we use a CG iterative solver with preconditioner).

## 4 Numerical Experiment

We verify the accuracy of the temporal discretization scheme and spatial discretization scheme using a 3D Electromagnetic time domain problem with high contrast in permeability for which an analytic solution is known. The theoretical solution is a 'steady state' solution at a single frequency, containing a source term that continuously injects current. We use the steady solution at $t = 0$ as initial state, and start integrating from there in the time domain. The computational domain $\mathcal{D} = \{0 \le x_1 \le 1, 0 \le x_2 \le 1, 0 \le x_3 \le 1\}$ is bounded by PEC boundaries. Let

$$h(\mathbf{x},t)=\frac{\sin(\omega t)}{\mu(\mathbf{x})\omega} \text{ and } g(\mathbf{x},t)=\sigma(\mathbf{x})\cos(\omega t)-\varepsilon(\mathbf{x})\omega\sin(\omega t)$$

The source density distributions be:

**Table 1:** Configuration of the four sub-domains

| $\mathscr{D}_i$ | Definition of sub-domains | $\varepsilon_r$ | $\sigma$ | $\mu_r$ |
|---|---|---|---|---|
| $\mathscr{D}_0$ | $0 \leq x_1 < 0.5, 0 \leq x_2 < 0.5, 0 \leq x_3 \leq 1$ | 1 | 0 | 1000 |
| $\mathscr{D}_1$ | $0.5 \leq x_1 < 1, 0 \leq x_2 < 0.5, 0 \leq x_3 \leq 1$ | 1 | 0 | 1 |
| $\mathscr{D}_2$ | $0 \leq x_1 < 0.5, 0.5 \leq x_2 < 1, 0 \leq x_3 \leq 1$ | 1 | 0 | 1 |
| $\mathscr{D}_3$ | $0.5 \leq x_1 < 1, 0.5 \leq x_2 < 1, 0 \leq x_3 \leq 1$ | 1 | 0 | 10 |

$$\mathbf{J}^{\text{ext}}(\mathbf{x},t) = [-2\pi^2 h(\mathbf{x},t) - g(\mathbf{x},t)]\sin(\pi x_1)\sin(\pi x_2)\mathbf{i}_3,.$$

The exact field strengths are:

$$\begin{aligned}\mathbf{E}(\mathbf{x},t) &= \sin(\pi x_1)\sin(\pi x_2)\cos(\omega t)\mathbf{i}_3,\\ \mathbf{H}(\mathbf{x},t) &= -\pi h(\mathbf{x},t)\sin(\pi x_1)\cos(\pi x_2)\mathbf{i}_1 + \pi h(\mathbf{x},t)\cos(\pi x_1)\sin(\pi x_2)\mathbf{i}_2.\end{aligned}$$

The angular frequency $\omega$ is chosen to be $2\pi 10^9 \mathrm{rad}/s$ corresponding to a source frequency of 1GHZ. The configuration will be computed for 10 wave periods ($0 \leq t \leq 10^{-8} s$). The whole domain is divided into four homogeneous sub-domains defined in Tab. 1. We compute this configuration with the SIFE method based on hybrid elements and the weighted ($w = 2 \times 10^{-3}$) Galerkin method based on nodal elements (see [9]). The computational domain is discretized with an interface conforming tetrahedron mesh (5853 nodes and 30208 tetrahedrons). Series of experiments have been done with different time step sizes. Note that, the contrast only exists for the magnetic field strength. Therefore, *discontinuity nodes* are used only when interpolating magnetic field strength on the material interfaces; since the electric field strength is totally continuous. *Discontinuity nodes* are not used for interpolating the electric field strength. As shown in Fig.2, the SIFE method based on hybrid elements has second order accuracy in time even in the presence of high contrast. The accuracy plots show that the SIFE method produces the correct result while the standard packages do not.

## 5 Conclusions

The SIFE method holds considerable promise to solve three dimensional time domain electromagnetic problems, in which high contrasts between different types of materials exist and irregular structures are present. Its accuracy and stability has now been demonstrated and verified with numerical experiments. With these basic properties being established, we are now working on improvements of the computational properties of the method in terms of numerical complexity and versatility.

**Acknowledgements** The authors gratefully acknowledge the support from the Technology Foundation STW under the MICES project. The authors hereby acknowledge the fact that Prof. A.T. de

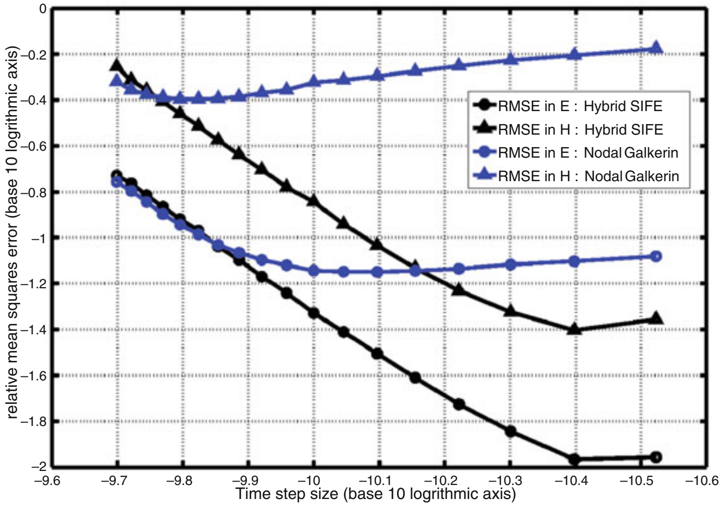


**Fig. 2:** BICG-stable + nest-dissection reordering + ICC(4) is used for the SIFE method, BICG-stable + nest-dissection reordering + ILU(4) is used for the weighted Galerkin's method. The accuracy of these iterative solvers is set to be $10^{-20}$

Hoop provided the central idea (the use of edge elements) of the paper and gratefully thank him for his contributions to the MICES project.

## References

1. Sheng, Z., Remis, R.F., Dewilde, P.: A least-squares implementation of the field integrated method to solve time domain electromagnetic problems. Computational Electromagnetics in Time-Domain, 2007. CEM-TD 2007. Workshop on pp. 1–4 (15-17 Oct. 2007). DOI 10.1109/CEMTD.2007.4373514
2. Sheng, Z., Dewilde, P.M., Remis, R.F.: The field integrated method to solve time domain electromagnetic problems. In: IEEE/ProRISC workshop on Circuits, Systems and Signal Processing, Veldhoven (NL), IEEE, (November 2007)
3. Lager, I.E.: Finite element modelling of static and stationary electric and magnetic fields. PhD thesis, Delft University of Technology (1996)
4. Mur, G.: The finite-element modeling of three-dimensional electromagnetic fields using edge and nodal elements. IEEE tranactions on antennas and propagation **41**(7) (1993)
5. Jorna, P.: Integrated field equations methods for the computation of electromagnetic fields in strongly inhomogeneous media. PHD thesis, TUDelft (2005)
6. Schoberl, J.: Netgen - an advancing front 2d/3d-mesh generator based on abstract rules. Comput.Visual.Sci **1**, 41–52 (1997)
7. van der Kolk, K., van der Meijs, N.: On the implementation of a 3-dimensional delaunay-based mesh generator. In: G. Ciuprina; D. Ioan (Ed.), SCEE 2006 Book of Abstracts, Sinaia, RO, pp. 171-172, 2006. ISBN: 978-973-718-520-4. (2006)
8. Jiang, B.N.: The Least-Squares Finite Element Method: Theory and Applications in Computational Fluid Dynamics and Electromagnetics (Scientific Computation). Springer (2006)
9. Sitapati, K.: Mixed-field finite element computations. PhD thesis, Virginia Polytechnic Institute and State University (2004)

# Reduced Basis Method for Electromagnetic Field Computations

Jan Pomplun and Frank Schmidt

**Abstract** We explain the reduced basis (RB) method applied to electromagnetic field computations with the finite element method. Rigorous numerical simulations for practical applications often become very time consuming. The RB method allows to split up the solution process of an e.g. geometrically parameterized problem into an expensive offline and a fast online part. For an actual simulation only the fast online part is evoked.

We apply the RB method to the rigorous simulation of light scattering from a parameterized phase shift mask.

## 1 Introduction

The finite element method (FEM) has been successfully applied to a large number of nano-optical problem classes [1, 2]. The computation time of a single FEM simulation however can become very long especially for 3D problems. Design, optimization and inverse problems usually involve a large number of such simulations having an underlying layout with a few varying geometrical parameters. The RB method can be applied to this setup, see [3] and subsequent citations. In the offline step the underlying model is solved rigorously several times for different values of the geometrical parameters. These solutions build the reduced basis. The full parameterized problem is projected onto the RB which results in a significant reduction of the problem size. In the online step the reduced problem is solved.

Jan Pomplun, Frank Schmidt
Zuse Institute Berlin, Takustrasse 7, 14195 Berlin, Germany, e-mail: pomplun@zib.de, frank.schmidt@zib.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_12,

## 2 Reduced Basis Method

The governing equations describing the propagation of light are Maxwell's equations. In many practical applications like computational lithography the time-harmonic form is the appropriate description:

$$\nabla \times {\mu_p}^{-1} \nabla \times \mathbf{E} - \omega^2 \varepsilon_p \mathbf{E} = 0, \tag{1}$$

where $\mu_p$ and $\varepsilon_p$ are the permeability and permittivity. For the discretization with finite elements we need the corresponding weak formulation. Therefore we multiply (1) with a test function $v$ integrate over the domain of interest $\Omega$ and do a partial integration of the curl curl integral. The weak form of Maxwell's equations then reads:
Find $\mathbf{E} \in X = H(curl, \Omega)$ such that $\forall v \in X$:

$$a(v, \mathbf{E}; p) = \int\limits_{\Omega} (\nabla \times v)\, {\mu_p}^{-1}\, (\nabla \times \mathbf{E}) - \omega^2 \int\limits_{\Omega} v\, \varepsilon_p\, \mathbf{E} = Bound.Terms = f(v) \tag{2}$$

Since we consider geometrically parameterized problems the material distribution described by the permittivity $\varepsilon_p$ and permeability $\mu_p$ is parameter dependent which is denoted by the subscript $p$. Figure 1 shows the phase mask example and the intensity of the electric field computed with finite elements. The geometrical parameters $p = \{d_1, d_2, d_3\}$ are the width of the absorber openings and their distance.

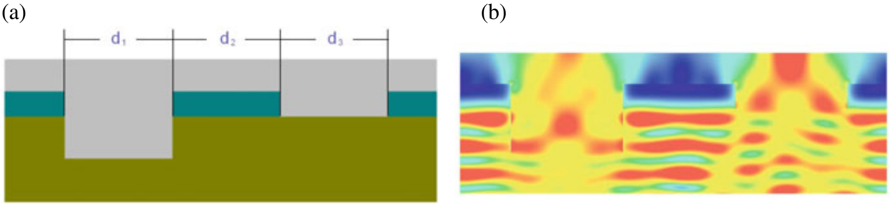


**Fig. 1: a** Parameterized phase shift mask for reduced basis computations. **b** Intensity of electric field obtained from FEM computation

Discretizing (2) with the finite element method leads to a linear system of equations [4] with parameter dependent coefficients and solution:

$$A_p u_p = \mathrm{f}. \tag{3}$$

In practice the number of unknowns (dimension of $u_p$) can be up to several millions. Now suppose that $u_p$ stays on a low dimensional sub manifold of the complete solution space if we vary the parameters $p$. Then it is reasonable to solve (3) only on this subspace, i.e. the full system can be projected onto a reduced basis $U$ of the low dimensional sub manifold. Usually the reduced basis is built by a number of so called snapshot solutions, which are rigorous solutions of (3) for different parameter values $p_i$. The reduced basis is computed in the offline step. The projected system reads:

$$\left[U^H A_p U\right] \lambda_p = U^H \mathrm{f} \tag{4}$$

which is a system of dimension $N$ ($\propto 100$) much smaller than the original problem. After solving (4) in the online step the RB solution can be obtained by: $\hat{u}_p = U\lambda_p$.

## 3 Affine Parameter Dependence

The expensive steps performing a reduced basis computation following (4) for an actual parameter set $\hat{p}$ is the assembling of $A_{\hat{p}}$ and the projection step onto the RB $U$. These steps can also be performed offline if the matrix $A_p$ has an affine parameter dependence defined by:

$$A_p = \sum_{i=1}^{M} \Theta_i(p) A^{(i)} \quad \Longrightarrow \quad \left[U^H A_p U\right] = \sum_{i=1}^{M} \Theta_i(p) \left[U^H A^{(i)} U\right]. \tag{5}$$

The matrices $\left[U^H A^{(i)} U\right]$ can be assembled and projected offline. The reduced basis computation then no longer depends on the number of unknowns of the finite element computation but only on the dimension of the reduced basis and the number of terms $M$ in the affine expansion of the Matrix $A_p$. For finite elements on a rectangular grid such an affine parameter dependence can be found easily [5]. But even for more general settings an affine dependence can be constructed.

For fixed reference values $p_{ref}$ of the parameters we have a reference configuration and a reference formulation (2) of our problem. We denote by $Q(x,y,z;p)$ the mapping of our reference onto a new configuration and by $J(x,y,z;p)$ its Jacobian. We have $Q(x,y,z;p_{ref}) = Id$. On the transformed domain (2) reads:

$$\begin{aligned} a(v,\mathbf{E};p) = & \int_{\Omega} (\nabla \times v) \left[\frac{1}{|J|} J^T \mu_{p_{ref}}{}^{-1} J\right] (\nabla \times \mathbf{E}) \\ & -\omega^2 \int_{\Omega} v \left[|J|\, J^{-1} \varepsilon_{p_{ref}} J^{-T}\right] \mathbf{E} = B.T. \end{aligned} \tag{6}$$

Now we assume that we can divide the domain $\Omega$ into disjunct open sets and that the Jacobian $J(x,y,z;p)$ is piecewise constant on these sets, i.e.:

$$\begin{aligned} \Omega &= \bigcup_{i=1}^{L} \overline{\Omega_i}\,, \quad \Omega_i \bigcap \Omega_j = \emptyset, \text{ for } i \neq j\,, \\ J(x,y,z;p) &= \sum_{i=1}^{L} J_i(p)\, \chi_{\Omega_i}, \end{aligned} \tag{7}$$

where $\chi_{\Omega_i}$ is the characteristic function of $\Omega_i$. Now we define the tensors:

$$\begin{aligned}
S(p) &= \frac{1}{|J|} J^T \mu^{-1} J = \sum_{i=1}^{L} \frac{1}{|J_i(p)|} J_i(p)^T \mu^{-1} J_i(p)\, \chi_{\Omega_i} \\
&= \sum_{i=1}^{L} S^i(p)\, \chi_{\Omega_i}, \\
M(p) &= |J|\, J^{-1} \varepsilon_{p_{ref}} J^{-T} = \sum_{i=1}^{L} |J_i(p)|\, J_i(p)^{-1} \varepsilon_{p_{ref}} J_i(p)^{-T}\, \chi_{\Omega_i} \\
&= \sum_{i=1}^{L} M^i(p)\, \chi_{\Omega_i}
\end{aligned} \tag{8}$$

and insert them into (6). E.g. for the stiffness integral we find:

$$\begin{aligned}
&\sum_{i=1}^{L} S^i_{11}(p) \int_{\Omega_i} (\nabla \times v) \begin{pmatrix} 1\,0\,0 \\ 0\,0\,0 \\ 0\,0\,0 \end{pmatrix} (\nabla \times \mathbf{E}) + S^i_{12}(p) \int_{\Omega_i} (\nabla \times v) \begin{pmatrix} 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,0 \end{pmatrix} (\nabla \times \mathbf{E}) + \\
&\cdots + S^i_{33}(p) \int_{\Omega_i} (\nabla \times v) \begin{pmatrix} 0\,0\,0 \\ 0\,0\,0 \\ 0\,0\,1 \end{pmatrix} (\nabla \times \mathbf{E}),
\end{aligned} \tag{9}$$

the mass term is decomposed accordingly. This leads to an affine decomposition of the bilinear form (6):

$$a(v, \mathbf{E}; p) = \sum_{i=1}^{M} \Theta_i(p) a_i(v, \mathbf{E}). \tag{10}$$

Discretizing this expression we get an affine matrix expansion according to (5). In total the matrix expansion has $12L$ terms for 3D Maxwell's equations (2 times 6 entries from the symmetric tensors $S^i(p)$, $M^i(p)$). For a 2D domain with arbitrary polarization of the incident light this reduces to $8L$ for TM or TE polarization in 2D to $4L$.

From the point of implementation the piecewise constant Jacobian is constructed by transforming a coarse triangulation of the reference domain. The mapping of a triangle onto another triangle is linear and its Jacobian constant. The mapping of a coarse triangulation onto another coarse triangulation has therefore a piecewise constant Jacobian. Figure 2 shows two grids of the phase mask which can be mapped onto each other by such transformation. They are topologically equivalent. In this example we have $L = 37$ which gives us $M = 297$ terms in the matrix expansion ($37 \cdot 8 + 1$ for the domain which is independent of all parameters). Note that $L$ is smaller than the number of triangles shown in Fig. 2 because several triangles can have the same parameter dependent Jacobian. For three dimensional computational domains a transformation with piecewise constant Jacobian can be constructed mapping tetrahedrons onto tetrahedrons.

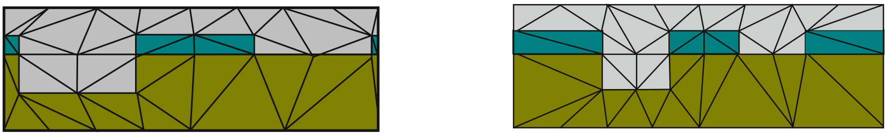


**Fig. 2:** Topologically equivalent grids which can be mapped by piecewise linear transformation onto each other

## 4 Construction of the Reduced Basis

An important question is how to construct the reduced basis $U$. A simple and effective strategy is to compute so called snapshots which build the reduced basis. These are solutions $u_{p_i}$ of the original problem for different values of the parameters $p_i$:

$$U = [u_{p_1}, u_{p_2}, \ldots, u_{p_N}]\,. \tag{11}$$

A basis made of such snapshots is called a Lagrange reduced basis. It is also possible to include derivatives of the snapshots with respect to geometrical parameters which gives so called Taylor reduced bases [3]. The question arises for which values of the parameters the snapshots should be computed to get a good approximation quality of the reduced basis. A simple strategy is a random choice of the parameters [5]. However this might become inefficient if the randomly chosen snapshots do not add new information to the basis. Especially for 3D computations which can take several hours we want to compute a minimum number of snapshots in the offline phase.

For the usage of optimized sampling strategies we define a so called training sample [3] by choosing a large number $n_{train} \propto 1000$ of candidate snapshot parameters from our parameter space: $P^{train} = \{p^1, \ldots, p^{n_{train}}\}$. Now we want to construct a reduced basis of dimension $N$ (typically much smaller than $n_{train}$) which gives our reduced basis a good approximation quality over the the whole parameter space. An optimal basis could be obtained computing all snapshots in the training sample and then performing a singular value decomposition (SVD) of the basis, i.e. the computation of a proper orthogonal decomposition (POD) basis. However this would involve $n_{train}$ solution runs of the forward problem which is often not feasible.

A suboptimal method is the choice of the snapshot parameters with a greedy algorithm [3]. The first snapshot parameter of the training sample $P^{train}$ is chosen randomly. Then at step $N$ the greedy algorithm appends to the current sample the parameter in $P^{train}$ which is least well approximated by the current reduced basis. Since every parameter in the large set $P^{train}$ is tested this has to be done fast using an a posteriori error estimator $\Delta(p)$. The parameter set $p_j$ which is estimated to have the largest error is then included into the current reduced basis. This assures that a maximum of new information is added to the current basis. An algorithm presented in [3] involves the computation of the Riesz representation of the residual of the reduced basis solution and the determination of the coercivity/inf-sup constant of the underlying elliptic problem. Here we directly use the energy norm of the Riesz representation of the residual as an error criterion which we shortly explain. If we have a reduced basis solution $\hat{u}$ for the solution $u$ of (2), the residual is defined as a linear functional on $X$:

$$r(v;p) = f(v) - a(v,\hat{u};p)\,,\ r(\cdot;p) \in X' \tag{12}$$

According to the Riesz representation theorem $r \in X'$ can be represented as an element of $\hat{e} \in X$:

$$(v,\hat{e}(p)))_X = r(v;p) = f(v) - a(v,\hat{u}(p);p) \tag{13}$$

which gives us a variational formulation for the computation of $\hat{e}$. The energy norm gives us the error criterion:

$$\Delta(p) = \sqrt{(\hat{e}(p),\hat{e}(p))_X} = ||r(\cdot;p)||_{X'}. \tag{14}$$

Eq. (13) shows what has to be computed at each parameter set of the training sample. For $\hat{u}$ a reduced basis computation with the current reduced basis has to be performed which is independent on the original size of the finite element problem. Also using the affine decomposition of the linear form $a(\cdot,\cdot;p)$ according to (10) the discretized variational problem (13) can be solved before determining $\Delta(p)$ [3].

Fig. 3(b) shows numerically that for the greedy reduced basis construction $\Delta(p)$ offers a good approximation to the energy norm of the error of the reduced basis solution, i.e. $e = u - \hat{u}$.

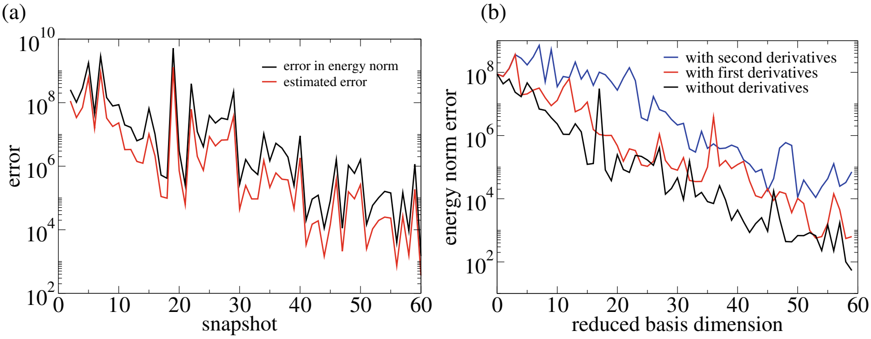


**Fig. 3: a** Comparison of error in energy norm and estimated error $\Delta(p)$ (14) for chosen snapshots in greedy construction of reduced basis. **b** Convergence of reduced basis solution for increasing dimension of reduced basis using no, first and second derivatives. The error was averaged over an ensemble of 125 solutions

## 5 Results

The reduced basis algorithms were implemented into the finite element package JCMsuite developed by the Zuse Institute Berlin and JCMwave for numerical solution of Maxwell's equations [1,2,6]. Here we focus on a 2D scattering problem for a phase shift mask with a incident plane wave. The reference values for the geometrical parameters were $d_1 = d_2 = d_3 = 380\,nm$. We varied the three parameters by $\pm 40nm$ which corresponds to a 20% variation. The training sample $P^{train}$ over this parameter domain consisted of a $9x9x9$ Cartesian grid. A reduced basis of dimension $N = 60$ was computed with the greedy algorithm.

Once the basis is computed the actual online computations can be performed. To check the convergence we computed reduced basis solutions over an ensemble (different then the training sample) and compared them with the exact finite element solution. As error measures we used the energy norm of the error of the complete field and the error of the lowest diffraction modes which are important output quantities in computational lithography. Figures 3 and 4(a) and (b) show the convergence of the reduced basis solution for an ensemble and a single parameter. We see that important quantities of interest like diffraction modes can be computed with a relative error smaller than 0.1% already with a reduced basis of dimension 60. The computational time for this example is of the order $\propto 10\,\mathrm{ms}$. If in application the offline computational time also becomes important and has to be minimized one can include derivatives of the snapshots in the reduced basis. E.g. if (3) is solved by LU-decomposition of $A_p$ the solution for the derivative with respect to a parameter $p_i$ is obtained by:

$$\partial_{p_i} u_p = -A_p^{-1} \left(\partial_{p_i} A_p\right) u_p. \tag{15}$$

I.e. we only have to perform the forward backward substitution with a new right hand side which is the derivative of the system matrix applied to the primary solution. Higher derivatives are obtained equivalently. Figure 3(b) shows the convergence of the reduced basis solution for an ensemble of 125 parameter values. The reduced basis dimension is 60. In the first case the reduced basis consists of 60 snapshots. In the second case 15 snapshots with their three first parameter derivatives are used and in the last case 6 snapshots with their three first and six second derivatives are used. This greatly reduces offline computational time but also leads to slightly slower convergence of the reduced basis solution.

## 6 Conclusions and Outlook

We applied the reduced basis method to an electromagnetic scattering problem namely transmission of light through a geometrically parameterized phase shift mask. We have described how to obtain an affine decomposition of the Maxwell system and studied convergence.

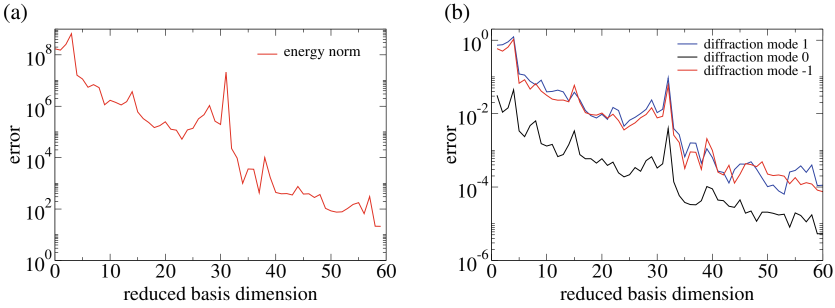


**Fig. 4: a** Convergence of reduced basis solution in energy norm. **b** Convergence of lowest diffraction modes of reduced basis solution

After a costly offline assembling phase of the reduced basis the projected Maxwell system can be solved very fast ($\propto 10\,\text{ms}$) in the online step for different geometrical parameters. The reduced basis solution convergences towards the exact solution for a small dimension ($\propto 100$) of the reduced system over a large parameter domain.

This offers the possibility to apply the reduced basis method to design, optimization and inverse problems in computational electromagnetics.

Future work will focus on the application of the reduced basis method to 3D examples relevant in practice which also include parameter dependent exterior domains. Furthermore rigorous error estimators for the output of interest have to be developed to ensure the reliability of the reduced basis solutions.

## References

1. Burger, S., Köhle, R., Zschiedrich, L., Gao, W., Schmidt, F., März, R., Nölscher, C.: Benchmark of FEM, Waveguide and FDTD Algorithms for Rigorous Mask Simulation. In: J.T. Weed, P.M. Martin (eds.) Photomask Technology, vol. 5992, pp. 378–389. Proc. SPIE (2005)
2. Pomplun, J., Burger, S., Schmidt, F., Scholze, F., Laubis, C., Dersch, U.: Metrology of EUV masks by EUV scatterometry and finite element analysis. In: Photomask and NGL Mask Technology XV, vol. 7028, p. 24. Proc. SPIE (2008)
3. Patera, A., Rozza, G.: Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations, 1. edn. MIT Pappalardo Graduate Monographs in Mechanical Engineering (2006)
4. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford University Press (2003)
5. Zhu, Z., Schmidt, F.: An efficient and robust mask model for lithography simulation. In: Proc. SPIE, vol. 6925, p. 126 (2008). URL http://arxiv.org/cond-mat/0510656
6. Enkrich, C., Wegener, M., Linden, S., Burger, S., Zschiedrich, L., Schmidt, F., Zhou, C., Koschny, T., Soukoulis, C.M.: Magnetic metamaterials at telecommunication and visible frequencies. Phys. Rev. Lett. **95**, 203,901 (2005). URL http://arxiv.org/cond-mat/0504774

# Using Nudg++ to Solve Poisson's Equation on Unstructured Grids

Christian Rüdiger Bahls, Gisela Pöplau, and Ursula van Rienen

**Abstract** In this paper we explore the viability of using nodal discontinuous Galerkin methods as implemented in the software library Nudg++ [1] to compute the space charge field of an electron or positron bunch. We will use a benchmark problem to evaluate this method by comparing results from this scheme with solutions obtained analytically.

## 1 Introduction

Being known from the early 70's [2] Discontinuous Galerkin finite element methods (DG-FEM) have only recently become popular as a method for the numerical solution of partial differential equations arising in computational fluid dynamics or computational electromagnetism.

The objective of this paper is to apply DG-FEM to the calculation of space charge fields of particle bunches. The efficient approximation of these fields is one of the primary focuses of beam dynamics simulations as needed for the design of future accelerator facilities like the XFEL [3] or the ILC [4].

One possible approach to these computations is Particle-in-Cell (PIC), especially the Particle-Mesh-method, which calculates the potential in the rest-frame of the bunch by means of the solution of Poisson's equation.

Though being aware that DG-FEM are not the most efficient scheme to solve elliptic Partial Differential Equations – Finite Difference schemes, FIT and FEM come to mind – they have advantages in terms of parallelizability and locality of approximation. Also there are promising advances in PIC methods [5].

We will use the class of nodal DG methods provided by the library Nudg++ [1]. Nudg++ is a C++ implementation of the Matlab-scripts found in [6] and is being developed in collaboration by Nigel Nunn, Tim Warburton, Nico Gödel and others.

Christian Rüdiger Bahls, Gisela Pöplau, Ursula van Rienen
Universität Rostock, A.-Einstein-Str. 2, 18051 Rostock, Germany, e-mail: christian.bahls@uni-rostock.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_13,

## 2 A Bit of Theory

### 2.1 Poisson's Equation

To estimate the electric potential of a particle bunch we aim to approximate the solution $u(x)$ of Poisson's equation in the domain $\Omega \subseteq \mathbb{R}^3$:

$$-\Delta u(x) = f(x), \quad \forall x \in \Omega. \tag{1}$$

For the moment we will use Dirichlet boundary conditions on $\partial\Omega_D = \partial\Omega$:

$$u(x) = g_D(x), \quad \forall x \in \partial\Omega_D. \tag{2}$$

In the context of space charge calculations $u(x)$ denotes the potential and $f(x)$ the source term $f(x) = \rho(x)/\varepsilon_0$ with the charge density $\rho(x)$ and the vacuum permittivity $\varepsilon_0$.

### 2.2 Approximation by Local Polynomials

We approximate $\Omega$ by $K$ non-overlapping simplices. On each of these elements $D^k$ the local solution $u^k(x)$ will be expressed as a polynomial $\tilde{\mathsf{u}}^k(x) \in \mathsf{P}_N$ up to order $N$. On a 3-dimensional simplex this polynomial will have $N_p$ degrees of freedom:

$$N_p = \frac{(N+1)\cdot(N+2)\cdot(N+3)}{6}. \tag{3}$$

In the so called nodal representation on the simplex $D^k$ (using interpolating Lagrange polynomials $l_i^k(x)$ as a basis) these degrees of freedom will be identified with the values of the local solution $u^k(x)$ at the collocation points $x_i^k$ [7]:

$$\tilde{\mathsf{u}}^k(x) = \sum_{i=1}^{N_p} \mathsf{u}^k(x_i^k) l_i^k(x), \quad \forall x \in D^k. \tag{4}$$

The global solution $\mathsf{u}(x)$ on the discretized domain $\tilde{\Omega}$ (the union of all elements $D^k$) will then be approximated by the piecewise $N$-th order polynomial function $\tilde{\mathsf{u}}(x) \in V$:

$$\mathsf{u}(x) \approx \tilde{\mathsf{u}}(x) = \bigoplus_{k=1}^{K} \tilde{\mathsf{u}}^k(x). \tag{5}$$

The function space $V$ is the direct sum of the space of polynomials of order $N$:

$$\tilde{\mathsf{u}}(x) \in V = \bigoplus_{k=1}^{K} \mathsf{P}_N(D^k). \tag{6}$$

## 2.3 Weak Formulation of DG-FEM

To get a suitable weak formulation of (1) we rewrite it as a system of first-order equations:

$$-\nabla\cdot\mathbf{q}(x) = \mathsf{f}(x), \quad \mathbf{q}(x) = \nabla\,\mathsf{u}(x), \quad \forall x \in \Omega. \tag{7}$$

We do now seek approximate solutions $\tilde{\mathsf{u}}(x) \in V$ and $\tilde{\mathbf{q}}(x) \in U = V^3$, such that:

$$-\int_{\tilde{\Omega}} \nabla\cdot\tilde{\mathbf{q}}\phi = \int_{\tilde{\Omega}} \mathsf{f}\,\phi, \quad \int_{\tilde{\Omega}} \tilde{\mathbf{q}}\cdot\pi = \int_{\tilde{\Omega}} \nabla\tilde{\mathsf{u}}\cdot\pi, \quad \forall(\phi,\pi)\in V\times U: \tag{8}$$

Using integration by parts we recover the following weak DG-formulation:

$$\int_{\tilde{\Omega}} \tilde{\mathbf{q}}\cdot\nabla\phi - \sum_{k=1}^{K}\oint_{\partial D^k} \mathbf{n}\cdot\mathbf{q}^*\,\phi = \int_{\tilde{\Omega}} \mathsf{f}\,\phi, \tag{9a}$$

$$\int_{\tilde{\Omega}} \tilde{\mathbf{q}}\cdot\pi = \sum_{k=1}^{K}\oint_{\partial D^k} \mathbf{n}\cdot\pi\;\mathsf{u}^* - \int_{\tilde{\Omega}} \tilde{\mathsf{u}}\,\nabla\cdot\pi. \tag{9b}$$

The functions $\mathsf{u}^*(x)$ and $\mathbf{q}^*(x)$ are called numerical fluxes [8]. They are needed because for a consistent scheme we have to determine a value at the discontinuities at element interfaces.

## 2.4 Invertibility and Numerical Fluxes

The numerical fluxes $\mathsf{u}^*$ and $\mathbf{q}^*$ have to be chosen. At first look a reasonable choice could be the average $\{\!\{.\}\!\}$, which is known as the central flux:

$$\mathsf{u}^* = \{\!\{u\}\!\} = \frac{u^- + u^+}{2}, \quad \mathbf{q}^* = \{\!\{\mathbf{q}\}\!\} = \frac{\mathbf{q}^- + \mathbf{q}^+}{2}. \tag{10}$$

Though when we have a look at the spectrum of the resulting operator we find that it has a non-trivial kernel (as depicted in Fig. 1).

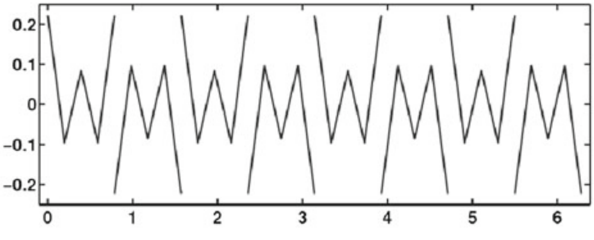


**Fig. 1:** Kernel of a 1D DG-Laplacian (order $N = 4$, $K = 8$ Elements)

**Table 1:** Choice of numerical fluxes for elliptic BVP's [6]

| | $\mathrm{u}^*(x)$ | $\mathbf{q}^*(x)$ |
|---|---|---|
| Stabilized Central flux | $\{\tilde{\mathrm{u}}\}$ | $\{\tilde{\mathbf{q}}\} - \tau[\![\tilde{\mathrm{u}}]\!]$ |
| Local Discontinuous (LDG) flux | $\{\tilde{\mathrm{u}}\} + \beta \cdot [\![\tilde{\mathrm{u}}]\!]$ | $\{\tilde{\mathbf{q}}\} - \beta[\![\tilde{\mathbf{q}}]\!] - \tau[\![\tilde{\mathrm{u}}]\!]$ |
| Interior Penalty (IPDG) flux | $\{\tilde{\mathrm{u}}\}$ | $\{\nabla\tilde{\mathrm{u}}\} - \tau[\![\tilde{\mathrm{u}}]\!]$ |

To get an invertible operator we have to penalize the jumps $[\![.]\!]$ at the interfaces:

$$[\![u]\!] = \mathbf{n}^- u^- + \mathbf{n}^+ u^+, \quad [\![\mathbf{q}]\!] = \mathbf{n}^- \cdot \mathbf{q}^- + \mathbf{n}^+ \cdot \mathbf{q}^+. \tag{11}$$

This leads to the major choices for the fluxes $\mathrm{u}^*(x)$ and $\mathbf{q}^*(x)$ as listed in Table 1.

In the rest of this paper we will be using the Interior Penalty flux.

## 2.5 Consistency and Coercivity

Using the Interior Penalty (IPDG) flux and the strong DG-formulation:

$$-\int_{\tilde{\Omega}} \nabla \cdot \tilde{\mathbf{q}}\, \phi + \sum_{k=1}^{K} \oint_{\partial D^k} \mathbf{n} \cdot (\mathbf{q}^* - \tilde{\mathbf{q}})\, \phi = \int_{\tilde{\Omega}} \mathrm{f}\, \phi, \quad \forall \phi \in V, \tag{12a}$$

$$\int_{\tilde{\Omega}} \tilde{\mathbf{q}} \cdot \pi = \sum_{k=1}^{K} \oint_{\partial D^k} \mathbf{n} \cdot \pi\, (\mathrm{u}^* - \tilde{\mathrm{u}}) + \int_{\tilde{\Omega}} \nabla \tilde{\mathrm{u}}\, \pi, \quad \forall \pi \in U, \tag{12b}$$

we can recover a primal scheme (without resorting to an auxiliary variable $\mathbf{q}$). Using Galerkin's method we can get a system of linear equations similar to the usual FEM formulation:

$$\mathbf{S}\, \tilde{\mathrm{u}} = \mathbf{M} \tilde{\mathrm{f}}. \tag{13}$$

The stiffness matrix $\mathbf{S}$ is symmetric (Fig. 8). If the penalty parameter $\tau$ is chosen large enough then the resulting operator will be coercive [6]. It can also be shown that this scheme is consistent with the classical solution [6].

The inhomogeneous Dirichlet boundary condition (2) is enforced by setting the numerical flux $\mathrm{u}^*(x)$ to $g_D(x)$ on $\partial\tilde{\Omega}_D$. It will than become part of the right hand side. Together with the coercivity of $\mathbf{S}$ we get a unique solution and the scheme converges to the solution of Poisson's equation.

## 2.6 Numerical Solution

The resulting system of equations can efficiently be solved by Cholesky decomposition (when small enough) or a preconditioned Conjugate Gradient algorithm (the matrix $\mathbf{S}$ is symmetric positive definite). As preconditioner Nudg++ provides an incomplete Cholesky decomposition with threshold. For comparison we also implemented an incomplete Cholesky decomposition with zero fill-in.

# 3 Using Nudg++

## 3.1 The Benchmark Problem

A common benchmark problem for the calculation of the potential and the electric field of a charged particle bunch is a uniformly charged sphere or ellipsoid. We will use it because the solution is known analytically.

For simplification we chose a uniformly charged sphere of radius $R = 1$ m with charge $Q = 4\pi C$ inside a spherical domain of radius 2 m. We will be using the analytical solution at the boundary as the inhomogeneous Dirichlet boundary condition. We then compare the result with the exact solution given by [9]:

$$\mathsf{u}(x) = \begin{cases} \dfrac{Q}{4\pi\varepsilon_0 R} \cdot \left(\dfrac{3}{2} - \dfrac{\|r(x)\|^2}{2R^2}\right) & \text{for } \|r(x)\| \leq R, \\ \dfrac{Q}{4\pi\varepsilon_0 \|r(x)\|} & \text{otherwise.} \end{cases} \tag{14}$$

In a first experiment we used the automatic mesh generator NETGEN [10] (bundled with Nudg++) to triangulate the sphere. The result (as can be seen in Fig. 2) was not very satisfying so we defined a series of three concentric spheres to force the mesh generator to generate a finer grid around the discontinuity in the charge-distribution (Fig. 3).

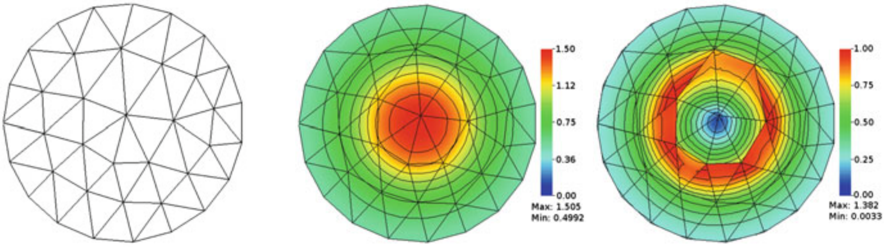


**Fig. 2:** Generated grid ($K = 668$) sliced in the $x$–$y$-plane, potential and **E**-field; order $N = 3$

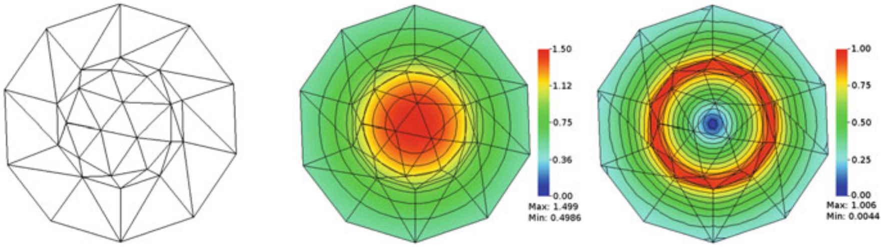


**Fig. 3:** Optimized grid ($K = 536$) sliced in the $x$–$y$-plane, potential and **E**-field; order $N = 3$

### 3.2 Adaptive Grid Refinement

The class ROHOP3D implements in three dimensions the grid-refinement described for two dimensions in [11]. For this it uses the oscillation of the locally approximated function on the element and its edges. The successive refinement reduces the error in the potential (Fig. 4 left-to-right). Though the maximum error in the **E**-field stagnates (Fig. 5). The Tables 2 and 3 in the appendix contain more details.

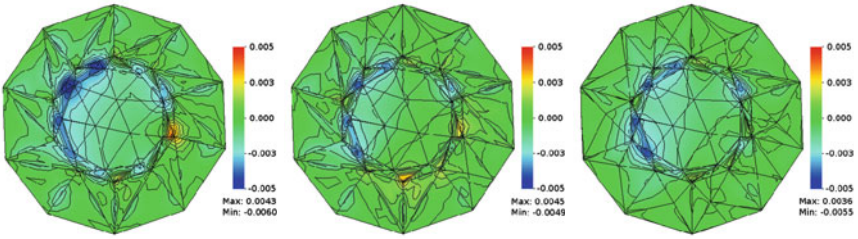


**Fig. 4:** Successive Grid refinement resulting in error reduction for the electric potential; order $N = 3$

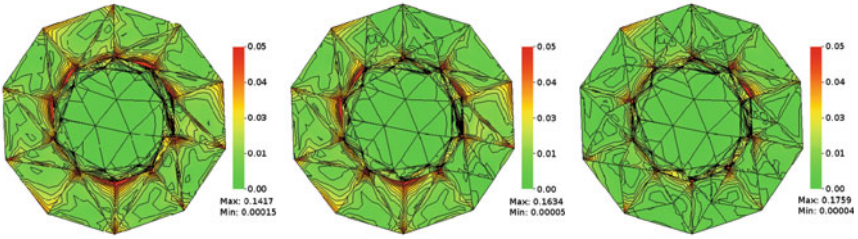


**Fig. 5:** Error reduction for the **E**-field is not optimal, order $N = 3$

Increasing the polynomial order does not increase the convergence rate of the refinement (Fig. 6). We assume this to be Gibbs phenomenon (Fig. 7). This was less pronounced for polynomial interpolation of even order (Fig. 6).

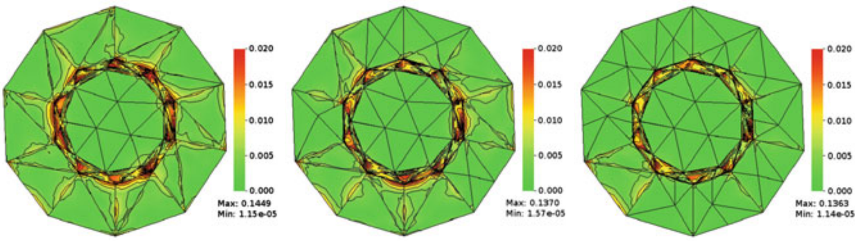


**Fig. 6:** Smaller initial error for polynomial order $N = 4$, but maximum error in **E**-field stagnates

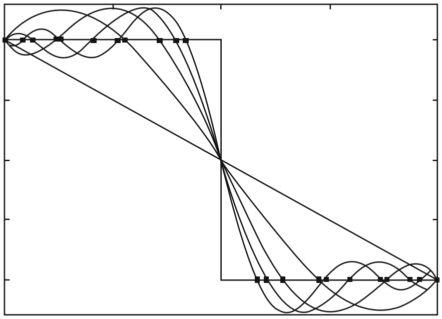


**Fig. 7:** Gibbs phenomenon: using higher order Polynomials does not improve the quality of the approximation, the maximum error does not decrease

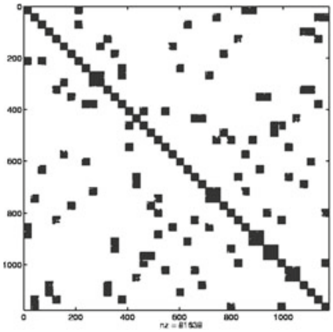


**Fig. 8:** Sparsity pattern of a 2D IPDG-operator ($N = 6$, $K = 42$, $N_p = 28$)

To check the consistency of Nudg++'s IPDG implementation we used the divergence of the analytical solution for the **E**-field (using the DG-differential operators) as a right-hand side and compared the numerical solution resulting from this system with the analytical solution. The error in the **E**-field is of the same magnitude as it is for the original system.

We also tried filtering the coefficients of the higher order polynomials, it did not improve the convergence rate. Another possibility is a smoother charge distribution:

$$\rho(x) = \frac{1 - \tanh(\sigma(\|r(x)\|^2 - R^2))}{2}. \tag{15}$$

Using an appropriate parameter $\sigma$ this gives optimal convergence in the energy norm. Though we loose the ability to compare with an analytical solution.

## 4 Conclusion

Even for very coarse grids the IPDG-method gives a very good approximation of the potential of the charged sphere. For a better approximation of the discontinuous charge distribution the grid should have a curved interface at the discontinuity. Perhaps a Local Discontinuous Galerkin flux should be tried.

Another possible improvement for the IPDG method in Nudg++ would be the implementation of an open boundary condition, possibly Robin's boundary condition. This way the computational domain could be shrunken considerably.

**Acknowledgements** This work is supported by the DFG under contract number RI 814/18-1. and partly by DESY, Hamburg.

# Appendix

**Table 2:** Tabulated data ($N = 3$, $N_p = 20$, starting with an automatically generated grid)

| $l$ | $K$ | $K \cdot N_p$ | $\|\cdot\|_{\tilde{\Omega}}$ | osc$^2$ | $\|\cdot\|_\infty$ | $\|\nabla \cdot\|_\infty$ | Time (s) |
|---|---|---|---|---|---|---|---|
| 1 | 668 | 13360 | 7.83e-02 | 5.795 | 0.044 | 0.427 | 2.27 |
| 2 | 804 | 16080 | 4.24e-02 | 2.326 | 0.038 | 0.289 | 2.91 |
| 3 | 1017 | 20340 | 2.59e-02 | 1.549 | 0.039 | 0.308 | 3.96 |
| 4 | 1323 | 26460 | 1.78e-02 | 1.014 | 0.033 | 0.316 | 5.54 |
| 5 | 1548 | 30960 | 1.20e-02 | 0.621 | 0.036 | 0.333 | 6.63 |
| 6 | 1943 | 38860 | 0.85e-02 | 0.394 | 0.031 | 0.293 | 8.71 |
| 7 | 2288 | 45760 | 0.62e-02 | 0.239 | 0.014 | 0.238 | 10.89 |

**Table 3:** Tabulated data ($N = 3$, $N_p = 20$, starting with the hand-optimized grid)

| $l$ | $K$ | $K \cdot N_p$ | $\|\cdot\|_{\tilde{\Omega}}$ | osc$^2$ | $\|\cdot\|_\infty$ | $\|\nabla \cdot\|_\infty$ | Time (s) |
|---|---|---|---|---|---|---|---|
| 0 | 536 | 10720 | 3.23e-03 | 0.084 | 0.0056 | 0.073 | 1.71 |
| 1 | 765 | 15300 | 2.38e-03 | 0.072 | 0.0047 | 0.082 | 2.94 |
| 2 | 1071 | 21420 | 1.76e-03 | 0.057 | 0.0055 | 0.091 | 4.42 |
| 3 | 1542 | 30840 | 1.28e-03 | 0.043 | 0.0049 | 0.094 | 7.03 |
| 4 | 1940 | 38800 | 0.97e-03 | 0.032 | 0.0049 | 0.093 | 9.46 |
| 5 | 2614 | 52280 | 0.72e-03 | 0.019 | 0.0039 | 0.085 | 13.43 |
| 6 | 3177 | 63540 | 0.58e-03 | 0.013 | 0.0030 | 0.079 | 17.13 |

# References

1. Nunn, N., Warburton, T.: Nudg++: a nodal unstructured Discontinuous Galerkin framework. Online document (2008). URL http://www.nudg.org/. Cited Oct 10 2008
2. Reed, W., Hill, T.: Triangular mesh methods for the neutron transport equation. Tech. report LA-UR-73-479, Los Alamos Scientific Laboratory (1973)
3. DESY, Hamburg, Germany: Das europäische Röntgenlaserprojekt XFEL (2005)
4. Various: What is the ILC? (2008). URL http://www.linearcollider.org/
5. Jacobs, G.B., Hesthaven, J.S.: Implicit-explicit time integration of a high-order particle-in-cell method with hyperbolic divergence cleaning. submitted to Elsevier (2008)
6. Hesthaven, J., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications, *Texts in Applied Mathematics*, vol. 54. Springer Verlag, New York (2008)
7. Warburton, T.: An explicit construction for interpolation nodes on the simplex. Journal of Engineering Mathematics **56**, 247–262 (2006)
8. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, D.: Discontinuous galerkin methods for elliptic problems
9. Pöplau, G., van Rienen, U.: A self-adaptive multigrid technique for 3D space charge calculations. IEEE Transactions on Magnetics **44**(4), to appear (2008)
10. Schöberl, J.: Netgen - automatic mesh generator. Online document (2008). URL http://www.mathcces.rwth-aachen.de/netgen/. Cited Oct 10 2008
11. Hoppe, R.H.W., Kanschat, G., Warburton, T.: Convergence analysis of an adaptive interior penalty discontinuous galerkin method. SIAM J. Numer. Anal. p. to appear (2008)

# Magnetic Force Calculations Applied to Magnetic Force Microscopy

Thomas Preisner and Wolfgang Mathis

**Abstract** In IC failure analysis the detection of currents is often used to indicate the presence of a defective device. One method used for this analysis is the Magnetic Force Microscopy (MFM). Employing this technique measurement errors often occur as for instance due to heterogeneous magnetic tip coatings, fabrication/abrasion errors of the MFM tips and vibrations during a MFM scanning process. Hence, in this work a theoretical model of the MFM was developed to verify and improve the results of laboratory MFM measurements. Therefore a scanning process is simulated and different force calculation methods are implemented and compared with each other in order to obtain the total magnetic force acting on the cantilever as well as the local magnetic force densities.

## 1 Introduction

Due to technical advances in the development of integrated circuits a reduction of the dimensions of electronic devices and structures is feasible. Consequently, the IC failure analysis, which makes use of occurring currents as a possible evidence for defective devices, becomes more complex. The magnetic field caused by these currents can be detected by using sensitive techniques such as the Magnetic Force Microscopy (MFM). By using this method a magnetic coated tip is mounted underneath a micrometer scaled cantilever, which scans over a magnetic field inducing sample surface. The magnetic interactions between the tip and the sample surface causes an attracting or repulsing force acting on the cantilever. This deflection is detected by a laser beam, which is focused onto the cantilever topside and reflected towards a segmented photo diode. With this technique it is possible to draw conclusions about the sample magnetizations or currents, depending on the degree of the

Thomas Preisner, Wolfgang Mathis
Institut für Theoretische Elektrotechnik, Leibniz Universität Hannover, Appelstr. 9A, 30167 Hannover, Germany, e-mail: preisner@tet.uni-hannover.de, mathis@tet.uni-hannover.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_14,

deflection. Further information about the scanning process and a detailed physical explanation are reported in [1, 2]. With respect to the IC failure analysis, in [3] it is demonstrated that the MFM technique is applicable for the detection of currents down to $1\mu A$. In order to improve these measurements and to overcome several possible error sources, as for example different kinds of tip geometries and physical properties of the magnetic coating or hysteresis influences of soft magnetic sample materials, a MFM model was developed by using the finite element method (FEM). In this paper the first results of these studies concerning the approach of modeling a MFM scanning process are presented. Therefore, different force calculation methods like the Virtual Work Principle and the Maxwell Stress Tensor have been implemented and compared with each other. In fact, both methods are often used to obtain the total force of an object under investigation, but in the case of a permanent magnet the force distributions strongly differ from each other [4, 5]. For this reason the Virtual Work Principle was furthermore implemented in our model in such a manner as it is shown in [4, 5], to obtain the appropriate physical local forces.

## 2 Numerical Formulation

In order to develop an applicable theoretical model of the MFM it is necessary to describe the causes for the magnetic fields. Therefore two different sources must be considered, the current density $\mathbf{J}$ and the material magnetization $\mathbf{M}$. For this kind of problem the fundamental expression is the well known curl-curl equation

$$\nabla \times \frac{1}{\mu}(\nabla \times \mathbf{A}) = \nabla \times \frac{\mu_0}{\mu}\mathbf{M} + \mathbf{J}, \tag{1}$$

whereas $\mathbf{A}$ is the magnetic vector potential, $\mu$ is the material permeability and $\mu_0$ is the permeability of free space. For the purpose to ensure the uniqueness of the magnetic vector potential and to make the solution of the coupled system numerically stable [6], the Coulomb gauge is added in such a manner to (1).

$$-\nabla \frac{1}{\mu} \nabla \cdot \mathbf{A} = 0 \tag{2}$$

Applying the method of weighted residuals and using Gauss law and a vector identity, the weak formulation can be obtained

$$\int\limits_{\Omega} \left[ \left( \frac{1}{\mu} \nabla \times \mathbf{A} \right)^T (\nabla \times \omega) + \frac{1}{\mu}(\nabla \cdot \mathbf{A})(\nabla \cdot \omega) \right] \mathrm{d}\Omega - \int\limits_{\Gamma} \omega \times \left( \frac{1}{\mu} \nabla \times \mathbf{A} \right) \mathrm{d}\Gamma$$
$$- \int\limits_{\Gamma} \left( \frac{1}{\mu} \nabla \cdot \mathbf{A} \right) \omega \, \mathrm{d}\Gamma = \int\limits_{\Omega} \left[ \frac{\mu_0}{\mu} \mathbf{M}^T (\nabla \times \omega) + \omega^T \mathbf{J} \right] \mathrm{d}\Omega - \int\limits_{\Gamma} \omega \times \frac{\mu_0}{\mu} \mathbf{M} \, \mathrm{d}\Gamma, \tag{3}$$

where $\omega$ is the vector weighting function. By using the Galerkin method and solving the coupled system of equations, the occurring magnetic induction can be found by $\mathbf{B} = \nabla \times \mathbf{A}$.

In order to describe the MFM scanning process with a theoretical model not only a high-precision field calculation is necessary, but also the forces acting on the cantilever must be obtained. Several works are dealing with the relation between occurring errors of the field calculation and the error propagation obtaining the occurring forces [7,8]. Thus, force calculations should be handled with care. Up until now they are still a topic of interest in research. Many different methods have been developed, but a calculation technique with a sufficiently high accuracy valid for every possible experimental configuration is still missing. Just like different results of authors show [9, 10], the solution is rather dependent on the problem under investigation. Thereby, the most commonly used methods are the equivalent source methods, the Maxwell Stress Tensor (MST) and the Virtual Work Principle (VWP). In the presented paper, the latter ones are used to calculate the forces acting on the cantilever and are compared with each other.

### 2.1 Maxwell Stress Tensor

The classical approach for the MST is to replace the ferromagnetic material in the region of interest by a distribution of currents, such that the external field is not altered [11]. Derivable from Lorentz volume force density, the MST can be obtained as

$$\mathbf{T} = \frac{1}{\mu_0} \begin{bmatrix} B_x^2 - \frac{1}{2}|\mathbf{B}|^2 & B_x B_y & B_x B_z \\ B_y B_x & B_y^2 - \frac{1}{2}|\mathbf{B}|^2 & B_y B_z \\ B_z B_x & B_z B_y & B_z^2 - \frac{1}{2}|\mathbf{B}|^2 \end{bmatrix}. \tag{4}$$

Using (4), the occurring magnetic force can be computed by an integration of the divergence of the Maxwell Stress Tensor $\mathbf{T}$ over a domain $\Omega$

$$\mathbf{F} = \int_{\Omega} \nabla \cdot \mathbf{T}\, \mathrm{d}\Omega = \int_{\Gamma} \mathbf{T}\, \mathrm{d}\Gamma, \tag{5}$$

which can be transformed to an integral over the enclosing surface by applying Gauss law. However, previous works have shown that the force computed with the MST heavily depends on the integration surface enclosing the body under investigation as it is shown for example in [12]. In order to calculate the total magnetic force a summation of the computed local stresses at all points of the bounding surface is needed. However, especially for a subsequent structural analysis it should be noted that as the MST is based on an equivalent mathematical model, the computed local stresses have no physical meaning [4,5].

### 2.2 Virtual Work Principle

The VWP, introduced by J.L. Coulomb [13], is based on the energy law and the principle of a virtual displacement of the considered body. Then, the total magnetic force can be calculated by the derivation of the magnetic energy or co-energy, while keeping the flux or current constant [10]. In a permanent magnet the energy formulation is

$$W = \int\limits_{\Omega} \int_{B_r}^{B} \mathbf{H}^T \,\mathrm{d}B \,\mathrm{d}\Omega = \frac{1}{2\mu_0} \int\limits_{\Omega} (\mathbf{B} - \mathbf{B}_r)^T (\mathbf{B} - \mathbf{B}_r) \,\mathrm{d}\Omega, \tag{6}$$

where $\mathbf{B}_r$ is the remanent induction. A derivation of the energy $W$ in one direction $i$ leads to the corresponding force $F_i$ in this direction. In a finite element approach the domain $\Omega$ is divided into a set of subdomains. A local displacement of a node $k$ leads to a variation of the energy in all elements surrounding this node. This yields to a nodal force which can be obtained by solving (7) at the elements $e$ corresponding to a node $k$ in a direction $i$

$$\begin{aligned} F_{ik} = -\sum_{e_k} \Bigg[ & \int\limits_{\Omega_{e_k}} \frac{(\mathbf{B} - \mathbf{B}_r)^T}{\mu_0} \mathscr{J}^{-1} \frac{\delta \mathscr{J}}{\delta s_i} \mathbf{B} \,|\mathscr{J}|\, \mathrm{d}\Omega_{e_k} \\ & + \int\limits_{\Omega_{e_k}} \frac{(\mathbf{B} - \mathbf{B}_r)^T (\mathbf{B} - \mathbf{B}_r)}{2\mu_0} \frac{\delta |\mathscr{J}|}{\delta s_i} \mathrm{d}\Omega_{e_k} \Bigg], \end{aligned} \tag{7}$$

where $s_i$ is the virtual displacement in the direction $i$ and $\mathscr{J}$ is the Jacobian matrix, which gives a relationship between the local and global coordinate systems. Then, the total force is given by a summation of the local forces at all nodes (in the following named as VW1). But due to the simplified energy expression (6) the whole energy of the permanent magnet, particularly the stored energy during the nonlinear magnetization process, is incorrectly described. Hence, similar to the local stresses obtained by the MST, the calculated local forces of VW1 are also physically not appropriate [4, 5].

### 2.3 Local Interaction Forces

It was already suggested before that for structural analyses or even magnetostriction phenomena a proper solution of the local forces is needed. In a theoretical consideration the total force solution can be decomposed into two different parts [4]. The first ones are called the intrinsic forces. These are the obtained forces of a single permanent magnet in air without other ambient influences. Thereby, the considered energy of this permanent magnet is the stored energy of the magnetization process. The second ones are the interaction forces, which arise from an external magnetic field. Unlike the intrinsic energy, the occurring interaction energy is well expressed with (6), while a linear rigid model can be assumed for the permanent magnet. With respect to the structural analysis and the design of electrical and mechanical coupled devices, the interaction forces are the relevant ones anyway. For the interaction force evaluation the intrinsic forces of the permanent magnet have to be withdrawn from the forces obtained by (6) and (7), respectively. Therefore (8) has to be solved

$$
\begin{aligned}
F_{interaction,ik} = F_{ik} - F_{intrinsic,ik} = -\sum_{e_k}\Bigg[ & \int_{\Omega_{e_k}} \frac{(\mathbf{B}-\mathbf{B}_r)^T}{\mu_0}\, \mathscr{J}^{-1}\frac{\delta \mathscr{J}}{\delta s_i}\mathbf{B}\,|\mathscr{J}|\,\mathrm{d}\Omega_{e_k} \\
& - \int_{\Omega_{e_k}} \frac{(\mathbf{B}_{air}-\mathbf{B}_r)^T}{\mu_0}\, \mathscr{J}^{-1}\frac{\delta \mathscr{J}}{\delta s_i}\mathbf{B}_{air}\,|\mathscr{J}|\,\mathrm{d}\Omega_{e_k} \\
- \int_{\Omega_{e_k}} \frac{\mathbf{B}^T\,(2\mathbf{B}_r-\mathbf{B})}{2\mu_0}\,\frac{\delta\,|\mathscr{J}|}{\delta s_i}\,\mathrm{d}\Omega_{e_k} & + \int_{\Omega_{e_k}} \frac{\mathbf{B}_{air}^T\,(2\mathbf{B}_r-\mathbf{B}_{air})}{2\mu_0}\,\frac{\delta\,|\mathscr{J}|}{\delta s_i}\,\mathrm{d}\Omega_{e_k}\Bigg], \qquad (8)
\end{aligned}
$$

where $\mathbf{B}_{air}$ is the magnetic induction of the single magnet in air. This equation allows the local force computation with respect to a permanent magnet (in the following named as VW2).

## 3 MFM Model

Describing the scanning process of a MFM theoretically, one has to solve a classical multiscale problem. In an ideal case, the tip has to be atomically sharp. In real terms, manufactured tips have a radius of the tip apex down to a few nanometers while the height is thousand times larger. Hence, for modeling the tip apex a very fine mesh is used while a rough one is taken for the outer regions. As an example for the three dimensional model used in this work, a current carrying u-shaped microconductor is considered which is scanned by a cantilever holding the magnetic coated tip underneath (Fig. 1). In this approach a conically shaped tip with an angle of $\alpha = 28°$

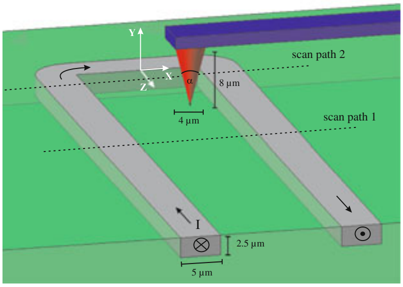


**Fig. 1:** Configuration of the 3D MFM model

was assumed. This tip consists of a cobalt-chromium compound, which is orthogonally magnetized with respect to the sample surface. Analogous to [14], the value of the magnetization in $y$-direction was set to $M_y = -749\frac{kA}{m}$. The thickness of this magnetic coating is equal to $50nm$. The microconductor features a width of $5\mu m$, a thickness of $2.5\mu m$ and carries a current of $1mA$. Furthermore, the two dashed lines shown in Fig. 1 denote two different scan paths of the MFM cantilever.

## 4 Numerical Results

In the case of scan path 1 the $x$-component of the lateral force is illustrated in Fig. 2a. The lateral force $F_x$ is shown in a span of approximately $60\mu m$, calculated with each of the three described force calculation methods. The MST and VW1 approaches are almost identical. The third method (VW2) differs slightly from these results, especially around the force peaks. A possible reason for this behavior is, that the VW2 approach is numerically negatively effected by the further terms being considered in (8), because of the discretization error of $\mathbf{B}$. Due to the mechanical properties of the cantilever, the lateral force $F_x$ does not play an important role during laboratory measurements. The force of interest is the attracting or repulsing force $F_y$ and the force $F_z$, which is responsible for the torsion of the cantilever. In Fig. 2b the normal force with respect to the sample surface along scan path 1 is shown and in Fig. 2c the lateral force $F_z$ along scan path 2. Both forces are located in a $pN$-range. Due to the direction of the current density, there is a negatively directed magnetic induction between the parallel segments of the microconductor and a decreasing value of $B_y$ outside of these segments. These fields lead to a dominating negative force between the wires (Fig. 2b). Along scan path 2 a dominating lateral force is clearly noticeable between the conductor segments. Considering the total force calculation, all different kinds of implemented methods are in good agreement with each other. Furthermore, the curve progression concurs with the reported measurement results in [3].

As it was mentioned in the introduction, measurement errors could occur due to fabrication errors, heterogeneous tip coatings and small vibrations during a scanning

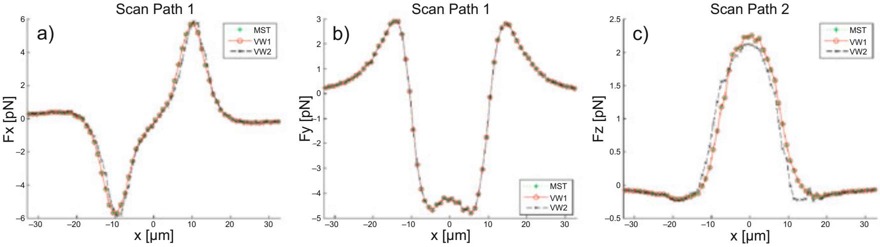


**Fig. 2:** Total force: **a** $F_x$ along scan path 1, **b** $F_y$ along scan path 1, **c** $F_z$ along scan path 2

process. But especially for the consideration of soft magnetic sample materials, the tip magnetization can negatively influence the magnetic properties of these sample materials and even change the direction of the magnetic domains. Thus, it is necessary to simulate the overall scanning process of the configuration under investigation in order to diminish these occurring errors. For this purpose a simulated MFM image of the whole scanning process is shown in Fig. 3a. The position of the microconductor is denoted by the gray lines. In order to improve the 3D MFM model, the mechanical deflection must be taken into account. For this reason the local force densities on the magnetic coating, which are calculated with (8), have to be considered for a detailed subsequent structural analysis (Fig. 3b). Thereby, the cantilever is located between the segments of the u-shaped microconductor. At this position the magnetic interactions between the tip and the sample lead to an attracting force acting on the cantilever. This behavior is depicted by the local force density vectors and an absolute value illustration (Fig. 3b). As expected, the dominating local force density can be found near the tip apex.

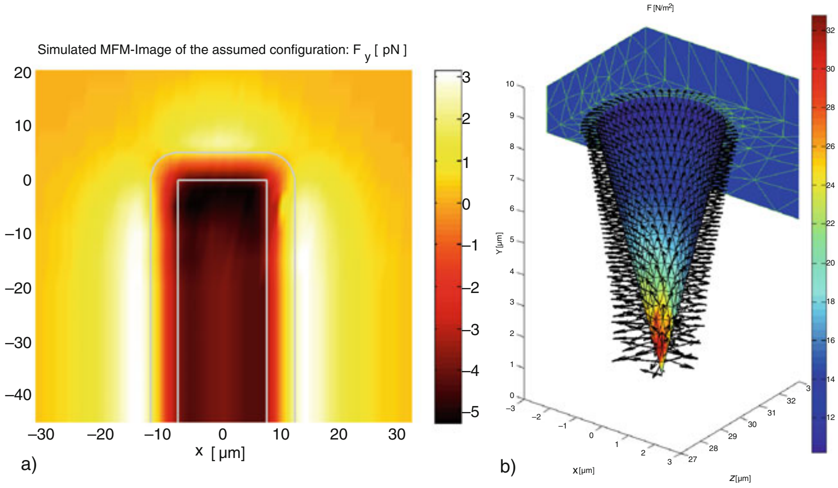


**Fig. 3: a** Simulated MFM image, **b** Local force densities on the magnetic coated tip

## 5 Conclusions

In the presented paper a three dimensional model based on a finite element approach of a magnetic force microscope was presented. In this model a scanning process of a MFM over a u-shaped microconductor was investigated theoretically. Therefore, based on a precise magnetic field calculation, it was shown at different scan paths of the assumed configuration that every implemented force calculation method is applicable to obtain the total force. The resulting total forces are in good agreement to each other. Furthermore, to compare the theoretical considerations with laboratory MFM measurements, an overall scanprocess was numerically investigated. As the Maxwell Stress Tensor and Virtual Work Principle are able to calculate the total force on a body under investigation, these methods are unsuitable to obtain the local force densities on a permanent magnet with physical meaning in the manner described in this paper. Therefore, a third method based on the Virtual Work Principle was implemented as it was reported in [4]. The resulting local interaction force densities on the magnetic coating seems to be appropriate with respect to the physical behavior.

**Acknowledgements** The authors would like to thank M. Greiff for his contribution in the implementation of the force calculation methods and for many valuable discussions.

## References

1. Martin, Y., Wickramasinghe, H.K.: Magnetic imaging by “force microscopy” with 1000 A resolution. Appl. Phys. Lett., **50**, 1455–1457 (1987)
2. Wadas, A., Güntherodt, H.J.: Lateral resolution in magnetic force microscopy. Application to periodic structures. Phys. Lett. A, **146**, 277–280 (1990)
3. Pu, A., Rahman, A., Thomson, D.J., Bridges, G.E.: Magnetic force microscopy measurement of current on integrated circuits. In: Proceedings of the 2002 IEEE Canadian Conference on Electrical and Computer Engineering, CCECE 2002, pp. 439-444 (2002)
4. De Medeiros, L.H., Reyne, G., Meunier, G., Yonnet, J.P.: Distribution of electromagnetic force in permanent magnets. IEEE Trans. Magn., **34**, 3012–3015 (1998)
5. De Medeiros, L.H., Reyne, G., Meunier, G.: About the distribution of forces in permanent magnets. IEEE Trans. Magn., **35**, 1215–1218 (1999)
6. Preis, K., Bardi, I., Biro, O., Magele, C., Renhart, W., Richter, K.R., Vrisk, G.: Numerical analysis of 3D magnetostatic fields. IEEE Trans. Magn., **27**, 3798–3803 (1991)
7. McFee, S., Lowther, D.A.: Towards Accurate and Consistent Force Calculation in Finite Element Based Computational Magnetostatics. IEEE Trans. Magn., **5**, 3771–3773 (1987).
8. Müller, W.: Comparison of different methods of force calculation. IEEE Trans. Magn., **26**, 1058–1061 (1990)
9. Chun, Y., Lee, J.: Comparison of magnetic levitation force between a permanent magnet and a high temperature superconductor using different force calculation methods. Physica C, **372**, 1491–1494 (2002)
10. De Medeiros, L.H., Reyne, G., Meunier, G.: Comparison of Global Force Calculations on Permanent Magnets. IEEE Trans. Magn., **34**, 3560–3563, (1998)
11. Salon, S.J., Slavik, C.J., DeBortoli, M.J., Reyne, G.: Analysis of Magnetic Vibrations in Rotating Electric Machines. In: S. Ratnajeevan and H. Hoole (eds.) Finite Elements, Electromagnetics and Design. Elsevier, Amsterdam (1995)
12. Benhama, A., Williamson, A.C., Reece, A.B.J.: Computation of electromagnetic forces from finite element field solutions. In: Proceedings of the 3rd International Conference on Computation Electromagnetics, pp. 247–252. Bath, UK (1996)
13. Coulomb, J.L.: A methodology for the determination of global electromechanical quantities from a finite element analysis and its application to evaluation of magnetic forces, torques and stiffness. IEEE Trans. Magn., **19**, 2514–2519 (1983)
14. Carl, A., Lohau, J., Kirsch, S., Wassermann, E.F.: Magnetization reversal and coercivity of magnetic force microscopy tips. J. Appl. Phys., **89**, 6098–6104 (2001)

# Relativistic High Order Particle Treatment for Electromagnetic Particle-In-Cell Simulations

Martin Quandt, Claus-Dieter Munz, and Rudolf Schneider

**Abstract** A recently developed high order field solver for the complete Maxwell equations provides all information needed by a new relativistic particle push method based on a truncated Taylor series expansion up to the desired order of convergence. The property and capability of this approach is demonstrated for different numerical experiments.

## 1 Introduction

The electrical behavior of technical systems like microwave devices is substantially influenced by a flow of charged particles forming a non-neutral plasma inside. A detailed understanding of the phenomena caused by this plasma requires the solution of the Maxwell-Vlasov equations for realistic configurations. An attractive numerical technique to do this is the Particle-in-Cell (PIC) method. In essence, the basic idea of the PIC approach can be summarized as follows: At each time step the electromagnetic fields are obtained by the numerical solution of the full set of the nonstationary Maxwell equations, where different kind of methods like finite volume [1] or discontinuous Galerkin schemes of free selectable order of convergence are applied. Note that the Maxwell part of this solver comprises additionally a purely hyperbolic divergence correction mechanism [2] to ensure the constrain of charge conservation during the simulation. Subsequently, these fields are interpolated to the actual locations of the charged plasma particles which are then pushed by the Lorentz force and redistributed in phase space according to the usual laws of

Martin Quandt, Claus-Dieter Munz
Institut für Aerodynamik und Gasdynamik, Universität Stuttgart, Stuttgart, Germany, e-mail: quandt@iag.uni-stuttgart.de, munz@iag.uni-stuttgart.de

Rudolf Schneider
Forschungszentrum Karlsruhe, Institut für Hochleistungsimpuls und Mikrowellentechnik, Karlsruhe, Germany, e-mail: rudolf.schneider@ihm.fzk.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_15,

dynamics. Afterwards, the particles have to be located with respect to the computational grid in order to assign the contribution of each charge to the changed charge and current density to the nodes of the mesh. These densities are the sources for the Maxwell equations of the subsequent iteration cycle which finally guarantee a self-consistent computation of the interaction of the electromagnetic fields with the charged plasma particles. The recent development of high order Maxwell solvers for electromagnetic wave propagation offers the possibility to construct high order PIC algorithms for the numerical solution of the Maxwell-Vlasov equations. In this context the central challenge is the high order computation of the phase space coordinates of the plasma particles. In the present paper we introduce a new particle treatment based on a Taylor series expansion (TSE) of the phase space variables in time up to the selected order of accuracy of the field solver. The important requirement to establish this high order phase space coordinates calculation is that all necessary spatial as well as temporal derivatives of the electromagnetic fields for each particle are known from the Maxwell solver. This knowledge of the high order derivatives is extensively used to obtain the TSE of the phase space coordinates of the charge as it is explained below.

In the next section the formulation of the governing equations is given and the numerical approximation is discussed. In section 3 we demonstrate the capability and reliability of the high order particle (HIOP) procedure by means of different numerical simulation experiment. Finally, a conclusion and a short outlook is given in section 4.

## 2 Governing Equations and Numerical Approximation

### *2.1 Equation of Motion for Charged Particles*

The general solution of the Vlasov equation is given by its characteristics

$$\frac{\mathrm{d}}{\mathrm{d}t}(m\mathbf{U}) = \mathbf{F}_L(\mathbf{v},\mathbf{x},t)\,, \quad \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{v} \tag{1}$$

with the Lorentz force

$$\mathbf{F}_L = q\,[\mathbf{E}(\mathbf{x},t) + \mathbf{v}\times\mathbf{B}(\mathbf{x},t)] \tag{2}$$

acting on charge $q$ with mass $m$, where $\mathbf{E}$ and $\mathbf{B}$ denote the external applied or/and self electric field and magnetic induction, respectively. The velocity of the charged particle $\mathbf{v}$ is related to the space component of the 4-velocity $\mathbf{U}$ according to [3]

$$\mathbf{v} = \hat{\gamma}\,\mathbf{U}(t)\,, \quad \hat{\gamma}(\mathbf{U}) = \left(1 + \frac{\mathbf{U}\cdot\mathbf{U}}{c^2}\right)^{-1/2} \tag{3}$$

with the inverse relativistic factor $\hat{\gamma}$, where $c$ is the speed of light. As a consequence of the latter relation, the phase space coordinates $(\mathbf{v},\mathbf{x})$ may be regarded as a function of $\mathbf{U}(t)$. For the sake of convenience we rewrite the first equation in (1) with (2) and (3) to obtain Newton's equation of motion for charged particles,

$$\dot{\mathbf{U}} = \frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} = \mathscr{E}(\mathbf{x},t) + \hat{\gamma}\,\mathbf{U} \times \mathscr{B}(\mathbf{x},t)\ , \tag{4}$$

where $q/m$ is absorbed in the electromagnetic fields, i.e. $\mathscr{E} = \frac{q}{m}\mathbf{E}$ and $\mathscr{B} = \frac{q}{m}\mathbf{B}$. Clearly, this expression for the acceleration of the charge also depends on the relativistic velocity (the space component of the 4-velocity), position and time: $\dot{\mathbf{U}} = \dot{\mathbf{U}}(\mathbf{U},\mathbf{x},t)$. Observe further from the latter relation that $\mathbf{U}\cdot\dot{\mathbf{U}} = \mathbf{U}\cdot\mathscr{E}$ holds. The total temporal derivative occurring in (1) and (4) is defined in the present context by

$$\frac{\mathrm{d}}{\mathrm{d}t}(.) = \mathscr{D}(.) := \{D_c + D_U\}\,(.)\ , \tag{5}$$

where the convective derivative $D_c = \frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla_x$ acts on space and time dependent quantities while $D_U = \dot{\mathbf{U}}\cdot\nabla_U = \sum_j \dot{U}_j \frac{\partial}{\partial U_j}$ acts only on velocity dependent expressions.

## 2.2 Numerical Approximation of Particle Phase-Space Coordinates

To obtain a numerical approximation of the phase space coordinates $(\mathbf{v},\mathbf{x})$ with the same order of accuracy (say $\mathscr{K}$) as the field solution from the Maxwell solver, we first perform a truncated Taylor expansion in time up to order $\mathscr{K}$ of the particle velocity according to

$$\mathbf{v}(t) = \sum_{\kappa=0}^{\mathscr{K}} \frac{(t-t_0)^\kappa}{\kappa!} \left[\mathscr{D}^{(\kappa)}\left(\hat{\gamma}\mathbf{U}\right)\right]_{t_0}\ , \tag{6}$$

where (3) is used and $\mathscr{D}^{(\kappa)}(.) = \mathrm{d}^\kappa/\mathrm{d}t^\kappa(.)$ has to be computed at the initial time $t = t_0$. Then, the integration of the latter expansion over the interval $[t_0, t]$ yields in an obvious way the position of the charged particle. Note that the expansion coefficient $\left[\mathscr{D}^{(0)}\left(\hat{\gamma}\mathbf{U}\right)\right]_{t_0}$ in (6) is nothing else than $\mathbf{v}_0 = \mathbf{v}(t_0)$. What remains now to do is to compute simply the $\kappa th$ derivative of $\hat{\gamma}\mathbf{U}$ by applying the operator (5) at $t = t_0$. This is, in principle, a straightforward task but it implies cumbersome and lengthy calculations because the operator $\mathscr{D}(.)$ itself depends on velocity, space and time. However, some auxiliary relation can be found, for instance, we can ascertain from (3) and (4) that

$$\mathscr{D}\left(\hat{\gamma}^n\right) = -\frac{n}{c^2}\,\hat{\gamma}^{n+2}\,\mathbf{U}\cdot\mathscr{E}\ , \tag{7}$$

and, clearly, higher order derivatives of $\hat{\gamma}^n$ can be recursively determined from $\mathscr{D}^{(m)}(.) = \mathscr{D}^{(m-1)}\left(\mathscr{D}(.)\right)$. Moreover, starting from $\mathscr{D}^{(\kappa)}\left(\hat{\gamma}\right) = -\frac{1}{c^2}\mathscr{D}^{(\kappa-1)}\left(\hat{\gamma}^3\,\mathbf{U}\cdot\mathscr{E}\right)$ the derivatives of $\hat{\gamma}$ greater than one may be successively obtained from

$$\mathscr{D}^{(\kappa)}(\hat{\gamma}) = -\frac{1}{c^2}\sum_{\nu=0}^{\kappa-1}\frac{(\kappa-1)!}{(\nu)!(\kappa-\nu-1)!}\mathscr{D}^{(\kappa-1-\nu)}\left(\hat{\gamma}^3\right)\mathscr{D}^{(\nu)}(\mathbf{U}\cdot\mathscr{E}), \tag{8}$$

where now the higher order derivatives of the relativistic velocity as well as the electric field are required. Furthermore, for the computation of the field derivatives the commutator relation

$$[D_c, D_U](.) = D_c\left(\dot{U}_k\right)\frac{\partial}{\partial U_k}(.) - \dot{U}_k\left[\frac{\partial}{\partial U_k}(\hat{\gamma}U_j)\right]\frac{\partial}{\partial x_j}(.), \tag{9}$$

alleviate the task, where the usual summation convention is adopted. The consistent application of the latter relations (7) to (9) is advantageous and the use of a computer algebra system like Maple is helpful.

# 3 Numerical Results

In this section we present results from three different simulation experiments which demonstrate the property and capability of the proposed HIOP approach based on Taylor series expansion of the phase space coordinates in time. Besides direct comparison between numerical and analytical results, we are also interested in the efficiency and accuracy of the proposed HIOP treatment. For this we compute at the end of the simulation time $t_e$ for a given number of discretization points $\delta$ the discrete $L_2$-error or Euclidian norm according to $eN(\mathbf{q},\delta) = ||\mathbf{q}_{num} - \mathbf{q}_{ana}||_2^{(\delta)}$, where $\mathbf{q}_{num}$ and $\mathbf{q}_{ana}$ are the numerical and analytical value of a certain quantity at $t = t_e$. Furthermore, from this norm it is possible to estimate the effective or experimental order of convergence by $EOC = -\log\left(\frac{eN(\mathbf{q},\delta_1)}{eN(\mathbf{q},\delta_0)}\right)/\log\left(\frac{\delta_1}{\delta_0}\right)$, where $\delta_0$ and $\delta_1$ denote the reference and refined time interval discretization, respectively.

## 3.1 Non-relativistic Test Problem

In the first numerical example we consider the non-relativistic motion of a charged particle ($q = m = 1$) in a spatial constant electric field where the magnetic induction is set equal to zero. Each component of the applied oscillating electric field has the form $\mathscr{E}_i(t) = \mathscr{E}_0 \sin(\omega_i t + \phi_0)$, i=1,2,3, where the amplitude $\mathscr{E}_0$ and the phase shift $\phi_0$ in all coordinate directions are fixed equal to one and $(2\pi)^{-1}$, respectively, and the frequencies are chosen to be $\omega_1 = 2\pi$, $\omega_2 = 2/3\pi$ and $\omega_3 = 3/2\pi$. Clearly, by construction this problem decouples and the equation of motion (4) can be immediately integrated. The analytic solution of the phase space coordinates are given by

$$v_i(t) = -\frac{\mathscr{E}_0}{\omega_i}\cos(\omega_i t + \phi_i) + c_{1,i}\,, \quad x_i(t) = -\frac{\mathscr{E}_0}{\omega_i^2}\sin(\omega_i t + \phi_i) + c_{1,i} t + c_{2,i}$$

where the integration constants $c_{1,i}$ and $c_{2,i}$ are determined from the initial values $v_i(t_0)$ and $x_i(t_0)$ at time $t = t_0$, respectively. The Lissajou trajectory depicted in Fig. 1 is obtained by plotting the particle coordinates $y = x_2(t)$ over $z = x_3(t)$ at the end of the simulation period $t_e = 10\,T$ with a periodic time $T = 2\,\pi$. It is obvious from this figure that the numerical result obtained from a formal 6*th* order TSE scheme (filled circles) is in very good agreement with the analytical solution (full line). To get information about the effective – also called design – order of the Boris leap-frog scheme (red line with symbol x) as well as of the HIOP treatment for $\mathscr{K} = 2, \ldots, 5$ (black lines with symbols), we plot in Fig. 2 the discrete Euclidian norm for $\mathbf{q} = \mathbf{v}$ versus the number of discretization points $\delta$. The slopes of the curves in this double-log scale presentation nicely reveal that the design order of the leap-frog and the TSE schemes agree very well with the formal order, for instance, we obtain 6.02 for the formal 6*th* order accurate TSE scheme.

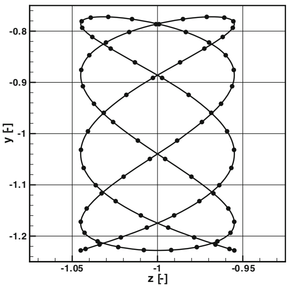


**Fig. 1:** Non-relativistic analytic particle motion (*line*) and the numerical solution (*dots*) after 10 periods calculated with a formal 6*th* order TSE scheme

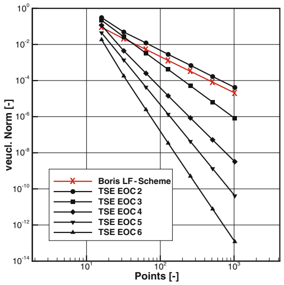


**Fig. 2:** Euclidian error norm for $\mathbf{q} = \mathbf{v}$ versus the number of discretization points for the leap-frog and five different TSE schemes

## 3.2 Relativistic Particle Motion in B-Fields

In the following example we consider the relativistic motion of a positron in the xy plane for a constant magnetic induction ($\mathscr{B} = \frac{qB_z}{m}\,\mathbf{e}_3$, $B_z = 0.1$ Vs/m$^2$). The components of the initial velocity vector are set to $v_{01} = 0.6\,c$ and $v_{02} = v_{03} = 0$ which corresponds to a Lorentz factor of $\gamma_0 = 1.25$. Furthermore the final simulation time is fix to $T_s = 10\,(2\pi/\omega)$ where $\omega = \frac{qB_z}{m\gamma_0}$. Due to the fact that the energy is conserved ($\gamma(t) = \gamma_0$) in this case, the integration of Newton's equation (4) can be straightforward performed [3]. A first simulation result (filled circles) computed with a formal 3*rd* order TSE scheme is seen in Fig. 3. There, a snapshot of the particle position recorded after 10 periods is plotted together with the analytical solution (full line). This 3*rd* order result is a somewhat surprising compared to that of the 2*nd* Boris leap-frog solution (not shown here) which is very close to the analytical one. The reason for the better approximation property of the Boris scheme may be traced back to the fact that this approach explicitly take into account the special form of the Lorentz force (2). However, increasing the order of the Taylor expansion up to $\mathcal{K} = 5$, the numerical solution after 10 periods is in nearly perfect agreement with the analytical result, as seen in Fig.4. The effective order of convergence study for the present relativistic case is given in Table 1 and 2. There, a formal $3rd$ (Table 1) and $5th$ (Table 2) order accurate TSE scheme is investigated for $\mathbf{q} = \mathbf{x}$ (left) and $\mathbf{q} = \mathbf{v}$ (right column). A closer inspection of these tables reveals an excellent agreement between the formal and design order of the Taylor expansion methods.

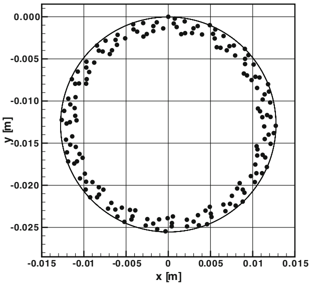


**Fig. 3:** Deviation of numeric solution (*dots*) with a formal 3rd order TSE scheme for the particle position compared to the analytic solution (*line*) after 10 periods with a total points resolution of 160

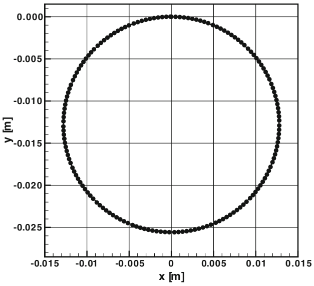


**Fig. 4:** Numerical determined particle position (*dots*) after 10 periods compared to the analytic "circle" is plainly reduced with the same resolution of points

**Table 1:** Maximum deviation of **x** and **v** to analytic solution on example 2 calculated with TSE method with a formal order 3 for different point resolutions

| Points | $eN(\mathbf{v})$ | EOC | $eN(\mathbf{x})$ | EOC |
|---|---|---|---|---|
| 40 | 2.104e+1 | | 5.121e-1 | |
| 80 | 2.099e+0 | 3.33 | 5.318e-2 | 3.27 |
| 160 | 2.324e-1 | 3.18 | 5.998e-3 | 3.15 |
| 320 | 2.728e-2 | 3.09 | 7.107e-4 | 3.08 |
| 640 | 3.303e-3 | 3.05 | 8.644e-5 | 3.04 |

**Table 2:** EOC with Euclidian norm for **x** and **v** obtained from a formal $5th$ order TSE calculation for example 2 for different point resolutions

| Points | $eN(\mathbf{v})$ | EOC | $eN(\mathbf{x})$ | EOC |
|---|---|---|---|---|
| 40 | 1.256e+0 | | 2.784e-2 | |
| 80 | 1.537e-1 | 3.03 | 3.295e-3 | 3.08 |
| 160 | 5.517e-3 | 4.80 | 1.178e-4 | 4.81 |
| 320 | 1.753e-4 | 4.98 | 3.744e-6 | 4.98 |
| 640 | 5.464e-6 | 5.00 | 1.167e-7 | 5.00 |

## *3.3 Particle Motion in Space-Time Dependent Electric Field*

In the previous two examples the electromagnetic field was constant or depends only on time. For the simulation experiment discussed in the following, we consider a one dimensional, non-relativistic test problem where the electric field represents a wave propagating along the x-axis with given frequency $\omega$ and wavenumber $k$ (see also [4]). The evolution of phase space coordinates ($v$, $x$) of the particle is obtained from

$$\dot{v} = \frac{\mathrm{d}v}{\mathrm{d}t} = \mathscr{E}_0 \sin\left(\omega t - kx\right) \ , \quad \dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t} = v \ , \tag{10}$$

where $\mathscr{E}_0 = (qE_0)/m$ is a constant. In the reference wave frame, given by the transformation

$$\xi = kx - \omega t \ , \quad \eta = \frac{k}{\omega} v - 1 \tag{11}$$

there exists an analytic solution which permits a characterization of the long term stability of the integration process. According to the fact that $\mathrm{d}\xi = \mathrm{d}t\,(kv - \omega)$ holds, the integration of the velocity equation yields immediately the result

$$\eta^2(t) = \eta_0^2 + 2Y^2\left[\cos\xi\,(t) - \cos\xi_0\right] \ , \tag{12}$$

where the abbreviation $Y^2 = (k\mathscr{E}_0)/\omega^2$ is introduced and $\eta_0 = \eta\,(t_0)$ and $\xi_0 = \xi\,(t_0)$ denote the initial data of the new variables ($\eta$, $\xi$) at time $t = t_0$. The solutions of the latter equation can be split into two different areas. The first area consists of all pairs of initial data of ($\eta_0$, $\xi_0$) which always satisfy $\eta^2(t) \geq 0$. In the reference wave frame these pairs are located on or outside of the separatrix which is characterized by $\eta_0 = \pm\sqrt{2Y^2\,[1 + \cos\xi_0]}$ and leads to an open trajectory. In the interior of this separatrix there exists some range of $\xi\,(t)$ for which $\eta^2(t) < 0$ and, consequently, the particle is trapped on a closed trajectory. Ideally, a particle will trace the trapped orbit indefinitely, however, loss of accuracy and stability of the integration procedure will lead to departures of the exact trajectory. As proposed in [4], it is convenient to study the accuracy and stability of the integrator near the condition $\eta_0^2 - 2Y^2\cos\xi_0 = 0$, which is also plotted in the Figures 5 and 6 for orientation. Note, if a particle enter into this regime indicates that the integration scheme is not area preserving. The long term stability of an integration scheme corresponds to the deviation to a closed circle. With the so-called stochasticity parameter [4] $K = (\omega\,\Delta t)^2\,Y^2$ a stability limit is defined up to which the integration scheme leads to a stable closed orbit solution. For the numerical simulation results depicted in Fig. 5 and 6 $K = 0.2$, the particle is initialized at $\eta_0 = 1.5$ and $\xi_0 = 0$ and the simulation is performed for 10000 iteration cycles. The numerical result obtained with

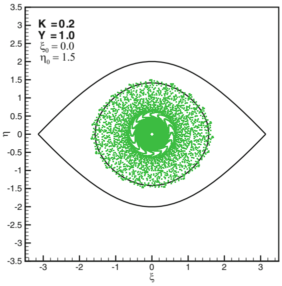


**Fig. 5:** Damped particle trajectory obtained with a third order TSE approach

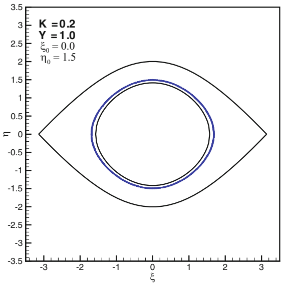


**Fig. 6:** Particle trajectory obtained with a formal fifth order Taylor series expansion

the 3*rd* order TSE scheme is depicted in Fig. 5. We observe that the trajectory deviates clearly from a closed orbit and is damped over the time to the center of the separatrix. Increasing the Taylor expansion up to order $\mathscr{K} = 5$ the scheme catch the problem, resulting in a stable closed orbit solution as it is seen in Fig. 6. The shown particle orbit is identical with this one computed with the area preserving classical leap-frog scheme presented in [4].

## 4 Conclusion and Outlook

The phase space coordinates of charged particles driven by the Lorentz force are numerically computed up to sixth order by a new high order particle (HIOP) method based on truncated Taylor series expansion (TSE) in time. Numerical results obtained from three simulation experiments clearly demonstrate the great potential of the proposed TSE approach. For both non-relativistic and relativistic test cases the numerical TSE results for $\mathscr{K} \geq 5$, are in very good agreement with the available analytic solutions. The capability of the TSE schemes is also proved in the complicated test case of non-linear electromagnetic field. Furthermore, we observe from experimental order of convergence studies that the design order of all schemes are very close to the formal order of the proposed approach. The test stage of the standalone HIOP solver draw to a close and the module should be applied as an attractive alternative to the Boris leap-frog solver in the existing Maxwell-Vlasov module in near future. Clearly, this accounts for a multitude of numerical standard tests to enhance the status to a verified method for scientific application of the new TSE approach and to establish an attractive high order alternative to the second order classical leap-frog method.

**Acknowledgements** We gratefully acknowledge the Landesstiftung Baden-Württemberg who funded the development within the program “Modeling and Simulation on High Performance Computers” from 2003–2005 and the Deutsche Forschungsgemeinschaft (DFG) for funding within the project “Numerische Modellierung und Simulation hochverdünnter Plasmaströmungen”.

## References

1. T. Schwartzkopff, F. Lörcher, C.-D. Munz, and R.Schneider, *Arbitrary High Order Finite-Volume Methods for Electromagnetic Wave Propagation*, Computer Physics Communications, 174:689–703, 2006.
2. C.-D. Munz, P. Omnes, R. Schneider, E. Sonnendrücker and U. Voß, *Divergence correction techniques for Maxwell solvers based on a hyperbolic model*, J. Comput. Phys., 161:484–511, 2000.
3. J.D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1999.
4. V. Fuchs and J.P. Gunn, *On the Integration of Equations of Motion for Particle-in-Cell Codes*, J. Comput. Phys., 214:299–315, 2006.

# A Statistical Characterization of Resonant Electromagnetic Interactions with Thin Wires: Variance and Kurtosis Analysis

O.O. Sy, M.C. van Beurden, B.L. Michielsen, J.A.H.M. Vaessen, and A.G. Tijhuis

**Abstract** A statistical characterization of random electromagnetic interactions affected by resonances is presented. It hinges on the analysis of the variance and the kurtosis to assess the intensity of the resonances. The method is illustrated by the study of a randomly varying thin wire modeled by a Pocklington integral equation.

## 1 Introduction

Interactions between electronic devices and electromagnetic sources in their environment are of prime importance in EMC models for design or maintenance studies. A convenient way to model such interactions is based on the multi-port models of both the electronic components and the interconnect networks making up the complete system. In principle, both types of multi-port models need extensions, in the form of Thévenin or Norton sources, accounting for the presence of exterior sources of electromagnetic fields. In practice, the sources added to the interconnect subsystem are the dominant ones because of the greater geometrical size of the printed wirings compared to the size of the electronic devices. This is even more so when exterior cables come into play.

The range of validity of these models depends on their ability to accurately represent an ensemble of configurations. For non-resonant systems, the study of a few configurations provides a good picture of the overall interaction. However, for resonant phenomena, the number of configurations needed can increase drastically. Instead, a stochastic approach yields a more suitable quantitative and qualitative

O.O. Sy, M.C. van Beurden, J.A.H.M. Vaessen, A.G. Tijhuis
Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands, e-mail: o.o.sy@tue.nl, m.c.v.beurden@tue.nl, j.a.h.m.vaessen@tue.nl, a.g.tijhuis@tue.nl

B.L. Michielsen
ONERA, 2, av E. Belin, 31055 Toulouse Cedex, France, e-mail: bastiaan.michielsen@onera.fr

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_16,

model. Stochastic methods are frequently used in fields as diverse as rough-surface scattering problems [1] and Mode-stirred-Chamber theory [2]. In EMC, random models have been applied to undulating thin-wire setups modeled by transmission-line theory [3], [4], or by integral equations [5]. In all these cases the aim is to quantify the uncertainty of the response parameters, or "observables", by their average and variance. Although these statistics provide bounds for the observable, they do not inform on the presence of extreme values beyond these bounds.

Estimating the probability that an observable will have values beyond a certain distance from the average is important in "risk assessment". Reliable estimates need a good approximation of the entire probability distribution, which is generally impossible to obtain. Gaussian distributions can be fitted by looking only at the first two moments and therefore provide easy estimates. The next few moments are *qualitative* indicators of the suitability of such fits [6]. This paper shows that the kurtosis should be investigated to identify significant deviations from the Gaussian distribution near "risky" resonance conditions.

The outline of this paper is as follows. Section 2 describes the general setup which involves the integral-equation model of a thin wire over a ground-plane. A random parametrization of the problem in Section 3 allows for the definition of the statistical moments of interest, viz. the average, the variance and the kurtosis. All these moments are computed by a sparse-grid quadrature rule, which efficiently handles integrals over multi-dimensional domains. The importance of these moments in characterizing electromagnetic interactions is illustrated in Section 4 through the example of a roughly undulating transmission line illuminated by a plane wave.

## 2 Deterministic Configuration

The purpose of this paper is to show that in electromagnetic interaction configurations with stochastic geometries, the value distribution of observables shows a peculiar behavior near resonance conditions which necessitates the computation of higher order moments, like the kurtosis, before a reliable interpretation of the results can be established. For that purpose, we choose a simple one-port system, consisting of a perfectly conducting wire $S_\alpha$ over a ground plane, in an incident plane wave $E^i$, as shown in Figure 1. The vector $\alpha$ gathers all the variables controlling the geometry of the wire. The electromagnetic coupling itself is observed through the equivalent Thévenin voltage source $V_e(\alpha)$ induced at the port of $S_\alpha$ and defined as

$$V_e(\alpha) = -\frac{1}{I_0}\int_{S_\alpha} j_\alpha \cdot E^i, \tag{1}$$

where $j_\alpha$ is the current distribution flowing on the device in absence of $E^i$, when a current source $I_0$ is applied at the port of the wire [7]. This current $j_\alpha$ follows by solving a frequency-domain electric-field integral equation (EFIE) representing the wire in a transmitting state [5]. The resonances appear at frequencies where a wave,

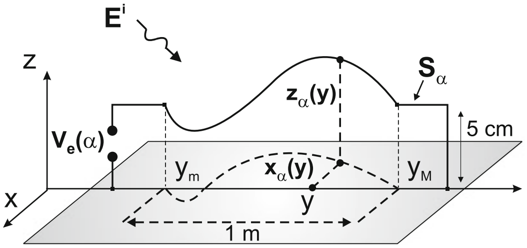


**Fig. 1:** Undulating thin-wire over a PEC plane

propagating along the waveguide formed by the wire and the ground plane, becomes resonant due to the boundary conditions at the wire extremals.

In spite of its simplicity, this configuration, derived from an EMC benchmark [8], is representative for a large class of interaction problems, for example the common-mode interference appearing at the connection of a power cable to a printed circuit board or certain types of wire antenna problems.

## 3 Random Parameterization

When an ensemble $A$ of configurations is considered, computing $V_e(\alpha)$ for each element $\alpha$ of $A$ can be very costly numerically. Instead, the variations of $\alpha$ in $A$ are viewed as random according to a *known* distribution $p_\alpha$. The voltage $V_e(\alpha)$ then becomes a random variable, with statistical moments, such as its mean $\mathbb{E}[V_e]$ and its standard deviation $\sigma[V_e]$, defined as

$$\mathbb{E}[V_e] = \int_A V_e(\alpha')p_\alpha(\alpha')\,\mathrm{d}\alpha', \tag{2}$$

$$\sigma[V_e] = \sqrt{\mathbb{E}[|V_e|^2] - |\mathbb{E}[V_e]|^2} \geq 0. \tag{3}$$

The standard deviation $\sigma[V_e]$ is a positive parameter measuring, in volts, the spread of $V_e$ *around* $\mathbb{E}[V_e]$, as can be seen from Chebychev's inequality [9].

Extreme values of $V_e$, at least $4\sigma[V_e]$ away from $\mathbb{E}[V_e]$, are accounted for by the *kurtosis* $\kappa[|V_e|]$, which is a dimensionless positive moment defined as

$$\kappa[|V_e|] = \mathbb{E}\left[\left(\frac{|V_e| - \mathbb{E}[|V_e|]}{\sigma[|V_e|]}\right)^4\right] \geq 0. \tag{4}$$

Gaussian random variables, which have approximately 95% of their values within a distance of $2\sigma$ to their average, have a kurtosisof 3. Hence, the higher the value of $\kappa[|V_e|]$ above 3, the more occurrences of $V_e$ with very large magnitude are to be expected.

Equation (2) shows that all the statistical moments are defined by integrals involving a known integrand which depends on $V_e$, and over the same support $A$. These integrals can therefore be computed numerically by quadrature rules. Moreover, a significant gain in computation time is achieved by re-using the same samples of $V_e$ to compute the different integrals in Equations (2)-(4).

The most straightforward generalization to integration over higher dimensional spaces, consists in using the Cartesian tensor product of a univariate quadrature rule. However, this leads to a "curse of dimensionality" [10], i.e. exponentially growing numbers of grid points and hence prohibitive numbers of evaluations. Moreover, such Cartesian product rules are not isotropic, i.e. in certain directions of a $d$-dimensional space, the accuracy is of much higher degree than in other directions.

Algorithms, such as Sparse grid (SG) methods, have been found which allow for the elimination of grid points while preserving exact integrals of polynomials up to a given degree in any direction. As such SG methods can be regarded as multidimensional generalizations of Gaussian-type integration rules defined in one dimension. For integrals over moderately dimensioned spaces ($d \leq 10$), the convergence rate of the SG rule is faster than a Monte-carlo approach. In addition, SG rules take advantage of the smoothness of the integrand, unlike Monte-Carlo rules [11]. In this paper, a SG rule is employed which starts from a 1D Clenshaw-Curtis quadrature rule and applies Smolyak's algorithm to build the multidimensional quadrature rule [12].

## 4 Results

With reference to Figure 1, a roughly undulating thin wire is studied with a geometry defined as

$$x_\alpha(y) = \alpha_1 \sin(5\pi y), \qquad z_\alpha(y) = 5 + \alpha_2 \sin(9\pi y) \qquad \text{in cm.} \tag{5}$$

The vector of amplitudes $\alpha = (\alpha_1, \alpha_2)$ has independent and uniformly distributed components in the domains $A_1 = A_2 = [-3;3]$ cm. The average geometry therefore corresponds to the straight wire $S_0$. The incident field is a $\theta$-polarized plane wave with an amplitude of 1 V.m$^{-1}$, and propagating in the direction $\theta_i = 45°$, $\phi_i = 0°$.

A single computation of the induced voltage amounts to 0.1 second. All the statistical moments are computed for 50 frequencies between 100 MHz and 500 MHz, with a relative error below 1%. The number of function evaluations ranges from $N_{min} = 321$ ($\equiv$ 32 seconds) at regular frequencies, to $N_{max} = 7169$ ($\equiv$ 12 minutes) at resonance frequencies, with an average of $N_{av} = 3782$ values per frequency ($\equiv$ 6 minutes). This appreciable performance is primarily dictated by the integral defining $\kappa[|V_e|]$, as it converges slower than $\sigma[V_e]$, which itself converges slower than $\mathbb{E}[V_e]$.

## 4.1 Average $\mathbb{E}[V_e]$ and Standard Deviation $\sigma[V_e]$

First, the voltage $V_e(0)$ corresponding to the average configuration is compared to the average of the voltage $\mathbb{E}[V_e]$. In a perturbation-like approach, $V_e(0)$ would be considered as the average of $V_e$, and local expansions would be performed around $V_e(0)$ to represent the global variations of $V_e$ [5]. Figure 2 points out the clear differences between $|V_e(0)|$ and $|\mathbb{E}[V_e]|$, mainly concerning the position of their extrema. These discrepancies back the need to take the true variations of $S_\alpha$ into account when computing the statistics of $V_e$. The effect of the variations of $S_\alpha$ on $V_e$ is also indicated by the standard deviation which is depicted in Figure 2. At regular frequencies, $\sigma[V_e]$ is of the order of 30 mV, but increases by several orders of magnitude around the resonance frequencies. This plot reveals three resonance regions with increasing widths viz. $\mathscr{R}_1 \approx [175;215]$ MHz, $\mathscr{R}_2 \approx [295;350]$ MHz and $\mathscr{R}_3 \approx [415;480]$ MHz. The intensity of the resonances decreases with the frequency: The peaks of $\sigma[V_e]$ go from 16.120 V in $\mathscr{R}_1$, and 2.227 V in $\mathscr{R}_2$ to 0.666 V in $\mathscr{R}_3$.

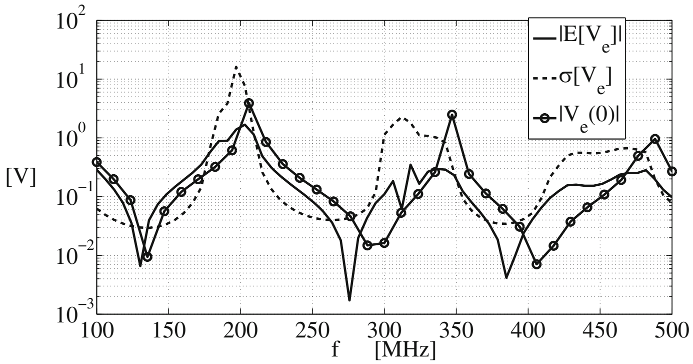


**Fig. 2:** $|V_e(0)|$ (*circled line*), $|\mathbb{E}[V_e]|$ (*dashed line*) and $\sigma[V_e]$ (*solid line*) vs frequency

High values of $\sigma[V_e]$ indicate a high physical variability of $V_e$ around its average $\mathbb{E}[V_e]$. However, the increased uncertainty of $V_e$ can be caused either by a smooth distribution of $V_e$ around $\mathbb{E}[V_e]$, or, by the presence of a few very large samples of $V_e$ coexisting with a cluster of samples around $\mathbb{E}[V_e]$. The distinction between these two cases is possible thanks to the analysis of $\kappa[|V_e|]$.

## 4.2 Kurtosis $\kappa[|V_e|]$

The kurtosis $\kappa[|V_e|]$ is displayed in Figure 3 together with the standard deviation $\sigma[V_e]$. Since $\kappa[|V_e|]$ is seldom equal to 3, the assumption of a Gaussian distribution of $V_e$ is generally inaccurate.

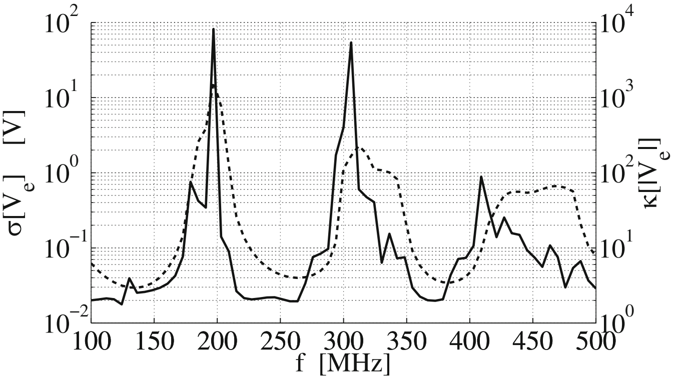


**Fig. 3:** $\sigma[V_e]$ (*dashed line*) and $\kappa[|V_e|]$ (*solid line*) vs frequency

The behavior of $\kappa[|V_e|]$ roughly follows that of $\sigma[V_e]$. Nevertheless, $\kappa[|V_e|]$ provides a finer characterization of $V_e$ than $\sigma[V_e]$ as it reveals the different types of sample distributions within a single resonance region. In $\mathscr{R}_2$ for instance, between 295 MHz and 320 MHz, $\sigma[V_e]$ rises from 40 mV to 2.277 V indicating an increase in the physical uncertainty of $V_e$. However, the variations of $\kappa[|V_e|]$ reveal that the effect of the extreme samples is mainly dominant at 306 MHz where $\kappa[|V_e|] = 5415$. Between 330 MHz and 350 MHz, in spite of a high value of $\sigma[V_e] \approx 1$V, $\kappa[|V_e|]$ drops below 15, thereby highlighting a smoother distribution of $V_e$ around $\mathbb{E}[V_e]$. A similar analysis can be conducted in $\mathscr{R}_1$ and $\mathscr{R}_3$.

### 4.3 Comparison with Deterministic Samples

To confirm the observations based on the analysis of Figure 3, $10^4$ deterministic samples have been computed at the frequencies specified in Tables 1 and 2. These samples are normalized as follows

$$V_n = \frac{V_e - \mathbb{E}[V_e]}{\sigma[V_e]}, \qquad \text{with } \mathbb{E}[V_n] = 0 \quad \text{and} \quad \sigma[V_n] = 1. \tag{6}$$

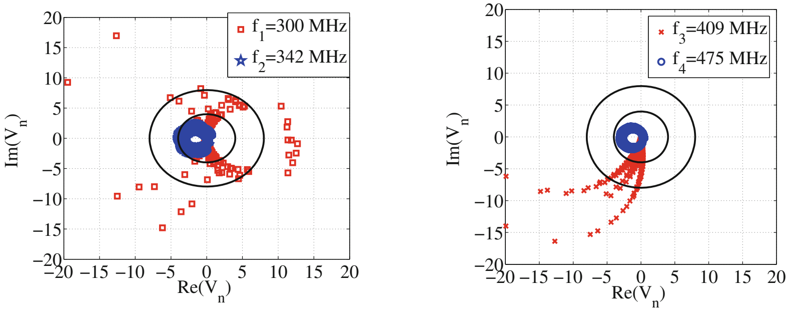
The statistical properties of the normalized samples can thus be compared on a common ground. In Figures 4a and 4b, concentric circles are shown, which correspond to the normalized samples with distances of $4\sigma[V_e]$ and $8\sigma[V_e]$ from $\mathbb{E}[V_e]$.

First, two frequencies $f_1$=300 MHz and $f_2$=342 MHz are considered in the resonance domain $\mathscr{R}_2$. As can be seen in Table 1, $\sigma[V_e]$ has comparable values at the two frequencies, with $\sigma[V_e]_{f_1} > \sigma[V_e]_{f_2}$. Nonetheless $\kappa[|V_e|]$ is two orders of magnitude larger at $f_1$ than at $f_2$.

**Table 1:** Statistical moments at given frequencies in $\mathscr{R}_2$

| | $\mathbb{E}[V_e]$ | $\sigma[V_e]$ | $\kappa[\lvert V_e\rvert]$ |
|---|---|---|---|
| $f_1 = 300$ MHz | 0.025 -j 0.107 V | 1.131 V | 402 |
| $f_2 = 342$ MHz | -0.239 +j 0.160 V | 0.822 V | 7 |

The normalized samples depicted in Figure 4a confirm that the samples $V_e$ are statistically more dispersed at $f_1$ than at $f_2$: at $f_1$ $V_e$ takes extreme values up to $20\sigma[V_e]$ away from $\mathbb{E}[V_e]$, whereas at $f_2$, all the samples are within $5\sigma[V_e]$ of $\mathbb{E}[V_e]$.



(a)

(b)

**Fig. 4:** Normalized samples $V_n$ for $f_1$ and $f_2$ (**a**) and for $f_3$ and $f_4$ (**b**)

Next, the resonance domain $\mathscr{R}_3$ is analyzed at the two frequencies $f_3 = 409$ MHz and $f_4 = 475$ MHz as detailed in Table 2. The standard deviation $\sigma[V_e]_{f_3}$ is more than 7 times smaller than $\sigma[V_e]_{f_4}$, thus the physical dispersion of $V_e$ is more intense at $f_4$. Conversely $\kappa[\lvert V_e\rvert]_{f_3}$ is approximately 30 times larger than $\kappa[\lvert V_e\rvert]_{f_4}$ which implies a much wider statistical spread at $f_3$. These predictions are confirmed in Figure 4b: all the samples corresponding to $f_4$ are clustered within $4\sigma[V_e]$ of $\mathbb{E}[V_e]$, unlike the samples at $f_3$, which can lie more than 20 $\sigma[V_e]$ away from the average. The non-negligible statistical uncertainty of $V_e$ at $f_3$, indicated by $\kappa[\lvert V_e\rvert]$, could not have been foreseen by the sole study of $\sigma[V_e]$.

**Table 2:** Statistical moments at given frequencies in $\mathscr{R}_3$

| | $\mathbb{E}[V_e]$ | $\sigma[V_e]$ | $\kappa[\lvert V_e\rvert]$ |
|---|---|---|---|
| $f_3 = 409$ MHz | - 0.004 -j 0.065 V | 0.093 V | 89 |
| $f_4 = 475$ MHz | - 0.232 +j 0.096 V | 0.632 V | 3 |

## 5 Conclusion

The results obtained for the varying thin-wire setup have revealed situations where, for high as well as low values of the standard deviation, a highly unsymmetrical distribution of the values around the average appears. Such cases are correctly signalled by high values of the kurtosis. Estimation of the probability of system failure conditions in such situations must therefore account for significant deviations from the Gaussian distribution. These statistical indicators can be determined numerically by quadrature rules such as a sparse-grid rule which outperforms a Monte-Carlo rule, for integrations over domains having moderate dimensions (below 10). A hierarchy has been observed in the computation of the statistical moments, as the average converges faster than the variance which, in turn, converges faster than the kurtosis. The analyses of the standard deviation and of the kurtosis are complementary: the variance is useful in a dimensioning process as it measures the physical variations of the voltage, whereas the kurtosis is valuable in a protection stage to foretell extreme values of the response parameter, which could damage the receiving device.

**Acknowledgements** This work is funded by the Dutch Ministry of Economic Affairs, in the Innovation Research Program (IOP) number EMVT 04302. The authors are also thankful to Dr. John Burkardt for his assistance in implementing the sparse-grid quadrature rule.

## References

1. Brown, G.S.: A stochastic Fourier transform approach to scattering from perfectly conducting randomly rough surfaces. IEEE Trans. Ant. Prop. **AP-30**(6), 1135–1144 (1982)
2. Hill, D.: Plane wave integral representation for fields in reverberation chambers. IEEE Trans. EMC. **40**(3), 209–217 (1998). DOI 10.1109/15.709418
3. Bellan, D., Pignari, S.: A probabilistic model for the response of an electrically short two-conductor transmission line driven by a random plane wave field. IEEE Trans. EMC. **43**(2), 130–139 (2001). DOI 10.1109/15.925532
4. Michielsen, B.L.: Probabilistic modelling of stochastic interactions between electromagnetic fields and systems. Comptes Rendus de l'Académie des sciences: Physique **7**, 543–559 (2006)
5. Sy, O., Vaessen, J., van Beurden, M., Tijhiuis, A., Michielsen, B.: Probabilistic study of the coupling between deterministic electromagnetic fields and a stochastic thin-wire over a pec plane. In: Proc. International Conference on Electromagnetics in Advanced Applications ICEAA 2007, pp. 637–640 (2007). DOI 10.1109/ICEAA.2007.4387382
6. De Roo, R., Misra, S., Ruf, C.: Sensitivity of the kurtosis statistic as a detector of pulsed sinusoidal rfi. IEEE Trans. on Geoscience and Remote Sensing **45**(7), 1938–1946 (2007). DOI 10.1109/TGRS.2006.888101
7. Michielsen, B.L.: A new approach to electromagnetic shielding. In: Proc. Int. Zürich EMC Symp. 1985, pp. 509–514 (1985)
8. Mrozynski, G., Schulz, V., Garbe, H.: A benchmark catalog for numerical field calculations with respect to emc problems. In: Proc. IEEE Int. EMC Symp., vol. 1, pp. 497–502 (1999)
9. Feller, W.: An introduction to probability theory and its applications. Wiley and sons (1971)
10. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numerica pp. 1–123 (2004)
11. Krommer, A.R., Ueberhuber, C.W.: Computational Integration. SIAM (1998)
12. Gerstner, T., Griebel, M.: Numerical integration using sparse grids. Numerical Algorithms **18 (24)**(3-4), 209–232 (1998)

# Part II
# Circuit Simulation

# Introduction to Part II

Janne Roos

This introduction gives some background to *circuit simulation* in general and provides a short overview of the 15 papers that follow.

Before proceeding further, it is pointed out that Section 1 of the first paper, the invited paper by Dautbegovic, gives a nice two-page introduction to circuit simulation. Thus, the reader may first want to read that section before returning to the text at hand, which completes the introduction to circuit simulation.

Most industrial analog circuit simulators support some or all of the following simulation modes, or analysis methods: DC operating point, transient, AC, distortion, noise, oscillator, single/multi-tone harmonic balance (HB), envelope-following transient analysis, etc. Although these analysis methods have been constantly developed during the last years, or even decades, there remain, still, many problems to be solved and extensions to be developed. In fact, it seems that this may be a never-ending story: ever more powerful analysis methods and computers — the latter being mainly built from integrated circuits (ICs) — will be needed to simulate the operation of ever more complex ICs.

A common way to deal with complexity is to model, simulate, and design the ICs in a top-down (and bottom-up) manner using several levels of abstraction like system, circuit, and device levels. In real-life industrial IC design flows, the seamless integration between different abstraction levels is challenging, both from the scientific (e.g., criteria for stable co-simulation) and the technical (e.g., interoperability between different tools and file formats) point of view. Limiting the discussion to the interface between the circuit and system levels, one way to proceed is to create measurement- or simulation-based behavioral models for entire circuit blocks, like power amplifiers, and to use these behavioral models for system-level simulations.

In Part II of this book, the first 10 papers mainly deal with the aforementioned circuit-level analysis methods, ranging from DC analysis to envelope-following

Janne Roos
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland, e-mail: janne.roos@tkk.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_17,

transient analysis. The remaining five papers deal with behavioral modeling of circuit blocks for efficient system-level simulation.

The invited paper by Dautbegovic discusses the potential exploitation of wavelets in circuit simulation. The key wavelet property is the capability of a simultaneous time and frequency representation of a signal. This is interesting, since one problem in circuit simulation is efficient time/frequency-domain representation of signals as well as the related signal sampling and transformation methods. Possible application areas could be, e.g., transient, HB, and envelope-following transient analysis, as well as mixed analog–digital simulation.

The contribution by Feldmann et al. proposes Schur-complement techniques for local handling of inner equations in compact models of semiconductor devices. The goal is to reduce the size of the model stamp in the resulting modified nodal analysis (MNA) equations. The approach was implemented in SPICE3 for DC, transient, AC, and noise analysis, and it was tested with ring-oscillator circuits.

The next three papers are all related to transient analysis.

The paper by Iwata et al. presents a hybrid analysis for nonlinear circuits leading to differential-algebraic equations (DAEs) with index at most one. This is desirable, since the common approach for using MNA to formulate the circuit equations may lead to DAEs with a higher index, thus causing problems for the numerical integration of DAEs.

Next, Christoffersen presents a new approach for transient analysis of nonlinear circuits. The circuit equations are formulated as functions of incident and reflected waves at the device ports. One interesting feature of the new approach is that all the required Newton iterations can be performed locally for each nonlinear device.

Transient noise analysis considers the noise effects that are due to certain random phenomena. Römish et al. discuss simultaneous step-size and path control for efficient transient noise analysis. Numerical experiments with a small industrial test circuit illustrate the practical relevance of the theoretical findings.

The paper by Rahkonen deals with distortion analysis, illustrating the use of a term-wise AC Volterra analysis tool that can plot the relevant distortion tones as vector sums of all important contributions. As an example, the tool is used to study the nonlinear distortion behavior in a fully differential amplifier when driven either with single-ended or balanced input signals.

The next three papers deal with large-signal steady-state analysis of nonlinear circuits and certain closely related topics.

The paper by Gourary et al. focuses on oscillator phase-noise analysis. General phase and frequency transfer functions are derived for frequencies that are close to the harmonics of the oscillator fundamental frequency. In the derivation, the equations of the linear(ized) time-varying approach in the context of the single-tone HB analysis are applied.

When a nonlinear circuit is excited by a periodic waveform and when some circuit parameters, like capacitor values, are described by random variables, the resulting DAEs have infinitely many periodic solutions. Pulch applies generalized polynomial chaos (gPC) to approximately resolve the stochastic model. In particular, failure probabilities are determined using the approximation from gPC.

In the contribution by Brachtendorf et al., the conventional trigonometric polynomials of single-tone or multi-tone HB are replaced by cubic or exponential splines in order to improve the approximation of sharp changes in the (quasi)periodical waveforms. It is shown that the amount of coding effort necessary on the top of an existing HB implementation is negligible.

The last paper on circuit-level analysis methods, the one by Xu and Condon, is related to envelope-following transient analysis. The so-called Devil's staircase of an injection-locked frequency divider (ILFD) is simulated using the multiple-phase-condition envelope-following method proposed. The locking range of the ILFD is determined from the Devil's staircase.

The last five papers of Part II discuss behavioral modeling of components and circuit blocks for the speeding up or enablement of higher-level simulation, optimization, statistical analysis, sensitivity analysis, etc.

The invited paper by Zhang and Zhang provides a tutorial overview on artificial neural networks (ANNs) and dynamic neural networks (DNNs) for behavioral modeling of RF/microwave components and circuits. The paper discusses ANN-based modeling, ANN-structure selection, ANN training, and the use of the trained ANN models in circuit simulation and design. Two illustrative application examples are given: automated model generation for embedded-passive modeling and use of DNNs for behavioral modeling of nonlinear circuits and systems.

The contribution by De Tommasi et al. presents a method for behavioral modeling of low-noise amplifiers based on transistor-level simulations. The surrogate modeling (SUMO) Matlab toolbox along with its adaptive sampling and modeling loops is utilized. Particular attention is paid to appropriate model-accuracy evaluation and modeling settings.

Li and Huang use a design-of-experiment setup, a 3-D field solver, and a second-order response-surface model (RSM) for behavioral modeling of the capacitances of a thin-film transistor liquid-crystal display (TFT-LCD). The RSM developed enables efficient sensitivity analysis and design optimization of the TFT-LCD capacitances with respect to the design parameters.

The work by Neitola and Rahkonen presents a data-based behavioral modeling scheme for the switched-capacitor (SC) integrator settling error. The method relies on the SC integrator transient simulation followed by the tabulation of the settling error. After this, the resulting settling-error table can be used in a delta-sigma A/D converter behavioral model as a lookup table.

The last paper by Rahkonen proposes speed-up techniques for time-domain system simulations using Matlab or Simulink behavioral models of sample-driven systems. First, the spectral effects of small time-skew errors are estimated. Then, state-space models of linear circuits are used to predict the circuit response, without the need for intermediate time steps.

# Wavelets in Circuit Simulation

Emira Dautbegovic*

*Invited speaker at the SCEE 2008 conference

**Abstract** Wavelet theory is a relatively recent area of scientific research, with a very successful application in a broad range of problems such as image, audio and signal processing, numerical analysis, electromagnetic scattering, data compression and denoising, stohastics, mathematics and physics, (bio)medicine, astronomy and many more. The key wavelet property contributing to its success in such a variety of disciplines is the capability of a simultaneous time and frequency representation of a signal embedded within a multi-resolution analysis (MRA) framework. The potential exploitation of this property for next-generation, wavelet-based techniques for analog circuit simulation is discussed in this paper.

## 1 Circuit Simulation

Analog circuit simulation is a standard industry approach to verify an integrated circuit (IC) design at the transistor level before committing it to the expensive manufacturing process. An Electronic Design Automation (EDA) suite takes the circuit description originating from a designer's draft or fabrication data files, and automatically generates a network description in form of a text file called netlist, which describes circuit elements (resistors, capacitors, transistors, voltage and current sources, etc.) and their connections. Then a circuit simulator (SPICE and its derivatives), an integral part of an EDA suite, parses this input and translates it to a data format reflecting the underlying mathematical model of the system. This is done by applying the basic physical laws (energy and charge conservation) onto network topology and taking the characteristic equations for the network elements into account. The most used "translation" approach is the charge/flux oriented modified nodal analysis (MNA) [1], which yields a mathematical model in the form of an initial-value problem of differential-algebraic equations (DAEs):

Emira Dautbegovic
Qimonda, 81726 Munich, Germany, e-mail: emira.dautbegovic@qimonda.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_18,

$$\mathbf{A}\frac{\mathrm{d}\mathbf{q}(\mathbf{x})}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}) = \mathbf{b}(t). \tag{1}$$

The matrix **A** is called an incidence matrix and, in general, is singular. **x** is the vector of node potentials and specific branch currents. **q** is the vector of charges and fluxes. **f** comprises static contributions, while **b** contains the contributions of independent sources. A numerical solution to (1) is found using the Newton's method in combination with implicit time integration schemes and sparse matrix techniques.

Instead of describing the system with a minimal set of unknowns, the mathematical modeling of an electric network via the charge/flux oriented MNA approach aims to preserve the topological structure of the network [1], thus enabling a physical interpretation of simulation results by a user. Next, this approach preserves information on charge/flux conservation, a crucial property of many analog circuits like charge pumps, switched capacitor filters, etc. Furthermore, the charge/flux formulation enables more realistic modeling of nonlinear capacitors and inductivities. In addition, (1) is suitable for the usage of special integrator schemes such as multistep methods (BDF-Gear, Trapezodial rule) and it does not require second partial derivatives of charges resp. fluxes, which are usually not available in standard circuit simulation packages and may not even exist due to the lack of smoothness in modern transistor models. On the other hand, in general (1) is a stiff system, i. e. it involves characteristic time constants that differ by several orders of magnitude, which is a serious hindrance to obtaining accurate results in a reasonable amount of CPU time. In addition, this representation suffers from poor smoothness properties of modern transistor models [2], which are struggling to describe complex physical processes with the smallest possible set of mathematical equations. Furthermore, if more general models for network elements are utilized or refined models are used to include second order and parasitic effects, an ill-conditioned problem may arise and very special care must be taken to avoid divergence while finding a numerical solution to (1).

Today modern industrial analog circuit simulators are facing two serious challenges: qualitative and quantitative [1,3,4]. The *qualitative challenge* is highlighted when simulating circuits containing mixed analog-digital parts. At present there is no standardized framework within which is possible to simulate efficiently a mixed analog-digital circuit. Analog circuits to be simulated are often multitone oscillatory circuits, with widely separated carrier and modulation tones. A high-frequency carrier forces a small timestep while a low-frequency modulation forces a long simulation interval, resulting in unacceptable long simulation times even for moderately-sized RF circuits. Under the assumptions that the circuit behavior is periodic or at most quasi-periodic and that its frequency spectrum contains only a small number of frequencies, the multitone oscillatory circuits may be efficiently simulated using a specialized RF simulator based on either the frequency-domain Harmonic Balance or the time-domain Shooting algorithm [5]. However, a digital subpart in the circuit introduces a substantial amount of high-frequency components and the efficiency of these specialized solvers diminishes, if they can be applied at all. Hence the current approach to an IC design is to simulate the analog RF front-end in a specialized RF simulator, while the rest of the circuit is designed employing standard circuit simulation techniques. Due to this separation during the design process, subparts of mixed analog-digital circuits are usually not realized on the same die in order to keep spurious couplings between them as small as possible, since they cannot be easily characterized in a common simulation environment. However, with the trend towards ever-decreasing chip size, integration of analog and digital circuit parts on the same die is eminent and new simulation tools that can support these mixed designs are urgently needed.

The *quantitative challenge* lies in the simulation of extremely large circuits featuring several millions transistors, e. g. memory chips. The sheer size of the underlying MNA representation of such large circuits yields simulations that can last weeks, even longer than a month. Or they simply cannot be performed due to extreme memory and computational requirements. To cope with this situation, designers are forced to aggressively simplify these very large circuits and simulate only the most critical parts, an approach which is error prone. Or they use so called fast-SPICE simulators, which utilize speed-up techniques such as table look-up models, circuit partitioning, event-driven algorithms, hierarchical and parallel computations, etc. In this manner a fast-SPICE simulator is able to achieve a speed up of factor 1000 in comparison to a standard circuit simulator but at the price of reduced accuracy (usually as high as 3–5%), a mismatch that sometimes leads to sub-optimal designs and failure of produced ICs, thus necessitating expensive re-design cycles.

## 2 Introduction to Wavelets

Wavelet theory emerged during the 20th century from the study of Calderon-Zygmund operators in mathematics, the study of the theory of subband coding in engineering and the study of renormalisation group theory in physics. The common foundation for the wavelet theory was laid down at the end of the 80’s and beginning of the 90’s by work of Daubechies [6, 7], Morlet and Grossman [8], Donoho [9], Coifman [10], Meyer [11], Mallat [12] and others. Today wavelet-based algorithms are already in productive use in a broad range of applications [11–18], such as image and signal compression (JPEG2000 standard, FBI fingerprints database), speech recognition), numerical analysis (solving operator equations, boundary value problems), stohastics, smoothing/denoising data, physics (molecular dynamics, geophysics, turbulence), medicine (heart-rate and ECG analysis, DNA analysis) to name just a few. Recent approaches [19–23] to the problem of multirate envelope simulation indicate that wavelets could also be used to address the qualitative challenge by a development of novel wavelet-based circuit simulation techniques capable of an efficient simulation of a mixed analog-digital circuit.

A wavelet is a waveform of finite duration, with zero average value. Its shape is usually irregular and asymmetric, unlike sines and cosines in Fourier series representation. Nevertheless, just like sines and cosines in the classical Fourier expansion, wavelets may be used as basis functions for a wavelet expansion to represent electrical signals. The wavelet basis is formed via translations and dilations of a single wavelet function $\psi(x)$, called *mother wavelet*, according to

$$\psi_{s,\tau}(x) = s^{-1/2}\,\psi\left(\frac{x-\tau}{s}\right), \tag{2}$$

where $(s,\tau) \in R^{+} \times R$. All wavelets from a specific basis are shifted (parameter $\tau$) and dilated/compressed (by factor $s$) versions of this mother wavelet. The translation parameter $\tau$ is responsible for the localization in time of a corresponding wavelet. The scaling or resolution parameter $s$, usually called the scale, is generally understood as the frequency inverse. Therefore, the high scale (resolution) corresponds to low frequencies or a global view of the signal and low scale (resolution) corresponds to high frequencies or a detailed view of the signal. The factor $s^{-1/2}$ is used for energy normalization across different scales. From (2) it is clear that a wavelet basis intrinsically supports a *simultaneous* time-frequency representation of a signal, where the translation parameter $\tau$ is responsible for the time localization and the scaling parameter $s$ for localization in the frequency domain. One particular wavelet property should be noted at this point: with wavelets it is not possible to exactly know a single frequency that exists at a single time instance, rather it is possible only to know what*frequency bands* exist at what *time intervals* [24].

There are numerous types of wavelets, each with different sets of features. Wavelets are usually grouped in wavelet families, according to several properties such as the support of wavelet and scaling functions, the number of vanishing moments, the symmetry, the regularity, existence of a scaling function $\phi$, the orthogonality and biorthogonality, existence of explicit expression and others [13]. Some of the most famous wavelets families include: Haar, Daubechies, spline, biorthogonal, Morlet, Mexican hat, symlet, coiflet, Meyer, Bessel, Cauchy, Gaussian, etc.

Transforms involving wavelets can roughly be divided into three classes: continuous (CWT), discretised (DWT) and multi-resolution based (MRA). Contrary to the name, DWT is a continuous-time transform, as is CWT. The discreteness here refers to the fact that discrete wavelets are not continuously scalable and translatable functions but can only be scaled and translated in discrete steps determined by some integers $(j,k)$. For example, a *discrete* wavelet suggested by Daubechies [7] is

$$\psi_{j,k}(x) = 2^{-j/2}\,\psi(2^{-j}x-k). \tag{3}$$

DWT in combination with MRA is a very efficient transform with its linear computational complexity $\mathcal{O}(N)$, it is even more efficient than the Fast Fourier Transform (FFT) with its $\mathcal{O}(N\log N)$ complexity. Against the background of the circuit simulation, MRA is of particular interest and it will be further explored in more details.

### 2.1 Multi-resolution Analysis

Formally defined, a *multi-resolution analysis (MRA)* in $L^2(R)$ is a set of closed subspaces $V_s$ with $s \in Z$ such that the following five properties are satisfied [25]

1. $\dots V_{-1} \subset V_0 \subset V_1 \subset \dots \subset L^2(R)$, that is $V_s \subset V_{s+1}$ for all $s \in Z$
2. $\bigcup_{s=-\infty}^{+\infty} V_s$ is dense in $L^2(R)$; and in addition $\bigcap_{s=-\infty}^{+\infty} V_s = \{0\}$
3. $f(t) \in V_s$ iff $f(2t) \in V_{s+1}$
4. if $f(t) \in V_0$, then $f(t-k) \in V_0$ for all $k \in Z$
5. $\exists$ scaling function $\phi(t) \in V_0$, so that set $\{\phi(t-k) \mid k \in Z\}$ is a Riesz basis of $V_0$

The first (structural) property states that subspaces $V_s$ in MRA are nested and the information at the resolution level $s$ is entirely included in the information at higher resolution level $s+1$. The second (resolution) property states that the $V_s$, $s \in Z$, cover $L^2(R)$, i. e. the approximation approaches any signal in the entire initial space $L^2(R)$ as more details are added, i. e. resolution goes to infinity. On the other hand, as more and more details are removed, i. e. resolution gets coarser, only constant functions are left. In a limit, only the zero function remains, since the functions are squarely integrable. The third (dilation) property states that all $V_s$ are scaled (dilated) versions of the central space $V_0$. The fourth (translation) property states that translation of $f(t)$ for some $k$ does not change its resolution, i. e. $V_0$ is integral translation-invariant. From the properties 3 and 4 it directly follows that if a function $f$ is in $V_0$, then its scaled and translated version $f(2^j t - k)$ is in $V_j$, i. e. if $f(t) \in V_0$, then $f(2^j t - k) \in V_j$ for all $k \in Z$. Finally, the fifth property states that similarly to the function $e^{j\omega t}$ in Fourier analysis, there exists one function $\phi(t)$ which generates the basis functions for all $V_s$. More precisely, if we define $\phi_{s,k} = 2^{s/2}\phi(2^s t - k)$, then $\{\phi_{s,k}(t)\}_{k \in Z}$ forms a Riesz basis of $V_s$.

To obtain the required resolution in a representation of an arbitrary signal, a sequence of scaling function expansions with wavelets of successively higher resolutions are used within the MRA. Interestingly, only *one* scaling function $\phi(t)$, called father wavelet, and *one* wavelet function $\psi(t)$, called mother wavelet, are needed to construct complete basis sets for systems of function spaces.

## *2.2 The Wavelet Expansion*

Let us now consider a wavelet expansion embedded in the MRA framework. We start by considering an electrical signal as a combination of a smooth background and fluctuations superimposed on it, as is done for electrical field representation [26]. At a given resolution level $s$ the signal is approximated in $V_s$ by ignoring all the fluctuations above this level in $V_k$ with $k > s$. Let $f_s(t) \in V_s$ denote the approximation of a signal $f(t)$ at given level $s$. In order to get better approximation, the level is increased to $s+1$ and a new approximation is obtained by adding the details, denoted as $d_s(t)$ to the approximation on previous level, i. e.

$$f_{s+1}(t) = f_s(t) + d_s(t). \tag{4}$$

Equation (4) means that at the resolution level $s+1$ a signal $f(t)$ is approximated with $f_s(t)$ in the scale subspace $V_s$ and $d_s(t)$ in the detail subspace $W_s$. The scale subspace $V_s$ consists of functions that contain the signal information down to scale $2^{-s}$. The members of the detail subspace $W_s = V_{s+1} \ominus V_s$ are differences $d_s(t) = f_{s+1}(t) - f_s(t)$ and it comprises the additional information regarding details on scales between $2^{-s}$ and $2^{-(s+1)}$. For best approximation in terms of $V_s$ the difference $d_s(t) = f_{s+1}(t) - f_s(t)$ should be orthogonal to $f_s(t)$. This is convenient to assume but not necessary. Assuming orthogonality means that $W_s \perp V_s$ and

$$V_{s+1} = W_s \oplus V_s = W_s \oplus W_{s-1} \oplus V_{s-1} = \ldots = \sum_{i=0}^{i=S} W_{s-i} \oplus V_{s-S}. \tag{5}$$

Furthermore, any two detail spaces at different resolutions are orthogonal, and the detail space $W_s$ is orthogonal to an approximation space $V_{s'}$, only when $s > s'$, i. e. when the detail space is at a higher resolution level.

If the improvement of approximation (4) was continued to infinity, the original signal $f(t)$ would be recovered as:

$$f(t) = f_s(t) + \sum_{j=s}^{\infty} d_j(t). \tag{6}$$

Hence an arbitrary electrical signal expanded as a summation of scaling and wavelet basis functions may be denoted in a hierarchical manner as:

$$f(t) = \sum_{i=-\infty}^{s} c_i \phi_i(t) + \sum_{j=s}^{+\infty} \sum_{k=-\infty}^{+\infty} d_{j,k} \psi_{j,k}(t). \tag{7}$$

The first term in (7) is the projection of $f(t)$ into the scaling subspace $V_s$. It corresponds to a coarse approximation of $f(t)$ at a previously selected resolution level $s$. The second term consists of projections of $f(t)$ into the wavelet subspaces $W_k$.

In practical computations only finite sums can be used and hence the sums in (7) must be truncated. In general, we are interested in the behavior of the circuit over a certain finite time interval of length $L$. This implies that the upper limit of a sum in the first term (index $i$) and the inner sum of the second term (index $k$) would naturally depend on the interval considered, i. e. the parameter $L$. The outer sum of the second term (index $j$) defines the number of levels of detail that are to be taken into account, and hence the resolution level of the approximation will be defined by the upper boundary of this sum. For example, a finite approximation of an electrical signal over the time interval $[0, L]$ on a $J^{th}$ resolution level could be denoted as:

$$f(t) \approx \sum_{i=0}^{2^s L-1} c_i \phi_i(t) + \sum_{j=s}^{(J-1)} \sum_{k=0}^{(2^j L-1)} d_{j,k} \psi_{j,k}(t) \tag{8}$$

At each resolution level $j$ there are $2^j L$ basis functions, thus there are in total $(2^{J-s})L$ wavelet coefficients to be computed. In addition, there are $2^s L$ coefficients corresponding to scaling functions at a resolution level $s$. Hence the total number of coefficients in a finite wavelet expansion (8) over the interval $[0, L]$ on a $J^{th}$ resolution level sums up to $2^J L$. For *efficient* computations the resolution level $s$ should be chosen so that the coarse level is satisfied for most values of $t$ and more details, i. e. wavelets, are added only at the points where they are needed to capture the abrupt signal fluctuations.

## 3 Wavelets in Circuit Simulation

Recent investigations into the use of wavelets in simulation of electronic circuits [19–23] have shown that these intrinsic properties make wavelets a natural candidate for a successful successor of time-domain (e. g. transient analysis, shooting analysis) and/or frequency domain (e. g. Harmonic Balance analysis) paradigms used in circuit simulation today. For example, Zhou and Cai propose the use of the wavelet collocation method in the time-domain [19] and the frequency domain [27] circuit simulation of mostly-linear circuits. For the computation of periodic steady state Soveiko and Nakhla [20,28] advocate a wavelet technique in combination with the Harmonic Balance approach, while Li et al. [29] use wavelet balance method. Christoffersen and Steer [21] used wavelets for transient circuit simulation within a state-variable based approach. Dautbegovic and Condon [22] use multitime partial differential equations (MPDE) in combination with wavelets for efficient simulation of multirate nonlinear RF circuits. Although valuable as a proof-of-concept, unfortunately these algorithms are still not mature enough to be used in industrial design flows.

We propose a wavelet expansion (8) embedded in the MRA framework as an approach to take when developing wavelet-based circuit simulation techniques. Consider the electrical signal depicted in Fig. 1, which is a typical time-domain output signal of a ring oscillator featuring a large amount of digital content. It can be considered as a "sum" of a digital signal and some irregular analog fluctuations. To describe such a signal efficiently, some sort of an adaptive approximation is needed. In such approximation an expansion of an electrical signal in those intervals where the signal varies smoothly and slowly should be simple and with as little degrees of freedom as possible, but whose resolution could be easily increased in places where the signal changes quickly and abruptly. For example, the smooth part could be represented by the low-resolution expansion of the signal, capturing the *average* signal behavior. A quickly changing part or *details* can only be captured by high-resolution components.

The wavelet expansion (8) is exactly the kind of the adaptive approximation that we are looking for. Embedded in the MRA framework, scaling functions can be used for an expansion of an electrical signal at a lower resolution level in those intervals where the signal varies smoothly and slowly, but in places where signal changes are quick and abrupt more details (i. e. wavelets) should be added. Therefore, the approximation effort is considerably reduced since only the "troublesome" regions are treated on a high-resolution level (i. e. with a larger number of coefficients), while smooth regions described on lower levels are captured by a smaller set of (possibly only) scaling coefficients. Compared to time-domain transient analysis,

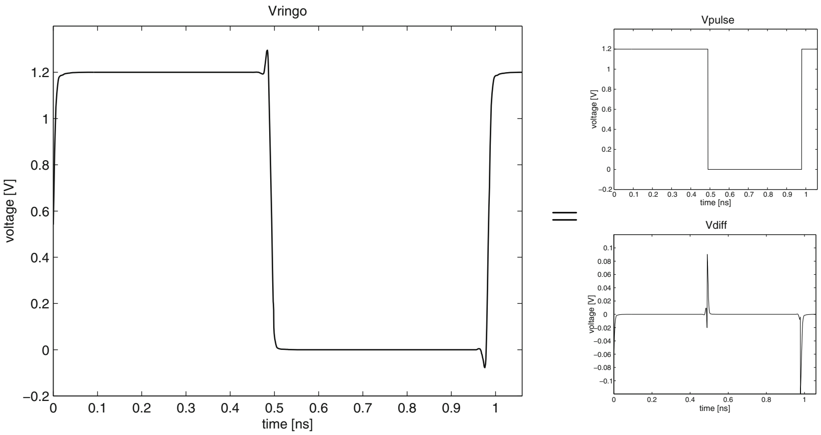


**Fig. 1:** An output voltage of a 1 GHz ring oscillator

taking fewer coefficients for the wavelet expansion in smooth regions is analogous to taking fewer time-steps during the transient analysis in intervals in which no large changes in signals are detected.

## 3.1 Advantages of the Wavelet-Based Approach in Circuit Simulation

Let us now explore particularly advantageous properties of the wavelet expansion against the target application of circuit simulation.

**Time-Frequency Representation.** The truncated wavelet expansion (8) may be written in general form as $f(t) = \sum_{I \in \mathscr{I}} \mathbf{a}_I(f)\ \Psi_I$, where $\Psi_I$ comprises all scaling and wavelet basis functions and $\mathbf{a}_I$ are the corresponding expansion coefficients on a finite index set $\mathscr{I} \leftrightarrow (j,k)$. In fact, these basis functions are generated by scaling (determined by the value of $j$) and translating (determined by the value of $k$) a single function $\psi$, i. e. $\psi_{j,k} = 2^{j/2}\ \psi(2^j t - k)$. Such an expansion associates with a function $f$, the array of coefficients $\mathbf{a} = \{\mathbf{a}_I(f)\}_{I \in \mathscr{I}}$ as is the case for the classical expansions. However, the coefficients $\mathbf{a}_I$ convey very detailed information on $f$ due to the structure of $\mathscr{I}$ [30]. Each $\mathscr{I}$ comprises two-fold information on time (spatial) location encoded by $k$ and information on scale, determined by $j$. Furthermore a scale is closely related to a frequency band and can be thought of as its inverse. Therefore, *each coefficient in a wavelet expansion (8) carries simultaneously both the time-domain and the frequency-domain information.*

**Adaptive Resolution.** In contrast to approximating the function $f$ of a given operator equation on some mesh (of fixed highest resolution), wavelet based schemes aim to determine its representation with respect to a basis [30]. This means that during the solution process, wavelet based algorithms will track only those coefficients in the unknown array **a** that are the most significant for approximating $f$ with as few as possible degrees of freedom. This property contributes immensely towards the efficiency of such algorithms.

In addition, an adaptive resolution equips a wavelet expansion with a natural way for an easy trade-off between required accuracy and reasonable simulation time. If the amplitude of a fast-changing fluctuation is below the noise-floor or the design process is in its early stages, when a designer is interested only in an average behavior of a designed IC, fluctuations above certain pre-defined cut-off level can be neglected. While a-priori definition of this cut-off level can be tricky with standard approaches, with wavelets it is a trivial task of setting the required resolution level $s$.

Furthermore, if the approximation is not satisfactory, we can continue with progressively increasing the resolution level, thus adding finer resolution details to the signal. Theoretically, by continuing this process to infinity resolution level, the signal will be exactly recovered just like for example in case of Taylor series expansion in the time domain or Fourier expansion in the frequency domain.

**Mixed Analog-Digital Simulation.** As briefly discussed in Section 1, at present there is no simulation framework (neither in the time nor in the frequency domain) in which a mixed analog-digital circuit can be efficiently simulated. The reason for this is a considerable approximation effort needed to capture a signal corresponding to one circuit part type when simulated in a simulator suitable for the other circuit type. For example, when a digital signal is to be simulated in a frequency-domain analog simulator, well suited for the analog RF front-end simulations, an extremely large number of Fourier coefficients is needed to accurately describe falling/rising edges of a digital signal. This is due to the poor time-domain localization property of the frequency-domain Fourier representation. In contrast, only a small number of coefficients corresponding to appropriately chosen scaling functions should be needed to approximate the signal well everywhere except in short intervals of sharp transitions. For those and only for those short intervals, additional coefficients corresponding to wavelet functions at higher resolution levels are needed to obtain equivalent or better accuracy to the Fourier representation, but at significantly reduced computation cost.

**Validity Range.** A Taylor expansion places strong demands on the regularity of $f$ such as analyticity, while wavelet expansion is typically valid for a much larger class of functions such as squarely integrable ones. This means that it is only required that the series on the right-hand side of (7) converges in the corresponding norm. Consequently the space of functions describing an electrical signal only needs to be a space of squarely integrable functions. Hence, a wavelet expansion has a potential to reduce negative influence of poor smoothness of transistor models on numerical convergence. However, this can only be confirmed after extensive testing on the existing industry models is performed within a working prototype of a wavelet-method.

### 3.2 Challenges of Wavelet-Based Algorithms

The foreseen advantages of the use of wavelet-based techniques in circuit simulation highlighted in Section 3.1 give us a solid justification for investing efforts for developing wavelet-based algorithms. However before an industry-wide exploitation of these techniques is possible, the following issues need to be addressed.

**Size of the Wavelet Expansion.** For a numerically effective wavelet method it is crucial to setup near-optimal wavelet expansions, so that only a small number of wavelet coefficients is needed for a signal representation. Unlike with the Fourier basis, in which the shape of a basis function is predefined and cannot be changed, wavelet basis functions can have many shapes, varying from smooth to highly irregular. A wavelet algorithm can be setup without having a priori knowledge on the type of the wavelet basis set to be used for signal representation. In fact, if a user has some previous insights about the expected results, drawn upon experience or on some prior simulation results, then a suitable wavelet set may be chosen prior to simulation start, as one of simulation parameters. For example, a smooth wavelet set could be chosen for ICs involving smoother functions and more irregular ones for digital-like signals. Matching a wavelet basis set to a signal shape to reduce the number of needed expansion coefficients is analogous to choosing the appropriate base frequency in the Fourier expansion to describe periodic signals with a minimum set of coefficients corresponding to the expected maximum harmonic in a signal's spectrum prior to the HB computations. In addition, an adaptive selection of expansion time points as well as both hard- and soft-thresholding techniques [31–34] can help to further decrease the size of a system to be solved.

**Numerical Considerations.** Even with a near-optimal selection of the wavelet basis the total number of wavelet coefficients is still very large; it equals the number of circuit variables times the number of coefficients in the chosen wavelet expansion for each node. For an efficient wavelet method a critical issue is how to store and invert a huge but relatively sparse Jacobian matrix arising from a Newton method applied to solve this nonlinear system. The investigations are ongoing into a setup of wavelet Jacobian in a block-diagonal form, which does not require storing the complete Jacobian at any point and is also easy to invert. Furthermore, one needs to be aware that significant matrix conditioning problems can arise due to a poor smoothness of MOSFET models (modeling problem) as well as solving higher-index DAEs (topological problem) and take appropriate care to minimize their negative influence on a solution process.

**Applicability and Functional Considerations.** The Harmonic Balance algorithm is an efficient tool for analyzing *periodic* or at most quasi-periodic circuits, unfortunately its use on any other type of circuits is a priori excluded. No such limitation is envisaged with wavelet based techniques and they are universal in the sense that they may be applied to any type of circuits. However, it is obvious that for pure sinusoidal signals there cannot exist a wavelet basis that is better than a Fourier basis in which a single expansion coefficient is needed to completely describe the signal.

But since the periodicity is not excluded from wavelet expansions, a wavelet basis can be found, such that it minimizes this expansion inefficiency and takes a small penalty when simulating pure sinusoidal circuits for the sake of generality.

Next, assuming that the previously mentioned challenges are successfully resolved and a wavelet solution is obtained, the question of the interpretation of these qualitatively new results arises. Wavelets are a powerful analysis tool but what can we conclude from a just performed wavelet analysis to enable a more robust design? An important point to enable faster adoption of wavelet based techniques in wider design community, governed by time- and frequency-domain specifications, is the derivation of a hopefully simple connection of wavelet-domain results to time- and frequency-domain design specifications.

## 4 Conclusion

With an ever-shrinking size and ever-increasing demand on functional complexity of a modern IC chip, a fast and scalable circuit simulation is a key design and verification approach in semiconductor industry. But increasing difficulties that current industrial circuit simulators are facing today, in particular in a simulation of mixed analog-digital circuit as well as circuits featuring millions of active devices, have highlighted the need for a novel approach to circuit simulation.

Intrinsic properties make wavelets a natural candidate for a successful successor of time- and frequency-domain paradigms used in circuit simulation today. This paper has discussed the advantages of wavelet expansions, which can be well utilized in circuit simulation, but also pointed out the challenges that must be resolved before an industry-wide acceptance and utilization of wavelet-based methods occurs. However, the expected benefits of a wavelet-based simulation engine, both in quantitative terms (efficient simulation of mixed-signal circuits) as well as qualitative terms (analyzing electrical signals with resolutions adapted to a problem at hands), is well worth allocating effort in a bid to develop the next-generation circuit simulators capable of answering industrial challenges of tomorrow.

**Acknowledgements** This work has been done within the ICESTARS (http://www.icestars.eu), a project supported by the EU Seventh Research Framework Project (FP7) with the grant number 214911.

## References

1. Günther, M., Feldmann, U., ter Maten, J.: Modelling and discretization of circuit problems. In: W. Schilders, E. ter Maten (eds.) Numerical Analysis in Electromagnetics, Spec. Vol. of Handbook of Numerical Analysis, vol. XIII, pp. 523–659. Elsevier, Amsterdam (2005)
2. Vladimirescu, A., Charlot, J.J.: Challenges of MOS analog circuit simulation with SPICE. Proc. IEE Colloquium on SPICE: Surviving Problems in Circuit Evaluation pp. 9/1–9/5 (1993)

3. Denk, G.: Circuit simulation for nanoelectronics. In: A. Anile, G. Al, G. Mascali (eds.) Scientific Computing in Electrical Engineering, vol. 9. Springer, Berlin Heidelberg (2006)
4. Kundert, K.: Why SPICE won't cut it for analog anymore. Online document (1999). URL http://www.designers-guide.org/Perspective/end-of-spice.pdf. Cited 10 Sep 2008
5. Kundert, K.: Simulation methods for RF integrated circuits. Proceedings of IEEE/ACM international conference on computer-aided design pp. 752–765 (1997)
6. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Communications on Pure and Applied Mathematics **41**, 909–996 (1988)
7. Daubechies, I.: Ten lectures on wavelets. SIAM (1992)
8. Grossmann, A., Morlet, J.: Decomposition of hardy functions into square integrable wavelets of constant shape. SIAM Journal of Mathematical Analysis **15**, 723–736 (1984)
9. Donoho, D.: Unconditional bases are optimal bases for data compression and for statistical estimation. Applied and computational harmonic analysis **1**, 100–115 (1993)
10. Coifman, R., Wickerhauser, M.: Entropy based algorithms for best basis selection. IEEE Trans. on Information Theory **38**, 713–718 (1992)
11. Meyer, Y.: Wavelets: Algorithms and Applications. SIAM (1993)
12. Mallat, S.: A wavelet tour of signal processing. Academic Press (1998)
13. Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.M.: Wavelet Toolbox 4. Mathworks (2008)
14. Young, R.: Wavelet theory and its applications. Kluwer Academic Publishers (1993)
15. Pan, G.: Wavelets in electromagnetics and device modelling. Wiley-Interscience (2003)
16. Vetterli, M., Kovacevic, J.: Wavelets and subband coding. Prentice Hall (1995)
17. Kaiser, G.: A friendly guide to wavelets. Birkhäuser (1994)
18. Le Maitre, O., Najm, H., Ghanem, R., Knio, O.: Multi-resolution analysis of wiener-type uncertainty propagation schemes. J. Comput. Phys. **197**, 502–531 (2004)
19. Zhou, D., Cai, W.: A fast wavelet collocation method for high-speed circuit simulation. IEEE Trans. Circuit and Systems **46**, 920–930 (1999)
20. Soveiko, N., Gad, E., Nakhla, M.: A wavelet-based approach for steady-state analysis of nonlinear circuits with widely separated time scales. IEEE Microwave and Wireless Components Letters **17**, 451–453 (2007)
21. Christoffersen, C., Steer, M.: State-variable-based transient circuit simulation using wavelets. IEEE Microwave and Wireless Components Letters **11**, 161–163 (2001)
22. Dautbegovic, E., Condon, M., Brennan, C.: An efficient nonlinear circuit simulation technique. IEEE Transactions on Microwave Theory and Techniques **53**, 548–555 (2005)
23. Bartel, A., Knorr, S., Pulch, R.: Wavelet-based adaptive grids for multirate partial differential-algebraic equations. submitted to Appl. Numer. Math. (2007)
24. Polikar, R.: The wavelet tutorial. Online document (2001). URL http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.htm. Cited 16 May 2004
25. Chen, Y., Cao, Q., Mittra, R.: Multiresolution time domain scheme for electromagnetic engineering. Wiley (2005)
26. Sarkar, T., Salazar-Palma, M., Wicks, M.: Wavelet applications in engineering electromagnetics. Artech House (2002)
27. Wang, J., Zeng, X., Cai, W., Chiang, C., Tong, J., Zhou, D.: Frequency domain wavelet method with GMRES for large-scale linear circuit simulations. Proc. ISCAS **5**, 321–324 (2004)
28. Soveiko, N., Nakhla, M.: Wavelet harmonic balance. IEEE Microwave and Wireless Components Letters **15**, 384–386 (2003)
29. Li, X., Hu, B., Ling, X., Zeng, X.: A wavelet balance approach for steady-state analysis of nonlinear circuits. Proc. ISCAS **3**, 73–76 (2001)
30. Dahmen, W.: Wavelet methods for PDEs—some recent developments. J. Comput. Appl. Math **128**, 133–185 (1999)
31. Jansen, M.: Noise reduction by wavelet thresholding. Springer (2001)
32. Vidakovic, B.: Nonlinear wavelet shrinkage with bayes rules and bayes factors. Journal of the American Statistical Association **93**, 173–179 (1998)
33. Donoho, D., Johnstone, I.: Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association **90**, 1200–1224 (1995)
34. Nason, G.: Wavelet shrinkage using cross-validation. J. R. Statist. Soc. B. **58**, 463–479 (1996)

# On Local Handling of Inner Equations in Compact Models

Uwe Feldmann, Masataka Miyake, Takahiro Kajiwara,
and Mitiko Miura-Mattausch

**Abstract** The burden of solving inner equations in compact models of semiconductor devices (such as transistors) is often shifted to the host circuit simulator. Schur complement techniques for local handling of these equations may help to reduce the size of the model stamp, which – depending on the host simulator – may have a positive impact on CPU time and memory needs. Some practical aspects of applying these concepts in compact modeling are discussed. A formulation is presented which accounts for the specific way of model evaluation in circuit simulation. It can be realized in a standard code for flat model evaluation by adding a software shell around the model core function itself.

First tests with an advanced high voltage MOS model demonstrate the feasibility of this approach in terms of accuracy, iterations and runtimes.

## 1 Introduction

Conventional modeling of semiconductor devices for circuit simulation aims at establishing explicit formulas for all relevant device features, which finally end up with the branch currents and node charges for describing device characteristics as a function of applied bias voltages. Unfortunately, this modeling paradigm is difficult to maintain for describing effects like nonquasistatic behavior or selfheating in high frequency or high voltage applications. State-of-the-art models therefore exhibit an increasing number of implicit model internal equations, either by introducing intrinsic circuit nodes, or by using auxiliary (nonlinear and/or differential) equations for more accurate description of device behavior. These equations are often exported to the host simulator. This is easy to implement, in general, but also may give rise to

Uwe Feldmann, Masataka Miyake, Takahiro Kajiwara, Mitiko Miura-Mattausch
HiSIM Research Center, Hiroshima-University, 1-3-1 Kagamiyama, Higashi-Hiroshima 739-0046, Japan, e-mail: uwe.feldmann@online.de, miyake-m053223@hiroshima-u.ac.jp, ancient-future@hiroshima-u.ac.jp, mmm@hiroshima-u.ac.jp

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_19,

a huge model stamp. So it may suffer from the overlinear complexity of the sparse solver in memory and CPU time due to increasing filling in the sparse pattern, unless sophisticated ordering strategies or hierarchical solver concepts are employed. Furthermore, robustness and efficiency (parallel processing!) are to a large extent simulator dependent.

Our objective is to implement compact models with local solver concepts, such that the device internal equations are as far as possible hidden from the host simulator. The benefits are obvious: Small model stamps, reduced simulator dependence, and possibly higher efficiency due to higher degree of locality and parallelism. Of course, there are also risks: Higher expense per model evaluation, convergence problems, and much higher efforts for model development. The first risk should be compensated by better efficiency of the sparse solver, which has to solve smaller systems. To avoid the second risk, we pursue only hierarchical versions of a single level Newton method, with a special focus to make sure that convergence is the same as if the equations were exported to the host simulator. And finally, the third risk can be reduced by the concept to perform the local handling in some kind of intermediate software layer between the model itself and the host simulator. The second and the third item distinguish this approach from previous attempts to apply local solver concepts for model evaluation [1].

The techniques presented here are not new at all, cf. [2,3]. Even fully hierarchical circuit simulators like PSTAR from NXP Semiconductors (former part of Philips) have been built on these principles [4,5], and are successfully applied in industrial practice. However, to our knowledge theses methods have not yet been used more globally in compact model development: All of the widespread bipolar [6–8] and MOS models [9–11] and even the standard MOS model PSP [12] offer only flat versions for download, or they employ local iteration loops in a multi level Newton setting for solving internal equations [11]. This contribution is focused on standard Newton methods. First, Schur complement techniques are shortly reviewed. Then some special aspects of their usage in SPICE like simulators will be discussed. A simple MOSFET model serves as an example. Finally, first results for an actual implementation of the high voltage MOS model HiSIM_HV [13] are presented.

## 2 Review of Schur Complement Techniques

We formulate the problem for the DC and (after time discretization) Transient Analysis: Solve the nonlinear coupled system

$$\begin{aligned} \mathbf{f}_i(\mathbf{x}_i, \mathbf{x}_m) &= \mathbf{0} \\ \mathbf{f}_m(\mathbf{x}_i, \mathbf{x}_m) &= \mathbf{0} \end{aligned} \tag{1}$$

where index $i$ denotes the model internal equations and variables, and $m$ denotes the outer equations from Modified Nodal Analysis and network variables[1]. Application

[1] For the ease of notation we consider only one single device and its contribution to the circuit.

of a single level Newton method yields the linear system for the Newton corrections $\triangle\mathbf{x}_i$, $\triangle\mathbf{x}_m$:

$$\begin{pmatrix} \frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_i} & \frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_m} \\ \frac{\partial\mathbf{f}_m}{\partial\mathbf{x}_i} & \frac{\partial\mathbf{f}_m}{\partial\mathbf{x}_m} \end{pmatrix} \begin{pmatrix} \triangle\mathbf{x}_i \\ \triangle\mathbf{x}_m \end{pmatrix} = - \begin{pmatrix} \mathbf{f}_i \\ \mathbf{f}_m \end{pmatrix} \tag{2}$$

Assuming regularity of $\frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_i}$, we can take the upper equation of (2) to express $\triangle\mathbf{x}_i$ by $\triangle\mathbf{x}_m$:

$$\triangle\mathbf{x}_i = -\left(\frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_i}\right)^{-1}\left(\mathbf{f}_i + \frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_m}\triangle\mathbf{x}_m\right) \tag{3}$$

Substitution of (3) in the lower equation of (2) yields:

$$\left[\frac{\partial\mathbf{f}_m}{\partial\mathbf{x}_m} - \frac{\partial\mathbf{f}_m}{\partial\mathbf{x}_i}\left(\frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_i}\right)^{-1}\frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_m}\right]\triangle\mathbf{x}_m = -\left[\mathbf{f}_m - \frac{\partial\mathbf{f}_m}{\partial\mathbf{x}_i}\left(\frac{\partial\mathbf{f}_i}{\partial\mathbf{x}_i}\right)^{-1}\mathbf{f}_i\right] \tag{4}$$

Compared to the flat case, (4) contains complementary entries both in the matrix and in the right-hand side, which account for the contributions of the inner equations to the outer system, and are referred to as Schur complements. Once a solution $\triangle\mathbf{x}_m$ for the outer system (4) has been calculated, the Newton correction for the inner system can be computed from (3).

Note that the method is in spite of the hierarchical formulation (3), (4) a standard Newton method. So convergence will not be affected, as long as the inner or outer variables are not subjected to different damping strategies or initial guess calculation.

## 3 Aspects of Implementation in Compact Modeling

Before applying Schur's techniques in compact modeling, we have to take some specialities of many SPICE like simulators into account:

- The linear system is established directly for $\mathbf{x}^{\text{new}} = \mathbf{x} + \triangle\mathbf{x}$ rather than for $\triangle\mathbf{x}$.
- Model evaluation is usually based on branch quantities $\mathbf{u}$ rather than on the mix of node voltages and branch currents which are considered in standard Modified Nodal Analysis: $\mathbf{f} = \mathbf{f}(\mathbf{u}(\mathbf{x}))$
- Often a branch oriented limiting $\mathbf{u} \overset{\text{limiter}}{\rightarrow} \tilde{\mathbf{u}} = \mathbf{u}^{\text{old}} + \lambda \cdot (\mathbf{u} - \mathbf{u}^{\text{old}})$ is employed, where $\mathbf{u}^{\text{old}}$ is the previous Newton iterate, $\lambda$ is a damping parameter, $0 < \lambda \leq 1$, and $\mathbf{f}(\mathbf{u})$ is approximated by $\mathbf{f}(\mathbf{u}) \approx \mathbf{f}(\tilde{\mathbf{u}}) + \left.\frac{\partial\mathbf{f}}{\partial\mathbf{u}}\right|_{\mathbf{u}=\tilde{\mathbf{u}}} (\mathbf{u} - \tilde{\mathbf{u}})$.
- Some kinds of analyses (Noise, Sensitivity) make use of the adjoint approach.

For the last item, the Schur concept has to be applied for the transposed system [14]. We skip the details here, and focus on the first three items. To this end we introduce a set of branch quantities

$$\mathbf{u}_i = \mathbf{B}_{ii}\mathbf{x}_i + \mathbf{B}_{im}\mathbf{x}_m \qquad \mathbf{u}_m = \mathbf{B}_{mi}\mathbf{x}_i + \mathbf{B}_{mm}\mathbf{x}_m \tag{5}$$

with constant incidence matrices $\mathbf{B}_{ii}$, $\mathbf{B}_{im}$, $\mathbf{B}_{mi}$, $\mathbf{B}_{mm}$, such that[2]

$$\mathbf{f}_i = \mathbf{f}_i(\mathbf{u}_i) \qquad \mathbf{f}_m = \mathbf{f}_m(\mathbf{u}_m).$$

Due to the special structure of (5), we get after some algebraic manipulations from (3), (4):

$$\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\mathbf{x}_i^{\text{new}} = -\underbrace{\left(\mathbf{f}_i(\tilde{\mathbf{u}}_i) - \frac{\partial \mathbf{f}_i}{\partial \mathbf{u}_i}\tilde{\mathbf{u}}_i\right)}_{\mathbf{f}_i^{\text{eq}}} - \frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_m}\mathbf{x}_m^{\text{new}} \tag{6}$$

$$\begin{aligned}\left(\frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_m} - \frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i}\underbrace{\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1}\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_m}}_{-\mathbf{P}}\right)\mathbf{x}_m^{\text{new}} &= \\ -\underbrace{\left(\mathbf{f}_m(\tilde{\mathbf{u}}_m) - \frac{\partial \mathbf{f}_m}{\partial \mathbf{u}_m}\tilde{\mathbf{u}}_m\right)}_{\mathbf{f}_m^{\text{eq}}} &+ \frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i}\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1}\mathbf{f}_i^{\text{eq}}\end{aligned} \tag{7}$$

To give an interpretation for $P$, we note that the first equation of (1) defines a relation $\mathbf{x}_i = \mathbf{x}_i(\mathbf{x}_m)$, whose derivative can be computed from implicit differentiation:

$$\frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_m} = -\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1}\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_m}$$

So, as long as (1) is not exactly solved, $\mathbf{P}$ can be considered as approximate for $\frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_m}$, and the matrix on the left side of (7) is an approximation for the admittance matrix of the model. The second term on the right-hand side of (7) propagates the defect of the inner model equations onto the outer terminals.

For the implementation we resolve (6) for $\mathbf{x}_i^{\text{new}}$, and introduce another intermediate quantity $\mathbf{x}_i^{\text{eq}} \overset{\text{def}}{=} -\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1}\mathbf{f}_i^{\text{eq}}$ to get finally

$$\mathbf{x}_i^{\text{new}} = \mathbf{x}_i^{\text{eq}} + \mathbf{P}\mathbf{x}_m^{\text{new}} \tag{8}$$

$$\left(\frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_m} + \frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i}\mathbf{P}\right)\mathbf{x}_m^{\text{new}} = -\mathbf{f}_m^{\text{eq}} - \frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i}\mathbf{x}_i^{\text{eq}} \tag{9}$$

Equation (9) defines the quantities from which the host simulator can compute a new Newton iterate $\mathbf{x}_m^{\text{new}}$. If $\mathbf{x}_i^{\text{eq}}$ and $\mathbf{P}$ are stored over the Newton iterations then the internal variables $\mathbf{x}_i^{\text{new}}$ can be updated using (8), just before evaluating the model for the subsequent Newton step. Algorithm 1 describes the sequence of computational steps for “stamping” the model contributions into matrix and right-hand side. The Schur related parts are highlighted, while all other parts are more or less standard in any SPICE like simulator.

[2] Note that (5) includes the case of current controlled device characteristics.

**Algorithm 1** Loading the model contributions into the outer system

```
Gather $\mathbf{x}_m$ and $\mathbf{x}_i^{\mathrm{eq}}$, $\mathbf{P}$ from host simulator
Calculate $\mathbf{x}_i$ from $\mathbf{x}_i = \mathbf{x}_i^{\mathrm{eq}} + \mathbf{P}\mathbf{x}_m$
Apply branch limiting to get $\tilde{\mathbf{u}}$
Call model evaluation routine to get currents and charges, inclusive derivatives with respect to $\mathbf{u}$
Calculate numerical approximates for time derivatives of charges
Assemble $\mathbf{f}_m^{\mathrm{eq}} = \mathbf{f}_m - \frac{\partial \mathbf{f}_m}{\partial \mathbf{u}}\tilde{\mathbf{u}}$ and $\mathbf{f}_i^{\mathrm{eq}} = \mathbf{f}_i - \frac{\partial \mathbf{f}_i}{\partial \mathbf{u}}\tilde{\mathbf{u}}$
Assemble $\frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_m}$ and $\frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i}$, $\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_m}$, $\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}$ and calculate $\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1}$
Calculate and store $\mathbf{x}_i^{\mathrm{eq}} = -\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1} \mathbf{f}_i^{\mathrm{eq}}$ and $\mathbf{P} = -\left(\frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_i}\right)^{-1} \frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_m}$
Setup and add right-hand side $(-\mathbf{f}_m^{\mathrm{eq}} - \frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i} x_i^{\mathrm{eq}})$ and matrix stamp $(\frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_m} + \frac{\partial \mathbf{f}_m}{\partial \mathbf{x}_i}\mathbf{P})$ to the outer system
```

It is worthwhile to note that all additional steps for realizing the Schur concept can be put into a compact software shell around the model evaluation; hence no major interaction with the host simulator is necessary beyond the standard, except from the task to store additional data across Newton iterations.

## 4 Example: A Simple MOS Model

If we want to handle the inner drain and source node voltages of the simple MOSFET model in Fig. 1 locally then

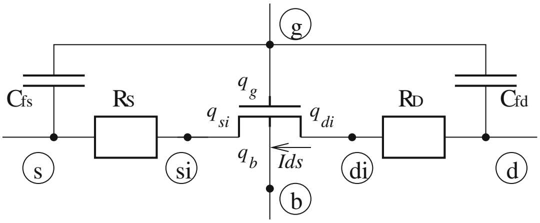


**Fig. 1:** MOSFET model with two internal nodes and outer fringing capacitances

$$\mathbf{x}_i = \begin{pmatrix} v_{\mathrm{di}} \\ v_{\mathrm{si}} \end{pmatrix} \qquad \mathbf{f}_i = \begin{pmatrix} 1/R_{\mathrm{D}}(v_{\mathrm{di}} - v_{\mathrm{d}}) + I_{\mathrm{ds}} + \dot{q}_{\mathrm{di}} \\ 1/R_{\mathrm{S}}(v_{\mathrm{si}} - v_{\mathrm{s}}) - I_{\mathrm{ds}} + \dot{q}_{\mathrm{si}} \end{pmatrix}$$

$$\mathbf{x}_m = \begin{pmatrix} v_{\mathrm{d}} \\ v_{\mathrm{g}} \\ v_{\mathrm{s}} \\ v_{\mathrm{b}} \end{pmatrix} \qquad \mathbf{f}_m = \begin{pmatrix} 1/R_{\mathrm{D}}(v_{\mathrm{d}} - v_{\mathrm{di}}) + \dot{q}_{\mathrm{fd}} \\ \dot{q}_{\mathrm{g}} - \dot{q}_{\mathrm{fd}} - \dot{q}_{\mathrm{fs}} \\ 1/R_{\mathrm{S}}(v_{\mathrm{s}} - v_{\mathrm{si}}) + \dot{q}_{\mathrm{fs}} \\ \dot{q}_{\mathrm{b}} \end{pmatrix}$$

where

$$\begin{aligned} I_{\mathrm{ds}} &= I_{\mathrm{ds}}(v_{\mathrm{di}} - v_{\mathrm{si}}, v_{\mathrm{g}} - v_{\mathrm{si}}, v_{\mathrm{b}} - v_{\mathrm{si}}) \\ q_k &= q_k(v_{\mathrm{di}} - v_{\mathrm{si}}, v_{\mathrm{g}} - v_{\mathrm{si}}, v_{\mathrm{b}} - v_{\mathrm{si}}) \quad k = \{\mathrm{di}, \mathrm{g}, \mathrm{si}, \mathrm{b}\} \end{aligned}$$

$$q_{\text{fd}} = C_{\text{fd}} \cdot (v_{\text{d}} - v_{\text{g}})$$
$$q_{\text{fs}} = C_{\text{fs}} \cdot (v_{\text{s}} - v_{\text{g}})$$

Alternatively, a formulation can be used which is even applicable for vanishing $R_{\text{D}}$ and/or $R_{\text{S}}$:

$$\mathbf{f}_i = \begin{pmatrix} v_{\text{di}} - v_{\text{d}} + R_{\text{D}} \cdot (I_{\text{ds}} + \dot{q}_{\text{di}}) \\ v_{\text{si}} - v_{\text{s}} + R_{\text{S}} \cdot (-I_{\text{ds}} + \dot{q}_{\text{si}}) \end{pmatrix} \qquad \mathbf{f}_m = \begin{pmatrix} I_{\text{ds}} + \dot{q}_{\text{di}} + \dot{q}_{\text{fd}} \\ \dot{q}_{\text{g}} - \dot{q}_{\text{fd}} - \dot{q}_{\text{fs}} \\ -I_{\text{ds}} + \dot{q}_{\text{si}} + \dot{q}_{\text{fs}} \\ \dot{q}_{\text{b}} \end{pmatrix}$$

This example shows that the inner equations are not restricted to Kirchhoff's current law. This gives valuable freedom for model development and is a clear advantage of hierarchical concepts in modeling.

## 5 Current State and First Results

For the high voltage MOS model HiSIM_HV [13] from Hiroshima University a test implementation using the Schur concept was done in parallel with a flat implementation. If all model features (selfheating, nonquasistatic behavior NQS, parasitic gate and bulk resistance network) are activated then the model has up to 10 internal equations; Table 1 shows the matrix stamp for the terminal node voltages $(d, g, s, b)$ and the internal variables $(d_i, g_i, s_i, b_i, d_b, s_b, t)$; the stamp for the NQS equations is omitted. For the Schur implementation not all of the inner equations are included yet; however, this is planned for the future. In the maximal case a reduction of the matrix stamp from 86 entries down to 16 entries can be expected.

The implementation was done in SPICE3 for DC/Transient, AC and Noise Analysis and tested with several HV_MOS ring oscillators of different size. Unfortunately, large HV_MOS circuits were not yet available for testing. Each circuit was simulated without inner series resistances $R_{\text{D}}$, $R_{\text{S}}$, as well as with $R_{\text{D}} \neq 0$, $R_{\text{S}} = 0$ and with $R_{\text{D}} \neq 0$, $R_{\text{S}} \neq 0$. In the third case the flat stamp into the matrix consists of 32 entries, while the Schur version yields a stamp of 16 entries. Some results of a comparison with a flat implementation using the same model evaluation function are given in Table 2.

The second column contains the number of MOSFETs, and the third column shows the number of circuit equations seen by SPICE. Of course, this number is not affected by activation of series resistances if the internal equations are handled locally. The waveforms match perfectly in any case, however the total number of Newton iterations shown in the fourth column are slightly different. This is mostly due to nonidentical branch limiting procedures and tolerance settings in our implementations. Even if the codes are completely synchronized, there might be small differences due to differing pivot handling of the sparse solver. In accordance with our expectations, there is however no significant impact of the Schur method on the global convergence, nor on the number of timesteps.

**Table 1:** Reducing the matrix stamp for the HiSIM_HV model by using the Schur concept

$$\begin{array}{c|ccccccccccc}
 & d & d_i & g & g_i & s & s_i & b & b_i & d_b & s_b & t \\
\hline
d & x & x & & x & x & & & x & x & & x \\
d_i & x & x & & x & x & x & & x & & & x \\
g & & & x & x & & & & & & & \\
g_i & x & x & x & x & x & x & & x & & & x \\
s & x & & & x & x & x & & x & & x & x \\
s_i & x & x & & x & x & x & & x & & & x \\
b & & & & & & & x & x & & & \\
b_i & & x & & x & & x & x & x & x & x & x \\
d_b & x & & & & & & & x & x & & x \\
s_b & & & & & x & & & x & & x & x \\
t & x & x & & x & x & x & & x & & & x
\end{array}
\quad \overset{\text{Schur concept}}{\Longrightarrow} \quad
\begin{array}{c|cccc}
 & d & g & s & b \\
\hline
d & x & x & x & x \\
g & x & x & x & x \\
s & x & x & x & x \\
b & x & x & x & x
\end{array}$$

63 matrix entries + 23 matrix entries for NQS $\overset{\text{Schur concept}}{\Longrightarrow}$ 16 matrix entries

**Table 2:** SPICE results for some HV_MOS ring oscillators; left/right: flat/Schur version

| Circuit | MOS FETs | SPICE equations | Newton iterations | LOAD/iter [msec] | SOLVE/iter [$\mu$sec] |
|---|---|---|---|---|---|
| ringo51 $R_D, R_S = 0$ | 114 | 63/63 | 23954/23886 | 1.1/1.0 | 4/5 |
| $R_D \neq 0$ | | 177/63 | 25243/24570 | 1.2/1.2 | 17/3 |
| $R_D, R_S \neq 0$ | | 291/63 | 21799/21130 | 1.2/1.1 | 33/4 |
| ringo101 $R_D, R_S = 0$ | 214 | 113/113 | 25736/25762 | 2.1/2.0 | 9/9 |
| $R_D \neq 0$ | | 327/113 | 27127/26475 | 2.3/2.2 | 30/9 |
| $R_D, R_S \neq 0$ | | 541/113 | 23588/22902 | 2.3/2.2 | 55/7 |
| ringo101_2 $R_D, R_S = 0$ | 414 | 213/213 | 27999/29245 | 4.3/3.8 | 19/16 |
| $R_D \neq 0$ | | 627/213 | 28620/29034 | 4.7/4.1 | 62/19 |
| $R_D, R_S \neq 0$ | | 1041/213 | 26123/25929 | 4.8/4.1 | 122/15 |

The CPU time data in columns five and six of Table 2 are given per Newton iteration. LOAD comprises all steps of Algorithm 1 – inclusive solving the inner equations in case of the Schur version – while SOLVE accounts for the sparse solver time of SPICE. The CPU times are a bit noisy due to inaccurate timing measurement in SPICE, and lack of tuning the codes. In particular the LOAD time for ringo101_2 with $R_D, R_S = 0$ gives rise to suspect that the flat code is less tuned than the Schur code. One can however conclude that there is no significant slowdown of the Schur code, as long as the number of internal equations per device is small. Finally the overlinear increase of the SOLVE time in SPICE with the number of equations can be seen from the last column in Table 2. Although the SOLVE time is negligible here, it gets more important for large circuits, and may finally dominate total CPU time. This applies to both the flat and the Schur version, but its impact is more severe for the flat handling due to the much larger number of SPICE equations.

## 6 Summary

To be applicable for compact modeling, Schur complement methods for local solving of device internal equations must be adapted to the standardized techniques of model evaluation in circuit simulators. A formulation is given, which fits well into existing flat model evaluation procedures, and can be realized in a software shell around the model evaluation function itself. A test implementation of the HiSIM_HV MOS model was developed in SPICE, which can be compared with a flat standard implementation exporting all device internal equations to SPICE. First tests confirm that the Schur complement techniques can significantly reduce the size of the model stamps at no penalty with respect to convergence and accuracy, and at very moderate penalty with respect to the LOAD time, as long as the number of internal equations per device is small. It depends on the host simulator if this translates into a real benefit for the practical use of a model: Architecture of its model interface, the ordering and pivot selection mechanisms of its sparse solver, its degree of parallelization, etc. may have an impact. To enable more comprehensive tests, a Schur based version of the HiSIM_HV model will be put for download on the HiSIM website [11], in parallel to the officially released flat version.

## References

1. Braack, M., Feldmann, U., Wever, U.: About local iteration for calculating transistor characteristics in circuit simulation. In: Proc. ITG Diskussionsitzung ANALOG'94, 165–170, Bremen (1994).
2. Rabbat, N.B.G., Sangiovanni-Vincentelli, A.L., Hsieh, H.Y.: A multilevel Newton algorithm with macromodeling and latency for the analysis of large-scale nonlinear circuits in the time domain. IEEE Trans. Circ. Syst. CAS, **26**, 733–741 (1979).
3. Günther, M., Feldmann, U., ter Maten, E.J.W.: Modelling and discretization of circuit problems. In: Schilders, W.H.A., ter Maten, E.J.W. (eds.) Handbook Numerical Analysis vol. XIII, pp. 523–659, Elsevier Science, Amsterdam, (2005).
4. Wehrhahn, E.: Hierarchical circuit analysis. In: Proc. IEEE Int. Sym. Circ. Syst., ISCAS 89, vol. 1, pp. 701–704. Portland, May (1989).
5. Fijnvandraat, J.G., Houben, S.H.M.J., ter Maten, E.J.W., Peters, J.M.F.: Time domain analog circuit simulation. J. of Comp. and Appl. Math., **185**, 441–459 (2006),
6. HICUM homepage. URL http://www.iee.et.tu-dresden.de/iee/eb/hic_new/hic_intro.html. Cited Dec. 2008.
7. MEXTRAM homepage. URL http://www.nxp.com/models/bi_models/mextram/index.html. Cited Dec. 2008.
8. VBIC homepage. URL http://www.designers-guide.org/VBIC/index.html. Cited Dec. 2008.
9. BSIM4 homepage. URL http://www-device.eecs.berkeley.edu/~bsim3/bsim4_intro.html. Cited Dec. 2008.
10. EKV homepage. URL http://legwww.epfl.ch/ekv/. Cited Dec. 2008.
11. HiSIM homepage. URL http://home.hiroshima-u.ac.jp/usdl/HiSIM.html. Cited Dec. 2008.
12. SiMKit homepage. URL http://www.nxp.com/models/source/. Cited Dec. 2008.
13. M. Yokomichi, M. et al.: Laterally diffused metal oxide semiconductor model for device and circuit optimization. Jpn. J. Appl. Phys., **47**, 2560–2563 (2008).
14. Wehrhahn, E.: Hierarchical sensitivity analysis of circuits. In: Proc. IEEE Int. Sym. Circ. Syst., ISCAS 91, vol. 2, pp. 864–867. Singapore, May (1991).

# Hybrid Analysis of Nonlinear Time-Varying Circuits Providing DAEs with Index at Most One

Satoru Iwata, Mizuyo Takamatsu, and Caren Tischendorf

**Abstract** Commercial packages for transient circuit simulation are often based on the modified nodal analysis (MNA) which allows an automatic setup of model equations and requires a nearly minimal number of variables. However, it may lead to differential-algebraic equations (DAEs) with higher index. Here, we present a hybrid analysis for nonlinear time-varying circuits leading to DAEs with index at most one. This hybrid analysis is based merely on the network topology, which possibly leads to an automatic setup of the hybrid equations from netlists. Moreover, we prove that the minimum index of the DAE arising from the hybrid analysis never exceeds the index from MNA. As a positive side effect, the number of equations from the hybrid analysis is always no greater than that one from MNA. This suggests that the hybrid analysis is superior to MNA in numerical accuracy and computational effort.

## 1 Introduction

When modelling electric circuits for transient simulation, one has to regard Kirchhoff's laws for the network and the constitutive equations for the different types of network elements. They are originally based on the branch voltages and the branch

Satoru Iwata
Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan, e-mail: iwata@kurims.kyoto-u.ac.jp

Mizuyo Takamatsu
Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan, e-mail: mizuyo_takamatsu@mist.i.u-tokyo.ac.jp

Caren Tischendorf
Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Cologne, Germany, e-mail: tischendorf@math.uni-koeln.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_20,

currents existing in the network. They form the basis for all modelling approaches as for instance the popular modified nodal analysis (MNA).

Concerning the huge number of variables involved (all branch voltages and branch currents), one is interested in a reduced system reflecting the complete circuit behaviour that can be generated automatically. Whereas MNA focuses on a description depending mainly on nodal potentials, the hybrid analysis approach [1] here employs certain branch voltages and branch currents obtained from a construction of a particular *normal tree*.

A normal tree is a tree containing all independent voltage sources, no independent current sources, a maximal number of capacitive branches, and a minimal number of inductive branches. Normal trees have already been used in [2] for state approaches for linear RLC networks. The results have been extended in [3] for linear circuits containing ideal transformers, nullors, independent/dependent sources, resistors, inductors, capacitors, and, under a topological restriction, gyrators.

The hybrid analysis is a common generalization of the loop analysis and the cutset analysis. Kron [4] proposed the hybrid analysis in 1939, and Amari [5] and Branin [6] developed it further in 1960s. In contrast to MNA, the hybrid analysis retains flexibility in the selection of a normal tree, which can be exploited to find a model description that reduces the numerical difficulties.

The differential-algebraic equations (DAEs) arising from the hybrid analysis are called the *hybrid equations*. Recently, the analysis of the *index* of the hybrid equations has been developed. For linear time-invariant RLC circuits, it is shown in [7] that the index of the hybrid equations never exceeds one, while MNA often results in a DAE with index two. Moreover, [7] gives a structural characterization of circuits with index zero. For linear time-invariant electric circuits which may contain dependent voltage/current sources, an algorithm for finding an optimal hybrid analysis which minimizes the index of the hybrid equations was proposed in [8].

For nonlinear time-varying circuits, this paper shows that the index of the hybrid equations is at most one, and gives a structural characterization for the index being zero, which is an extension of the results in [7]. By this structural characterization, we prove that the minimum index of the hybrid equations does not exceed the index of the DAE arising from MNA (cf. [9–11]). Here, we follow the hybrid analysis approach in [8] but use projection techniques (cf. [10]) in order to prove the index results for general nonlinear time-varying circuit systems.

The organization of this paper is as follows. In Section 2, we describe nonlinear time-varying circuits. We present the procedure of the hybrid analysis in Section 3. We analyze the hybrid equation system in Section 4, and characterize its index in Section 5. All the technical proofs omitted in this paper can be found in [12].

## 2 Nonlinear Time-Varying Circuits

Here, we consider nonlinear time-varying circuits composed of resistors, conductors, inductors, capacitors, and voltage/current sources.

We denote the vector of branch currents by $\mathbf{i}$, and the vector of branch voltages by $\mathbf{u}$. The vector of currents through independent voltage sources, independent current sources, capacitors, inductors, resistors, conductors, controlled current sources, and controlled voltage sources are denoted by $\mathbf{i}_V$, $\mathbf{i}_J$, $\mathbf{i}_C$, $\mathbf{i}_L$, $\mathbf{i}_R$, $\mathbf{i}_G$, $\mathbf{i}_{S_J}$, and $\mathbf{i}_{S_V}$. Similarly, the vector of voltages are denoted by $\mathbf{u}_V$, $\mathbf{u}_J$, $\mathbf{u}_C$, $\mathbf{u}_L$, $\mathbf{u}_R$, $\mathbf{u}_G$, $\mathbf{u}_{S_J}$, and $\mathbf{u}_{S_V}$. The physical characteristics of elements determine *constitutive equations*. Independent voltage and current sources simply read as

$$\mathbf{u}_V = \mathbf{v}_s(t) \quad \text{and} \quad \mathbf{i}_J = \mathbf{j}_s(t). \tag{1}$$

Capacitors and inductors can be modelled by

$$\mathbf{i}_C = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}(\mathbf{u}_C,t) \quad \text{and} \quad \mathbf{u}_L = \frac{\mathrm{d}}{\mathrm{d}t}\phi(\mathbf{i}_L,t). \tag{2}$$

Moreover, we assume that conductors and resistors are described by $\mathbf{i}_G = \mathbf{g}(\mathbf{u}_G,t)$ and $\mathbf{u}_R = \mathbf{r}(\mathbf{i}_R,t)$. Finally, let the controlled sources be given in the form of $\mathbf{i}_{S_J} = \gamma(\mathbf{i}_{S_V},\mathbf{u}_{S_J},t)$ and $\mathbf{u}_{S_V} = \rho(\mathbf{i}_{S_V},\mathbf{u}_{S_J},t)$.

A square matrix $U$ is called *positive definite* if $\mathbf{x}^\top U\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$. In this paper, we assume the following conditions.

**Assumption 1** *The capacitance matrix C, the conductance matrix G, the resistance matrix R, the inductance matrix L, and the controlled source matrix S given by*

$$C = \frac{\partial \mathbf{q}}{\partial \mathbf{u}_C}, \quad G = \frac{\partial \mathbf{g}}{\partial \mathbf{u}_G}, \quad R = \frac{\partial \mathbf{r}}{\partial \mathbf{i}_R}, \quad L = \frac{\partial \phi}{\partial \mathbf{i}_L}, \quad \text{and} \quad S = \begin{pmatrix} \frac{\partial \rho}{\partial \mathbf{i}_{S_V}} & \frac{\partial \rho}{\partial \mathbf{u}_{S_J}} \\ \frac{\partial \gamma}{\partial \mathbf{i}_{S_V}} & \frac{\partial \gamma}{\partial \mathbf{u}_{S_J}} \end{pmatrix}$$

*are all positive definite.*[1]

Introducing $\mathbf{u}_Y := \begin{pmatrix} \mathbf{u}_G \\ \mathbf{u}_{S_J} \end{pmatrix}$, $\mathbf{u}_Z := \begin{pmatrix} \mathbf{u}_R \\ \mathbf{u}_{S_V} \end{pmatrix}$, $\mathbf{i}_Y := \begin{pmatrix} \mathbf{i}_G \\ \mathbf{i}_{S_J} \end{pmatrix}$, $\mathbf{i}_Z := \begin{pmatrix} \mathbf{i}_R \\ \mathbf{i}_{S_V} \end{pmatrix}$, $\mathbf{f}(\mathbf{i}_Z,\mathbf{u}_Y,t) := \begin{pmatrix} \mathbf{g}(\mathbf{u}_G,t) \\ \gamma(\mathbf{i}_{S_V},\mathbf{u}_{S_J},t) \end{pmatrix}$, and $\mathbf{h}(\mathbf{i}_Z,\mathbf{u}_Y,t) := \begin{pmatrix} \mathbf{r}(\mathbf{i}_R,t) \\ \rho(\mathbf{i}_{S_V},\mathbf{u}_{S_J},t) \end{pmatrix}$, we find

$$\mathbf{i}_Y = \mathbf{f}(\mathbf{i}_Z,\mathbf{u}_Y,t), \quad \mathbf{u}_Z = \mathbf{h}(\mathbf{i}_Z,\mathbf{u}_Y,t) \tag{3}$$

and the matrix $\begin{pmatrix} \frac{\partial \mathbf{h}}{\partial \mathbf{i}_Z} & \frac{\partial \mathbf{h}}{\partial \mathbf{u}_Y} \\ \frac{\partial \mathbf{f}}{\partial \mathbf{i}_Z} & \frac{\partial \mathbf{f}}{\partial \mathbf{u}_Y} \end{pmatrix}$ to be positive definite because of Assumption 1.

Let $\Gamma = (W,E)$ be the connected network graph with vertex set $W$ and edge set $E$. An edge in $\Gamma$ corresponds to a branch that contains one element in the circuit. For a consistent model description, $\Gamma$ contains no cycles consisting of independent

[1] Assuming the controlled source matrix $S$ to be positive definite is very restrictive and usually not fulfilled when controlled sources are considered alone. However, controlled sources are often used to describe certain transistor behaviour. Considering the whole static behavior of a transistor (e.g. including bulk resistances) as a controlled source may lead to a positive definite matrix $S$.

voltage sources only and no cutsets consisting of independent current sources only. We split $E$ into $E_y$ and $E_z$, i.e., $E_y \cup E_z = E$ and $E_y \cap E_z = \emptyset$. A partition $(E_y, E_z)$ is called an *admissible partition*, if $E_y$ includes all the independent voltage sources, all the capacitors, all the conductors as well as all the controlled current sources, and $E_z$ includes all the independent current sources, all the inductors, all the resistors as well as all the controlled voltage sources.

We call a spanning tree $T$ of $\Gamma$ a *reference tree* if $T$ contains all the edges of the independent voltage sources, no edges of the independent current sources, and as many edges in $E_y$ as possible. Note that a reference tree $T$ may contain some edges in $E_z$. A reference tree is called *normal* if it contains as many edges corresponding to capacitors and as few edges corresponding to inductors as possible. The co-tree of $T$ is denoted by $\overline{T} = E \setminus T$. The hybrid equations are determined by an admissible partition $(E_y, E_z)$ and a reference tree $T$, which is not necessarily normal. For the sake of simplicity, we adopt a normal reference tree throughout this paper.

With respect to a normal reference tree $T$, we further split $\mathbf{i}$ and $\mathbf{u}$ into

$$\mathbf{i} = (\mathbf{i}_V, \mathbf{i}_C^\tau, \mathbf{i}_Y^\tau, \mathbf{i}_Z^\tau, \mathbf{i}_L^\tau, \mathbf{i}_C^\lambda, \mathbf{i}_Y^\lambda, \mathbf{i}_Z^\lambda, \mathbf{i}_L^\lambda, \mathbf{i}_J)^\top \text{ and } \mathbf{u} = (\mathbf{u}_V, \mathbf{u}_C^\tau, \mathbf{u}_Y^\tau, \mathbf{u}_Z^\tau, \mathbf{u}_L^\tau, \mathbf{u}_C^\lambda, \mathbf{u}_Y^\lambda, \mathbf{u}_Z^\lambda, \mathbf{u}_L^\lambda, \mathbf{u}_J)^\top,$$

where the superscripts $\tau$ and $\lambda$ designate the tree $T$ and the co-tree $\overline{T}$. With respect to a normal reference tree $T$, the vector valued function $\mathbf{f}$ is also split into $\mathbf{f}^\tau$ and $\mathbf{f}^\lambda$. This means $\mathbf{i}_Y^\tau = \mathbf{f}^\tau(\mathbf{i}_Z, \mathbf{u}_Y, t)$ and $\mathbf{i}_Y^\lambda = \mathbf{f}^\lambda(\mathbf{i}_Z, \mathbf{u}_Y, t)$. Similarly, we split $\mathbf{h}$, $\mathbf{q}$, and $\phi$.

By the definition of a normal reference tree, the *fundamental cutset matrix* $K$ is given by

$$K = \begin{array}{c} \begin{array}{cccccccccc} \mathbf{i}_V & \mathbf{i}_C^\tau & \mathbf{i}_Y^\tau & \mathbf{i}_Z^\tau & \mathbf{i}_L^\tau & \mathbf{i}_C^\lambda & \mathbf{i}_Y^\lambda & \mathbf{i}_Z^\lambda & \mathbf{i}_L^\lambda & \mathbf{i}_J \end{array} \\ \left(\begin{array}{cccccccccc} I & 0 & 0 & 0 & 0 & A_{VC} & A_{VY} & A_{VZ} & A_{VL} & A_{VJ} \\ 0 & I & 0 & 0 & 0 & A_{CC} & A_{CY} & A_{CZ} & A_{CL} & A_{CJ} \\ 0 & 0 & I & 0 & 0 & 0 & A_{YY} & A_{YZ} & A_{YL} & A_{YJ} \\ 0 & 0 & 0 & I & 0 & 0 & 0 & A_{ZZ} & A_{ZL} & A_{ZJ} \\ 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & A_{LL} & A_{LJ} \end{array}\right) \end{array}.$$

Then *Kirchhoff's current law* (KCL) may be written as $K\mathbf{i} = \mathbf{0}$. *Kirchhoff's voltage law* (KVL) provides $K^\perp \mathbf{u} = \mathbf{0}$ with $K^\perp$ being the *fundamental loop matrix*

$$K^\perp = \begin{array}{c} \begin{array}{cccccccccc} \mathbf{u}_V & \mathbf{u}_C^\tau & \mathbf{u}_Y^\tau & \mathbf{u}_Z^\tau & \mathbf{u}_L^\tau & \mathbf{u}_C^\lambda & \mathbf{u}_Y^\lambda & \mathbf{u}_Z^\lambda & \mathbf{u}_L^\lambda & \mathbf{u}_J \end{array} \\ \left(\begin{array}{cccccccccc} -A_{VC}^\top & -A_{CC}^\top & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 \\ -A_{VY}^\top & -A_{CY}^\top & -A_{YY}^\top & 0 & 0 & 0 & I & 0 & 0 & 0 \\ -A_{VZ}^\top & -A_{CZ}^\top & -A_{YZ}^\top & -A_{ZZ}^\top & 0 & 0 & 0 & I & 0 & 0 \\ -A_{VL}^\top & -A_{CL}^\top & -A_{YL}^\top & -A_{ZL}^\top & -A_{LL}^\top & 0 & 0 & 0 & I & 0 \\ -A_{VJ}^\top & -A_{CJ}^\top & -A_{YJ}^\top & -A_{ZJ}^\top & -A_{LJ}^\top & 0 & 0 & 0 & 0 & I \end{array}\right) \end{array}.$$

## 3 Hybrid Analysis

In this section, we describe the procedure of the hybrid analysis. The idea is to use all constitutive equations such that the equations $K\mathbf{i} = \mathbf{0}$ and $K^\perp \mathbf{u} = \mathbf{0}$ provide a system depending on $\mathbf{u}_C^\tau$, $\mathbf{u}_Y^\tau$, $\mathbf{i}_Z^\lambda$, and $\mathbf{i}_L^\lambda$ only. The details are described in [12]. The second and third line of $K\mathbf{i} = \mathbf{0}$ as well as the third and fourth line of $K^{\perp}\mathbf{u} = \mathbf{0}$ provide us the *hybrid equations* (or *hybrid equation system*)

$$\begin{aligned}
-A_{CZ}^{\top}\mathbf{u}_C^{\tau} - A_{YZ}^{\top}\mathbf{u}_Y^{\tau} - A_{ZZ}^{\top}\mathbf{h}^{\tau} + \mathbf{h}^{\lambda} &= A_{VZ}^{\top}\mathbf{v}_s(t),\\
-A_{CL}^{\top}\mathbf{u}_C^{\tau} - A_{YL}^{\top}\mathbf{u}_Y^{\tau} - A_{ZL}^{\top}\mathbf{h}^{\tau} - A_{LL}^{\top}\frac{\mathrm{d}}{\mathrm{d}t}\phi^{\tau} + \frac{\mathrm{d}}{\mathrm{d}t}\phi^{\lambda} &= A_{VL}^{\top}\mathbf{v}_s(t),\\
A_{CY}\mathbf{f}^{\lambda} + A_{CZ}\mathbf{i}_Z^{\lambda} + A_{CL}\mathbf{i}_L^{\lambda} + \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}^{\tau} + A_{CC}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}^{\lambda} &= -A_{CJ}\mathbf{j}_s(t),\\
\mathbf{f}^{\tau} + A_{YY}\mathbf{f}^{\lambda} + A_{YZ}\mathbf{i}_Z^{\lambda} + A_{YL}\mathbf{i}_L^{\lambda} &= -A_{YJ}\mathbf{j}_s(t),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{q} &= \mathbf{q}(\mathbf{u}_C^{\tau}, A_{VC}^{\top}\mathbf{v}_s(t) + A_{CC}^{\top}\mathbf{u}_C^{\tau}, t), \quad \phi \ = \ \phi(-A_{LL}\mathbf{i}_L^{\lambda} - A_{LJ}\mathbf{j}_s(t), \mathbf{i}_L^{\lambda}, t),\\
\mathbf{f} &= \mathbf{f}(-A_{ZZ}\mathbf{i}_Z^{\lambda} - A_{ZL}\mathbf{i}_L^{\lambda} - A_{ZJ}\mathbf{j}_s(t), \mathbf{i}_Z^{\lambda}, \mathbf{u}_Y^{\tau}, A_{VY}^{\top}\mathbf{v}_s(t) + A_{CY}^{\top}\mathbf{u}_C^{\tau} + A_{YY}^{\top}\mathbf{u}_Y^{\tau}, t),\\
\mathbf{h} &= \mathbf{h}(-A_{ZZ}\mathbf{i}_Z^{\lambda} - A_{ZL}\mathbf{i}_L^{\lambda} - A_{ZJ}\mathbf{j}_s(t), \mathbf{i}_Z^{\lambda}, \mathbf{u}_Y^{\tau}, A_{VY}^{\top}\mathbf{v}_s(t) + A_{CY}^{\top}\mathbf{u}_C^{\tau} + A_{YY}^{\top}\mathbf{u}_Y^{\tau}, t).
\end{aligned}$$

The procedure of the hybrid analysis is as follows.

1. The values of $\mathbf{u}_V$ and $\mathbf{i}_J$ are obvious from (1).
2. Compute the values of $\mathbf{i}_Z^{\lambda}$, $\mathbf{i}_L^{\lambda}$ and $\mathbf{u}_C^{\tau}$, $\mathbf{u}_Y^{\tau}$ by solving the hybrid equations.
3. Compute the values of $\mathbf{i}_Z^{\tau}$, $\mathbf{i}_L^{\tau}$ from the fourth and fifth line of $K\mathbf{i} = \mathbf{0}$ (KCL) and $\mathbf{u}_C^{\lambda}$, $\mathbf{u}_Y^{\lambda}$ from the first and second line of $K^{\perp}\mathbf{u} = \mathbf{0}$ (KVL) by substituting the values obtained in Steps 1–2.
4. Compute the values of $\mathbf{u}_Z^{\tau}$, $\mathbf{u}_Z^{\lambda}$, $\mathbf{u}_L^{\tau}$, $\mathbf{u}_L^{\lambda}$, and $\mathbf{i}_C^{\tau}$, $\mathbf{i}_C^{\lambda}$, $\mathbf{i}_Y^{\tau}$, $\mathbf{i}_Y^{\lambda}$ by substituting the values obtained in Steps 1–3 into (2) and (3).
5. Compute the values of $\mathbf{i}_V$ and $\mathbf{u}_J$ by substituting the values obtained in Steps 1–4 into the first line of KCL and the fifth line of KVL.

All operations in Steps 3–5 are substitutions and differentiations of the obtained solutions. Consequently, the numerical difficulty is determined by the index of the hybrid equation system. Higher index variables as known from MNA do not appear in the hybrid equation system. In this paper, we prove that the hybrid equation system has index at most one. The proof relies on the *tractability index* concept for DAEs with the use of a projector based analysis.

## 4 Hybrid Equations with Properly Stated Leading Term

Consider a DAE in the form of

$$A\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(\mathbf{x}(t),t) + \mathbf{b}(\mathbf{x}(t),t) = \mathbf{0}. \tag{4}$$

Let $A$ be an $m \times n$ matrix. We define $D(\mathbf{x},t) := \dfrac{\partial \mathbf{d}(\mathbf{x},t)}{\partial \mathbf{x}}$, $B(\mathbf{x},t) := \dfrac{\partial \mathbf{b}(\mathbf{x},t)}{\partial \mathbf{x}}$, and $M(\mathbf{x},t) := AD(\mathbf{x},t)$. A matrix $P$ satisfying $P^2 = P$ is called a *projector*.

**Definition 1 ([13, Definition 2.1]).** The equation (4) is a *DAE with properly stated leading term* if the size of $D(\mathbf{x},t)$ is $n \times m$, $\ker A \oplus \operatorname{im} D(\mathbf{x},t) = \mathbb{R}^n$ holds for all $\mathbf{x}$ and $t$ from the definition domain, and there is an $n \times n$ projector function $P(t)$ continuously differentiable with respect to $t$ such that $\ker P(t) = \ker A$, $\operatorname{im} P(t) = \operatorname{im} D(\mathbf{x},t)$, and $\mathbf{d}(\mathbf{x},t) = P(t)\mathbf{d}(\mathbf{x},t)$.

A DAE with properly stated leading term (4) arises in circuit simulation via analysis methods such as MNA [14]. A DAE with properly stated leading term was first introduced in [15]. The analysis of such DAEs has been developed in [14, 16–19].

Obviously, the DAE (4) represents a regular ODE if and only if the matrix $M(\mathbf{x},t)$ is nonsingular for all $\mathbf{x}$ and $t$ of the definition domain. In this case we say that the DAE (4) has index 0. In the case of a singular matrix $M(\mathbf{x},t)$ for all $\mathbf{x}$ and $t$, the DAE (4) contains algebraic equations. Furthermore, one may have to differentiate certain part of the system to get a solution. A simple criteria for the absence of this problem is given by the tractability index 1 condition (see [13], Theorem 4.3).

**Definition 2 ([13, Definition 3.3]).** The DAE (4) is *regular with index 1* on their definition domain if $M(\mathbf{x},t)$ is singular and $\ker D(\mathbf{x},t) \cap \{\mathbf{z} \in \mathbb{R}^m |\ B(\mathbf{x},t)\mathbf{z} \in \operatorname{im} M(\mathbf{x},t)\} = \{\mathbf{0}\}$ for all $(\mathbf{x},t)$ of the definition domain.

*Remark 1 ([20, Remark 4.6]).* A DAE (4) is regular with index 1 if and only if the matrix $M(\mathbf{x},t) + B(\mathbf{x},t)Q(\mathbf{x},t)$ is nonsingular for all $\mathbf{x}$ and $t$ with a projector $Q(\mathbf{x},t)$ satisfying $\operatorname{im} Q(\mathbf{x},t) = \ker M(\mathbf{x},t)$.

We rewrite the hybrid equation system as a DAE with properly stated leading term. A *reflexive generalized inverse* of a matrix $A$ is a matrix $A^-$ which satisfies $AA^-A = A$ and $A^-AA^- = A^-$. Let us define

$$A = \begin{pmatrix} 0 & 0\,0 & 0 \\ -A_{LL}^\top & I\;0 & 0 \\ 0 & 0\;I & A_{CC} \\ 0 & 0\,0 & 0 \end{pmatrix}, \quad \mathbf{d}(\mathbf{x},t) = A^-A \begin{pmatrix} \phi^\tau(-A_{LL}\mathbf{i}_L^\lambda - A_{LJ}\mathbf{j}_s(t), \mathbf{i}_L^\lambda, t) \\ \phi^\lambda(-A_{LL}\mathbf{i}_L^\lambda - A_{LJ}\mathbf{j}_s(t), \mathbf{i}_L^\lambda, t) \\ \mathbf{q}^\tau(\mathbf{u}_C^\tau, A_{VC}^\top\mathbf{v}_s(t) + A_{CC}^\top\mathbf{u}_C^\tau, t) \\ \mathbf{q}^\lambda(\mathbf{u}_C^\tau, A_{VC}^\top\mathbf{v}_s(t) + A_{CC}^\top\mathbf{u}_C^\tau, t) \end{pmatrix},$$

$$\mathbf{x}(t) = \begin{pmatrix} \mathbf{i}_Z^\lambda \\ \mathbf{i}_L^\lambda \\ \mathbf{u}_C^\tau \\ \mathbf{u}_Y^\tau \end{pmatrix}, \quad \mathbf{b}(\mathbf{x},t) = \begin{pmatrix} -A_{VZ}^\top\mathbf{v}_s(t) - A_{CZ}^\top\mathbf{u}_C^\tau - A_{YZ}^\top\mathbf{u}_Y^\tau - A_{ZZ}^\top\mathbf{h}^\tau + \mathbf{h}^\lambda \\ -A_{VL}^\top\mathbf{v}_s(t) - A_{CL}^\top\mathbf{u}_C^\tau - A_{YL}^\top\mathbf{u}_Y^\tau - A_{ZL}^\top\mathbf{h}^\tau \\ A_{CY}\mathbf{f}^\lambda + A_{CZ}\mathbf{i}_Z^\lambda + A_{CL}\mathbf{i}_L^\lambda + A_{CJ}\mathbf{j}_s(t) \\ \mathbf{f}^\tau + A_{YY}\mathbf{f}^\lambda + A_{YZ}\mathbf{i}_Z^\lambda + A_{YL}\mathbf{i}_L^\lambda + A_{YJ}\mathbf{j}_s(t) \end{pmatrix}.$$

This gives the hybrid equation system in the form of (4). Under Assumption 1, the hybrid equation system (4) is shown to be a DAE with properly stated leading term.

## 5 Index of Hybrid Equations

In this section, we show that the index of the hybrid equations is at most one, and give a structural criteria for hybrid equations with index zero. We now introduce the *Resistor-Acyclic condition* for admissible partition $(E_y, E_z)$.

**[Resistor-Acyclic condition]**

- Each conductor and controlled current source in $E_y$ belongs to a cycle consisting of independent voltage sources, capacitors, and itself.
- Each resistor and controlled voltage source in $E_z$ belongs to a cutset consisting of inductors, independent current sources, and itself.

Consider the graph $\tilde{\Gamma}$ obtained from $\Gamma = (W,E)$ by contracting all edges of independent voltage sources and capacitors and deleting all edges of inductors and independent current sources. The Resistor-Acyclic condition means that $\tilde{\Gamma}$ is acyclic [7].

**Theorem 1.** *Under Assumption 1, the index of the hybrid equations is at most one for any admissible partition $(E_y,E_z)$ and normal reference tree $T$. Moreover, the index is zero if and only if an admissible partition $(E_y,E_z)$ satisfies the Resistor-Acyclic condition.*

Here we present only a sketch of the proof. Details are given in [12]. Computation of the matrix $M(\mathbf{x},t)+B(\mathbf{x},t)Q(\mathbf{x},t)$ leads to (omitting the arguments)

$$M+BQ=\begin{pmatrix} B_Z & 0 & 0 & -A_{YZ}^\top+B_H \\ * & M_L & 0 & * \\ * & 0 & M_C & * \\ A_{YZ}+B_F & 0 & 0 & B_Y \end{pmatrix} \quad \text{for} \quad Q=\begin{pmatrix} I & & & \\ & 0 & & \\ & & 0 & \\ & & & I \end{pmatrix}.$$

Here, $M_L$ and $M_C$ are nonsingular and $\begin{pmatrix} B_Z & -A_{YZ}^\top+B_H \\ A_{YZ}+B_F & B_Y \end{pmatrix} = \begin{pmatrix} 0 & -A_{YZ}^\top \\ A_{YZ} & 0 \end{pmatrix} + \begin{pmatrix} B_Z & B_H \\ B_F & B_Y \end{pmatrix}$ is a sum of a positive semidefinite and a positive definite matrix. These properties imply $M+BQ$ to be nonsingular. Since $AD(\mathbf{x},t) = A\begin{pmatrix} L & 0 \\ 0 & C \end{pmatrix} A^\top$ holds, also the second statement of Theorem 1 is clear.

By Theorem 1, we can prove that the minimum index of the hybrid equations never exceeds the index of the DAE arising from MNA for nonlinear time-varying circuits without controlled voltage/current sources.

*Remark 2.* A simple algorithm for finding the optimal admissible partition is given in [7]. See [7, Examples 4.13–4.14] for circuit examples, which trace the procedure of the hybrid analysis and make comparisons between the hybrid analysis and MNA.

*Remark 3.* For nonlinear time-varying circuits composed of resistors (all modelled as conductances), inductors, capacitors, and voltage/current sources, the dimension of the hybrid equation system is no greater than that one for the MNA system. This is because $\dim(\mathbf{u}_C^\tau, \mathbf{u}_Y^\tau) < n$ for $n$ being the number of nodes of the circuit, $\dim \mathbf{i}_L^\lambda$ is not greater than the number of inductors in the system, and $\dim \mathbf{i}_Z^\lambda$ is not greater than the number of (controlled) voltage sources of the system.

**Acknowledgements** The first author was supported by a Grant-in-Aid for Scientific Research from the Japan Society of Promotion of Science. The second author was supported by the

Global COE "The Research and Training Center for New Development in Mathematics." The third author was supported by the European Union within the framework of the project "Integrated Circuit/EM Simulation and design Technologies for Advanced Radio Systems-on-chip" (FP7/2008/ICT/214911).

## References

1. Iri, M.: Applications of matroid theory. In: Mathematical Programming — The State of the Art, pp. 158–201. Springer-Verlag, Berlin (1983)
2. Bryant, P.R.: The order of complexity of electrical networks. Proceedings of the Institution of Electrical Engineers, Part C **106**, 174–188 (1959)
3. Reißig, G.: Extension of the normal tree method. International Journal of Circuit Theory and Applications **27**, 241–265 (1999)
4. Kron, G.: Tensor Analysis of Networks. John Wiley and Sons, New York (1939)
5. Amari, S.: Topological foundations of Kron's tearing of electric networks. RAAG Memoirs **3**, 322–350 (1962)
6. Branin, F.H.: The relation between Kron's method and the classical methods of network analysis. The Matrix and Tensor Quarterly **12**, 69–115 (1962)
7. Takamatsu, M., Iwata, S.: Index characterization of differential-algebraic equations in hybrid analysis for circuit simulation. International Journal of Circuit Theory and Applications (to appear)
8. Iwata, S., Takamatsu, M.: Index minimization of differential-algebraic equations in hybrid analysis for circuit simulation. Mathematical Programming (to appear)
9. Reißig, G.: The index of the standard circuit equations of passive RLCTG-networks does not exceed 2. Proc. ISCAS'98, IEEE International Symposium on Circuits and Systems **3**, 419–422 (1998)
10. Estévez Schwarz, D., Tischendorf, C.: Structural analysis of electric circuits and consequences for MNA. International Journal of Circuit Theory and Applications **28**, 131–162 (2000)
11. Encinas, A.J., Riaza, R.: Tree-based characterization of low index circuit configurations without passivity restrictions. International Journal of Circuit Theory and Applications **36**, 135–160 (2008)
12. Iwata, S., Takamatsu, M., Tischendorf, C.: Hybrid analysis of nonlinear time-varying circuits providing DAEs with index at most one. METR 2008-37, University of Tokyo (2008). URL http://www.keisu.t.u-tokyo.ac.jp/research/techrep/data/2008/METR08-37.pdf
13. März, R.: Nonlinear differential-algebraic equations with properly formulated leading term. Tech. Rep. 01-3, Department of Mathematics, Humboldt-Universität zu Berlin (2001). URL http://www.mathematik.hu-berlin.de/publ/pre/2001/P-01-3.ps
14. Higueras, I., März, R., Tischendorf, C.: Stability preserving integration of index-1 DAEs. Applied Numerical Mathematics **45**, 175–200 (2003)
15. Balla, K., März, R.: A unified approach to linear differential algebraic equations and their adjoint equations. Zeitschrift für Analysis und ihre Anwendungen **21**, 783–802 (2002)
16. Higueras, I., März, R.: Differential algebraic equations with properly stated leading terms. Computers and Mathematics with Applications **48**, 215–235 (2004)
17. März, R.: The index of linear differential algebraic equations with properly stated leading term. Results in Mathematics **42**, 308–338 (2002)
18. März, R., Riaza, R.: Linear differential-algebraic equations with properly stated leading term: Regular points. Journal of Mathematical Analysis and Applications **323**, 1279–1299 (2006)
19. Riaza, R., März, R.: Linear index-1 DAEs: Regular and singular problems. Acta Applicandae Mathematicae **84**, 29–53 (2004)
20. Voigtmann, S.: General linear methods for integrated circuit design. Ph.D. thesis, Humboldt-Universität zu Berlin (2007)

# Transient Analysis of Nonlinear Circuits Based on Waves

Carlos Christoffersen

**Abstract** A new approach for transient analysis of nonlinear circuits is presented. The circuit equations are formulated as functions of incident and reflected waves at the device ports. Only one large matrix decomposition is necessary if time step is constant. The proposed method is parallelizable, allows straightforward inclusion of complex nonlinear device models and has better convergence properties compared to existing methods. Simulation results are provided to demonstrate the approach.

## 1 Introduction

Transient simulation of circuits using wave quantities has been previously proposed mainly in the framework of Wave Digital Filter (WDF) theory [1]. References [2–8] are some examples. WDF are discrete structures that mimic an analog reference circuit. The reference circuit is not required to be a filter and thus WDF theory can be applied to model any circuit. A good introduction of WDF concepts for the purpose of circuit simulation can be found in Reference [6]. The basic idea is to formulate equations in terms of wave quantities at the ports of each device. The network topology is represented by means of *adaptors*, which is the name for the scattering matrix representing port junctions. WDF preserve losslessness and passivity of the reference circuit, and are less sensitive to parameter quantization than discretizations based on voltage and current [7]. References [2, 3, 5, 7–10] show how nonlinear devices (both algebraic and dynamic) can be represented in terms of waves. In general it is not possible to avoid delay-free loops (DFLs) in circuits with more than one nonlinear port and thus some iterative method is required to solve the equations. References [9, 10] propose methods to eliminate DFLs created by multiple nonlinearities. These works basically propose to simultaneously compute

Carlos Christoffersen
Department of Electrical Engineering, Lakehead University, 955 Oliver Road, Thunder Bay, ON, Canada P7B 5E1, e-mail: c.christoffersen@ieee.org

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_21,

(or pre-compute) all possible reflections given all possible incident waves from all nonlinear ports. That kind of approach is useful only when the number of nonlinear ports is small. Reference [8] proposes a method to eliminate DFLs for circuits containing nonlinear inductors. This method relies on an estimation of the inductor current at each time step and can not be applied for algebraic nonlinearities. In Reference [6] WDF are used to simulate power electronic circuits with nonlinear devices treated as switches. A relaxation approach was proposed by Meerkötter *et al.* [3] to eliminate DFLs. This approach always converges for circuits that contain passive nonlinear devices that are also locally passive. Local passivity implies that the spectral radius of the device small-signal scattering parameter matrix is less than one (note transistors are not locally passive). Reference [4] further investigates this approach and shows that it is possible to cut DFLs and split the computation in independent blocks suitable for parallel processing without losing convergence.

The wave-based transient analysis approach presented in this work solves DFLs through an iterative procedure that has better convergence properties than previous works. Another advantage of the proposed approach is that allows easy inclusion of complex multiport nonlinear devices formulated in the Kirchhoff domain. Only one large (sparse) matrix decomposition is required for a given time step size. This paper is organized as follows: the formulation of the method and some of its properties are presented in Section 2, followed by simulation examples in Section 3. Conclusions and directions for further research are given in Section 4.

## 2 Proposed Method

In the following we will assume that all sources have some internal resistance and all remaining devices are passive as shown in Fig. 1. Let $n$ be the total number of ports and $m$ the number of nonlinear ports. For each port, we associate a characteristic impedance equal to $Z_j$. Since a time domain method is being considered, $Z_j$ is pure real. The voltage and current at Port $j$ can be expressed in terms of power waves as follows:

$$v_j = \sqrt{Z_j}(a_j + b_j) \tag{1}$$

$$i_j = \frac{a_j - b_j}{\sqrt{Z_j}}\,, \tag{2}$$

where $a_j$ and $b_j$ are the incident and reflected power waves at Port $j$ as seen from the devices as shown in Fig. 1. The circuit topology defines the relationship between the vector of incident and reflected waves, $\mathbf{a}$ and $\mathbf{b}$, respectively. Let $\mathbf{Q}$ and $\mathbf{B}$ be the full cut-set and loop-set matrices for a given tree in the circuit. The vectors of all port currents ($\mathbf{i}$) and port voltages ($\mathbf{v}$) satisfy

$$\mathbf{Qi} = \mathbf{0} \tag{3}$$

$$\mathbf{Bv} = \mathbf{0}\,. \tag{4}$$

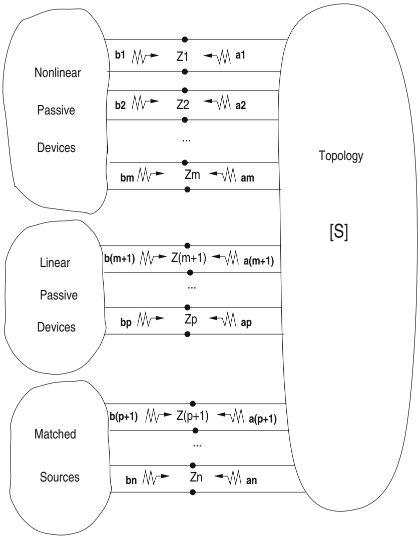


**Fig. 1** Circuit partition. Linear and nonlinear devices are assumed to be passive and sources are assumed matched to the impedances of their respective ports. The interconnection (topology) is represented by a scattering parameter matrix

Combining (1) and (2) with (3) and (4) the following system of $n$ equations is obtained:

$$\begin{bmatrix} \mathbf{QD}^{-1} \\ \mathbf{BD} \end{bmatrix} \mathbf{a} = \begin{bmatrix} \mathbf{QD}^{-1} \\ -\mathbf{BD} \end{bmatrix} \mathbf{b}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix that has the square root of the reference impedances ($\sqrt{Z_j}$) in its diagonal. Matrices $\mathbf{Q}$ and $\mathbf{B}$ are sparse and thus $\mathbf{a}$ can efficiently be obtained for large circuits if $\mathbf{b}$ is known. This equation defines the scattering matrix that represents the circuit topology,

$$\begin{bmatrix} \mathbf{a}_N \\ \mathbf{a}_L \\ \mathbf{a}_S \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_{33} \end{bmatrix} \begin{bmatrix} \mathbf{b}_N \\ \mathbf{b}_L \\ \mathbf{b}_S \end{bmatrix}, \tag{6}$$

where $\mathbf{a}_N$, $\mathbf{a}_L$ and $\mathbf{a}_S$ are the vectors of waves incident to nonlinear devices, linear devices and sources, respectively and $\mathbf{b}_N$, $\mathbf{b}_L$ and $\mathbf{b}_S$ are similarly defined for the reflected waves. The trapezoidal rule is used for time discretization as is usual in the WDF literature [6]. Reference resistances at sources and linear devices are chosen such that there are no DFL from the device back to the network [6]. Thus $\mathbf{b}_L$ is constant for one time step as all loops contain at least one delay.

### *2.1 Nonlinear Models*

Nonlinear devices, both static and dynamic, will cause DFLs. Reflections are calculated in this work using Newton's method. For example, suppose the current in a nonlinear device is given by a nonlinear function, $i(v_j)$. An error function, $e(a_j)$, is defined

$$e(a_j) = \sqrt{Z_j}\, i(\sqrt{Z_j}(a_j + b_j)) - a_j + b_j = 0\,. \tag{7}$$

This can easily be generalized for multi-port nonlinear devices, both static and dynamic. Newton iterations are performed independently for each nonlinear device and require a small Jacobian matrix factorization. One advantage of this approach is that it allows straightforward treatment of complex nonlinear models.

## 2.2 *Iterative Method*

Vector $\mathbf{b}_S$ is known since it is forced by the matched sources. The nonlinear equation to be solved is thus

$$\mathbf{b}_N = \mathbf{F}(\mathbf{S}_{11}\mathbf{b}_N + \mathbf{a}_0)\,, \tag{8}$$

with $\mathbf{a}_0 = \mathbf{S}_{12}\mathbf{b}_L + \mathbf{S}_{13}\mathbf{b}_S$ being constant for a given time instant and $\mathbf{F}$ is a nonlinear vector function that represents the contribution of nonlinear passive devices. The following iterative fixed-point scheme is proposed:

$$\mathbf{b}_N^{k+1} = (\mathbf{I} - \mathbf{K}^k)\mathbf{b}_N^k + \mathbf{K}^k\mathbf{F}(\mathbf{S}_{11}\mathbf{b}_N^k + \mathbf{a}_0)\,, \tag{9}$$

where the $k$ superscript denotes the iteration number, $\mathbf{I}$ is an identity matrix and $\mathbf{K}^k$ is an $m \times m$ matrix that may be constant or updated at each iteration as described in the following subsection.

### 2.2.1 Convergence Analysis

Assume for now that $\mathbf{K}^k = \mathbf{I}$. In that case (9) is equivalent to just propagating reflections along the DFLs in the circuit (*i.e.*, plain relaxation). It can be shown that iterations converge to the desired solution if the spectral radius of $\mathbf{J}_F\mathbf{S}_{11}$ is less than one, where $\mathbf{J}_F$ is the Jacobian matrix of $\mathbf{F}$. This condition is satisfied if all nonlinear devices are locally passive (*e.g.*, diodes) and in this case convergence is global. Unfortunately convergence is not guaranteed if devices such as transistors are present in the circuit. Consider now a scalar ($\gamma$) between 0 and 1 and let

$$\mathbf{K}^k = \gamma\mathbf{I}\,. \tag{10}$$

Equation (9) becomes essentially equivalent to the formulation proposed in Reference [4]. A good selection of $\gamma$ may improve convergence properties compared to plain relaxation but this modification is not enough to guarantee convergence in the presence of locally active devices.

Local convergence can be obtained by setting

$$\mathbf{K}^k = \left(\mathbf{I} - \mathbf{J}_F^k \mathbf{S}_{11}\right)^{-1} . \tag{11}$$

Equation (9) becomes then equivalent to Newton's method. Note that $\mathbf{J}_F$ is a block-diagonal matrix with each block being the small-signal scattering matrix for each nonlinear device. Although it is possible to re-order (5) and (9) in order to perform one sparse matrix decomposition per iteration, this is more expensive than the backward substitution of the relaxation approach.

One possibility to avoid the factorization of a large matrix at each iteration is to make $\mathbf{K}^k$ artificially sparse. If elements of $\mathbf{S}_{11}$ in (11) are set to zero to obtain the same block-diagonal pattern as in $\mathbf{J}_F$, the resulting $\mathbf{K}^k$ matrix is also block-diagonal. This requires explicit knowledge of $\mathbf{S}_{11}$ which in turn requires $m$ backward substitutions of the decomposed matrix originating from (5), but this only need to be performed once. The resulting iterative scheme is known as Newton-Jacobi algorithm [11]. Local convergence is no longer guaranteed but it will be shown in Section 3 that this modification is an improvement over the relaxation approach.

### 2.2.2 Reflected Power Considerations

An interesting property of the pure relaxation approach ($\mathbf{K}^k = \mathbf{I}$) is that due to the nature of power waves iterations are guaranteed not to diverge to infinity, even if the initial guess is far from the solution. This is demonstrated as follows: The total reflected power at each iteration is given by $|\mathbf{b}_N^{k+1}|^2$, where the bars denote the Euclidean Norm, and is bounded by

$$|\mathbf{b}_N^{k+1}|^2 < P_{\max} + L|\mathbf{S}_{11}\mathbf{b}_N^k + \mathbf{a}_0|^2 , \tag{12}$$

here $L$ is a scalar less than one (since nonlinear devices are passive) and $P_{\max}$ is the maximum power that can be sent from the nonlinear devices to the network during one time step and depends on the total energy stored in nonlinear capacitors and inductors. If the upper bound is propagated from the first iteration the following result is obtained:

$$\lim_{k\to\infty} |\mathbf{b}_N^{k+1}|^2 < \frac{1}{1-L}\left(P_{\max} + L|\mathbf{a}_0|^2\right) . \tag{13}$$

An upper bound can also be found if $\mathbf{K}^k$ is chosen as in (10). This property can be extended to the Newton-Jacobi approach if $\mathbf{K}^k$ is chosen as follows:

$$\mathbf{K}^k = \alpha\gamma\mathbf{I} + (1-\alpha)\mathbf{K}_{\mathrm{NJ}}^k , \tag{14}$$

where $\mathbf{K}_{\mathrm{NJ}}^k$ is the block diagonal matrix used for the Newton-Jacobi approach and $\alpha$ is the lowest scalar between 0 and 1 such that $\mathbf{b}_N^{k+1}$ satisfies (12). For circuits with locally passive nonlinear devices this choice tends to provide both global convergence when initial guess is far from the solution and faster convergence rate near the solution as $\alpha$ becomes 0.

In summary, iterations are performed according to (9) using $\mathbf{K}^k$ as defined in (14). Thus the proposed method is an hybrid between relaxation and Newton-Jacobi. Steffensen [12] updates,

$$b_j^{k+1} = b_j^{k-2} + \frac{(b_j^{k-1} - b_j^{k-2})^2}{b_j^k - 2b_j^{k-1} + b_j^{k-2}} \,, \tag{15}$$

are occasionally used along with regular iterations to accelerate convergence. All calculations in (9) and (15) can be performed locally for each nonlinear device and thus could be computed in parallel. The only communication between processors at each iteration is to evaluate the product $\mathbf{S}_{11}\mathbf{b}_N^k$ which requires the reflections from all nonlinear devices.

## 3 Numerical Example

The circuit shown in Fig. 2 contains both static and dynamic nonlinearities included in the diode model. Since diodes are locally passive, convergence is guaranteed in this case. The circuit parameters are: $C = 4\ \mu$F, $R_S = 50\ \Omega$, $R_L = 5$ k$\Omega$. The source is sinusoidal with a peak of 3 V (later increased to 200 V) and a frequency of 500 Hz. The diode parameters are $I_S = 1$ fA, $N = 1$, $C_j = 100$ nF, $M_j = 0.5$, $V_j = 1$ V and $F_C = 0.5$. Tolerance was set to $10^{-8}$. A transient simulation with a duration of 8 ms and a time step equal to 50 $\mu$s was performed. Figure 2 also shows a comparison of simulation results obtained with the proposed algorithm and Spice.

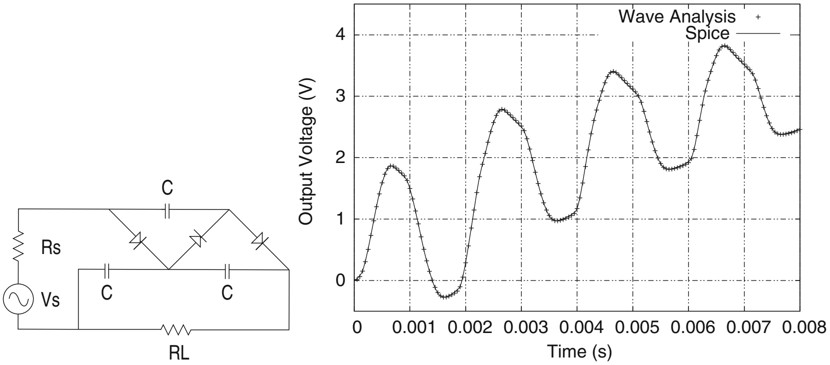


**Fig. 2:** Nonlinear circuit and comparison of simulation results at load resistor with 3 V input

Table 1 summarizes the results for different input voltages, reference impedances at the diode ports. The number of time steps is the same for all simulations. The last column indicates how many iterations with $\alpha \neq 0$ (*i.e.*, limited reflected power) were performed. The first three rows show the results when the full Jacobian matrix

**Table 1:** Summary of simulation results

| Input (V) | Ref. impedance ($\Omega$) | Steffensen U. | Total iterations | Reflected power limit |
|---:|---:|---:|---:|---:|
| 3 | 50 | Disabled | [a]763 | 56 |
| 3 | 500 | Disabled | [a]806 | 15 |
| 200 | 500 | Disabled | [a]1950 | 1141 |
| 3 | 50 | 87 | [b]4897 | 0 |
| 3 | 500 | 271 | [b]3591 | 0 |
| 200 | 500 | Disabled | [b]42075 | 0 |
| 200 | 500 | 250 | [b]13836 | 0 |
| 3 | 50 | Disabled | 5225 | 1 |
| 3 | 50 | 62 | 4412 | 1 |
| 3 | 500 | Disabled | 2518 | 4 |
| 3 | 500 | 27 | 2483 | 4 |
| 200 | 500 | Disabled | 10234 | 1234 |
| 200 | 500 | 138 | 6852 | 953 |

[a] Regular Newton method + Power limit
[b] Pure relaxation

is used instead of a block-diagonal matrix. The choice of reference impedance does not significantly affect the convergence rate for that case, but this is not so when pure relaxation or the modified Newton-Jacobi method is used as seen in the remaining rows of the table. The most nonlinear cases (200 V input) show that both Steffensen updates and reflected power limitation are frequently used. Selective Steffensen updates improve convergence in all cases. It should be noted that sometimes the reflected power limitation reduces somewhat the convergence rate. However even when that happens the proposed approach converges faster than plain relaxation.

## 4 Conclusions and Discussion

The proposed method has the good numerical properties associated with the WDF approach and allows the inclusion of complex nonlinear device models. Equation (9) introduces a cause-effect relation in the power delivered to nonlinear devices that is not apparent in more classical approaches. That formulation along with selective Steffensen's updates, the treatment of nonlinear device models and the approach to prevent numerical divergence with Newton-Jacobi iterations are novel in this work. An admittedly simple nonlinear circuit was simulated to illustrate the performance of the proposed method. It was shown that the combination of the techniques proposed here improves the convergence rate compared with plain relaxation. As expected, Newton-Jacobi iterations do not converge as fast as regular Newton iterations. Newton-Jacobi could be faster however for large circuits as the cost of iterations does not grow as much with the circuit size. The performance of the proposed method with large scale circuits has yet to be evaluated.

Circuits with locally active devices (*e.g.*, transistors) can be handled by this method but then convergence is no longer guaranteed. Work in progress indicates that with some modifications it is possible to guarantee (at least local) convergence in that case. The analysis presented here is for fixed time step, but variable time step could be handled without additional matrix decompositions by considering variable reflections from linear devices, *i.e.* $\mathbf{b}_L$ changes with each iteration. There is an associated computational cost with this modification, but the main features of the method remain intact. Another important issue is the optimum selection of reference impedances at nonlinear device ports. This selection may have to be adaptive since the optimum values are dependent on the circuit state. Further research will also include the application of the ideas presented here for other types of circuit analysis such as Harmonic Balance, or Envelope Following Transient.

**Acknowledgements** This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

1. Fettweis, A.: Wave Digital Filters: Theory and Practice. IEEE Proceedings, **74**, 270–327 (1986)
2. Meerkötter, K. and Scholz, R.: Digital simulation of nonlinear circuits by wave digital filter principles. In: Proc. IEEE ISCAS '89, pp. 720–723. Portland, May 8–11 (1989)
3. Meerkötter, Felderhoff, T.: Simulation of nonlinear transmission lines by wave digital filter principles. In: Proc. IEEE ISCAS '92, pp. 875–879. San Diego, May 3–6 (1992)
4. Felderhoff, T.: Jacobi's method for massive parallel wave digital filter algorithm. In: Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing, pp. 1621–1624. Atlanta, May 7–10 (1996)
5. Felderhoff T.: A new wave description for nonlinear elements. In: Proc. IEEE ISCAS '96, pp. 221–224. Atlanta, May 12–15 (1996)
6. Fiedler, A., Grotstollen, H.: Simulation of power electronic circuits with principles used in wave digital filters. IEEE Trans. on Indutry Applications, **33**, 49–57. (1997)
7. Sarti, A., De Poli, G.: Towards nonlinear wave digital filters, IEEE Trans. on Signal Processing, **47**, 1654–1668. (1999)
8. Fränken, D.: Wave digital simulation of electrical networks containing nonlinear dynamical elements — a new approach. In: Proc. IEEE ISCAS 2000, pp. 535–538. Geneva, May 28–31 (2000)
9. Petrausch, S., Rabenstein, R.: Wave digital filters with multiple nonlinearities. In: Proc. of the 12th European Signal Processing Conf. (2004) http://www-nt.e-technik.uni-erlangen.de/LMS/publications/web/lnt2004_22.pdf.Cited8Sept.2008
10. Borin, G., De Poli, G., Rocchesso, D.: Elimination of delay-free loops in discrete-time models of nonlinear acoustic systems. IEEE Trans. on Speech And Audio Processing, **8**, 597–605. (2000)
11. Bahi, J. M., Contassot-Vivier, S., Couturier, R.: Parallel Iterative Algorithms, Chapman & Hall / CRC, Boca Raton London New York (2008)
12. Burden, R., Faires, J.: Numerical Analysis, Fifth Edition, Prindle, Weber & Schmidt, Boston (1993)

# Simultaneous Step-Size and Path Control for Efficient Transient Noise Analysis

Werner Römisch, Thorsten Sickenberger, and Renate Winkler

**Abstract** Noise in electronic components is a random phenomenon that can adversely affect the desired operation of a circuit. Transient noise analysis is designed to consider noise effects in circuit simulation. Taking noise into account by means of Gaussian white noise currents, mathematical modelling leads to stochastic differential algebraic equations (SDAEs) with a large number of small noise sources. Their simulation requires an efficient numerical time integration by mean-square convergent numerical methods. As efficient approaches for their integration we discuss adaptive linear multi-step methods, together with a new step-size and path selection control strategy. Numerical experiments on industrial real-life applications illustrate the theoretical findings.

## 1 Transient Noise Analysis in Circuit Simulation

In current chip design the decreasing feature sizes, high clock frequencies and low supply voltages cause several parasitic effect. As a consequence the signal-to-noise ratio decreases, i.e., the difference between the desired signal and noise is getting smaller. To address the signal-to-noise ratio the modelling and the simulation can be improved by taking the inner electrical noise into account. An important requirement for a transient noise simulation is the appropriate modelling of the noise

Werner Römisch
Inst. of Mathematics, Humboldt-Universität zu Berlin, 10099 Berlin, Germany, e-mail: romisch@math.hu-berlin.de

Thorsten Sickenberger (Corresponding author)
Dept. of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, UK,
e-mail: t.sickenberger@hw.ac.uk

Renate Winkler
Dept. of Mathematics, Bergische Universität Wuppertal, 42199 Wuppertal, Germany, e-mail: winkler@math.uni-wuppertal.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_22,

sources in the time domain. We consider two different sources of inner electrical noise, namely, thermal noise of resistors and shot noise of semiconductors. Thermal noise $i_{th}$ of resistors is caused by the thermal motion of electrons and is described by Nyquist's theorem. Shot noise $i_{shot}$ of *pn*-junctions, caused by the discrete nature of currents due to the elementary charge, is modelled by Schottky's formula and inherits noise intensities that depend on the deterministic currents (see e.g. [1,2]).

A noisy element is modelled as an additional stochastic current source in parallel to the original electronic element. The noise intensity is given by the physical characteristics and the noise models are added to the network equations. Combining Kirchhoff's current law with the element characteristics and using the charge-oriented formulation formally yields a stochastic differential-algebraic equation (SDAE) of the type (see e.g. [3,4])

$$A\frac{\mathrm{d}}{\mathrm{d}t}q(x(t))+f(x(t),t)+\sum_{r=1}^{m}g_r(x(t),t)\xi_r(t)=0\,, \tag{1}$$

where $A$ is a constant singular incidence matrix determined by the topology of the dynamic circuit parts, the vector $q(x)$ consists of the charges and the fluxes, and $x$ is the vector of unknowns consisting of the nodal potentials and the branch currents through voltage-defining elements. The term $f(x,t)$ describes the impact of the static elements, $g_r(x,t)$ denotes the vector of noise intensities (amplitudes) for the $r$-th noise source, and $\xi := (\xi_1,\ldots,\xi_m)^T$ is an $m$-dimensional vector of independent Gaussian white noise sources (see e.g. [1]).

Although this system (1) appears to be similar to a noise-free system, it requires a completely different mathematical background. A serious mathematical description begins by introducing the Brownian motion or the Wiener process that is caused by integrating the white noise "$W(t)=\int_0^t \xi(s)\mathrm{d}s=\int_0^t \mathrm{d}W(s)$" (see e.g. [5]). Problem (1) is then understood as a stochastic integral equation

$$Aq(X(s))\Big|_{t_0}^{t}+\int_{t_0}^{t}f(X(s),s)\,\mathrm{d}s+\sum_{r=1}^{m}\int_{t_0}^{t}g_r(X(s),s)\,\mathrm{d}W_r(s)=0,\;\; t\in[t_0,T]\,, \tag{2}$$

where the second integral is an Itô-integral, and $W$ denotes an $m$-dimensional Wiener process (or Brownian motion) given on the probability space $(\Omega,\mathscr{F},P)$ with a filtration $(\mathscr{F}_t)_{t\geq t_0}$. The solution is a stochastic process depending on the time $t$ and on the random sample $\omega$ where the argument $\omega$ is usually dropped. The value at fixed time $t$ is a random variable $X(t,\cdot)=X(t)$ - for a fixed realization of the driving Wiener process, the function $X(\cdot,\omega)$ is called a path of the solution. Due to the influence of the Gaussian white noise, typical paths of the solution are rough and nowhere differentiable.

In current chip design one has to deal with a large number of equations as well as of noise sources. Fortunately, the noise intensities are small compared to the other quantities which can be used for the construction of efficient numerical schemes.

The focus here is on efficient numerical methods to simulate sample solution paths [14], i.e., strong approximations of the solution of the arising large systems of SDAEs, since only such paths can reveal the phase noise. The calculation of hundreds or even a thousand solution paths are necessary for getting sufficient numerical confidence about the phase. Moreover, from the solution paths any other statistical data and measurements can be computed in a postprocessing step.

In this paper we present variable step-size two-step methods, in particular stochastic analogues of the trapezoidal rule and the two-step backward differentiation formula, see Section 2. The applied step-size control strategy is described in Section 3. Here we extensively use the smallness of the noise. In Section 4 new ideas for the control both of time and chance discretization are discussed. Test results that illustrate the performance of the presented methods are given in Section 5.

## 2 Adaptive Numerical Methods

The key idea to design methods for SDAEs is to force the iterates to fulfill the constraints of the SDAE at the current time-point. We consider stochastic analogues of methods that have proven very useful in the deterministic circuit simulation. Paying attention to the DAE structure, the discretization of the deterministic part (drift) is implicit, whereas the discretization of the stochastic part (diffusion) is explicit.

We consider stochastic analogues of the variable coefficient two-step backward differentiation formula (BDF$_2$) and the trapezoidal rule, where only the increments of the driving Wiener process are used to discretize the diffusion part. Analogously to the Euler-Maruyama scheme we call such methods multi-step Maruyama methods. The variable step-size BDF$_2$ Maruyama method for the SDAE (2) has the form (see [6] and, for constant step-sizes, e.g. [7])

$$\begin{aligned} A\frac{\alpha_{0,\ell}q(X_\ell)+\alpha_{1,\ell}q(X_{\ell-1})+\alpha_{2,\ell}q(X_{\ell-2})}{h_\ell} &+ \beta_{0,\ell}f(X_\ell,t_\ell) \\ +\alpha_{0,\ell}\sum_{r=1}^{m} g_r(X_{\ell-1},t_{\ell-1})\frac{\Delta W_r^\ell}{h_\ell} &- \alpha_{2,\ell}\sum_{r=1}^{m} g_r(X_{\ell-2},t_{\ell-2})\frac{\Delta W_r^{\ell-1}}{h_\ell} = 0, \end{aligned} \tag{3}$$

$\ell = 2,\dots,N$. Here, $X_\ell$ denotes the approximation to $X(t_\ell)$, $h_\ell = t_\ell - t_{\ell-1}$, and $\Delta W_r^\ell = W_r(t_\ell) - W_r(t_{\ell-1}) \sim N(0,h_\ell)$ on the grid $0 = t_0 < t_1 < \cdots < t_N = T$. The coefficients $\alpha_{0,\ell}, \alpha_{1,\ell}, \alpha_{2,\ell}, \beta_{0,\ell}$ depend on the step-size ratio $\kappa_\ell = h_\ell/h_{\ell-1}$ and satisfy the conditions for consistency of order one and two in the deterministic case. Let the coefficients of the scheme be normalized in such a way that $\alpha_{0,\ell} = 1$ for all $\ell$.

A correct formulation of the stochastic trapezoidal rule for SDAEs requires more structural information (see [8]). It should implicitly realize the stochastic trapezoidal rule for the so called inherent regular SDE of (2) that governs the dynamical components. Both the BDF$_2$ Maruyama method and the stochastic trapezoidal rule of Maruyama type have only an asymptotic order of strong convergence of $1/2$, i.e.,

$$\|X(t_\ell) - X_\ell\|_{L_2(\Omega)} := \max_{\ell=1,\dots,N}(E|X(t_\ell) - X_\ell|^2)^{1/2} \leq c \cdot h^{1/2}, \tag{4}$$

where $h := \max_{\ell=1,\dots,N} h_\ell$ is the maximal step-size of the grid. This holds true for all numerical schemes that include only information on the increments of the Wiener process. However, the noise densities given in Section 1 contain small parameters and the error behaviour is much better. In fact, the errors are dominated by the deterministic terms as long as the step-size is large enough [6,7].

In more detail, the error of the given methods behaves like $O(h^2 + \varepsilon h + \varepsilon^2 h^{1/2})$, when $\varepsilon$ is used to measure the smallness of the noise, i.e., $g_r(x,t) = \varepsilon \hat{g}_r(x,t)$, $r = 1,\dots,m$ where $\varepsilon \ll 1$. Thus we can expect order 2 behaviour if $h \gg \varepsilon$. Higher numerical effort for higher deterministic order pays off only if the noise is *very* small.

## 3 Local Error Estimates

The smallness of the noise allows us to construct special estimates of the local error terms, which can be used to control the step-size. We aim at an efficient estimate of the mean-square of dominating local errors by means of a sufficiently large number of simultaneously computed solution paths. This leads to an adaptive step-size sequence that is identical for all paths. For the drift-implicit Euler-Maruyama scheme this step-size control has been presented in [9], see also [1,4].

In [8, 10] the authors extended this strategy to stochastic linear multi-step methods with deterministic order 2 and provided a reliable error estimate. Let $\widetilde{L}_\ell$ approximate the dominating local error in $Aq(X_\ell)$ by

$$\widetilde{L}_\ell = c_\ell h_\ell \frac{2\kappa_\ell}{\kappa_\ell + 1} \Big[ f(X_\ell, t_\ell) - (\kappa_\ell + 1) f(X_{\ell-1}, t_{\ell-1}) + \kappa_\ell f(X_{\ell-2}, t_{\ell-2}) \Big], \qquad (5)$$

where $c_\ell$ is the error constant of the related deterministic scheme and $\kappa_\ell$ is the step-size ratio. The estimate (5) is based on already computed values of the drift term. Recall that $\widetilde{L}_\ell$ is a vector valued random variable as is the solution $X_\ell$. In dependence on the small parameter $\varepsilon$ and the step-size $h_\ell$ the $L_2$-norm of the local error behaves like $O(h_\ell^3 + \varepsilon h_\ell^{3/2} + \varepsilon^2 h_\ell)$. The term of order $O(h_\ell^3)$ dominates the local error behaviour as long as $h_\ell^3$ is much larger than $\varepsilon h_\ell^{3/2}$, i.e., $\varepsilon^{2/3} \ll h_\ell$. Under this condition also the expression $\|\widetilde{L}_\ell\|_{L_2}$ approximates the local error at time $t_\ell$.

Depending on the available information we will monitor different quantities to satisfy accuracy requirements,

I. control $\|(A + h_\ell \beta_{0,\ell} J_\ell)^{-1} \widetilde{L}_\ell\|_{L_2}$ to match a given tolerance for $X_\ell$,
II. control $\|\widetilde{L}_\ell\|_{L_2}$ to match a given tolerance for $Aq(X_\ell)$, or
III. control $\|A^- \widetilde{L}_\ell\|_{L_2}$ to match a given tolerance for $Pq(X_\ell)$.

Here $J$ is the Jacobian of the drift function $f$ w.r.t. the first variable, and $A^-$ denotes the pseudo inverse of $A$ with $A^- A = P$, where $P$ is a projector onto the dynamic components of $q(X_\ell)$ [11]. Since $(A/h_\ell + \beta_{0,\ell} J_\ell) = 1/h_\ell \cdot (A + h_\ell \beta_{0,\ell} J_\ell)$ is the Jacobian of the discrete scheme (3) this matrix (or a good approximation to it) and its factorization are usually available. In case of $M$ sampled paths, the $L_2$-norm in (I)–(III) is approximated by using the $M$ values $J_\ell^i$ and $\widetilde{L}_\ell^i$ $(i = 1,\dots,M)$ that use values $X_\ell^i$, $X_{\ell-1}^i$, and $X_{\ell-2}^i$ from the $i$th path. For example, in case (I) we use

$$\left\|(A+h_\ell\beta_{0,\ell}J_\ell)^{-1}\widetilde{L}_\ell\right\|_{L_2} \approx \left(\frac{1}{M}\sum_{i=1}^{M}\left|(A+h_\ell\beta_{0,\ell}J_\ell^i)^{-1}\widetilde{L}_\ell^i\right|^2\right)^{1/2} =: \hat{\eta}_\ell\,. \qquad (6)$$

Especially in circuit simulation the different ways of scaling the defect will enable us to control different quantities of the solution. In (I) the local error estimate is used unscaled to match a given tolerance based on a vector representing the charges and the fluxes of the electronic network. Considering the second case (II), the scaled error estimate can be used to match a given tolerance for the solution $X_\ell = (e, j_L, j_V)$ which represent the nodal potentials and some branch currents.

## 4 A Solution Path Tree Algorithm

In the analysis so far, we have considered a constant number $M$ of sample paths. These number influences the approximation of the solution as well as of the mean-square norm in (6). There we make an additional error, the so-called sampling error $\vartheta_\ell$, and the error expansions reads $\|\widetilde{L}_\ell\|_{L_2} = \hat{\eta}_\ell + \vartheta_\ell$, where $\hat{\eta}_\ell$ is the approximation of the dominating local error term based on the sample paths. The idea is to control also the number of sample paths using an estimate of $\vartheta_\ell$. This yields an approximate solution which consists of a tree of paths that is extended, reduced or kept fixed adaptively.

Our aim in tuning the number of paths is to balance the local error and the sampling error. Let $\text{STOL}_\ell$ be the tolerance for the sampling error $\vartheta_\ell$ at time $t_\ell$. One possibility is to calculated this tolerance as an approximation of the higher deterministic error term of order $O(h_\ell^4)$. We then derive the best number $M_\ell$ of paths

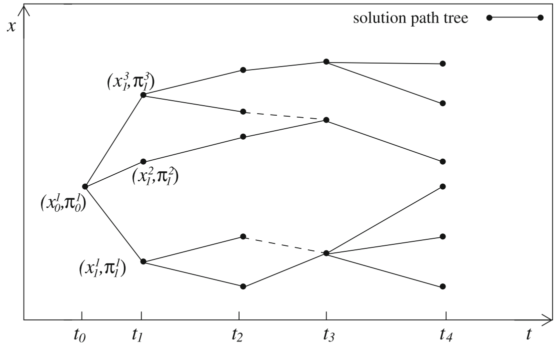


**Fig. 1:** A solution path tree: Variable time-points $t_\ell$, solution states $x_\ell^i$ and path weights $\pi_\ell^i$

by

$$M_\ell = \left\lfloor \frac{1}{\text{STOL}_\ell^2} \frac{\hat{\mu}_\ell^2 \cdot \hat{\sigma}_\ell^2}{\hat{\mu}_\ell^2 + \hat{\sigma}_\ell^2} \right\rfloor , \tag{7}$$

(see [4]), where $\hat{\mu}_\ell$ and $\hat{\sigma}_\ell^2$ are estimates of the mean and the standard deviation of the error estimate at time-point $t_\ell$, respectively. Here $\lfloor x \rfloor$ denotes the smallest integer greater or equal to $x$.

The best number of paths $M_\ell$ depends on the time-point $t_\ell$ and is realized by approximate solutions generated on a tree of paths that is extended, reduced or kept fixed adaptively. In [4, 12] the authors describe the construction of a solution path tree in detail. The method uses probabilities $\pi_\ell^i$ $(\ell = 1,\dots,N;\, i = 1,\dots,M_\ell)$ to weight the solution paths. Figure 1 gives an impression, how a solution path tree looks like. Here the dashed lines indicates the optimal redistribution of the weights after a reduction step (see [4] for a detailed description of the path tree generation).

At each time-step the optimal expansion or reduction problem is formulated by means of combinatorial optimization models. The path selection is modelled as a mass transportation problem in terms of the $L_2$-Wasserstein metric (see [13] in context of scenario reduction in stochastic programming). The algorithm has been implemented in practice. The results presented in the next section show its performance.

## 5 Numerical Results

Here we present numerical experiments for the stochastic BDF$_2$ applied to a test circuit examples. To be able to handle real-life problems, a slightly modified version of the schemes has been implemented in Qimonda's in-house analog circuit simulator TITAN. We consider a model of an inverter circuit with a MOSFET transistor, under the influence of thermal noise. The related circuit diagram is given in Figure 2. The MOSFET is modelled as a current source from source to drain that is controlled by the nodal potentials at gate, source and drain. The thermal noise of the resistor and of the MOSFET is modelled by additional white noise current sources that are shunt

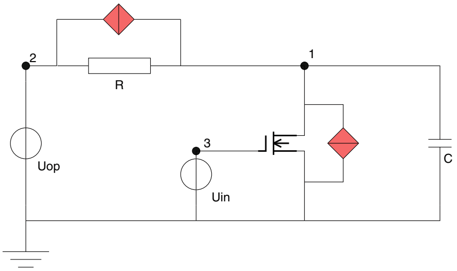


**Fig. 2:** Thermal noise current sources in a MOSFET inverter circuit marked by grey diamonds

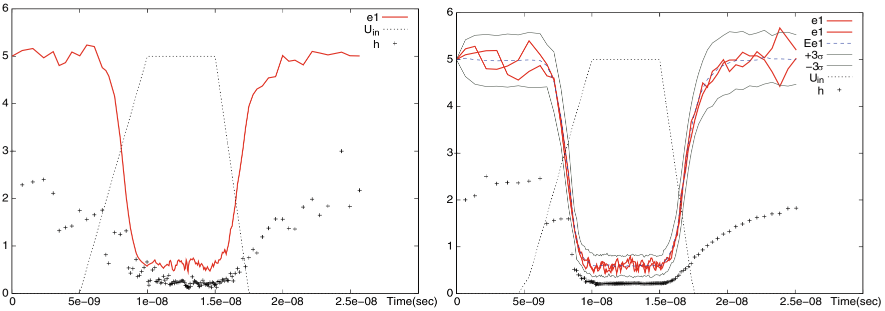


**Fig. 3:** Simulation results for the noisy inverter circuit:
*Left*: 1 path, 127 (+29 rejected) steps; *Right*: 100 paths, 134 (+11 rejected) steps

in parallel to the original, noise-free elements. To highlight the effect of the noise, we scaled the diffusion coefficient by a factor of 1000.

In Figure 3 we present simulation results, where we plotted the input voltage $U_{in}$ and values of the output voltage $e_1$ versus time. Moreover, the applied step-sizes, suitably scaled, are shown by means of single crosses. We compare the results for the computation of a single path (left) with those for the computation of 100 simultaneously computed solution paths (right). The additional solid lines show two arbitrarily chosen solution paths, the dashed line gives the mean of 100 paths and the outer thin lines the $3\sigma$-confidence interval (computed as a statistical estimate for the standard deviation) for the output voltage $e_1$. We observe that using the information of an ensemble of simultaneously computed solution paths smoothes the step-size sequence and considerably reduces the number of rejected steps, when compared to the simulation of a single path. The computational cost that is mainly determined by the number of computed (accepted $+$ rejected) steps is reduced.

Additionally we have applied the solution path tree algorithm to this example. The upper graph in Figure 4 shows the computed solution path tree together with the applied step-sizes which are used simultaneously for all path segments. The lower graph shows the simulation error (solid line), its tolerance (dashed line) and the used number of paths (marked by $\times$), vs. time. Here the tolerance is determined by an approximation of the deterministic local error of order $O(h^4)$ (see [10]) and the maximal number of paths was set to 250. The results indicate that there exists a region from nearly $t = 1 \cdot 10^{-8}$ up to $t = 1.5 \cdot 10^{-8}$ where we have to use much more than 100 paths. This is exactly the area in which the MOSFET is active and the input signal is inverted. Outside this region the algorithm proposes approximately 70 simultaneously computed solution paths.

Especially in circuit simulation the solution path tree algorithm provides an advantage. It helps the designer to identify critical noisy elements of the circuit. In this example the active MOSFET featuring nonlinear noise causes a high fluctuation in the local error estimate whereas the additive noise of the linear resistor behaves harmless.

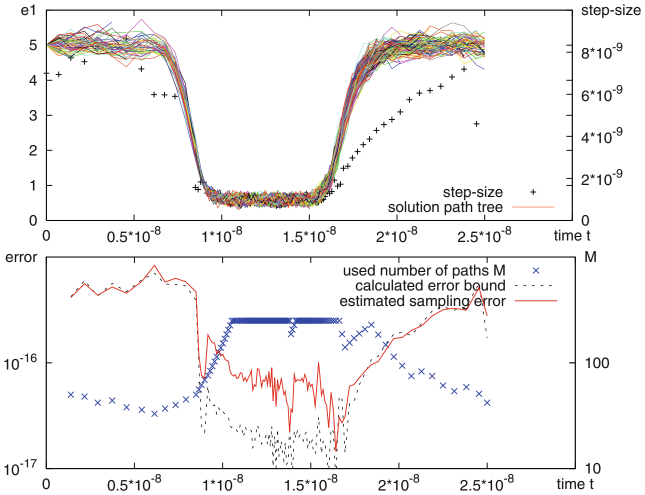


**Fig. 4:** Simulation results for the noisy inverter circuit: Solution path tree and step-sizes (*top*), sampling error, its error bound and the number of paths (*bottom*)

## References

1. Denk, G., Winkler, R.: Modeling and simulation of transient noise in circuit simulation. Math. and Comp. Modelling of Dyn. Systems, **13**(4), 383–394 (2007)
2. Sickenberger, T., Winkler, R.: Adaptive Methods for Transient Noise Analysis. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 161–166. Springer, Berlin (2007)
3. Winkler, R.: Stochastic differential algebraic equations of index 1 and applications in circuit simulation. J. Comput. Appl. Math., **157**(2), 477–505 (2003)
4. Sickenberger, T.: Efficient Transient Noise Analysis in Circuit Simulation. Ph.D. thesis Humboldt Universität zu Berlin, Logos Verlag, Berlin (2008)
5. Arnold, L.: Stochastic differential equations: Theory and Applications. Wiley, New York (1974)
6. Sickenberger, T.: Mean-square convergence of stochastic multi-step methods with variable step-size. J. Comput. Appl. Math., **212**(2), 300–319 (2008)
7. Buckwar, E., Winkler, R.: Multi-step methods for SDEs and their application to problems with small noise. SIAM J. Num. Anal., **44**(2), 779–803 (2006)
8. Sickenberger, T., Weinmüller, E., Winkler, R.: Local error estimates for moderately smooth problems: Part II – SDEs and SDAEs. To appear in BIT Numerical Mathematics
9. Römisch, W., Winkler, R.: Stepsize control for mean-square numerical methods for SDEs with small noise. SIAM J. Sci. Comp., **28**(2), 604–625 (2006)
10. Sickenberger, T., Weinmüller, E., Winkler, R.: Local error estimates for moderately smooth problems: Part I – ODEs and DAEs. BIT Numerical Mathematics, **47**(2), 157–187 (2007)
11. Estévez Schwarz, D., Tischendorf, C.: Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl., **28**, 131–162 (2000)
12. Römisch, W., Sickenberger, Th.: On generating a solution path tree for efficient step-size control. In preparation.
13. Dupačová J., Gröwe-Kuska, N., Römisch, W.: Scenario reduction in stochastic programming. Math. Program., Ser. A **95**, 493–511 (2003)
14. Higham, D.J.: An algorithmic introduction to numerical simulation of stochastic differential equations. SIAM Review, **43**, 525–546 (2001)

# Nonlinear Distortion in Differential Circuits with Single-Ended and Balanced Drive

Timo Rahkonen

**Abstract** This paper illustrates the use of term-wise Volterra analysis tool that can plot both IM3 tone and relevant 2nd order tones as vector sums of all important contributions. As an example the nonlinear distortion behaviour in a fully differential amplifier is studied, when driven either with single-ended or balanced input signal. It is shown that with single-ended drive a small tail-impedance of the differential pair generates 2nd-order distortion into the output of the first stage, and this mixes further to 3rd-order distortion in the 2nd-degree nonlinearity of the second stage.

## 1 Introduction

Nowadays most of the high-performance analog components have a fully balanced structure, i.e. they have differential input and output signals. This is beneficial for the performance, but complicates the design. The circuit now needs impedance matching and stability analysis of both the differential and common mode operation, and also mode conversions from common mode to differential and vice versa may be important. Also the nonlinear distortion behaves differently when driven with differential or single-ended signals. This is relevant from the measurement point of view, too. Many measuring instruments characterize balanced circuits by successive single-ended measurements, and it is important to understand the limitations of such measurements.

This paper aims to show the reasons for different distortion behaviour under different driving conditions. It also illustrates how a combination of vector sum plots of 3rd and 2nd order distortion tones can be used to obtain quite a detailed idea of what is really causing IM3 distortion, and how much of it is mixed from 2nd-order tones.

Timo Rahkonen
Department of Electrical and Information Engineering and Infotech Oulu, University of Oulu, Oulu, Finland, e-mail: timo.rahkonen@oulu.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_23,

# 2 Analysis Techniques

## 2.1 Linear Analysis

Circuit theory includes a lot of tricks for studying balanced circuits. For RF circuits, the formalism was presented by Bockelman and Eisenstadt [1]. They assumed that the balanced 2-port is characterized by single-ended measurements as a linear 4-port, and developed the linear matrix transformations needed to calculate differential and common mode behaviour. This technique is called mixed-mode presentation, and it converts a normal 4-port s-parameter matrix to four 2-port matrices, one describing the truly differential operation, one common mode behaviour, and two remaining ones the mode conversions between these two.

Bockelman presented his formulation to scattering parameters, but the idea can naturally be extended to any linear I-V parameters (see [2]), and all normal design methods can be used — for example, one can pick up the common-mode only presentation, and calculate the stability circles for the common mode matching impedances. However, this analysis technique assumes superposition to be valid and obviously does not operate with non-linear circuits.

## 2.2 Term-Wise Volterra Analysis

Volterra analysis has been used as a quick nonlinear analysis method, and it can also be used to study the different distortion mechanisms ([3], [4], [5]). In this paper, the program described in [6] is used. It is a standard AC Volterra analysis software, that stores all the different contributions of distortion in phasor form. Hence, distortion from each nonlinear device and mixing mechanism can be plotted separately, and their phase relations are immediately visualised. It is also capable of separating mixing results from different harmonic bands, and this makes it possible to track how much of the third-order distortion is actually generated by down or up-conversion from DC or 2nd harmonic bands.

Volterra analysis is based on polynomial device models, where frequency mixing is easy to calculate. The time-domain polynomial can be converted to frequency domain by replacing all signals by spectra and multiplications by convolution. The analysis starts by solving the linear voltages $V_1$ using normal AC analysis. Then, order by order, the already solved lower-order voltages are used to calculate the higher-order distortion currents $I_2, I_3, ..$ of the nonlinear circuit elements, and these are used as excitations to calculate the corresponding nonlinear voltages $V_2, V_3, ...$ Analysis is done in frequency domain, and in narrowband applications the signal spectra can be further expanded to different harmonic bands called baseband, fundamental, and second etc. harmonic bands (BB, FU, H2, H3, ...).

$$\begin{aligned} V &= V_1 + V_2 + V_3 + \ldots \\ V &= V_1 + V_{2BB} + V_{2H2} + V_{3FU} + V_{3H3} + \ldots \end{aligned} \tag{1}$$

Let us substitute this driving voltage $V$ to the frequency-domain version of a polynomial I-V function $i = \Sigma(K_i \cdot v^i)$, where $*$ means spectral convolution.

$$I = K_1 \cdot V + K_2 \cdot (V * V) + K_3 \cdot (V * V * V) + \ldots \tag{2}$$

By separating output currents of different orders $I_1, I_2, I_3$ (and the resulting voltages) we can study these as linear combinations of mixing results from different harmonic bands.

$$\begin{aligned} I_1 &= K_1 \cdot V_1 \\ I_2 &= K_1 \cdot V_2 + K_2 \cdot (V_1 * V_1) \\ I_3 &= K_1 \cdot V_3 + K_3 \cdot (V_1 * V_1 * V_1) + 2K_2 \cdot (V_1 * V_{2BB} + V_1 * V_{2H2}) \end{aligned} \tag{3}$$

Here, the linear terms $K_1 \cdot V_j$ are calculated by using the nonlinear terms as an excitation and solving the node voltages $V_j$. $V_{2BB}$ and $V_{2H2}$ are the second-order DC-band and 2nd harmonic band spectra, respectively. More specifically, signal at $2f_2$ will mix with $-f_1$ to frequency $2f_2 - f_1$, $2f_1$ to $2f_1 - f_2$ and $f_2 - f_1$ and $f_1 - f_2$ to upper and lower IM3, respectively. Altogether, the last row in (3) is the key of this study: The third-order distortion is shown to be caused by cubic nonlinearity $K_3 \cdot v^3$ and by square-law mixing from baseband and second harmonic band.

As an example, let us analyse the lower IM3 tone ($2f_1 - f_2$) at one collector of a BJT differential pair (see Fig. 2) driven by a two-tone test at frequencies $f_1$ and $f_2$. In Fig. 1, voltage phasors at the chosen frequencies ($2f_1 - f_2$ and $f_2 - f_1$) are plotted as vector sums. The naming of the vectors shows the name of the nonlinearity where the distortion current is generated (*gma*, *gmb* etc.), degree of the mixing nonlinearity ($Kxy$ is the coefficient of a $v_i^x v_o^y$ nonlinearity in a 2-dimensional polynomial), and the band where the signal is mixed from ($BB$ or $H2$). For example, $gmbk20V2BB$ is an IM3 term, that is generated by upconverting the $f_2 - f_1$ tone in the driving voltage of *gmb* via its square-law nonlinearity $k20$.

Fig.1a shows how the IM3 tone is built up as a sum of cubic nonlinearity ($gmak30$ and $gmbk30$) and mixing from envelope difference frequency ($gmak20V2BB$ and $gmbk20V2BB$) and 2nd harmonic ($gmak20V2H2$ and $gmbk20V2H2$), as predicted on the last row of (3). The square-law terms arise in the following way: The 2nd-degree nonlinearities of the BJTs generate frequency components at the DC and 2nd harmonic bands. Due to the high-impedance emitter feedback these signals sum up to the common emitter voltage and hence to the BE controlling voltages of the transistors, where they get further mixed to the IM3. This broadband feedback from 2nd-order products converts the expansive exp() shaped nonlinearity of a single BJT to a compressive tanh() shaped nonlinearity of a differential pair. Note, however, that the tanh() response is achieved only by having a high and broadband emitter impedance. Based on Fig.1a we could actually guess, that an inductive tail bias that has low impedance at DC (shorting the strong BB terms) and large impedance at the

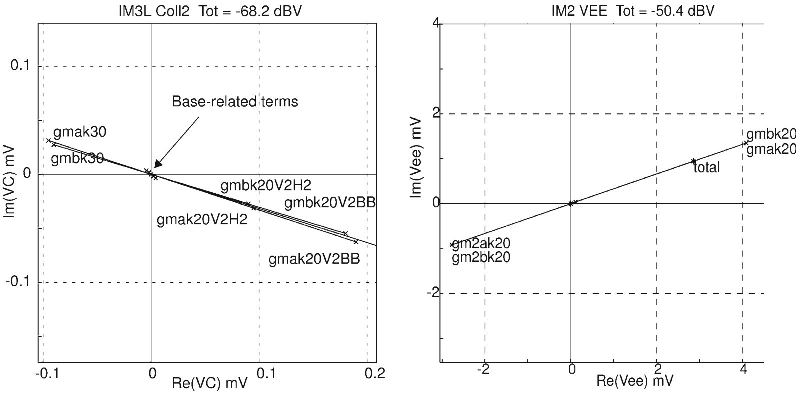


**Fig. 1:** Voltage phasor sums of **a** IM3 tone $2f_1{-}f_2$ at collector of transistor B, **b** envelope tone $f_2{-}f_1$ at the common emitter node

fundamental and 2nd harmonic bands would result in quite nice cancellation of the total IM3.

Studying the plot further, we note that the IM3 currents generated in both devices split 1:1 between devices A and B and sum up coherently in their collectors. The slight difference between *gmak*30 and *gmbk*30 is due to the current lost in the tail impedance $r_o$ (here $10\,\mathrm{M}\Omega$). Next we are most probably interested on what is causing the 2nd-order voltages, and for that purpose similar vector plots can be drawn for the BB and H2 tones $f_2$-$f_1$ and $2f_1$, too. To illustrate this, Fig. 1b shows the construction of the voltage at the common emitter point at the envelope frequency $f_2$-$f_1$. It is seen that the 2nd-order currents of both transistors are in the same phase, and sum up coherently to generate to the 2nd-order voltage seen as the dominant terms in plot a. Terms *gm*2*ak*20 and *gm*2*bk*20 do not appear in a single open-loop differential pair: they are results of opamp feedback, as discussed later.

## *2.3 Computational Complexity*

In general, Volterra analysis is based on linear AC analysis and is very quick to calculate. Compared to AC noise analysis, the number of frequency points is usually smaller but one needs to calculate the response to the controlling ports of all nonlinear sources. For multi-tone analysis, the program [6] calculates the spectral regrowth by numerical convolution, which – utilizing the symmetry of real spectra – is luckily quite easy to calculate. Further, the need to find separate solution of cubic and quadratic terms shown in (3) calls for three successive solutions with the same transfer functions but with different signal amplitudes. As a result, most of the computational time is spent in solving the higher-order node voltages.

## 3 The Balanced Opamp Circuit

As a more complicated example, a fully balanced BJT operational amplifier shown in Fig.2 is analysed to study the differences of single-ended and balanced driving. The mechanism to be illustrated is intuitively the following. It is well known that a common-mode signal easily generates even-order distortion. This can be seen e.g. in (4), that models a bipolar differential pair, where $I_0$ is the tail bias current and $r_o$ is the output impedance of the current source. In balanced mode $v_1 = -v_2$ and the bias current remains constant, resulting in a series expansion with odd powers only. However, with single-ended drive ($v_2 = 0$) the common mode variation also modulates the bias current, causing additional even-order distortion, that is normally not present in the tanh() response.

$$i(t) = (I_0 + \frac{v_1 + v_2}{2r_o}) \cdot \tanh(\frac{v_1 - v_2}{2V_t}) \tag{4}$$

The schematic in Fig.2 shows a resistively loaded differential pair *gma*, *gmb*, *RCA*, *RCB*, frequency compensation $C_c$, and a second stage *gm*2*a*,*gm*2*b*,*Roa*,*Rob*. Many balanced amplifiers have a common-mode control in the 2nd stage, but it was omitted here for simplicity. External feedback is provided by resistors *R*1*a*, *R*1*b*, *R*2*a*, *R*2*b*. In a single-ended mode one of the input signals is zeroed, and the amplitude of the remaining is doubled to maintain the same output amplitude. In the Volterra analysis, the gm elements are treated nonlinear: *gma*,*gmb* are modeled by a 3rd-degree Taylor expansion of the $\exp(v/Vt)$ response, and the 2nd stage device have a mild 2nd-degree nonlinearity but no cubic non-linearity. Also a small exponential nonlinearity was added across the BE junctions of transistors *gma* and *gmb* to model the effect of base currents.

The importance of the tail impedance is illustrated next in Fig. 3, where the amplifier is driven by a single-ended signal, and the tail impedance is varied. Both plots show the vector structure of the envelope signal $f_2 - f_1$ at the collector of transistor *gmB*. In Fig. 3a, the tail impedance is high ($10\,\mathrm{M}\Omega$), and it is seen that (despite of a single-ended drive) the 2nd-order currents *gmak*20 and *gmbk*20 from transistors A and B cancel each other. Hence, there is no net even-order distortion even in a single-ended output. Again, phasors *gm*2*ak*20 and *gm*2*bk*20 are coming from the 2nd stage by coupling from the amplifier outputs to the inputs via the external feedback network.

In plot b the tail impedance $r_o$ is reduced to $1\,\mathrm{k}\Omega$. Now the low tail impedance short-circuits the common emitter node and breaks up the coupling from device A to device B. Hence, at the collector of transistor B the 2nd-order distortion of transistor B only (*gmbk*20) is seen. This ruins the cancellation seen in plot a, resulting in notable amount of envelope and 2nd harmonic voltage at the output of the first stage.

The 2nd-order distortion appearing at the output of the first stage can further mix to IM3 in the quadratic nonlinearity of the 2nd stage. This is illustrated in Fig. 4 that studies the total IM3L in the differential output of the opamp, when the tail impedance is low. In plot 4a the circuit is driven differentially, and the IM2 distortion at the output of the first stage is very low. This is seen as very small IM3 terms

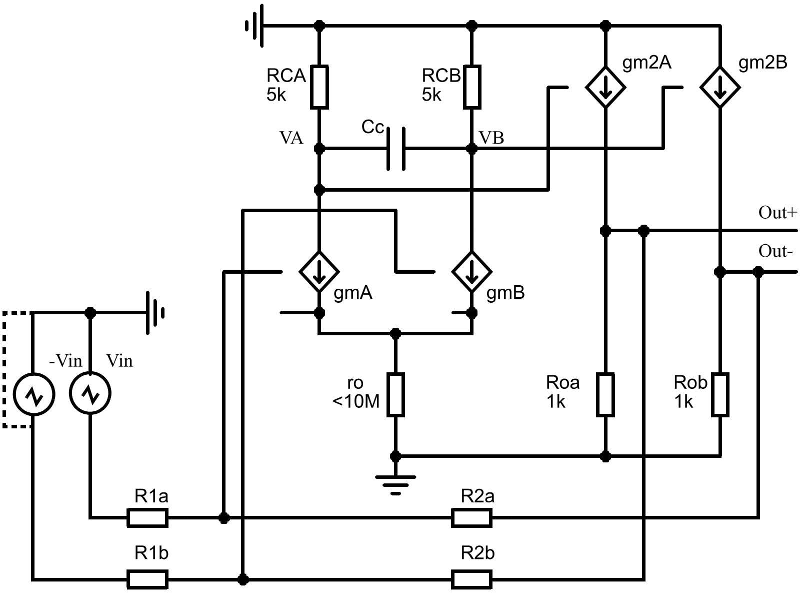


**Fig. 2:** Analysis model of a balanced amplifier

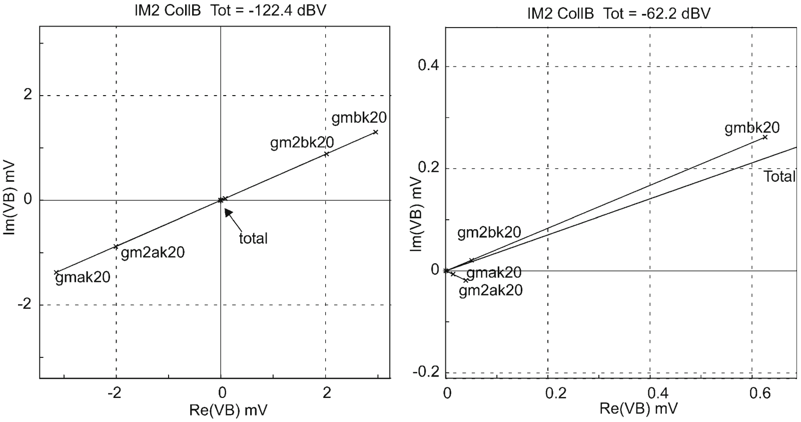


**Fig. 3:** 2nd-order envelope voltage at collector, when the differential pair is driven differentially and $r_o$ is **a** 10 M$\Omega$ and **b** $1k\Omega$

caused by the quadratic nonlinearity of the 2nd-stage transistors $gm2a$ and $gm2b$. When we switch to a single-ended drive in plot b, the BB-signal at the output of the first stage increases (at the simulated frequency, compensation capacitor attenuates the second harmonic) and mixes to IM3 in the 2nd-degree nonlinearity of the second stage. This appears as a strong $gm2aK20VBB$ term. This results in ca. 2 dB higher IM3 with the same output level, as summarised in Table 1.

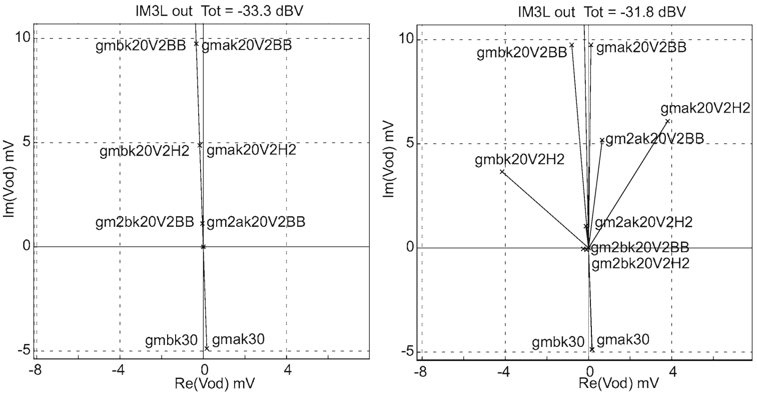


**Fig. 4:** Output IM3, when $r_o$ is low (1 k$\Omega$) and the circuit is driven **a** differentially, **b** single-endedly

**Table 1:** Comparison of different analysis setups (units dBV)

| Setup | In(fund) | Out diff(fund) | Out diff(IM3L) | CollectorB(IM2BB) |
|---|---|---|---|---|
| $r_o$=10 M$\Omega$, s-e | -17.1 | -17.9 | -34.2 | -108.7 |
| $r_o$=10 M$\Omega$, diff | -23.1 | -17.9 | -34.2 | -122.5 |
| $r_o$=1 k$\Omega$, s-e | -17.1 | -17.9 | -32.6 | -64.1 |
| $r_o$=1 k$\Omega$, diff | -23.1 | -17.6 | -34.0 | -73.2 |

# 4 Summary

This paper illustrated the use of an AC Volterra simulator, that handles each distortion contribution separately. It also treats different harmonic bands separately, hence making it possible to separate the amount of IM3 mixing from DC and 2nd harmonic bands. This can be used to find the reasons for frequency dependence (e.g. filtered 2nd harmonic response), or to experiment different linearisation schemes, and is very handy in studying various cancellation effects (or "sweet spots") often seen in the distortion behaviour of analog circuits. Similar precision in recognising the mixing from different frequency bands (but dealing with complete amplifiers and no individual sources within them) is aimed with the use of new X-parameters, described in [7].

To get a complete figure of 3rd-order distortion, it is very helpful to draw the following plots:

1. a vector plot showing the IM3 voltage in the node under study. IM3 will be drawn as a sum of all cubic terms and up and downconverted quadratic terms.
2. vector plots of the difference frequency $f_2 - f_1$ 2nd-order voltages in the controlling nodes of the nonlinearities that had strong K20V2BB terms in plot 1, and

3. similar plots for the relevant products mixing from 2nd harmonic voltages.

The above three types of plots together give a very detailed description of the distortion. Still, some thought may be needed, as the distortion products are already calculated through complete transfer functions, and especially closed feedback loops may cause surprising coupling effects. In the studied circuit, for example, the feedback of the opamp couples all the terms appearing in the output back to the input. In Fig. 1 it would have been technically correct to plot the IM3 at the collector of gmb and then the BB and H2 voltages in the controlling BE junction (and not in the common emitter node alone) of gmb. However, the output-input feedback causes such a cancellation of terms that would have been difficult to explain as a first example.

This paper also explained the differences in IM3 distortion in a fully balanced opamp when driven by single-ended and differential signals. It was illustrated, that the difference comes from the fact, that with a single-ended drive, the amount of 2nd order distortion at the output of the first amplifying stage is higher, and it will mix to IM3 in the quadratic nonlinearity of the 2nd stage. This effect can be minimised by improving the CMRR of the first stage by increasing the tail bias impedance of the differential pair. Interstage mixing can also be reduced by increasing the CMRR or by reducing the 2nd-degree nonlinearity of the second stage.

Note also that many low-voltage RF circuits often employ pseudo-differential circuits, where there is no tail bias at all. Such an amplifier essentially consists of two parallel and decoupled signal paths. As there is no coupling between the parallel amplifiers, their compression behaviour is independent of each other. However, single-ended drive leaves the other path completely idle, hence halving the output compression level.

## References

1. Bockelman, D.E., Eisenstadt, W.R.: Combined Differential and Common-Mode Scattering Parameters: Theory and Simulation. IEEE Trans. Microw. Theory and Techniques, **7/43**, 1530 – 1539 (1995)
2. Rahkonen, T., Kortekangas, J.: Mixed-mode parameter analysis of fully differential circuits. In: Proceedings of the 2004 International Symposium on Circuits and Systems, ISCAS 2004, Vancouver, Canada, May 23-26 2004, vol. 1, pp. 269-272 (2004)
3. Wambacq, P., Sansen, W.: Distortion Analysis of Analog Integrated Circuits. The Springer International Series in Engineering and Computer Science (1998)
4. Peng Li, Pileggi, L.T.: Efficient Per-Nonlinearity Distortion Analysis for Analog and RF Circuits. IEEE Trans. Computer-aided design of Int. Circ. Syst., **10/22**, 1297 – 1309 (2003)
5. Guoji Zhu: Sensitivity Analysis and Distortion Decomposition of Mildly Nonlinear Circuits. M.Sc. thesis, Waterloo University, Waterloo, Canada (2007). URL http://uwspace.uwaterloo.ca/handle/10012/2705
6. Heiskanen, A. Rahkonen, T.: 5th order electro-thermal multi-tone Volterra simulator with component-level output. In: Proceedings of the 2003 International Symposium on Circuits and Systems, ISCAS 2003, Bangkok,Thailand, May 25-28,2003, Vol. 4 pp. 612-615 (2003).
7. Verspecht, J., Root, D.E.: Polyharmonic distortion modelling. IEEE Microwave Magazine., **3/7**, 44 – 57 (2006)

# Evaluation of Oscillator Phase and Frequency Transfer Functions

M.M. Gourary, S.G. Rusakov, S.L. Ulyanov, M.M. Zharov, and B.J. Mulvaney

**Abstract** A general expression for the phase transfer functions of an oscillator for frequencies close to the harmonics of the oscillator fundamental is derived. Numerical testing and comparison with some known results are performed.

## 1 Introduction

The problem of the phase noise analysis of oscillators has been intensively investigated in the last decade. The rigorous theory of the phase noise in oscillators based on the nonlinear perturbation analysis has been developed in [1–5]. White and $f^{-\alpha}$ noise sources are taken into consideration that provides modeling of the thermal, shot and flicker physical noise. However, the design of modern RF circuits often requires the analysis of phase perturbations resulting from substrate/supply noise, which can be approximated by an arbitrary power spectral density (PSD). Furthermore, determining phase perturbations due to a deterministic excitation with a known spectrum can also be considered as a case of the phase noise analysis. These problems can be easily solved if the oscillator is represented by a linear dynamic system with the oscillator phase as the output variable. Then the phase transfer function (TF) in the frequency domain can be defined as a ratio of the phase modulation magnitude to the magnitude of the input sinusoidal excitation.

The linear model for the phase noise analysis based on the intuitive definition of the impulse sensitivity function is proposed in [6]. As shown in [7], this model can be obtained from the nonlinear phase macromodel [3] by ignoring the phase variable

M.M. Gourary, S.G. Rusakov, S.L. Ulyanov, M.M. Zharov
IPPM RAS, 3 Sovetskaya, Moscow, Russia, e-mail: gourary@ippm.ru, rusakov@ippm.ru

B.J. Mulvaney
Freescale Semiconductor Inc., 7700 W. Parmer Lane, Austin, TX, USA, e-mail: brian.mulvaney@freescale.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_24,

on the right-hand side of the nonlinear differential equation of the macromodel. It is demonstrated in [7] that both models predict the same phase noise characteristics.

The phase TF in the Laplace domain obtained by the linear model is derived in [8] in the form of the harmonic transfer matrix (HTM). Each matrix entry $(i,k)$ defines a scalar TF from the sideband frequency of the $k$-th harmonic $(k\omega_0 + \Delta\omega)$ of the excitation to the sideband frequency of the $i$-th harmonic $(i\omega_0 + \Delta\omega)$ of the phase waveform. Here $\omega_0$ is the oscillation frequency, $\Delta\omega$ is the offset frequency. Thus the HTM for the oscillator phase variable is similar to the HTM (representing frequency translation in the periodic AC mode) for the voltage output of a nonlinear circuit. The extension of the HTM definition to the phase output variable seems to be methodologically doubtful because this implies that the phase variation on the period is much less than the period itself. Hence the modulation frequency $i\omega_0 + \Delta\omega$, corresponding to the $i$-th row of the phase HTM, cannot be reasonably interpreted for nonzero $i$. Thus it is desirable to obtain the phase TF through rigorous analysis of the full system of equations of the oscillator circuit without a macromodel approximation.

In this paper we derive a general expression for the phase TF of an arbitrary oscillator by the application of the linear periodically time varying (LPTV) approach in the context of the harmonic balance (HB) technique, similar to periodic AC analysis. First, the analysis of the asymptotic behavior of the solution of the oscillator is performed for excitations with small amplitude and at a frequency close to that of the oscillation frequency. Then the solution transformed to the time domain is compared with the perturbation waveform of the phase modulated signal. This allows us to show that a small sinusoidal excitation produces the sinusoidal phase modulation up to the first order terms of offset frequency. Based on the evaluated magnitude of the phase modulation we derive our expression for the phase TF, including an expression for the instantaneous frequency TF. Numerical experiments performed with SPICE simulation confirm the correctness of the obtained expressions.

## 2 Background

The LPTV analysis is performed after the periodic steady-state (PSS) solution of an oscillator is obtained. The steady-state solution in the frequency domain involves the oscillator fundamental $\omega_0$ and the Fourier coefficients $X$ of the PSS waveforms $x(t)$ [9]. The vector $X$ consists of complex components $X_{kl}$, where $k$ and $l$ are the $k$th harmonic and the $l$th nodal indices, respectively. The LPTV model of the oscillator is similar to the HB system for a forced circuit [10] and we get a perturbation $\Delta X$ to the PSS solution that satisfies

$$J(\Delta\omega)\Delta X = B\,. \tag{1}$$

Here components $B_{kl}$ of the right-hand side (rhs) vector $B$ represent the harmonic signal with frequency $k\omega_0 + \Delta\omega$ applied to the $l$th circuit node. The vector $\Delta X$

defines the small signal solution, and $J(\Delta\omega)$ is a conversion matrix for the given frequency offset $\Delta\omega$,

$$J(\Delta\omega) = G + \mathrm{j}(\Omega + \Delta\omega E)C\,. \tag{2}$$

Here $G$, $C$ are block Toeplitz matrices of the harmonics of nodal conductances and capacitances, respectively. Furthermore $\Omega$ is a block-diagonal matrix of harmonic frequencies $\Omega = diag(\dots, -k\omega_0 E_N, \dots, 0E_N, \dots, k\omega_0 E_N, \dots)$ in which $E_N$ is the $N \times N$ identity matrix, where $N$ is the number of the circuit variables. Finally, $E$ is the identity matrix of the full size. The matrix (2) at zero offset coincides with the HB Jacobian matrix of the free running oscillator at the PSS solution $J_0 = G + j\Omega C$. Thus (2) can be written as follows:

$$J(\Delta\omega) = J_0 + \mathrm{j}\Delta\omega C\,. \tag{3}$$

The Jacobian matrix $J_0$ is singular. Hence there exists a right eigenvector $U$ and a left eigenvector $V$ associated with the 0 eigenvalue such that

$$J_0 U = 0, \quad V^T J_0 = 0\,. \tag{4}$$

The eigenvector $U$ is the HB approximation of the time derivatives of the PSS solution $u = \mathrm{d}x/\mathrm{d}t$ (i.e. the orbital derivative). The eigenvector $V$ is the HB approximation of the solution $v(t)$ of the adjoint system of equations (linearized at the PSS-solution). In [3, 11], $v(t)$ is called the “perturbation projection vector” (PPV). Usually $V$ is normalized such that

$$V^T C U = 1\,. \tag{5}$$

We assume that the kernel of $J_0$ is 1-dimensional. Thus (5) unambiguously defines the unique left eigenvector.

## 3 Asymptotic LPTV Analysis at Small Offset

To obtain the TF it is needed to consider only one nonzero component in the rhs vector. For the component corresponding to harmonic $k$ and node $l$

$$B = e^{(kl)}\,, \tag{6}$$

where the unit vector $e^{(kl)}$ selects component $k$,$l$. To analyze the solution at a small frequency offset we transform (1, 6) into an equivalent linear system with a nonsingular matrix by the method proposed in [12]. First, we form a new equation by left multiplying (1) by the vector $V$. Taking into account (3), (4), and (6), we obtain

$$V^T C \Delta X = \frac{V_{kl}}{\mathrm{j}\Delta\omega}\,. \tag{7}$$

Next, we replace the equation in (1) corresponding to the nonzero component of the excitation (6) by the obtained equation (7). To simplify notation, we assume that the equation to be replaced is the last one in (1). Thus we obtain the transformed equivalent linear system

$$\hat{J}(\Delta\omega)\Delta X = \frac{V_{kl}}{\mathrm{j}\Delta\omega}e^{(kl)} , \tag{8}$$

where $\hat{J}(\Delta\omega) = \begin{bmatrix} \overline{J}(\Delta\omega) \\ V^T C \end{bmatrix}$.

Here the non-square matrix $\overline{J}(\Delta\omega)$ results from $J(\Delta\omega)$ by omitting the last row. One can prove that $\hat{J}(\Delta\omega)$ is non-singular at $\omega = 0$ if $V_{kl} \neq 0$. Hence at a small frequency offset one can neglect the matrix dependence on the offset

$$\hat{J}(\Delta\omega) = \hat{J}_0 + O(\Delta\omega) \approx \hat{J}_0 = \begin{bmatrix} \overline{J_0} \\ V^T C \end{bmatrix} , \tag{9}$$

where $\hat{J}_0 = \hat{J}(0)$, $\bar{J}_0 = \bar{J}(0)$.

Consequently, the solution of (8) can be approximated by a "$1/\omega$" effect [6]

$$\Delta X(\Delta\omega) = \frac{V_{kl}}{\mathrm{j}\Delta\omega}\hat{J}_0^{\,-1} e^{(kl)} . \tag{10}$$

From the definition of $U$ in (4) and $\hat{J}_0$ in (9), we can conclude that the product $\hat{J}_0 U$ contains only one nonzero component corresponding to the replaced row

$$\hat{J}_0 U = V^T C U e^{(kl)} . \tag{11}$$

By the normalization (5) we obtain $\hat{J}_0^{\,-1} e^{(kl)} = U$, and (10) is transformed to

$$\Delta X(\Delta\omega) = \frac{V_{kl}}{\mathrm{j}\Delta\omega}U . \tag{12}$$

Each component $\Delta X_{mn}(\Delta\omega)$ of the vector (12) defines the magnitude of the harmonic $m\omega_0 + \Delta\omega$ at the $n$th node. So the time domain waveform at node $n$ is defined as the sum of all harmonics

$$\Delta x_n(t) = \sum_m \Delta X_{mn} \exp(\mathrm{j}(m\omega_0 + \Delta\omega)t) = \exp(\mathrm{j}\Delta\omega t)\sum_m \Delta X_{mn} \exp(\mathrm{j}m\omega_0 t) . \tag{13}$$

Substitution (12) into (13) yields

$$\Delta x_n(t) = \frac{V_{kl}}{\mathrm{j}\Delta\omega}\exp(\mathrm{j}\Delta\omega t)\sum_m U_{mn} \exp(\mathrm{j}m\omega_0 t) . \tag{14}$$

It has been pointed out that the $U_{mn}$ are the Fourier components of the time derivatives of the PSS solution. Hence the sum term in (14) represents the Fourier series of $\mathrm{d}x_n/\mathrm{d}t$. Thus

$$\Delta x_n(t) = \frac{V_{kl}}{\mathrm{j}\Delta\omega} \frac{\mathrm{d}x_n}{\mathrm{d}t} \exp(\mathrm{j}\Delta\omega t)\ . \tag{15}$$

This expression gives the perturbation waveform at the $n$th node resulting from the unit sinusoidal excitation $b_l = \exp(\mathrm{j}(k\omega_0 + \Delta\omega)t)$.

## 4 Phase Transfer Functions

The TF is determined by applying a sinusoidal excitation with a small magnitude $A$. The corresponding perturbation waveform is defined by (15) with multiplier $A$

$$\Delta x_n^A(t) = A \frac{V_{kl}}{\mathrm{j}\Delta\omega} \frac{\mathrm{d}x_n}{\mathrm{d}t} \exp(\mathrm{j}\Delta\omega t)\ . \tag{16}$$

To obtain the phase TF we also can consider the perturbation waveform due to the phase modulation of the PSS solution $x(t + \phi(t)/\omega_0)$, where $\phi(t)$ in radians is a sine waveform of magnitude $\Phi$

$$\phi(t) = \Phi \exp(\mathrm{j}\Delta\omega t)\ . \tag{17}$$

Assuming that the magnitude $\Phi$ is sufficiently small we can apply the first order Taylor expansion $x_n(t + \phi(t)/\omega_0) \approx x_n(t) + \dot{x}_n\phi(t)/\omega_0$, where $\dot{x}_n = \mathrm{d}x_n/\mathrm{d}t$. Then

$$\Delta x_n^{\Phi}(t) = x_n(t + \frac{\Phi}{\omega_0}\exp(\mathrm{j}\Delta\omega t)) - x_n(t) = \frac{\Phi}{\omega_0} \frac{\mathrm{d}x_n}{\mathrm{d}t} \exp(\mathrm{j}\Delta\omega t)\ . \tag{18}$$

Comparing the perturbation waveform induced by the excitation (16) with the waveform corresponding to the phase modulated PSS solution (18), one can see that the waveforms coincide at $\Phi = A\omega_0 V_{kl}/\mathrm{j}\Delta\omega$. This implies that a small sinusoidal excitation produces the sinusoidal phase modulation (17) with the magnitude $\Phi$ linearly dependent on the input magnitude $A$. Hence we can define the phase TF as $H_{kl}^{\phi} = \Phi/A$ that is evaluated by the expression

$$H_{kl}^{\phi}(\Delta\omega) = \frac{\omega_0}{\mathrm{j}\Delta\omega} V_{kl}\ . \tag{19}$$

The derivation of (19) is based on the non-singularity of the matrix $\hat{J}_0$ (9). It can be shown that $\hat{J}_0$ is singular if and only if $V_{kl} = 0$. In this case we obtain the zero phase TF that seems to be the correct result. Note that the expression for voltage perturbations (16) is incorrect at $V_{kl} = 0$ because the amplitude variations cannot be neglected.

Components (19) form a vector TF for the given node of the excitation. Using this vector TF we can evaluate the phase spectrum for the given spectrum of the excitation (deterministic or stochastic). Note that, unlike the matrix form of the

TF (HTM) [8], the vector TF involves only baseband frequency components of the phase spectrum.

## 5 Frequency Transfer Functions

It is well known that the phase modulated waveform can be considered as a frequency modulated waveform with the instantaneous angular frequency $\omega_{inst}(t)$ defined by the time derivative of $\omega_0 t + \phi(t)$. Thus the deviation of the instantaneous angular frequency from the fundamental for the phase modulation (17) is defined by

$$\Delta\omega_{\text{inst}}(t) = \omega_{\text{inst}}(t) - \omega_0 = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = \mathrm{j}\Delta\omega\Phi\exp(\mathrm{j}\Delta\omega t)\ . \tag{20}$$

It is seen that the frequency deviation (20) is a sine wave with magnitude $\delta\Omega = \mathrm{j}\Delta\omega\Phi$. Defining the frequency TF in the form $H_{kl}^{\omega} = \delta\Omega/A$, we obtain its value from (19, 20) as the transfer factor independent of the offset

$$H_{kl}^{\omega}(\Delta\omega) = \mathrm{j}\Delta\omega H_{kl}^{\phi}(\Delta\omega) = \omega_0 V_{kl}\ . \tag{21}$$

The behavior of the instantaneous frequency (20) can also be described in the following way. Under the excitation of magnitude $A$ the maximum deviation of the instantaneous frequency $\omega_{\text{inst}}$ from the fundamental $\omega_0$ can be presented as

$$\max_t |\omega_{\text{inst}}(t) - \omega_0| = |\delta\Omega| = |H_{kl}^{\omega}|A = \omega_0|V_{kl}|A\ . \tag{22}$$

The analysis of the injection locking mode of an oscillator gives the expression for the locking range [13] $\Delta\omega_{\text{lock}} = \omega_0|V_{kl}|A$ that coincides with the rhs of (22). Therefore

$$\max_t |\omega_{\text{inst}}(t) - \omega_0| = \Delta\omega_{\text{lock}}\ . \tag{23}$$

The evaluation of the TF $H_{kl}^{\phi}$, $H_{kl}^{\omega}$ can be easily implemented in a circuit simulator. The harmonics of the PPV can be obtained by the algorithm [14].

## 6 Numerical Experiments

Numerical experiments were performed with the SPICE transient simulation. The following oscillator circuits were used in numerical experiments: a Colpitts oscillator and a 3-stage CMOS ring oscillator. The oscillator circuits were analyzed under small sinusoidal excitations with frequencies near the oscillator fundamental. The instantaneous frequency was evaluated by postprocessing the output waveforms: $f(t_n) = 1/(x(t_{n+1}) - x(t_n))$. Here $t_n$ are time moments at which the waveform crosses its mean value with $\mathrm{d}x/\mathrm{d}t > 0$. For example, the instantaneous frequency waveforms for a Colpitts oscillator are presented in Fig. 1. Here the amplitude of the excitation was chosen to provide 1% locking range. Hence the offset 0.98% corresponds to the locking mode of the oscillator. Note that after the transient died out the waveform is a constant approximating the locking range. The offset 3% produces a sine waveform with a maximum deviation equal to the locking range. With a 10% offset value the maximum deviation is less than the locking range.

The results of the SPICE simulation along with theoretical evaluations by (23) are shown in Fig. 2- 3. One can see the dependency of the maximum relative deviation of the instantaneous frequency ($\frac{\max|\omega_{inst}(t)-\omega_0|}{\omega_0}$) as a function of the relative offset frequency ($\frac{\Delta\omega}{\omega_0}$).

At a small frequency offset the oscillator is locked, and the oscillation frequency is equal to the excitation frequency (sloping line in the figures). When the excitation frequency is outside the locking range the deviation, in accordance with (23), keeps a constant value. At larger frequency offsets expression (23) obtained from an asymptotic solution of the LPTV model is not valid. This explains the fact that simulated curves do not correspond with (23) at larger frequency offsets.

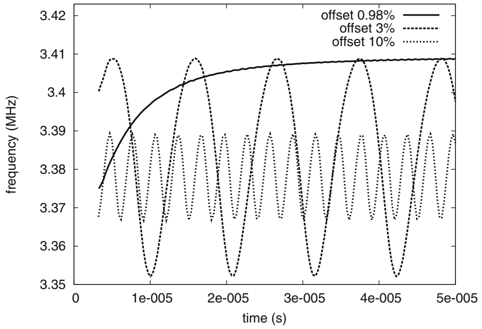


**Fig. 1** Example of the instantaneous frequency waveforms for the Colpitts oscillator obtained by postprocessing the simulated voltage waveforms. The oscillation frequency is 3.376 MHz

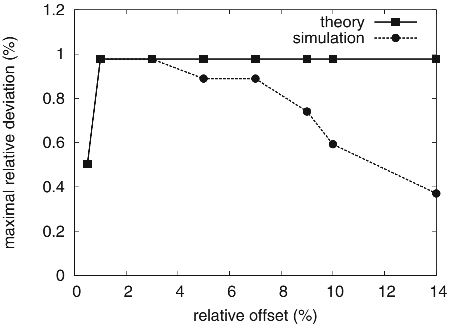


**Fig. 2** Dependency of the maximum frequency deviation on the input frequency offset for the Colpitts oscillator. The oscillation frequency is 3.376 MHz. The input amplitude is defined to provide a 1% locking range

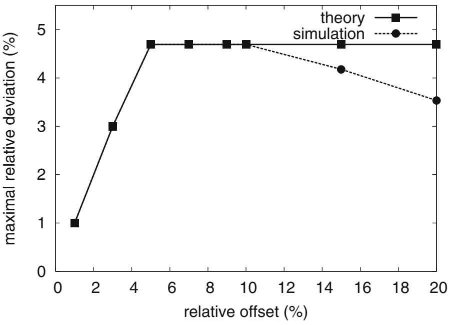


**Fig. 3** Dependency of the maximum frequency deviation on the input frequency offset for a CMOS ring oscillator. The oscillation frequency is 1.556 GHz. The input amplitude is defined to provide a 5% locking range

## References

1. Kaertner, F.X.: Determination of the correlation spectrum of oscillators with low noise. IEEE Trans. Microwave Theory Tech. **37**, 90–101, January (1989)
2. Kaertner, F.X.: Analysis of white and $f^{-\alpha}$ noise in electrical oscillators. Int. J. Circ. Theory Appl. **18**, 485–519, (1990)
3. Demir, A., Mehrotra, A., Roychowdhury, J.: Phase Noise in oscillators: A unifying theory and numerical methods for characterization. IEEE Trans. on Circuits and Systems – I **47**, 655–674, May (2000)
4. Demir, A.: Phase noise and timing jitter in oscillators with colored-noise sources. IEEE Trans. on Circuits and Systems-I: Fund. Theory and Applics. **49**(12), 1782–1791 (2002)
5. Maffezzoni, P.: Frequency-shift induced by colored noise in nonlinear oscillators, IEEE Trans. on Circuits and Systems-II: Express Briefs, **54**(16), 887–891 (2007)
6. Hajimiri, A.: A general theory of phase noise in electrical oscillators, IEEE J. of Solid-State Circuits **33**(2), 179–194 (1998)
7. Vanassche, P., Gielen, G., Sansen, W.: On the difference between two widely publicized methods for analyzing oscillator phase behavior. In: Proc. Int. Conf. Computer-Aided Design, San Jose, pp. 229–233 (2003).
8. Vanassche, P., Gielen, G., Sansen, W.: Time-varying, frequency-domain modeling and analysis of phase-locked loops with sampling phase-frequency detectors. In: Proc. Design, Automation and Test in Europe Conf., Munich, pp. 238–243 (2003)
9. Kundert, K.S., White, J.K., Sangiovanni-Vincentelli, A.: Steady-state methods for simulating analog and microwave circuits, Kluwer, Boston (1990)
10. Rizzoli, V., Mastri, F., Masotti, D.: General noise analysis of nonlinear microwave circuits by the piecewise harmonic balance technique. IEEE Trans. Microwave Theory Tech. **42**, 807–819, May (1994)
11. Günther, M., Feldmann, U., ter Maten, J.: Modelling and discretization of circuit problems, In: W.H.A. Schilders, E.J.W. ter Maten (Guest Eds), Handbook of Numerical Analysis, Vol. XIII, Special Volume on Numerical Methods in Electromagnetics, Elsevier North-Holland, pp. 523–658 (2005)
12. Gourary, M.M., Rusakov, S.G., Ulyanov, S.L., Zharov, M.M., Mulvaney, B.J., Gullapalli K.K.: New numerical technique for cyclostationary noise analysis of oscillators. In: Proc. of the 37th European Microwave Conference, Munich, pp. 1173–1176, October (2007).
13. Gourary, M.M., Rusakov, S.G., Ulyanov, S.L., Zharov, M.M., Mulvaney, B.J.: Analysis of oscillator injection locking by harmonic balance method. In: Proc. of Design, Automation and Test in Europe Conf., Munich, pp. 318–323, March (2008)
14. Demir, A., Long, D., Roychowdhury, J.: Computing phase noise eigenfunctions directly from steady-state Jacobian matrices. In: Int. Conf. Computer-Aided Design, San Jose, pp. 283–288, Nov. (2000)

# Polynomial Chaos for the Computation of Failure Probabilities in Periodic Problems

Roland Pulch

**Abstract** Numerical simulation of electric circuits uses systems of differential algebraic equations (DAEs) in general. We examine forced oscillators, where the DAE models involve periodic solutions. Uncertainties in physical parameters can be described by random variables. We apply the strategy of the generalised polynomial chaos (gPC) to resolve the stochastic model. In particular, failure probabilities are determined using the approximation from gPC. We present results of numerical simulations for a system of DAEs modelling a Schmitt trigger.

## 1 Introduction

Mathematical modelling of electric circuits yields time-dependent systems of ordinary differential equations (ODEs) or differential algebraic equations (DAEs), see [1]. The solutions consist of unknown node voltages and branch currents. Typically, the systems include many physical parameters like capacitances, inductances, resistances, etc. Assuming some uncertainties, we replace several parameters by random variables. Accordingly, the solution of the DAEs becomes a random process. The generalised polynomial chaos (gPC) provides techniques for solving the stochastic model approximately, see [2,3].

We consider forced oscillators, where a periodic boundary value problem of the DAEs results for each realisation of the parameters. A Galerkin approach yields a larger coupled system of DAEs for the finite representation in the polynomial chaos. Thus a periodic boundary value problem of the larger system has to be solved, which can be done by well-known techniques like shooting methods, finite difference schemes or harmonic balance, cf. [4]. Previous work on periodic problems of

Roland Pulch
Lehrstuhl für Angewandte Mathematik und Numerische Mathematik, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany, e-mail: pulch@math.uni-wuppertal.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_25,

ODEs or DAEs using the strategy of gPC is given in [5–7]. Moreover, a numerical solution of the coupled system from gPC can be used to determine failure probabilities of the problem, see [2, 8].

In this article, we consider the circuit of a Schmitt trigger, which converts an analogue input signal into a digital output signal. The mathematical model that is used represents the circuit as a system of DAEs. Assuming a random capacitance, we solve numerically the periodic problem of the corresponding gPC system from the Galerkin method, i.e., one random parameter appears. An according strategy and simulations with several random parameters are presented for ODEs in [5, 6]. We compute failure probabilities with respect to the behaviour of the output signal based on the gPC approximation using a common approach. Thereby, failure means that some result exceeds a reference value, which can be determined in a post-processing of a time integration. The numerical results illustrate the performance of the gPC expansions.

## 2 Problem Definition

We consider a system of DAEs in the form

$$A(\mathbf{p})\dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{f}(t,\mathbf{x}(t,\mathbf{p}),\mathbf{p}). \tag{1}$$

The matrix $A \in \mathbb{R}^{n\times n}$ and the right-hand side $\mathbf{f} : [t_0,t_1] \times \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}^n$ depend on parameters $\mathbf{p} = (p_1,\ldots,p_q)^\top$. Hence the solution $\mathbf{x} : [t_0,t_1] \times \mathbb{R}^q \to \mathbb{R}^n$ becomes also parameter-dependent. Let $\mathbf{p} \in Q$ for some relevant set $Q \subseteq \mathbb{R}^q$ of parameters. Typically, a parameter $p_j$ is included either in the matrix $A$ or in the right-hand side $\mathbf{f}$.

We assume that independent input signals in the right-hand side force a periodic solution for each parameter. Thus it holds $\mathbf{x}(t,\mathbf{p}) = \mathbf{x}(t+T,\mathbf{p})$ for all $t \in \mathbb{R}$ and each $\mathbf{p} \in Q$. where the period $T > 0$ is known from the input signals. We set $[t_0,t_1] = [0,T]$ in the following.

Let the chosen parameters exhibit some uncertainty. Consequently, we arrange random variables $\mathbf{p} : \Omega \to Q$ with respect to a probability space $(\Omega,\mathscr{A},P)$. We assume that each random variable $p_j$ exhibits a classical distribution like uniform, beta, Gaussian, etc. Consequently, the solution of the DAEs (1) becomes a random process $\mathbf{X} : [0,T] \times \Omega \to \mathbb{R}^n$. We are interested in the properties of the random process like expected values and variances or more sophisticated quantities. In particular, we will investigate failure probabilities in Section 4.

For a function $f : \mathbb{R}^q \to \mathbb{R}$ depending on the parameters, we denote the corresponding expected value (if exists) by

$$\langle f(\mathbf{p})\rangle = \int_\Omega f(\mathbf{p}(\omega))\;\mathrm{d}P(\omega) = \int_{\mathbb{R}^q} f(\mathbf{p})\rho(\mathbf{p})\;\mathrm{d}\mathbf{p}$$

using the probability density function $\rho : \mathbb{R}^q \to \mathbb{R}$. We apply this operation component-wise also to vector-valued or matrix-valued functions. The expected value implies the inner product $\langle f(\mathbf{p})g(\mathbf{p})\rangle$ for functions $f,g : \mathbb{R}^q \to \mathbb{R}$ with $f,g \in L^2$.

## 3 Generalised Polynomial Chaos

Now we examine the stochastic process solving the DAEs (1) with random parameters. Assuming finite second moments, the stochastic process exhibits the representation

$$\mathbf{X}(t,\mathbf{p}(\omega)) = \sum_{i=0}^{\infty} \mathbf{v}_i(t)\Phi_i(\mathbf{p}(\omega)), \tag{2}$$

see [2]. The functions $(\Phi_i)_{i\in\mathbb{N}}$ with $\Phi_i : \mathbb{R}^q \to \mathbb{R}$ represent a complete basis of multivariate polynomials. We apply an orthonormal basis, i.e., it holds $\langle \Phi_i \Phi_j \rangle = \delta_{ij}$ with the Kronecker-delta. The coefficient functions $\mathbf{v}_i : [0,T] \to \mathbb{R}^n$ are unknown a priori. The periodicity of the stochastic process $\mathbf{X}$ implies periodic coefficient functions with rate $T$, see [7].

We truncate the series (2) to achieve the finite representation

$$\mathbf{X}^m(t,\mathbf{p}(\omega)) = \sum_{i=0}^{m} \mathbf{v}_i(t)\Phi_i(\mathbf{p}(\omega)). \tag{3}$$

Approximations for expected values and variances are obtained component-wise by

$$\langle X_j^m(t,\mathbf{p})\rangle = v_{0,j}(t), \quad \mathrm{Var}(X_j^m(t,\mathbf{p})) = \sum_{i=1}^{m} v_{i,j}(t)^2 \qquad \text{for } j = 1,\ldots,n \tag{4}$$

with $\mathbf{X}^m = (X_1^m,\ldots,X_n^m)^\top$ and $\mathbf{v}_i = (v_{i,1},\ldots,v_{i,n})^\top$. The coefficient functions can be computed approximately by stochastic collocation, see [3, 9]. Alternatively, we construct a system of DAEs for the coefficient functions. Inserting the finite approximation (3) in the DAEs (1) yields the residual

$$\mathbf{r}(t,\mathbf{p}) \equiv A(\mathbf{p}) \left( \sum_{i=0}^{m} \dot{\mathbf{v}}_i(t)\Phi_i(\mathbf{p}) \right) - \mathbf{f}\left( t, \sum_{i=0}^{m} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \mathbf{p} \right).$$

Due to the Galerkin method, we demand that the residual is orthogonal to the space of the applied basis polynomials with respect to the inner product of $L^2$ in the probability space. It follows a larger coupled system of DAEs

$$\sum_{i=0}^{m} \langle \Phi_l(\mathbf{p})\Phi_i(\mathbf{p})A(\mathbf{p})\rangle \dot{\mathbf{v}}_i(t) = \left\langle \Phi_l(\mathbf{p})\, \mathbf{f}\left( t, \sum_{i=0}^{m} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \mathbf{p} \right) \right\rangle \tag{5}$$

for $l = 0,1,\ldots,m$ with the coefficient functions $\mathbf{v}_i(t)$ of (3) as unknowns. Although the solutions of (5) are not identical to the coefficients in (2), we apply the same symbol for convenience. A periodic boundary value problem of the system (5) has to be solved, which can be done by according numerical methods, see [4].

In models of electric circuits, the matrix $A$ typically includes parameters like capacitances $C$ and inductances $L$, for example. We assume the structure of parameter-dependence of the matrix $A$ motivated in [10], namely

$$A(\mathbf{p}) = A_0 + \sum_{j=1}^{q} \eta_j(p_j) A_j$$

with constant matrices $A_0, A_1, \ldots, A_q \in \mathbb{R}^{n \times n}$ and scalar functions $\eta_j : \mathbb{R} \to \mathbb{R}$. We can write the complete system using Kronecker products as

$$\left[ I_{m+1} \otimes A_0 + \left( \sum_{j=1}^{q} S_j \otimes A_j \right) \right] \dot{\mathbf{v}}(t) = \mathbf{F}(t, \mathbf{v}) \tag{6}$$

with $\mathbf{v} = (\mathbf{v}_0^\top, \mathbf{v}_1^\top, \ldots, \mathbf{v}_m^\top)^\top$, the identity matrix $I_{m+1} \in \mathbb{R}^{(m+1)\times(m+1)}$ and an abbreviation $\mathbf{F}$ for the right-hand side. The matrices $S_j$ are defined via

$$S_j = (\sigma_{li}^j) \in \mathbb{R}^{(m+1)\times(m+1)}, \quad \sigma_{li}^j := \langle \eta_j(p_j) \Phi_i(\mathbf{p}) \Phi_l(\mathbf{p}) \rangle.$$

In case of a single parameter and $\eta_1(p_1) \equiv p_1$, the constant matrix in the left-hand side of (6) becomes block-tridiagonal, since the matrix $S_1$ is tridiagonal due to the orthogonality of the basis polynomials.

## 4 Determination of Failure Probabilities

If the solution of the DAEs (1) exhibits specific critical values, the corresponding electric circuit may produce a failure. We describe the state of the solution via a function $g : [t_0, t_1] \times \mathbb{R}^n \to \mathbb{R}$, where $g \leq 0$ represents the undesired cases. For example, we define as failure that a component $x_j$ for a particular $j \in \{1, \ldots, n\}$ becomes smaller or larger than some threshold value $\theta \in \mathbb{R}$, i.e.,

$$g(t, \mathbf{x}(t, \mathbf{p})) \equiv x_j(t, \mathbf{p}) - \theta \quad \text{or} \quad g(t, \mathbf{x}(t, \mathbf{p})) \equiv -x_j(t, \mathbf{p}) + \theta. \tag{7}$$

In the general case, the failure probability at each time point reads

$$P_F(t) := \int_{\mathbb{R}^q} \chi(g(t, \mathbf{X}(t, \mathbf{p}))) \rho(\mathbf{p}) \, \mathrm{d}\mathbf{p} \qquad \text{with} \quad \chi(g) := \begin{cases} 0 \text{ for } g > 0, \\ 1 \text{ for } g \leq 0. \end{cases} \tag{8}$$

In a Monte-Carlo or quasi Monte-Carlo simulation, the integrals (8) are approximated using realisations $\mathbf{p}^k \in Q$ for $k = 1, \ldots, K$. For each realisation, a periodic boundary value problem of the DAEs (1) has to be solved. Alternatively, we apply the solution of the system (5) from the gPC. The computation of this solution can be more costly than the (quasi) Monte-Carlo simulation with same accuracy.

Nevertheless, the gPC solution includes more information and may be available from a previous simulation for another purpose. We insert the approximation (3) in the integral (8) and thus obtain

$$P_F(t) \doteq \int_{\mathbb{R}^q} \chi(g(t,\mathbf{X}^m(t,\mathbf{p})))\rho(\mathbf{p})\,\mathrm{d}\mathbf{p}. \tag{9}$$

The formulation (9) can be evaluated by (quasi) Monte-Carlo sampling again. Given a numerical solution for the coefficients $\mathbf{v}_0,\ldots,\mathbf{v}_m$, just polynomials have to evaluated in an approximation of (9), i.e., no further DAE systems have to be resolved, since we apply $\mathbf{X}^m(t,\mathbf{p})$ instead of $\mathbf{x}(t,\mathbf{p})$. Sophisticated techniques have been constructed for this purpose in case of parameters $\mathbf{p}$ with Gaussian distributions and/or small failure probabilities, see [2,8].

We consider w.l.o.g. the first case in (7). Since we examine periodic boundary value problems, the total probability of failure $\hat{P}_F \in [0,1]$ corresponds to the time-independent function

$$g(\mathbf{x}(\cdot,\mathbf{p})) = \left(\min_{t\in[0,T]} x_j(t,\mathbf{p})\right) - \theta. \tag{10}$$

Typically, this probability is computed by a discretisation $0 \le t_1 < \cdots < t_R < T$ and identification of the minimum value in the grid points.

## 5 Illustrative Example: Schmitt Trigger

We apply the circuit of a Schmitt trigger illustrated in Figure 1. The Schmitt trigger converts an analogue input signal $u_{\mathrm{in}}$ into a digital output signal $u_{\mathrm{out}}$. A mathematical modelling yields a system of DAEs (1) for five unknown node voltages with differential index 1, see [1]. More precisely, the system exhibits the form

$$A(C)\dot{\mathbf{u}} = \mathbf{f}(t,\mathbf{u}), \qquad \mathbf{u} : [t_0,t_1] \to \mathbb{R}^5.$$

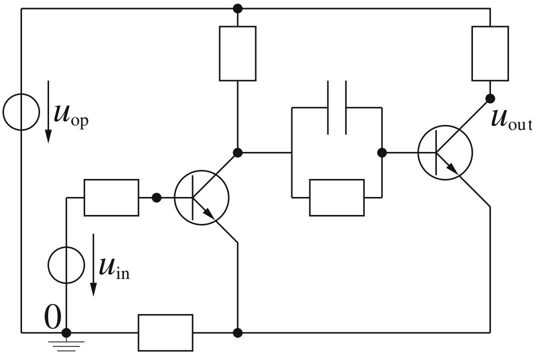
Figure 1 also shows the singular matrix $A$, which depends on the linear capacitance $C$ only. We use a sinusoidal input signal with period $T = 2$ ms. Thus all node voltages become periodic functions.

Let the capacitance be a random variable with uniform distribution in a certain interval. We consider two cases in the simulation, namely

$$\text{case (a)}: \;\; C \in [10^{-9}\,\mathrm{F}, 10^{-7}\,\mathrm{F}], \quad \text{case (b)}: \;\; C \in [1\cdot 10^{-10}\,\mathrm{F}, 2\cdot 10^{-10}\,\mathrm{F}].$$

The first case involves large uncertainties for demonstration and corresponds to the results displayed in Figure 2 and 3. The second case is more realistic and serves for the computation of failure probabilities only.

We solved all periodic boundary value problems via a finite difference method, see [4], using the unsymmetric difference formula of second order (BDF2) at



$$A(C) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & C & 0 & -C & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -C & 0 & C & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

**Fig. 1:** Schmitt trigger circuit (*left*) and capacitance matrix of the mathematical model (*right*)

equidistant distributed time points. Thereby, the same accuracy was demanded in each Newton iteration and we arranged 200 grid points.

We employ the strategy of gPC based on the representation (2), where the orthonormal basis functions are the Legendre polynomials in case of the uniform distribution. We discuss the periodic problems of the coupled systems of DAEs (5) for different orders $m$.

Figure 2 demonstrates the expected value and the standard deviation of the output voltage in case (a) with $m = 3$ calculated via (4). Moreover, three samples of the output voltage for specific values of the capacitance are given. Figure 3 shows the other coefficient functions of (3). Although the solutions are computed in $[0,T]$ only, the figures show the domain $[0,2T]$ for a better impression of the signals.

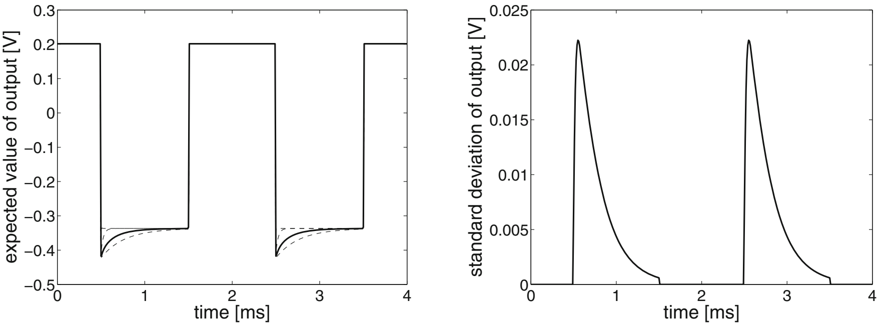


**Fig. 2:** Expected value (*left*, *solid line*) together with three samples for $C = 10^{-j}$ F with $j = 7,8,9$ (*left*, *dashed lines*) and standard deviation (*right*) of output $u_{\text{out}}$ computed by gPC with $m = 3$

In the case (a), we recognise that variations in the capacitance do not influence the upper value of the digital output signal. In particular, the standard deviation evidences the critical time intervals. Using $C = 10^{-10}$ F of case (b), the behaviour at

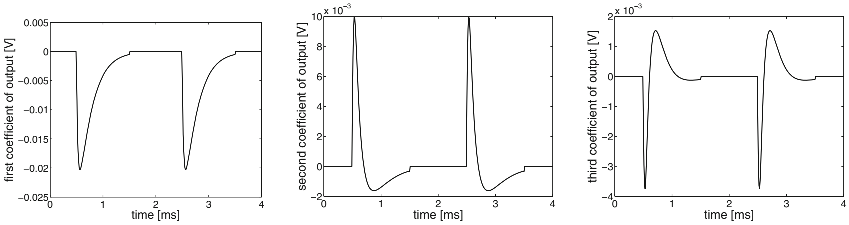


**Fig. 3:** Coefficient functions $v_1, v_2, v_3$ for output voltage $u_{\mathrm{out}}$ obtained by gPC with $m = 3$

the lower value becomes the same as at the upper value, which represents the desired behaviour. An overshoot appears for larger parameters $C > 10^{-10}$ F. However, this effect decreases again for even larger capacitances $C > 10^{-7}$ F.

To illustrate the convergence of the periodic coefficient functions in (2), we compute the corresponding maximal values within $[0,T]$ for a simulation using $m = 8$. Table 1 presents these maxima with respect to the output signal in case (a) as well as case (b). The other components exhibit a similar behaviour. We recognise the convergence of the gPC representation (2) in both situations. However, case (a) implies a much slower convergence due to the large range of the random parameter.

Next, failure probabilities are determined in this example. We demand that the periodic output voltage must not decrease below some threshold value, which corresponds to the definition (10). We arrange the threshold values $\theta = -0.415$ for case (a) and $\theta = -0.34$ for case (b). The corresponding total failure probability $\hat{P}_F$ is determined by the values in the grid points.

For large numbers of random parameters, (quasi) Monte-Carlo methods have to be used in solving (8). Since one random parameter is considered here (only $C$), we apply equidistant realisations $C_k$ for $k = 1, \ldots, K$, which represents the special case of a quasi Monte-Carlo technique. On the one hand, a reference solution of (8) is computed by solving $K = 10^4$ systems (1). On the other hand, we sample approximations (9) with $K = 10^3$ using solutions of (5) with different numbers $m$. Remark that the probability in (8) and (9) does not depend on time here, since (10) is observed. The results are shown in Table 2. We note that the approximation becomes more accurate for increasing orders $m$. In case (b), the usage of $m = 2$ already yields a sufficient result. However, a linear approximation ($m = 1$) is too rough, which indicates that the application of the gPC as a nonlinear approach is necessary.

**Table 1:** Maximum values of coefficients $v_i$ for output $u_{\mathrm{out}}$ in gPC simulation using $m = 8$

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Case (a) | $2 \cdot 10^{-2}$ | $1 \cdot 10^{-2}$ | $5 \cdot 10^{-3}$ | $3 \cdot 10^{-3}$ | $2 \cdot 10^{-3}$ | $1 \cdot 10^{-3}$ | $6 \cdot 10^{-4}$ | $3 \cdot 10^{-4}$ |
| Case (b) | $9 \cdot 10^{-4}$ | $1 \cdot 10^{-5}$ | $2 \cdot 10^{-7}$ | $5 \cdot 10^{-9}$ | $1 \cdot 10^{-10}$ | $2 \cdot 10^{-12}$ | $5 \cdot 10^{-14}$ | $1 \cdot 10^{-14}$ |

**Table 2:** Computed total failure probabilities from gPC system with $m = 1,\dots,8$ and reference solution from solving the original systems

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ref. |
|---|---|---|---|---|---|---|---|---|---|
| Case (a) | 0.940 | 0.759 | 0.775 | 0.802 | 0.821 | 0.830 | 0.830 | 0.820 | 0.8177 |
| Case (b) | 0.082 | 0.079 | 0.079 | 0.079 | 0.079 | 0.079 | 0.079 | 0.079 | 0.0786 |

# 6 Conclusions

We have applied the technique of the generalised polynomial chaos to periodic boundary value problems of DAEs with time-dependent input signals. The resulting larger coupled systems of DAEs are solved successfully for the electric circuit of a Schmitt trigger. Moreover, the computed numerical solution provides a cheap method to determine failure probabilities a posteriori. In the used examples, it follows that the accuracy of the achieved failure probabilities is adequate if the order of the polynomial chaos is chosen sufficiently high. The construction of techniques based on generalised polynomial chaos is feasible also for autonomous oscillators with a priori unknown periods, which will be part of further research.

**Acknowledgements** This work has been supported by the PostDoc programme of ‘Fachgruppe Mathematik und Informatik’ from Bergische Universität Wuppertal (Germany). The author is indebted to Prof. Robert D. Skeel (Purdue University, Indiana, U.S.A.) for helpful discussions.

# References

1. Kampowsky, W., Rentrop, P., Schmitt, W.: Classification and numerical simulation of electric circuits. Surv. Math. Ind. 2, 23–65 (1992).
2. Augustin, F., Gilg, A., Paffrath, M., Rentrop, P., Wever, U.: Polynomial chaos for the approximation of uncertainties: chances and limits. Euro. Jnl. of Applied Mathematics 19, 149–190 (2008).
3. Xiu, D.: Fast numerical methods for stochastic computations: a review. Comm. Comput. Phys. 5 (2-4), 242–272 (2009).
4. Günther, M., Feldmann, U., ter Maten, E.J.W.: Modelling and discretization of circuit problems. In: Schilders, W.H.A., ter Maten, E.J.W. (eds.), Handbook of Numerical Analysis, Vol. XIII: Numerical Methods in Electromagnetics, pp. 523-659, Elsevier, N. Holland (2005).
5. Lucor, D., Karniadakis, G.E.: Adaptive generalized polynomial chaos for nonlinear random oscillators. SIAM J. Sci. Comput. 26 (2), 720–735 (2004).
6. Lucor, D., Su, C.H., Karniadakis, G.E.: Generalized polynomial chaos and random oscillators. Int. J. Numer. Meth. Engng. 60, 571–596 (2004).
7. Pulch, R.: Polynomial chaos for analysing periodic processes of differential algebraic equations with random parameters. To appear in: Proc. Appl. Math. Mech.
8. Paffrath, M., Wever, U.: Adapted polynomial chaos expansion for failure detection. J. Comput. Phys. 226, 263–281 (2007).
9. Xiu, D., Hesthaven, J.S.: High order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. 27 (3), 1118–1139 (2005).
10. Pulch, R.: Polynomial chaos for linear DAEs with random parameters. Preprint, Bergische Universität Wuppertal (2008).

# Quasiperiodic Steady-State Analysis of Electronic Circuits by a Spline Basis

Hans Georg Brachtendorf, Angelika Bunse-Gerstner, Barbara Lang, Siegmar Lampe, and Ashish Awasthi

**Abstract** Multitone Harmonic Balance (HB) is widely used for the simulation of the quasiperiodic steady-state of RF circuits. HB is based on a Fourier expansion of the waveforms. Unfortunately, trigonometric polynomials often exhibit poor convergence properties when the signals are not quasi-sinusoidal, which leads to a prohibitive run-time even for small circuits. Moreover, the approximation of sharp transients leads to the well-known Gibbs phenomenon, which cannot be removed by an increase of the number of Fourier coefficients, because convergence is only guaranteed in the $L_2$ norm. In this paper we present alternative approaches based on cubic or exponential splines for a periodic or quasiperiodic steady state analysis. Furthermore, it is shown below that the amount of coding effort is negligible if an implementation of HB exists.

## 1 Introduction: System Equations Steady States

Depending on the topology of the circuit and the device constitutive equations the Modified Nodal Analysis (MNA) leads to a system of generally nonlinear differential-algebraic equations (DAEs) of first order of dimension $N$:

$$f(v(t),t) = i(v(t)) + \frac{\mathrm{d}}{\mathrm{d}t} q(v(t)) + b(t) = 0 \tag{1}$$

wherein $t \in \mathbb{R}$ is time and $0 \in \mathbb{R}^N$ the zero vector. Moreover $v : \mathbb{R} \to \mathbb{R}^N$ is the vector of the unknown node voltages and branch currents. $q : \mathbb{R}^N \to \mathbb{R}^N$ is the vector of charges and magnetic fluxes, $i : \mathbb{R}^N \to \mathbb{R}^N$ is the vector of sums of currents entering

Hans Georg Brachtendorf, Ashish Awasthi
University of Applied Science of Upper Austria, Hagenberg, Austria, e-mail: brachtd@fh-hagenberg.at, aawasthi@fh-hagenberg.at

Angelika Bunse-Gerstner, Barbara Lang, Siegmar Lampe
University of Bremen, Bremen, Germany

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_26,

each node and branch voltages. Furthermore $b(t) : \mathbb{R} \rightarrow \mathbb{R}^N$ is the vector of input sources.

Let $P(T)$ be the space of all $T$-periodic $x \in P(T) := \{x \mid x(t) = x(t+T)\}$ and $QP(T_1, T_2, \ldots, T_d)$ the space of all $d$-quasiperiodic and continuous functions. The Fourier expansions of periodic waveforms is

$$x(t) \in P(T) \Leftrightarrow x(t) = \sum_{k=-\infty}^{\infty} X(k) \cdot e^{\mathrm{j}\,k\,\omega_0 t} \tag{2}$$

and for quasiperiodic signals

$$x(t) = \sum_{k_1=-\infty}^{\infty} \cdots \sum_{k_d=-\infty}^{\infty} X(k_1, \ldots, k_d) \cdot e^{\mathrm{j}\,(k_1\,\omega_1 + \cdots + k_d\,\omega_d\,)\,t} \tag{3}$$

If the fundamental frequencies are incommensurable, it is guaranteed that they cannot be integer multiples of a common fundamental frequency. In practical applications, the number of fundamental frequencies is $d = 2$ to $d = 3$.

The input signals or stimuli $b(t)$ are quasiperiodic with typically two or three fundamental frequencies as well. For simplicity only the case $d = 2$ is considered here. The extension to an arbitrary number of fundamentals is straightforward.

In [1–3] it has been shown that a reformulation of the underlying ordinary DAE (1) into an appropriate partial DAE system eases the numerical treatment of the multitone problem. This method has been widely accepted by different research groups [4–8, 10, 11].

In [1] the following theorem was proven:

*Theorem:*

*Consider the system of ordinary differential-algebraic equations (1) with quasiperiodic stimulus*

$$b(t) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} B(k_1, k_2) \cdot e^{\mathrm{j}k_1\omega_1 t}\, e^{\mathrm{j}k_2\omega_2 t} \tag{4}$$

*and the partial DAE system*

$$f(\hat{v}(t_1,t_2); t_1, t_2) = i(\hat{v}(t_1,t_2)) + \frac{\partial}{\partial t_1} q(\hat{v}(t_1,t_2)) + \frac{\partial}{\partial t_2} q(\hat{v}(t_1,t_2)) + b(t_1,t_2) = 0 \tag{5}$$

*where the quasiperiodic stimulus $b(t_1, t_2)$ is given by*

$$b(t_1,t_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} B(k_1, k_2) \cdot e^{\mathrm{j}k_1\omega_1 t_1}\, e^{\mathrm{j}k_2\omega_2 t_2} \tag{6}$$

*with Fourier coefficients $B(k_1, k_2)$. Then*

$$v(t) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} V(k_1, k_2) \cdot e^{(\mathrm{j}k_1\omega_1 + \mathrm{j}k_2\omega_2)t} \tag{7}$$

*is a steady-state solution of the ordinary differential-algebraic equation (1), if and only if*

$$\hat{v}(t_1,t_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \hat{V}(k_1,k_2) \cdot e^{\mathrm{j}k_1\omega_1 t_1}\, e^{\mathrm{j}k_2\omega_2 t_2} \tag{8}$$

*is a steady-state solution of the partial differential-algebraic equation as well. The two solutions are related by $v(t) = \hat{v}(t,t)$ for all $t \in \mathbb{R}$ and the relation $V(k_1,k_2) = \hat{V}(k_1,k_2)$ holds.* ■

The theorem states, that a solution of the underlying ordinary DAE can be obtained along a characteristic of the partial DAE.

# 2 Short Summary of the Harmonic Balance Method

## 2.1 HB for Periodic Steady States

HB approximates the solution in a subspace, which is given by a finite number of Fourier coefficients

$$\mathscr{S} = \left\{ x \,\middle|\, x(t) = \sum_{k=-K}^{K} X(k) \exp\left(\mathrm{j}\,\frac{2\pi k}{T}\,t\right) \right\} \tag{9}$$

The device constitutive equations are given in most practical cases solely in the time domain, $i(v(t))$ and $q(v(t))$. HB circumvents the implementation problems in the way that the devices are evaluated on an equidistant grid or mesh at collocation points $t_i$ in the time domain. Employing the Discrete Fourier Transform (DFT) or its fast implementation the FFT transforms evaluated waveforms into the frequency domain.

Let $\mathscr{F}$ be the matrix of the DFT, $P := I_N \otimes \mathscr{F}$ a matrix which transforms all $N$ waveforms into the frequency domain and $P^{-1} = I_N \otimes \mathscr{F}^H$ its inverse ($\otimes$ is the Kronecker- or tensor product). The boundary value problem is discretized at $2K+1$ gridpoints $t_i$ and equidistant grid spacing $\Delta t = \frac{T}{2K+1}$. The transformation to and from the spectrum is given in matrix formulation by $X = Px$, $I = Pi$, $Q = Pq$, $B = Pb$. The time derivatives of a waveform are represented in the frequency domain by

$$\bar{\Omega}(\omega) := \omega \cdot \mathrm{diag}(-K,\ldots,K), \quad \omega = \frac{2\pi}{T} \tag{10}$$

and $\Omega := I_N \otimes \bar{\Omega}$. HB solves the algebraic system of equations

$$\begin{aligned} F(V) &= P\,i(P^{-1}V) + \mathrm{j}\,\Omega(\omega) \cdot P\,q(P^{-1}V) + Pb \\ &= I(V) + \mathrm{j}\,\Omega(\omega) \cdot Q(V) + B = 0 \end{aligned} \tag{11}$$

for the unknown vector of Fourier coefficients $V$. Equation (11) is an algebraic system of equations $F : \mathbb{C}^{(2K+1)N} \to \mathbb{C}^{(2K+1)N}$ for the unknowns $V$ which can be solved by Newton-like methods [1, 9]. The evaluation of the Jacobian is given i.e. in [1]. HB can be generalized to quasiperiodic steady states. For more details see i.e. [3].

## 3 Spline Interpolation

In the cases of sharp transients the convergence of the Fourier series is poor and only guaranteed in the $L_2$ norm.

Alternatively, in this section cubic and exponential splines are considered as an alternative to Fourier basis functions. Unlike Fourier series, the spline basis functions are only locally defined, therefore the approximation of sharp transients is significantly improved. We restrict here to an equidistant discretization. In that case the coding of the spline basis is simple when a HB simulator exists. Circuit designers are mainly interested in the spectrum of the waveforms. We show here further, that the spectrum can be directly evaluated from the coefficients of the spline approximation.

### 3.1 Cubic Spline Interpolation

**Definition 1 ([12]).** Let $t_0 < t_1 < \cdots < t_n$ be an ordered sequence of collocation points. The B-splines $\hat{y}_{ik}(t)$ of order $k$ for $k = 1, \ldots, n$ and $i = 1, \ldots, n-k$ are defined recursively by

$$\hat{y}_{i1}(t) := \begin{cases} 1 & \text{if} \quad t_i \le t < t_{i+1} \\ 0 & \text{elsewhere} \end{cases}, \; \hat{y}_{ik}(t) := \frac{t - t_i}{t_{i+k-1} - t_i}\,\hat{y}_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}}\,\hat{y}_{i+1,k-1}(t)$$

The cubic spline solves the variational problem [12, 13]

$$E[y] = \sum_{i=0}^{n-1} \int_{t_i}^{t_i+\Delta t} \big(\ddot{y}(t)\big)^2 \mathrm{d}t \tag{12}$$

We denote as $t_{i+1} - t_i =: \Delta t$ the grid spacing. The collocation points coincide therefore with HB based on a Fourier expansion. This eases the implementation as shown below.

The periodic waveform $x(t) = x(t+T)$ is approximated by a linear combination of weighted and shifted basis functions $\hat{y}(t)$, the shifts being integer multiples of the grid spacing $t_l = l \cdot \Delta t,\ l = 0, 1, \ldots$

$$y(t) = \sum_{l=0}^{2K} \hat{Y}(l) \cdot \hat{y}(t - t_l) \tag{13}$$

The unknown coefficients $\hat{Y}(l)$ are uniquely calculated by requiring that the error vanishes at the collocation points and the periodicity constraint of the signal waveform. One obtains the system of equations with $x(l) := x(t_l)$

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & & & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & & & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} \hat{Y}(0) \\ \vdots \\ \hat{Y}(2K) \end{bmatrix} = \begin{bmatrix} x(0) \\ \vdots \\ x(2K) \end{bmatrix} \tag{14}$$

The coefficients matrix is circulant, representing the periodicity constraint, the eigenvectors are therefore the column vectors of the DFT $z_k = \left[\exp\left(-\mathrm{j}k\cdot\frac{2\pi}{2K+1}K\right),\ldots,\exp\left(\mathrm{j}k\cdot\frac{2\pi}{2K+1}K\right)\right]^T$ with corresponding eigenvalues $\lambda_k = \frac{2}{3}+\frac{1}{3}\cos\left(k\cdot\frac{2\pi}{2K+1}\right)$, $-K \le k \le K$. For solving the underlying DAE, time derivatives of the approximation at the grid points are required. Introducing the operator $\nabla$ of the time-derivatives at the collocation points and the coefficient matrix of the DFT $\mathscr{F}$,

$$\mathrm{j}\bar{\Omega} = \mathscr{F}\,\nabla\,\mathscr{F}^{-1}$$

holds. The matrix $\bar{\Omega}$ is again a diagonal matrix

$$\bar{\Omega} = \operatorname{diag}\left(\omega(-K),\ldots,\omega(k),\ldots,\omega(K)\right) \tag{15}$$

The diagonal elements $\omega(k)$ are obtained from the eigenvalues by

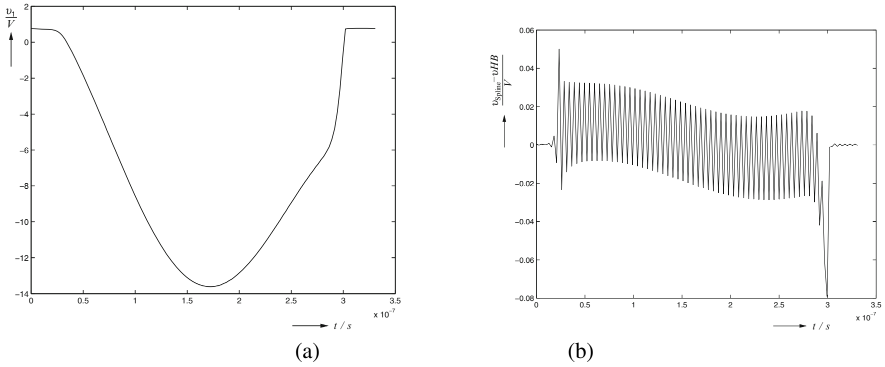
$$\omega(k) = \left(\frac{2K+1}{2\pi}\right)\omega_0\,\frac{3\,\sin\left(k\cdot\frac{2\pi}{2K+1}\right)}{2+\cos\left(k\cdot\frac{2\pi}{2K+1}\right)} \tag{16}$$

The additional coding load for (15, 16) is marginal, if a HB simulator exists because the sparsity structure of the $\bar{\Omega}$ matrix of the HB and spline methods are identical.

The coefficients $\hat{Y}(l)$ of the spline basis are of minor interest for circuit designers. Instead, the Fourier spectrum is of superior interest. From the spline approximation (5) one can calculate the Fourier spectrum exactly. Based on the DFT of the collocation points $X_k$ one gets the spectrum of the spline approximation by

$$\frac{Y_k}{X_k} = \frac{12}{z^4}\,\frac{(1-\cos z)^2}{2+\cos z}, \quad z = \frac{2\pi k}{2K+1}, \quad k \neq 0 \tag{17}$$

Figure 1(a) illustrates the simulated limit-cycle for a Colpitt oscillator. 128 equidistant gridpoints are taken and Fig. 1(b) shows the difference signal between the HB and the spline solution.



(a) (b)

**Fig. 1:** **a** Calculated limit cycle by a cubic spline approximation, **b** Comparison between the numerical solutions by cubic spline and trigonometric basis function

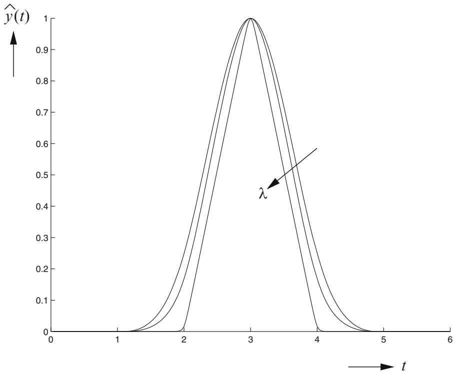


**Fig. 2:** Exponential spline basis functions for the parameter set $\lambda = 10^{-0,5}$, $10^{0,5}$ and $\lambda = 10^{1,5}$

## *3.2 Exponential Splines*

The exponential splines [12, 13] can be fit to the specific interpolation problem by a free parameter $\lambda$. They solve the variational problem

$$E[y] = \sum_{i=0}^{n-1} \int_{t_i}^{t_i+\Delta t} \left[ \left(\ddot{y}(t)\right)^2 + \frac{\lambda^2}{\Delta t^2} \left(\dot{y}(t)\right)^2 \right] \mathrm{d}t \tag{18}$$

The Fig. 2 depicts the exponential spline for different parameter values of $\lambda$.

The next steps are formally identical to the treatment of the cubic splines and are only summarized for brevity. One gets a diagonal matrix representing the collection of derivatives to time in the frequency domain by $\mathrm{j}\,\bar{\Omega} = \mathscr{F}\,\nabla\,\mathscr{F}^{-1} =$

j diag $\{\omega(-K),\dots,\omega(k),\dots,\omega(K)\}$ wherein the diagonal elements are expressed by

$$\omega(k)=2\left(\frac{2K+1}{2\pi}\right)\omega_0\,\frac{1}{\lambda^2}\,\frac{[\cosh\lambda-1]\sin\left(\frac{2\pi k}{2K+1}\right)}{c+\frac{2}{\lambda^3}\,[\sinh\lambda-\lambda]\cos\left(\frac{2\pi k}{2K+1}\right)} \tag{19}$$

## 3.3 Cubic Spline for Multitone Steady State Analysis

For keeping the derivation as simple as possible only the 2-tone case is considered here. The generalization to $d$-quasiperiodic steady states is simple, making use of the partial DAE formulation (5).

A two-dimensional spline basis function can be written as the product of one-dimensional basis functions, i.e.

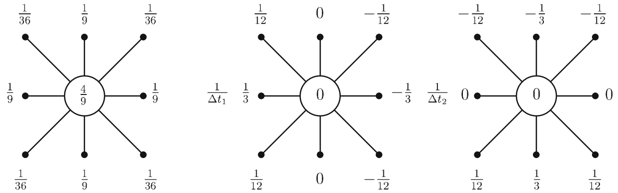
$$\hat{y}(t_1,t_2)=\hat{y}(t_1)\cdot\hat{y}(t_2) \tag{20}$$

Similarly to (13) one gets a system of equations for the coefficients $\hat{Y}_{l_1 l_2}$ of the spline interpolation

$$y(t_1,t_2)=\sum_{l_1,l_2\subset\mathbb{Z}}\hat{Y}_{l_1 l_2}\cdot\hat{y}(t_1-t_{l_1},t_2-t_{l_2}) \tag{21}$$

with pre-supposed periodic conditions in $t_1$ and $t_2$ and grid-points $t_{l_1}$ and $t_{l_2}$.

The Fig. 3 (left) illustrates how the coefficients of the matrix are obtained. Due to the periodicity in $t_1$ and $t_2$, the coefficient matrix is a hierarchical or nested circulant matrix, i.e. any block of the circulant matrix is itself a circulant matrix.



$\frac{1}{36}$ $\frac{1}{9}$ $\frac{1}{36}$
$\frac{1}{9}$ $\frac{4}{9}$ $\frac{1}{9}$
$\frac{1}{36}$ $\frac{1}{9}$ $\frac{1}{36}$
$\frac{1}{12}$ 0 $-\frac{1}{12}$
$\frac{1}{\Delta t_1}$ $\frac{1}{3}$ 0 $-\frac{1}{3}$
$\frac{1}{12}$ 0 $-\frac{1}{12}$
$-\frac{1}{12}$ $-\frac{1}{3}$ $-\frac{1}{12}$
$\frac{1}{\Delta t_2}$ 0 0 0
$\frac{1}{12}$ $\frac{1}{3}$ $\frac{1}{12}$

**Fig. 3:** Molecule for evaluating the coefficients of the spline interpolation (*left*) and its partial derivatives

Further, the partial DAE requires the sum of partial derivatives, which must be calculated for the spline interpolation at the collocation points. From a similar derivation (Fig. 3), one obtains the entries for the $\bar{\Omega}$ matrix

$$\omega(k_1,k_2)=\frac{\mu(k_1,k_2)}{\lambda(k_1,k_2)}=\frac{3}{\Delta t_1}\,\frac{\sin\left(\frac{2\pi k_1}{2K_1+1}\right)}{2+\cos\left(\frac{2\pi k_1}{2K_1+1}\right)}+\frac{3}{\Delta t_2}\,\frac{\sin\left(\frac{2\pi k_2}{2K_2+1}\right)}{2+\cos\left(\frac{2\pi k_2}{2K_2+1}\right)} \tag{22}$$

Please note that for multitone HB one gets $\omega(k_1,k_2) = k_1\,\omega_1 + k_2\,\omega_2$.

Again, the spectrum of the spline approximation can be evaluated from the trapezoidal method. The derivation is similar to the periodic case.

## 4 Conclusions

Cubic and exponential spline bases are an interesting alternative for simulating periodic and quasiperiodic steady states when sharp transients occur in the waveforms. The implementation effort is negligible when a code for the Harmonic Balance technique is available. The Fourier spectrum can easily be calculated from the spline approximation which is very important for electronic engineers.

## References

1. Brachtendorf, H.G.: Simulation des eingeschwungenen Verhaltens elektronischer Schaltungen. Shaker, Aachen (1994)
2. Brachtendorf, H.G.: On the relation of certain classes of ordinary differential algebraic equations with partial differential algebraic equations.zw Technical Report 1131G0-791114-19TM, Bell-Laboratories (1997)
3. Brachtendorf, H.G., Welsch, G., Laur, R., Bunse-Gerstner, A.: Numerical steady state analysis of electronic circuits driven by multi-tone signals. Electronic Engineering, **79**(2), 103–112, April (1996)
4. Pulch, R.: PDE techniques for finding quasi-periodic solutions of oscillators. Preprint 09, IWRMM, Universität Karlsruhe (2001)
5. Pulch, R., Günther, M.: A method of characteristics for solving multirate partial differential equations in radio frequency application. Preprint 07, IWRMM, Universität Karlsruhe, (2000)
6. Roychowdhury, J.: Efficient methods for simulating highly nonlinear multi-rate circuits. In: Proc. IEEE Design Automation Conf., pp 269–274 (1997)
7. Roychowdhury, J.: Analyzing circuits with widely separated time scales using numerical pde methods. IEEE Transactions on Circuits and Systems I - Fundamental Theory and Applications **48**, 578–594, May (2001)
8. Mei, T., Roychowdhury, J., Coffey, T., Hutchinson, S., Day, D.: Robust, Stable Time-Domain Methods for Solving MPDEs for Fast/Slow Systems. IEEE Transactions on Circuits and Systems I - Fundamental Theory and Applications (2004)
9. Kundert, K.S., Sangiovanni-Vincentelli, A.: Simulation of nonlinear circuits in the frequency domain. IEEE Trans. on CAS, No. 4, 521–535 (1986)
10. Pulch, R., Günther, M.: A method of characteristics for solving multirate partial differential equations in radio frequency application. Appl. Numer. Math., No. 42, 399–409 (2002)
11. Constantinescu, F., Nitescu, M., Enache, F.: 2D time domain analysis of nonlinear circuits using pseudo-envelope initialization. In: Proceedings of the 2nd International Conference on Circuits and Systems for Communication, Moscow (2004)
12. Deuflhard, P., Hohmann, A.: Numerische Mathematik I - Eine algorithmisch orientierte Einführung. de Gruyter (1993)
13. Stoer, J., Bulirsch, R.: Numerische Mathematik 2. Springer (1990)

# Accurate Simulation of the Devil's Staircase of an Injection-Locked Frequency Divider

Tao Xu and Marissa Condon

**Abstract** The Devil's Staircase of an Injection-Locked Frequency Divider (ILFD) is simulated in a novel and efficient manner in this contribution. In particular, the Multiple-Phase-Condition Envelope Following (MPCENV) method is employed. The locking range of the ILFD is then determined from the Devil's Staircase. The proposed method is applied to an LC oscillator based ILFD and the results are validated by comparison with experimental results.

## 1 Introduction

In general, Injection-Locked Frequency Dividers (ILFD) are used in the negative feedback of a frequency synthesizer as a prescaler to divide the frequency by a fixed number. In comparison with the traditional static [1] frequency dividers, the ILFDs consume less power but at the expense of a narrow locking range. Hence, the accurate determination of the locking range is important in the design of ILFDs.

The Devil's Staircase [2] was introduced as an *experimental* technique to measure the locking range of an ILFD. Since it requires expensive equipment and takes a long time, some analysis and simulation techniques were introduced to predict the locking range, for example, using expressions derived by Harmonic Balance analysis [3]. A simulation method was also shown using the Warped Multi-Time Scale Partial Differential Equation (WaMPDE) [4] technique. It was improved in [5] to increase the simulation speed. Here, a more accurate and efficient simulation technique is proposed, which utilises the Multiple-Phase-Condition Envelope Following (MPCENV) method to reproduce the Devil's Staircase.

Tao Xu, Marissa Condon
RF Modeling and Simulation Group, Research Institute for Networks and Communications Engineering (RINCE), School of Electronic Engineering, Dublin City University, Dublin 9, Ireland, e-mail: taoxu@eeng.dcu.ie, condonm@eeng.dcu.ie

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_27,

In Section 2, the background of the Devil's Staircase is introduced. In Section 3, the previous work with the WaMPDE to determine the locking range is described. The proposed MPCENV method for simulating the Devil's Staircase is introduced in Section 4. Both of the two simulation methods are applied to an LC oscillator based ILFD. The numerical results are validated by the experimental results in Section 5.

## 2 Background of the Devil's Staircase

The Devil's Staircase [2] is a method to visualize the locking range of an ILFD. Normally, the ILFD is considered as an oscillator with an injected external signal. In order to plot a Devil's Staircase, the frequencies of the oscillator and the injected signal must be varied relative to each other. In practice, it is easier and more accurate to adjust the injected frequency, $\omega_{\text{inj}}$, automatically. The output frequency of the ILFD, $\omega_{\text{o}}$, is then the only unknown variable. The Devil' Staircase [2] is obtained by plotting $\omega_{\text{inj}}/\omega_{\text{o}}$ against $\omega_{\text{inj}}$, as shown in Fig. 1.

The locking range can be measured from the Devil's Staircase diagram [2]. From the staircase diagram in Fig. 1, it is clear that there are lockings (flat regions) at division ratios of 2 and 4, as predicted experimentally in [6].

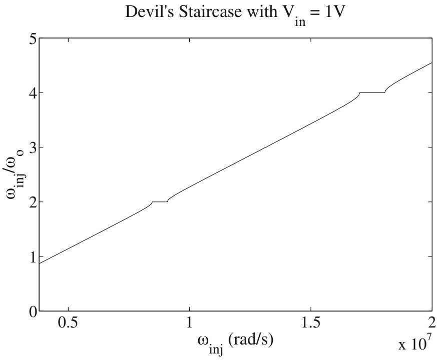


**Fig. 1:** Experimentally measured Devil's staircase diagram showing lockings at $\omega_{\text{inj}}/\omega_{\text{o}} = 2$ and $\omega_{\text{inj}}/\omega_{\text{o}} = 4$

## 3 Previous Work with the WaMPDE

Consider a fairly general nonlinear circuit which is described by:

$$\dot{x}(t) = f(x(t)) + b(t) \tag{1}$$

where $b(t)$ is the excitation vector, $x(t)$ are the state variables and $f$ is a nonlinear function.

The state variable is defined as

$$\begin{aligned} x(t) &= \hat{x}(\tau_1, \tau_2, \ldots, \tau_p, t) \\ &= \hat{x}(\phi_1(t), \phi_2(t), \ldots, \phi_p(t), t), \end{aligned} \tag{2}$$

where

$$\phi_i(t) = \int_0^t \omega_i(\tau_i)\,\mathrm{d}\tau_i. \tag{3}$$

$$\omega_i = \frac{\mathrm{d}\tau_i}{\mathrm{d}t} \tag{4}$$

The $(p+1)$-dimensional WaMPDE [7] corresponding to (1) is:

$$\sum_{i=1}^{p}\left(\omega_i(t)\frac{\partial \hat{x}}{\partial \tau_i}\right) + \frac{\partial \hat{x}}{\partial t} = f(\hat{x}) + \hat{b}(\tau_1, \ldots, \tau_p, t) \tag{5}$$

where $\tau_1, \ldots, \tau_p$ correspond to the warped time scales and $t$ is the real time-scale. $\hat{x}$ and $\hat{b}$ are multivariate functions of the $p+1$ time variables. Obviously, (1) can be solved from (2) and (3) after the solution of (5) is found.

In the case of the ILFD, (5) is rewritten in three dimensions as:

$$\omega_{\mathrm{o}}(t)\frac{\partial \hat{x}}{\partial \tau_1} + \omega_{\mathrm{inj}}\frac{\partial \hat{x}}{\partial \tau_2} + \frac{\partial \hat{x}}{\partial t} = f(\hat{x}) + b(t) \tag{6}$$

where $\tau_1$ and $\tau_2$ are the free-running oscillation time scale and the injected signal time scale, respectively, while $t$ denotes the real time. Note that both $\tau_1$ and $\tau_2$ are warped time scales to enable the slow variation of the local and injected frequency. $\omega_{\mathrm{o}}$ is output frequency of the oscillator and $\omega_{\mathrm{inj}}$ is the injected frequency.

From the experimental results on an ILFD performed in [2] and as shown in Fig. 1, it is noted that the relationship between $\omega_{\mathrm{inj}}/\omega_{\mathrm{o}}$ and $\omega_{\mathrm{inj}}$ is approximately linear between the locking intervals (the ILFD locks at multiples of its natural frequency - the $n$th locking range is described by $\omega_{\mathrm{inj}}/\omega_{\mathrm{o}} = n$). During the locking intervals, the slope is obviously zero. Consequently, two simulations are performed with two carefully selected input frequencies. The two selected $\omega_{\mathrm{inj}}$ are known not to lock the ILFD and to be below the lower limit of the particular $n$th locking range. In other words, (6) is solved twice to obtain the values of $\omega_{\mathrm{o}}$ corresponding to the two values of the manually picked $\omega_{\mathrm{inj}}$. From this, an estimate of the start of the $n$th locking range can be obtained. For instance, the start of the divide-by-two locking range is when $\omega_{\mathrm{inj}}/\omega_{\mathrm{o}} = 2$. Thus the slope the slope of the line connecting the two points determined from simulations is determined by

$$m_{\mathrm{below}} = \frac{2 - \left(\frac{\omega_{\mathrm{inj}}}{\omega_{\mathrm{o}}}\right)}{\omega_{\mathrm{start}} - \omega_{\mathrm{inj}}} \tag{7}$$

where $\omega_{\text{start}}$ is the lower limit of the locking interval and $\omega_{\text{o}}$ is the determined output frequency corresponding to one of the two selected input frequencies, $\omega_{\text{inj}}$. Hence, the lower limit of the locking range is

$$\omega_{\text{start}} = \omega_{\text{inj}} + \frac{2 - \left(\frac{\omega_{\text{inj}}}{\omega_{\text{o}}}\right)}{m_{\text{below}}} \tag{8}$$

The upper limit of the locking range can be obtained with a similar procedure. In this case, the two input frequencies are selected to determine the slope between the divide by 2 and the divide by 4 locking ranges.

$$2\omega_n < \omega_{\text{inj1}} < \omega_{\text{inj2}} < 4\omega_n \tag{9}$$

where $\omega_n$ is the natural frequency of the oscillator.

## 4 The Proposed Method with MPCENV

Here, a novel transient envelope following method, MPCENV, is proposed to determine the output frequencies corresponding to different input frequencies:

The circuit solution is assumed to be composed of fast oscillations whose amplitude and frequency vary much more slowly than the oscillations themselves. Let the period of the fast oscillation be $T$. In the case of oscillators, this will vary slowly. Let $T_{\text{env}}$ be the envelope time-step over which the response of the system can be extrapolated.

Consider Fig. 2. Let $x_0$ and $x_1$ be the state at $t_0 = t_S + T_{\text{env}}$ and $t_1 = t_S + T_{\text{env}} + T$, respectively, where $t_S$ is the ending time of the last envelope step. Using the implicit Euler method for stability purposes, the envelope following process is described by:

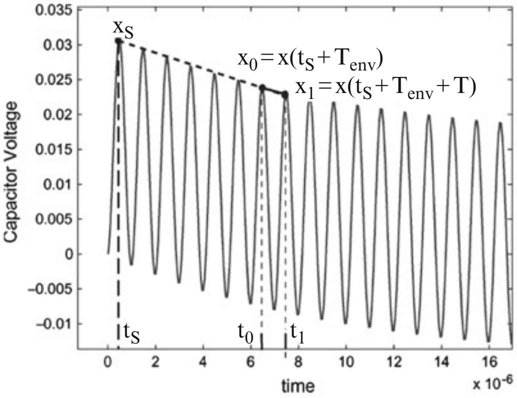


**Fig. 2:** Backward-Euler-based envelope-following method

$$\frac{x_1 - x_0}{T} = \frac{x_0 - x_S}{T_{\mathrm{env}}} \tag{10}$$

where $x_S = x(t_S)$ is known from a previous step, and $x_1$ is determined using the trapezoidal integration method from $t_0$ to $t_1$. This means that $x_1$ depends on $x_0$ and $T$. Note that apart from the circuit variables, there are two extra unknowns, $T$ and $T_{\mathrm{env}}$, since the period of the oscillator is always changing, and $T_{\mathrm{env}}$ has to vary in order to remain equal to an integer number of periods $T$. To solve for the extra unknowns, two further equations are required [8]:

$$\begin{cases} \dfrac{\mathrm{d}x_{0l}}{\mathrm{d}t} = 0 \\ \dfrac{\mathrm{d}x_{1l}}{\mathrm{d}t} = 0 \end{cases} \tag{11}$$

where $l$ denotes the $l_{th}$ state variable. The two derivative-based phase conditions (11) ensure that $x_{0l}$ and $x_{1l}$ are the peaks or troughs of a fast cycle. In practice, value-based constraints are better for numerical handling of certain circuits such as the ILFD:

$$\begin{cases} x_{0l} = c \\ x_{1l} = d \end{cases} \tag{12}$$

where $c$ and $d$ are constants.

Equations (10) and (12) are reorganized as a matrix and solved using the Newton–Raphson method [9]:

$$F = \begin{bmatrix} f_1(x_0, T, T_{\mathrm{env}}) \\ f_2(x_0, T, T_{\mathrm{env}}) \\ f_3(x_0, T, T_{\mathrm{env}}) \end{bmatrix} = \begin{bmatrix} (x_1 - x_0)T_{env} - (x_0 - x_s)T \\ x_{0l} - c \\ x_{1l} - d \end{bmatrix} = 0. \tag{13}$$

If the circuit has $n$ state variables, this system consists of $n+2$ equations with $n+2$ unknowns.

The Jacobian matrix corresponding to (13) is given by:

$$J = \begin{bmatrix} \frac{\mathrm{d}f_1}{\mathrm{d}x_0}, & \frac{\mathrm{d}f_1}{\mathrm{d}T}, & \frac{\mathrm{d}f_1}{\mathrm{d}T_{\mathrm{env}}} \\ \frac{\mathrm{d}f_2}{\mathrm{d}x_0}, & \frac{\mathrm{d}f_2}{\mathrm{d}T}, & \frac{\mathrm{d}f_2}{\mathrm{d}T_{\mathrm{env}}} \\ \frac{\mathrm{d}f_3}{\mathrm{d}x_0}, & \frac{\mathrm{d}f_3}{\mathrm{d}T}, & \frac{\mathrm{d}f_3}{\mathrm{d}T_{\mathrm{env}}} \end{bmatrix} = \begin{bmatrix} \frac{\partial x_1}{\partial x_0} T_{\mathrm{env}} - (T_{\mathrm{env}} + T) I_n, & \frac{\partial x_1}{\partial T} T_{\mathrm{env}} - (x_0 - x_s), & x_1 - x_0 \\ I_n|_l\,, & 0, & 0 \\ \frac{\partial x_1}{\partial x_0}|_l\,, & \frac{\partial x_1}{\partial T}|_l\,, & 0 \end{bmatrix} \tag{14}$$

where $I_n$ is an identity matrix of size $n \times n$, $I_n|_l$, $(\partial x_1/\partial x_0)|_l$ and $(\partial x_1/\partial T)|_l$ are the $l$th row of $I_n$, $\partial x_1/\partial x_0$ and $\partial x_1/\partial T$, respectively.

In this implementation, both $\partial x_1/\partial x_0$ and $\partial x_1/\partial T$ are derived using the trapezoidal integration method, as introduced in [10]. Set $x_r$ to be the state at $t_r$, where $t_0 \le t_{r-1} < t_r \le t_1$. Then

$$\frac{\mathrm{d}x_r}{\mathrm{d}x_0} = \left(1 - \frac{h}{2}\frac{\partial f(x)}{\partial x}\,|_{x_r}\right)^{-1} \left(1 + \frac{h}{2}\frac{\partial f(x)}{\partial x}\,|_{x_{r-1}}\right)\frac{\mathrm{d}x_{r-1}}{\mathrm{d}x_0}, \tag{15}$$

where $f(x)$ is the expression to represent the derivative of the circuit variables: $\dot{x} = f(x)$. The term $\mathrm{d}x_1/\mathrm{d}x_0$ can be found by repeatedly evaluating (15) from $t_0$ to $t_1$ with $\mathrm{d}x_0/\mathrm{d}x_0 = I$, where $I$ is an $n \times 1$ matrix with all ones.

In a similar manner, $\mathrm{d}x_1/\mathrm{d}T$ can be found by solving

$$\begin{aligned}\frac{\mathrm{d}x_n}{\mathrm{d}T} &= \left(1 - \frac{h}{2}\frac{\partial f(x)}{\partial x}\,|_{x_n}\right)^{-1} \\ &\quad \times \left[\left(1 + \frac{h}{2}\frac{\partial f(x)}{\partial x}\,|_{x_{n-1}}\right)\frac{\mathrm{d}x_{n-1}}{\mathrm{d}T} + \frac{x_n - x_{n-1}}{T}\right]\end{aligned} \tag{16}$$

starting from $\mathrm{d}x_0/\mathrm{d}T = 0$.

Then the system in (13) can be solved using the Newton-Raphson method [9]:

$$Z_{\text{new}} = Z - J^{-1}F \tag{17}$$

where $Z_{\text{new}}$ and $Z$ represent the current and previous states of all the variables, i.e., $Z = [x_0\ T\ T_{\text{env}}]^T$. In the case of the ILFD, $x_0$ represents the capacitance voltage and the inductance current, i.e., $x_0 = [V_C\ I_L]^T$.

As described in [2], the Devil's Staircase is a plot of $\omega_{\text{inj}}/\omega_0$ against $\omega_{\text{inj}}$. For simulation purposes, the injected frequency, $\omega_{\text{inj}}$ is increased from the minimum $\omega_{\text{inj}}$ with a fixed frequency step-size. $\omega_0$ is then determined from the MPCENV solution as:

$$\omega_0 = \frac{2\pi}{T}. \tag{18}$$

## 5 Case Study and Numerical Results

The LC oscillator-based ILFD is selected as an example. The schematic is shown in Fig. 3.

The governing equations are

$$\begin{cases} C\frac{\mathrm{d}V_C}{\mathrm{d}t} = I_L - (A + daV_{\text{inj}})V_C + \frac{A+daV_{\text{inj}}}{V_{\text{DD}}^2}V_C^3 \\ L\frac{\mathrm{d}V_C}{\mathrm{d}t} = -I_L - V_C \end{cases} \tag{19}$$

where $A$ and $da$ are the coefficients obtained from the negative resistance characteristic [6].

The Devil's Staircases obtained by simulation and experiment are shown in Fig. 4. The widths of the locking ranges agree when $\omega_{\text{inj}}/\omega_{\text{o}}$ is an even number [6] i.e., 2 and 4, as shown in Fig. 1 and 4. The staircase from MPCENV is almost the same as that from experiment, while the one from WaMPDE has an appreciably bigger difference. However, it should be accepted that this is a very basic method. It could be used to obtain an initial estimate.

Table 1 shows the locking ranges captured from the staircases. The difference between the MPCENV method and experimental results is less than 6%. Therefore, it is sufficiently accurate to predict the locking range when designing ILFDs of this type.

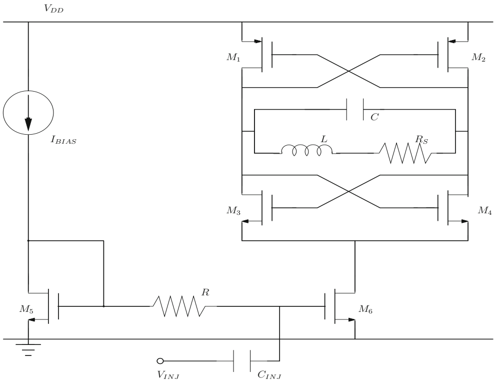


**Fig. 3:** Circuit schematic

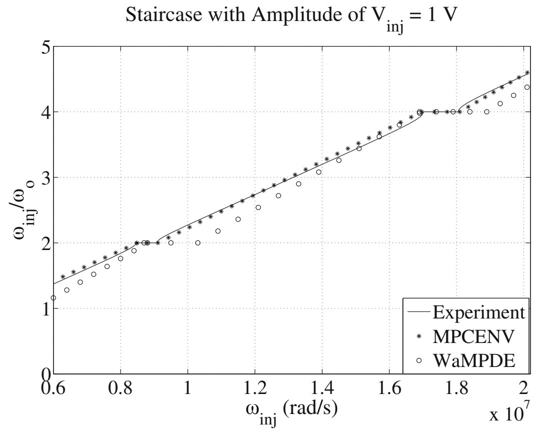


**Fig. 4:** The staircase obtained from simulation and experiment

**Table 1:** Locking range captured from staircases, where *M* and *E* represent *MPCENV* and *Experiment*

| | $\omega_{\text{inj}}/\omega_{\text{o}}$=2 | | $\omega_{\text{inj}}/\omega_{\text{o}}$=4 | |
|---|---|---|---|---|
| $V_{\text{inj}}$ | M (Mrad/s) | E (Mrad/s) | M (Mrad/s) | E (Mrad/s) |
| 1 V | 0.63 | 0.62 | 1.1 | 1.04 |
| 1.5 V | 0.88 | 0.92 | 1.48 | 1.51 |

# 6 Conclusions

A technique based on the Multiple-Phase-Condition Envelope Following (MPCENV) algorithm has been proposed for the determination of the locking range of ILFDs. The simulation technique is advantageous for design and analytical work as it is far less costly than experimental determination. Results for an LC-oscillator based ILFD confirm its efficacy.

**Acknowledgements** This material is based upon works supported by Science Foundation Ireland under Principal Investigator Grant No. 05/IN.1/I18.

# References

1. Singh, U., Green, M.: Dynamics of high-freuqncy CMOS dividers. In: Proceedings of International Symposium on Circuits and Systems, ISCAS 2002, vol.5, pp.421–424, Phoenix, USA, May (2002).
2. O'Neill, D., Bourke, D., Kennedy, M.P.: The devil's staircase as a method of comparing injection-locked frequency divider topologies. In: Proceedings of the European Conference on Circuit Theory and Design, ECCTD 2005, vol.3, pp.317–320, Cork, Ireland, Sept. (2005).
3. Ye, Z., Xu, T., Kennedy, M.P.: Locking range analysis for injection-locking frequency dividers. In: Proceedings of International Symposium on Circuits and Systems, ISCAS 2006, pp.4070–4073, Island of Kos, Greece, May (2006).
4. Christoffersen, C.E., Condon, M., Xu, T.: A new method for the determination of the locking range of oscillators. In: Proceedings of the 2007 European Conference on Circuit Theory and Design, ECCTD 2007, pp.575–578, Sevilla, Spain, Aug. (2007).
5. Xu, T., Condon, M.: An effective method for the determination of the locking range of an Injection-Locked Frequency Divider. In: Proceedings of Emerging Trends in Wireless Communications, pp.47–50, Dublin, Ireland, April (2008).
6. Xu, T., Ye, Z., Kennedy, M.P.: Mathematical analysis of injection-locked frequency dividers. In: Proceeding of International Symposium on Nonlinear Theory and its Applications, NOLTA 2006, pp.639-642, Bologna, Italy, Sept. (2006).
7. Narayan, O., Roychowdhury, J.: Analysing oscillators using multitime PDEs. IEEE Trans. on Circuits and Systems, vol.50, no.7, pp.894–903, July (2003).
8. Mei, T., Roychowdhury, J.: A time-domain oscillator envelope tracking algorithm employing dual phase conditions. IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, Vol. 27, No. 1, January (2008).
9. Langtangen, H.P.: Computational Partial Differential Equations: Numerical Methods and Diffpack Programming, Springer (1999).
10. Aprille, T.J., Trick, T.N.: Steady-state analysis of nonlinear circuits with period inputs. In: Proceedings of IEEE, vol. 60, no. 1, pp. 108C114, Jan. (1972).

# ANN/DNN-Based Behavioral Modeling of RF/Microwave Components and Circuits

Q.J. Zhang* and Lei Zhang

*Invited speaker at the SCEE 2008 conference*

**Abstract** This paper provides a tutorial overview of artificial neural network/dynamic neural network (ANN/DNN) for radio frequency (RF) and microwave modeling and design. We will describe neural network structures suitable for representing high-speed/high-frequency behaviors in components and circuits, ANN training exploiting RF/microwave device and circuit data, formulation of ANN/DNN for microwave component and circuit behavioral modeling, and use of ANN/DNN models for high-level RF/microwave simulation and design optimization.

## 1 Introduction

Artificial neural networks (ANNs) have gained recognition as an emerging vehicle in enhancing the effectiveness of computer-aided modeling and design of RF and microwave circuits and systems [1, 2]. ANNs can be trained to learn electromagnetic (EM)/circuit behaviors from component/circuit data. Trained ANNs can be used as fast and accurate models in high-level circuit and system simulation and optimization. They are able to improve speed, accuracy, and flexibility of microwave modeling and computer-aided design (CAD). This is made possible because of their established network structures, universal approximation property, and the ability to integrate with circuit knowledge [1]. The learning capabilities of ANN can also be used for enhancing the existing CAD models of passive and active components, and thereby extending our ability of describing component behaviors to be even closer towards reality.

ANN techniques and their applications have been applied in a variety of circuit modeling and design [3–6] such as modeling microstrip lines, vias, spiral inductors,

Q.J. Zhang, Lei Zhang
Department of Electronics, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K2C 1N5, e-mail: qjz@doe.carleton.ca, leizhang@doe.carleton.ca

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_28,

transistors, VLSI interconnects, coplanar waveguide discontinuities, printed antennas, and embedded passives for circuit synthesis, optimization, and yield analyses. Further advances in embedding microwave information into neural networks lead to new knowledge-based methods [7–11] for robust electrical modeling. There are increased initiatives for integration of neural network capabilities into circuit design and test processes. For example, recent reports include embedding neural networks in circuit optimization, statistical design, global modeling, computational electromagnetics, measurement standards, and nonlinear circuit and system level design. In addition, automated model generation algorithms [12] have been developed allowing systematic and computerized model creation by ANN, to complement existing human based approaches in RF and microwave modeling. This paper will describe the fundamentals of using neural networks for RF and microwave modeling and design, and highlight its recent applications.

# 2 ANN Approach for RF/Microwave Modeling

## *2.1 Introduction to ANN Based Modeling*

Let $\mathbf{x}$ represent an $Nx$-vector containing physical/electrical parameters of a microwave device. Let $\mathbf{y}$ represent an $Ny$-vector containing the responses of the device under consideration. The physics/EM relationship between $\mathbf{x}$ and $\mathbf{y}$ needs to be represented by a model. The theoretical model for this relationship may be unavailable, or computationally too intensive for online microwave design and repetitive optimization. The objective now is to develop a fast and accurate model that will represent the $\mathbf{x}$-$\mathbf{y}$ relationship efficiently.

In order to develop a neural model, we teach/train a neural network to learn the microwave problem. The external representation of a neural network model can be described as

$$\mathbf{y} = \mathbf{y}(\mathbf{x}, \mathbf{w}) \tag{1}$$

where the $\mathbf{x}$ and $\mathbf{y}$ are neural network inputs and outputs, and $\mathbf{w}$ are neural network internal parameters called weight parameters.

The internal representation of a neural network typically consists of neurons and the connections between neurons. Every connection has a corresponding weight parameter associated with it. Each neuron receives stimulus from other neurons connected to it, processes the information, and produces an output. Neurons that receive stimuli from outside the network are called input neurons while neurons whose outputs are externally used, are called output neurons. Neurons that receive stimuli from other neurons and whose outputs are stimuli for other neurons in the network are known as hidden neurons. Different neural network structures can be constructed by using different types of neurons and by connecting them differently.

There are several types of neural networks that can achieve the required modeling relationship. The most popular form of neural network is the multilayer perceptrons (MLP) structure [1]. As an example for a 3-layer perception shown in Fig. 1, the input layer neurons are neural network inputs $x_i$, $i = 1,2,\ldots,Nx$, and the output layer neurons are neural network outputs, $y_j$, $j = 1,2,\ldots,Ny$. Suppose the outputs of the hidden neurons are $z_k$, $k = 1,2\ldots,Nh$, where $Nh$ is the number of hidden neurons. The computation of the neural network from input $\mathbf{x}$ to output $\mathbf{y}$ is

$$z_k = \sigma(\sum_{i=1}^{Nx} w_{ki}^1 \cdot x_i + w_{k0}^1),\ i = 1,2,...,Nx \tag{2}$$

$$y_j = \sum_{k=1}^{Nh} w_{jk}^2 \cdot z_k + w_{j0}^2,\ j = 1,2,...,Ny \tag{3}$$

where $\sigma(\gamma) = \frac{1}{1+e^{-\gamma}}$. The term $w_{ki}^1$ (or $w_{jk}^2$) represents the weight parameter connecting the neurons between the input (or hidden) layer and the hidden (or output) layer, with $w_{k0}^1$ (or $w_{j0}^2$) being the bias.

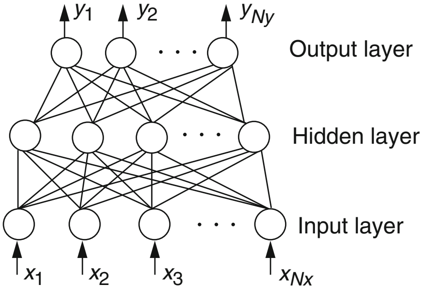


**Fig. 1** Illustration of a feed-forward 3-layer perceptron structure. Typically, the neural network consists of one input layer, one hidden layer, and one output layer

In general, the ANN-based modeling involves three major steps: (1) selection of an appropriate ANN structure, (2) ANN model training, and (3) use of the trained ANN model in simulation, optimization, and circuit design.

## *2.2 ANN Structure Selection*

Various types of neural network structures, such as multilayer perceptrons [1], radial basis function (RBF) networks [1], wavelet neural networks [1], recurrent neural networks (RNN) [13, 14], and dynamic neural networks (DNN) [15, 16], have been used for different modeling scenarios of RF and microwave applications.

The selection of an appropriate neural network structure normally starts by identifying the nature of the input-output relationship of a given application. The modeling of microwave components in the frequency domain is usually formulated with static parameters for neural network inputs and outputs. Such problems can be solved using MLP, RBF, and wavelet networks [1]. The most popular choice is the MLP. RBF and wavelet networks can be used when the microwave problem exhibits highly nonlinear and localized phenomena (e.g., sharp variations). Recent research in the area of microwave-oriented ANN structures leads to the time-domain ANN formulations, such as RNN [13,14] and DNN [15,16], for modeling the dynamic behavior of RF and microwave circuits/systems. Knowledge-based networks [7–11], which combine existing engineering knowledge (e.g., empirical/equivalent-circuit models) with neural networks, are also developed to achieve accurate model with less training data.

## 2.3 ANN Model Training

The most important step in neural model development is the neural network training [1]. ANN models cannot accurately represent the component/circuit behavior until they are trained by data, which is in the form of input-output sample pairs generated by either simulation or measurement.

Define **d** as a vector containing simulated or measured data of the output responses **y**. Then **x** and **d** become the input-output sample pairs called training data. Many samples of **x** are usually needed in the **x**-space to make the ANN a valid model in that range of **x**. The training error of the ANN model is defined as

$$E(\mathbf{w}) = \frac{1}{2}\sum_{k\in T_r} \|\mathbf{y}(\mathbf{x}_k,\mathbf{w}) - \mathbf{d}_k\|^2 \tag{4}$$

where $\mathbf{d}_k$ is the $k^{th}$ sample of the output in the training data, $\mathbf{y}(\mathbf{x}_k,\mathbf{w})$ is the neural network output for the $k^{th}$ sample of the input, i.e., $\mathbf{x}_k$, and $T_r$ is the index set of all training data. The purpose of neural network training is to adjust **w** such that the error function $E(\mathbf{w})$ is minimized. Training is an iterative process and is usually performed by optimization algorithms [1].

## 2.4 Use of the Trained ANN Models in Circuit Design

Once the ANN models have been trained and verified, they can then be incorporated into a microwave circuit simulator for circuit design and optimization [1–3]. ANN models can interconnect each other or connect to other components or models in the simulator to form a high-level circuit. During design, the circuit simulator passes input variables, e.g., gate length of a device or simulation frequency, to the ANN model, which then computes and returns the corresponding outputs, e.g., drain current or S-parameters, back to the simulator. ANN models achieve much-improved simulation efficiency while maintaining the same accuracy as the detailed and slow EM/physics model.

## 3 ANN/DNN Applications in RF/Microwave Modeling

In this section, ANN based modeling techniques are demonstrated through application examples. We will highlight two important ANN structures, i.e., the knowledge-based structure combining ANN and equivalent circuit, and the DNN. We illustrate different modeling scenarios such as linear/nonlinear component and circuit modeling, and time-domain and frequency-domain modeling.

### *3.1 Automated Model Generation for Embedded Passive Modeling*

This subsection demonstrates an example of knowledge-based neural network modeling where an existing equivalent circuit is combined with neural networks [6]. The neural network helps converting the equivalent circuit model into a parametric model with physical/geometrical variables, and improving the modeling accuracy by training the model to match electromagnetic data. Fig. 2(a) illustrates the geometry of an embedded capacitor. The knowledge-based model combining a user-defined equivalent circuit with an MLP neural network is shown in Fig. 2(b). The neural model has two inputs (length $l$ and dielectric permittivity $\varepsilon_{rcap}$) and two outputs (inductance $\mathbf{L}$ and capacitance $\mathbf{C}$), which are used in the user-defined equivalent circuit to produce the 2-port S-parameters to match the training data obtained from EM simulations [6]. The training data are the real and imaginary parts of the 2-port S-parameters (i.e. $\mathbf{S}_{11}$ and $\mathbf{S}_{21}$) at multiple frequencies and different geometries.

An advanced ANN training method, called automated model generation (AMG) [12], is used. It automates the model development from data generation to ANN training to achieve the required model accuracy with minimum amount of training data. An adaptive sampling algorithm is used during the training process to decide how many training data is needed, and how the data should be distributed/sampled in the training space. The neural network size (such as the number of hidden neurons) is determined during the training process according to a set of under-learning and over-learning criteria. A simulator driver in AMG is set up to drive *Sonnet-Lite* [17] to generate training and validation EM data. Once the data is generated, the lumped components $\mathbf{L}$ and $\mathbf{C}$ of the equivalent circuit are extracted from the EM data [6].

In this particular example, AMG training starts with an initial ANN with 5 hidden neurons. The number of hidden neurons increases to 12 to meet the user-defined accuracy of 1% ANN test error. The ANN learns the $\mathbf{L}$-$\mathbf{C}$ dependence on the input geometry through the automated training. The final combined model (combining the user-defined equivalent circuit and the trained ANN) results in an overall accuracy of 0.86% when compared to the EM data. The comparisons between the real part

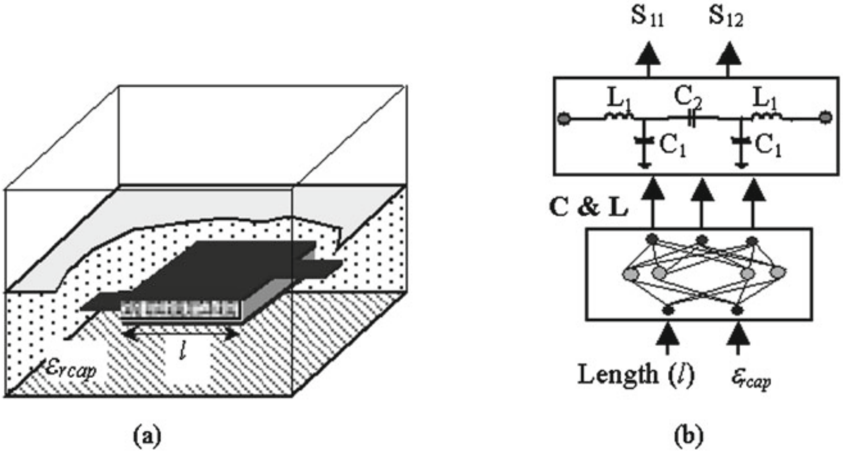


**Fig. 2:** Illustration of linear component modeling using ANN. **a** Geometry of an embedded capacitor and **b** the neural model combined with equivalent circuit

of $\mathbf{S}_{11}$ of the EM simulation and the results from the combined model for two sets of geometry inputs are shown in Fig. 3, confirming that the S-parameters can be accurately regenerated by the combined model.

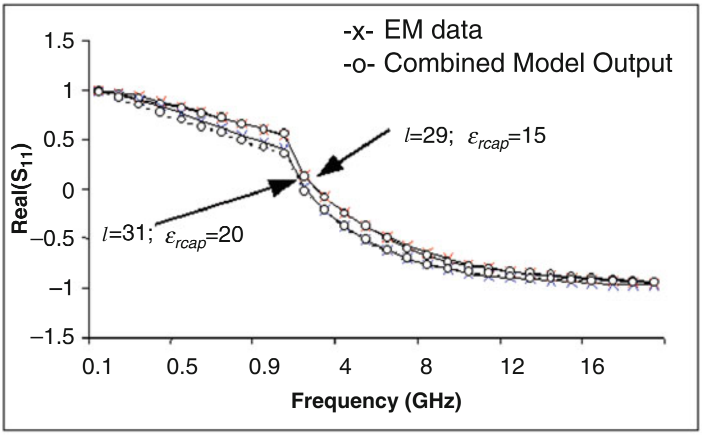


**Fig. 3:** Comparison of Real($\mathbf{S}_{11}$) of the original EM data (-x-) and the combined equivalent circuit and ANN model (-o-) for two sets of ($l$, $\varepsilon_{rcap}$) values

### 3.2 DNNs for Behavioral Modeling of Nonlinear Circuits/Systems

Behavioral modeling of nonlinear circuits and systems is significant with the increasing need for efficient CAD techniques in high-level and large-scale nonlinear microwave design. DNN is an advance in this area [15]. DNN is formulated in continuous time domain to represent the dynamic input-output signal relationship. The output signal at a time instance is a function of the input signal at that time, its time derivatives, as well as the time derivatives of the output signal itself.

Here we illustrate DNN and its use in nonlinear simulation through a direct broadcast satellite (DBS) receiver subsystem example [15]. The DBS consists of a mixer, a gain stage amplifier, and an output stage amplifier. The original DBS representation is a detailed transistor-level nonlinear circuit in Agilent ADS [18], as shown in Fig. 4(a). To develop fast behavioral models, we train 3 DNNs to model the mixer, the gain stage amplifier, and the output stage amplifier, respectively, as shown in Fig. 4(b).

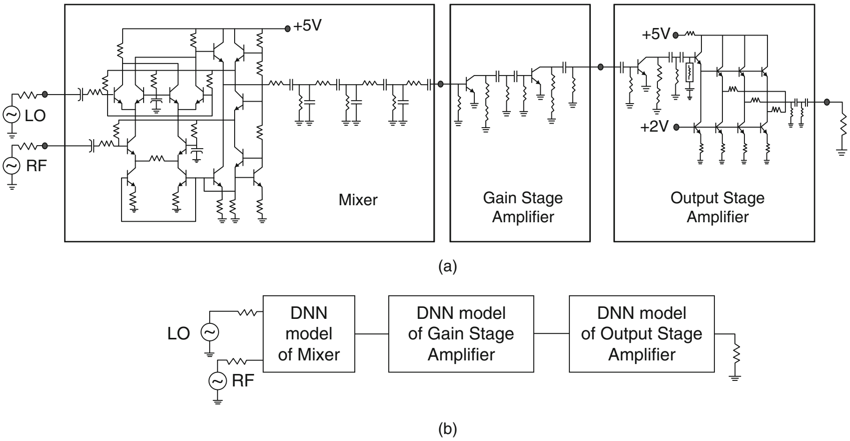


**Fig. 4:** DBS receiver subsystem implemented by **a** connecting original detailed equivalent circuits in ADS, and **b** connecting the trained DNN models

The structure of each DNN model in Fig. 4(b) can be illustrated by Fig. 5. The dynamic output voltage $v_{out}(t)$ is computed from a neural network function in terms of the dynamic input $v_{in}(t)$ and the higher order derivatives of both output and input signals as

$$V_{out}(t) = f_{ANN}(v_{out}^{(1)}(t),\ldots,v_{out}^{(n)}(t),v_{in}(t),v_{in}^{(1)}(t),\ldots,v_{in}^{(n)}(t)) \tag{5}$$

where $f_{ANN}$ represents a multilayer perceptron neural network [1], $v_{in}^{(i)}(t)$ and $v_{out}^{(j)}(t)$ represent the derivative inputs as $\mathrm{d}^i v_{in}/\mathrm{d}t^i$ $(i = 1,2,\ldots,n)$ and $\mathrm{d}^j v_{out}/\mathrm{d}t^j$ $(j = 1,2,\ldots,n)$, respectively, and $n$ denotes the dynamic order.

The DNNs are trained using large-signal data from original ADS-simulation data of the mixer and amplifiers [15]. The trained DNN models of the amplifiers and mixer can be conveniently incorporated into ADS to perform harmonic balance simulation of the DBS subsystem. Such implementation can be achieved by either using the circuit representation of the DNN or programming the HB representation of the DNN models [15]. The overall DBS subsystem solution using DNNs matches that of the original system as shown in Fig. 6, even though these obviously distorted signals were never used in training of any of the DNNs. The time for simulating the DBS system is reduced from 20.03 seconds using the original detailed-circuit simulation down to 4.87 seconds using the DNN based simulation.

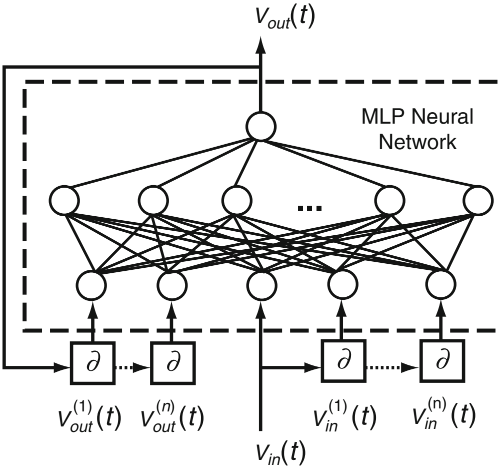


**Fig. 5:** Illustration of the input-output relationship of the DNN model structure

## 4 Conclusion

Neural network based RF/microwave modeling has been introduced. We highlighted major parts of neural modeling approaches including ANN structures, training, and application to RF/microwave modeling for linear/nonlinear components and circuits. Through application examples, we illustrated concepts of knowledge based neural networks, automated model generation, and DNN based dynamic modeling. The techniques have been applied to behavioral modeling of electromagnetic structures and nonlinear microwave circuits. The ANN/DNN approach helps to provide fast and accurate models for RF/microwave design, enhancing design quality and efficiency.

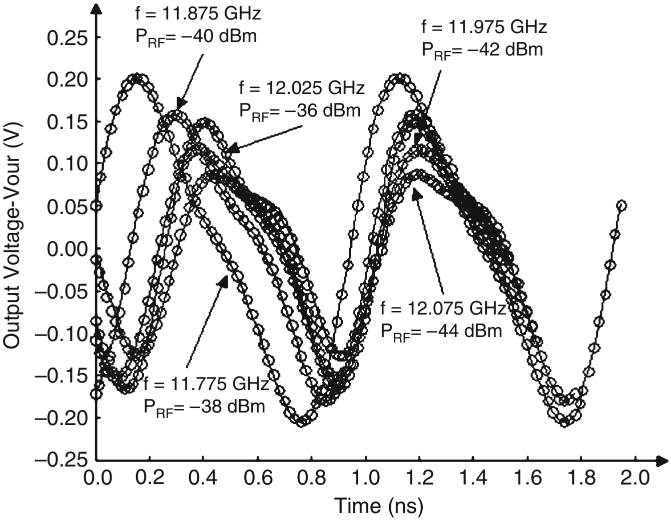


**Fig. 6:** Comparison of DBS subsystem output between system solutions using DNN models (*dashes*) and ADS simulations of original system (*circles*). Excellent agreement is achieved even though these nonlinear solutions were never used in training

# References

1. Zhang, Q.J., Gupta, K.C.: Neural Networks for RF and Microwave Design. Artech House, Boston, MA (2000)
2. Zhang, Q.J., Gupta, K.C., Devabhaktuni V.K.: Artificial neural networks for RF and microwave design: from theory to practice. IEEE Trans. Microw. Theory Tech., **51**, 1339–1350 (2003)
3. Burrascano, P., Fiori, S., Mongiardo, M.: A review of artificial neural networks applications in microwave computer-aided design. Int. J. RF and Microwave CAE, **9**, 158–174 (1999)
4. Rayas-Sanchez, J.: EM-based optimization of microwave circuits using artificial neural networks: the state-of-the-art. IEEE Trans. Microw. Theory Tech., **52**, 420–435 (2004)
5. Zaabab, H., Zhang, Q.J., Nakhla, M.: A neural network modeling approach to circuit optimization and statistical design. IEEE Trans. Microw. Theory Tech., **43**, 1349–1358 (1995)
6. Ding, X., Devabhaktuni, V.K., Chattaraj, B., Yagoub, M.C.E., Doe, M., Xu, J.J., Zhang, Q.J.: Neural network approaches to electromagnetic based modeling of passive components and their applications to high-frequency and high-speed nonlinear circuit optimization. IEEE Trans. Microw. Theory Tech., **52**, 436–449 (2004)
7. Watson, P.M., Gupta, K.C.: EM-ANN models for microstrip vias and interconnects in dataset circuits. IEEE Trans. Microw. Theory Tech., **44**, 2495–2503 (1996)
8. Wang, F., Zhang, Q.J.: Knowledge based neural models for microwave design. IEEE Trans. Microw. Theory Tech., **45**, 2333–2343 (1997)
9. Bandler, J.W., Ismail, M.A., Rayas-Snchez, J.E., Zhang, Q.J.: Neuromodeling of microwave circuits exploiting space-mapping technology. IEEE Trans. Microw. Theory Tech., **47**, 2417–2427 (1999)

10. Watson, P.M., Gupta, K.C., Mahajan, R.L.: Applications of knowledge-based artificial neural network modeling to microwave components. Int. J. RF and Microwave CAE, **9**, 254–260 (1999)
11. Zhang, L., Xu, J.J., Yagoub, M.C.E., Ding, R., Zhang, Q.J.: Efficient analytical formulation and sensitivity analysis of neuro-space mapping for nonlinear microwave device modeling. IEEE Trans. Microw. Theory Tech., **53**, 2752–2767 (2005)
12. Devabhaktuni, V.K., Chattaraj, B., Yagoub, M.C.E., Zhang, Q.J.: Advanced microwave modeling framework exploiting automatic model generation, knowledge neural networks and space mapping. IEEE Trans. Microw. Theory Tech., **51**, 1822–1853 (2003)
13. Fang, Y., Yagoub, M.C.E., Wang, F., Zhang, Q.J.: A new macromodeling approach for nonlinear microwave circuits based on recurrent neural networks. IEEE Trans. Microw. Theory Tech., **48**, 2335–2344 (2000)
14. Sharma, H., Zhang, Q.J.: Transient electromagnetic modeling using recurrent neural networks. In: Proceedings of the 2005 IEEE MTT-S International Microwave Symposium., pp. 1597–1600. Long Beach, LA, June 12–17 (2005)
15. Xu, J.J., Yagoub, M.C.E., Ding, R., Zhang, Q.J.: Neural-based dynamic modeling of nonlinear microwave circuits. IEEE Trans. Microw. Theory Tech., **50**, 2769–2780 (2002)
16. Cao, Y., Ding, R., Zhang, Q.J.: State-space dynamic neural network technique for high-speed IC applications: modeling and stability analysis. IEEE Trans. Microw. Theory Tech., **54**, 2398–2409 (2002)
17. Sonnet-Lite v.9.51, Sonnet Software Inc., Liverpool, NY, USA.
18. Advanced Design System 2006A, Agilent Technologies, Santa Rosa, CA, USA.

# Surrogate Modeling of Low Noise Amplifiers Based on Transistor Level Simulations

Luciano De Tommasi, Dirk Gorissen, Jeroen Croon, and Tom Dhaene

**Abstract** Although the behavior of several RF circuit blocks can be accurately evaluated via transistor-level simulations, the design space exploration is limited by the high computational cost of such simulations. Therefore, cheap-to-evaluate surrogate models of the circuit simulator are introduced. This paper presents some results of a feasibility study concerning the development of surrogate models of low noise amplifiers.

## 1 Introduction

A surrogate model is a cheap-to-evaluate replacement model of expensive, highly accurate, computer simulations (e.g. circuit simulations). The surrogate modeling approach can be used to increase the efficiency of design space exploration, what-if analyses, optimization and sensitivity analyses.

We aim to investigate the feasibility of surrogate modeling approach for RF circuit blocks. Accurate surrogate models of single RF and microwave components have been already developed (e.g. using ANNs [1]). In this work, we do not model a single device like a MOSFET, but a complete RF circuit block: a Low Noise Amplifier (LNA) [2] (Fig. 1). Other RF circuit blocks (e.g. mixers, VCOs, etc.) could be analyzed as well.

The behavior of an LNA is described by means of the admittance and noise functions, which are evaluated via accurate transistor-level simulations. In particular,

Luciano De Tommasi, Dirk Gorissen, Tom Dhaene
Ghent University – IBBT, Department of Information Technology (INTEC), Gaston Crommenlaan 8 Bus 201, 9050 Ghent, Belgium, e-mail: luciano.detommasi@ua.ac.be, dirk.gorissen@ugent.be, tom.dhaene@ugent.be

Jeroen Croon
NXP-TSMC Research Center, High Tech Campus 37, Post Box WY4-01, 5656 AE Eindhoven, The Netherlands, e-mail: jeroen.croon@nxp.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_29,

the characterization of weakly nonlinear LNA behavior demands several periodic steady state analyses, which are particularly time-consuming.

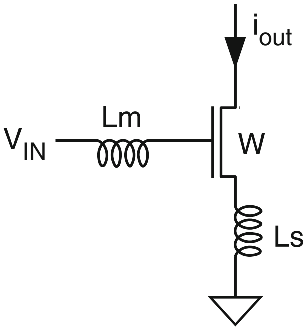


**Fig. 1:** A narrowband low noise amplifier

Hence, a complete LNA simulation typically requires one to two minutes, which is too long to effectively explore how performance figures scale with key circuit-design parameters, such as the dimensions of transistors, passive components, signal properties and bias conditions. Therefore, circuit simulations can be usefully replaced with an accurate surrogate model which is much cheaper to evaluate.

Several model types included in the SUMO Toolbox modeling software [3] were compared in [4] and [5], the goal being the selection of the best approximation of LNA describing functions. Such comparison exploited a first order analytical model of the LNA, which is faster to evaluate than circuit simulations, and at the same time satisfactorily reproduces the shape of the simulator outputs.

In [4] and [5] the error given by a surrogate model was computed comparing reference function (the analytical LNA model) and surrogate model over a dense grid of samples. However, when the reference functions are the outputs of expensive transistor level simulations, the error can only be estimated by comparing few samples. In this paper, such accuracy issues are discussed and results are compared with [4] and [5].

## 2 Software Environment

The surrogate modeling approach developed in this paper, is based on the SUrrogate MOdeling (SUMO) Matlab Toolbox, a plug-in based, adaptive tool that automatically tries to generate a surrogate model with the required accuracy within the time limits set by the user [6].

The SUMO Toolbox modeling flow is shown in Fig. 2. It is based on *adaptive modeling* and *adaptive sampling* loops.

The surrogate modeling process starts with the evaluation of an initial design (e.g. a Latin hypercube) which uniformly fills the design space (the number of samples is specified by the user). Based on this initial set of samples, one or more surrogate models are constructed. Adaptive modeling implies that a suitable optimization algorithm (e.g. genetic algorithm) is used to tune relevant model hyperparameters, in order to minimize the error between model and data. Model error is evaluated according to one or more measures and functions, see section (3). The models are then ranked according to their score, and the best model is selected.

In order to improve the accuracy, an adaptive sampling procedure drives the selection and simulation of new samples. Optimal selection of additional data samples, known as *reflective exploration*, is based on the best performing models and the behavior of the reference function. In this work we applied the *gradient-based* sample selector because it has shown good performances with the LNA modeling problem [10].

After each sampling iteration, an adaptive modeling iteration including the new samples is started, and the whole process repeats itself until one of the following three user-defined conditions is satisfied: (1) the user required accuracy has been met, (2) the maximum allowed number of samples has been reached, or (3) the maximum allowed modeling time has been exceeded.

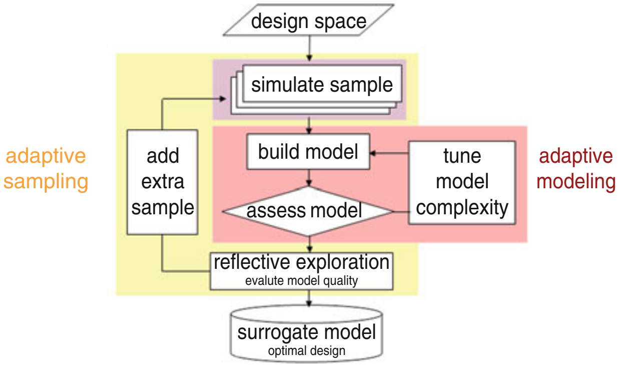


**Fig. 2:** Modeling flow followed by the SUMO toolbox software

The SUMO toolbox has been interfaced with Cadence SPECTRE simulator, which runs transistor level simulations, so providing samples of the LNA describing functions. As mentioned in section 1, the LNA is characterized through admittances and noise currents. Admittances are 52 complex functions. These include four admittances describing the linear behavior plus high order admittances describing the weakly-nonlinear transfer [8].

Noise functions include two real functions (input and output noise currents) and one complex function (correlation between input and output noise currents).

## 3 Model Accuracy Evaluation

Model accuracy evaluation involves the definition of an *error measure* and an *error function*. An error measure is a criterion to estimate the error, defining which samples contribute to the error and how they contribute (e.g. holdout, crossvalidation, etc). An error function is a mathematical expression of the error as function of the samples (e.g. root relative square error, mean relative square error, maximum relative error, average relative error, etc).

We use the Root-Relative-Square-Error (RRSE) to drive the model selection:

$$RRSE(\{y_i\},\{\tilde{y}_i\}) = \sqrt{\frac{\sum(y_i-\tilde{y}_i)^2}{\sum(y_i-\bar{y})^2}} \tag{1}$$

being $\{y_i\}$ the samples of the reference function, $\{\tilde{y}_i\}$ the samples of the surrogate model, $\bar{y}$ the mean of the reference function, $N$ the number of samples and $i = 1,\ldots,N$. In fact, our investigations showed that RRSE tends to produce the smoothest models with a reduced number of samples. However, it is worth to note that accuracy specifications may also be given in terms of other relative error functions, e.g., average relative error (2) or maximum relative error (3), rather than in terms of RRSE. In such case, these error functions can be evaluated to eventually check the accuracy of models.

$$ARE(\{y_i\},\{\tilde{y}_i\}) = \frac{1}{N}\sum\left|\frac{y_i-\tilde{y}_i}{y_i}\right| \tag{2}$$

$$MRE(\{y_i\},\{\tilde{y}_i\}) = max\left|\frac{y_i-\tilde{y}_i}{y_i}\right| \tag{3}$$

As for the error measures, we consider two different approaches used in statistical classification for accuracy estimation and model selection: *holdout* and *crossvalidation* [9]. Holdout involves the partitioning of the sample set in two subsets, respectively named *training set* and *validation set*. The training set is used to train the model (e.g. find the coefficients of a rational function or the weights of a neural network), whereas the error is computed only using the validation set samples. The bigger is the validation set, the higher the bias of error estimation. On the other hand, fewer validation samples lead to a wider confidence interval of the error estimation.

The advantage of $K$-fold crossvalidation (CV) over holdout is that all the samples are used for both training and validation. This means that the model is trained by exploiting all the available samples and its error is then estimated as follows. The sample set is divided into $K$ subsets (folds) of equal or similar size, and $K$ models are trained, each time leaving out one of the folds from training. The error of each model is computed by using only the omitted fold. Finally, the $K$ errors are averaged to produce the error estimation of the model trained with all the samples.

# 4 Modeling Settings

## 4.1 Accuracy

A target accuracy of RRSE $< 0.01$ is required. 5-folds crossvalidation estimation of RRSE is used as error measure driving the model selection process. In order to compare different error measures, RRSE is also estimated by means of the holdout method, with a validation set (VS) including the 20% of the total number of evaluated samples. The density of VS samples throughout the design space has to be as uniform as possible. This is achieved by means of a SUMO Toolbox optimization algorithm. In such way, a new VS is evaluated after each new sampling iteration. The VS estimation is not used to select the best model.

## 4.2 Model Type and Adaptive Modeling

Results in [4] [5] have shown that rational functions [7] and ANNs are the most suitable model types, respectively, for the admittance and noise functions of an LNA. Hence, we choose the same model types with the transistor level simulator.

Rational approximation implemented in SUMO toolbox exploits a function in the form[1]:

$$y(x_1, x_2) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{n-i} \alpha_{ij}(w_1 x_1)^i (w_2 x_2)^j}{\sum_{i=0}^{n} \sum_{j=0}^{n-i} \beta_{ij} f_{ij}(w_1 x_1)^i (w_2 x_2)^j} \tag{4}$$

where $f_{ij}$ are boolean flags with values in $\{0,1\}$, and $\{w_1, w_2\}$ are input weights.

The coefficients $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$ are computed by solving a linear least squares problem [7], whereas a genetic algorithm (GA) is used to optimize $\{w_1, w_2\}$, $\{f_{ij}\}$ and $n$. Constraints applied to the rational approximation are: $1 < w_1, w_2 < 40$ and $n \leq 100$. GA settings are: *population size* = 30, *crossover fraction* = 0.7, *maximum number of generations* = 10, *elite count* = 1, *stall generation limits* = 4 [11].

## 4.3 Sampling

Modeling starts with a 20-samples Latin Hyper Cube. Afterwards, gradient-based adaptive sampling is applied.

We remark that adaptive sampling has to be used because the LNA functions are almost flat in large regions of the input domain space (most of them have only one

[1] It is shown a function of two variables: $x_1$ and $x_2$.

or two resonance peaks). Therefore, an uniform sample distribution would waste a lot of samples in large uninteresting regions.

## 5 Results

Surrogate models of admittance parameters and noise currents have been generated with respect to the transistor width $W$ and the inductance $L_s$. Due to space limitation, only modeling of admittances can be discussed in this paper.

Target accuracy is reached with 155 samples. Modeling the analytical admittances $y_{11}$ and $y_{21}$ just required 24 samples [5]. However, analytical admittances are exactly rational functions of $W_n$ and $L_{sn}$, and this simplifies the rational modeling.

A surrogate modeling flow may include final accuracy assessment by computing the RRSE onto an independent set of samples (test set) [1]. A small set of samples is usually used (to keep low the cost of additional sample simulations). In this study, we aim to assess the modeling flow rather than the models, therefore we choose a big test set (TS): a uniform grid of 400 samples. It provides a more accurate RRSE estimation than CV.

Each admittance function has been labeled with an index between 1 and 52. The corresponding RRSEs (as given by CV, VS and TS) are displayed in Fig. 3. It can be seen that CV estimation of RRSE is lower than RRSE computed onto TS. The VS estimation gives a better agreement with the TS. The model exhibiting the highest TS-RRSE is the admittance with index 41 (Fig. 4 left side) and gives an $RRSE \cong 0.1$. For such model, the relative error:

$$RE(L_{sn}, W_n) = \left| \frac{y(L_{sn}, W_n) - \tilde{y}(L_{sn}, W_n)}{y(L_{sn}, W_n)} \right| \tag{5}$$

was also computed over the TS. The maximum relative error (3) is about 30%, with the maximum deviation (5) assumed near the main resonance peak.

Average relative error (ARE) (2) and maximum relative error (MRE) (3) have been computed for all the 52 admittance models. It is seen that models may give large MRE (and ARE) but small RRSE. For example, the model of admittance $y_{1100m0}$ gives: RRSE=0.00132 and, respectively for real and imaginary part, AREr=13.3 and AREi=0.028. Figure 5 shows the absolute (AE= $|y - \tilde{y}|$) and relative (5) errors. Relative error is very large in a region where the function is very close to zero.

## 6 Conclusions

Surrogate models of low noise amplifiers can be obtained with a satisfactory accuracy considering two input parameters. However, the use of 5-folds crossvalidation

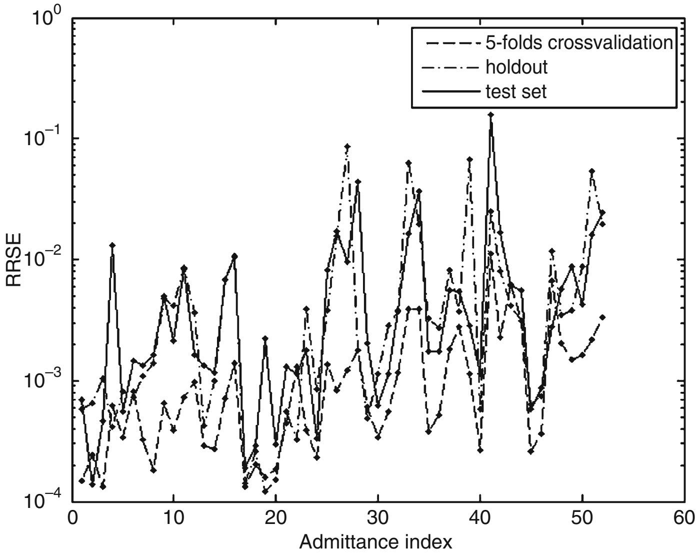


**Fig. 3:** Root-Relative-Square-Error of each admittance function estimated by 5-folds crossvalidation, holdout (validation set 20% of the samples) and a test set (20 × 20 uniform grid of samples)

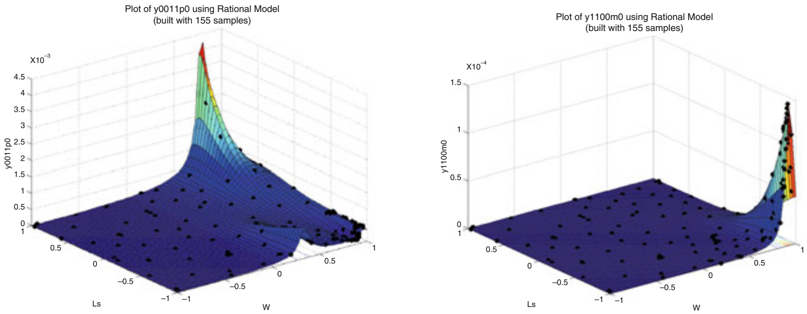


**Fig. 4:** *Left*: Admittance model with the highest TS-RRSE. *Right*: Admittance model with large MRE but small TS-RRSE

leads to an underestimation of the error. Therefore, more samples will be in practice needed to achieve the required level of accuracy.

Admittance functions assume values close to zero throughout large regions of inputs domain. In such regions, it may become very difficult to control the relative error. Hence, although ARE and MRE are more intuitive functions than RRSE, they are not suitable for accuracy specifications.

Further work will be aimed to increase the accuracy of surrogate models with more input parameters.

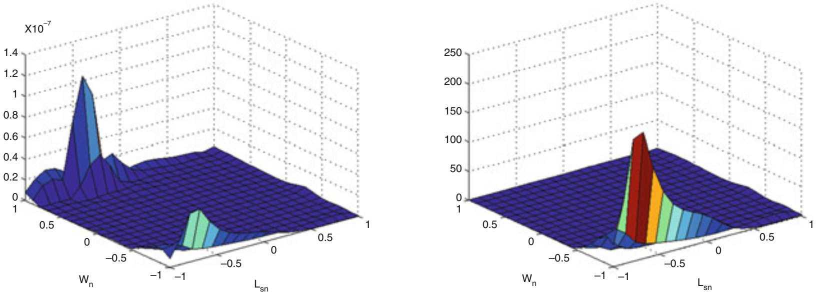


**Fig. 5:** Pointwise errors of model $y_{1100m0}$. *Left*: absolute. *Right*: relative

**Acknowledgements** This work was supported by the European Commission through the Marie Curie Actions of its Sixth Program under the contract number MTKI-CT-2006-042477.
The authors thank Dr. Joost Rommes (NXP semiconductors, Eindhoven) for several stimulating discussions.

## References

1. Zhang, Q.J., Gupta, K.C.: Neural Networks for RF and Microwave Design. Artech House, Boston - London (2000).
2. Lee, T.H.: The Design of CMOS Radio-Frequency Integrated Circuits (Second Edition). Cambridge University Press (2003).
3. Hendrickx, W., Gorissen, D., Dhaene, T.: Grid Enabled Sequential Design and Adaptive Metamodeling. In: Proceedings of the 2006 Winter Simulation Conference, WSC 2006, pp. 872–881. December 3–6 (2006).
4. Gorissen, D., De Tommasi, L., Croon, J., Dhaene, T.: Automatic Model Type Selection with Heterogeneous Evolution: An application to RF circuit block modeling. In: Proceedings of IEEE World Congress on Computational Intelligence, WCCI 2008, pp. 989–996. Hong Kong, China, June (2008).
5. Gorissen, D., De Tommasi, L., Crombecq, K., Dhaene, T.: Sequential Modeling of a Low Noise Amplifier with Neural Networks and Active Learning. Accepted for publication in Neural Computing and Applications, (2008)
6. The SUrrogate MOdeling Toolbox Wiki Page. URL http://www.sumowiki.intec.ugent.be/.
7. Hendrickx, W., Dhaene, T.: Sequential design and rational metamodelling. In: Proceedings of the 2005 Winter Simulation Conference, WSC 2005, pp. 290–298. December 4–7 (2005)
8. Croon, J.A., Leenaerts, D.M.W., Klaassen, D.B.M.: Accurate Modeling of RF Circuit Blocks: Weakly-Nonlinear Narrowband LNAs. In: Proceedings of the IEEE Custom Integrated Circuits Conference 2007, CICC 2007, pp 865–868. September 16–19 (2007)
9. Kohavi, R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, pp. 1137–1145. August 20–25 (1995)
10. Crombecq, K., A gradient based approach to adaptive metamodeling. Technical report, University of Antwerp. (2007)
11. The Mathworks: Genetic Algorithm and Direct Search Toolbox. URL http://www.mathworks.com/products/gads/.

# Computational Statistics Approach to Capacitance Sensitivity Analysis and Gate Delay Time Minimization of TFT-LCDs

Yiming Li and Hsuan-Ming Huang

**Abstract** In this paper, we practically implement a systematical method for thin-film transistor liquid-crystal display (TFT-LCD) design optimization and sensitivity analysis. Based upon a three-dimensional (3D) field solver and a Design of Experiments, we construct a second-order response surface model (RSM) to examine the capacitances' effect on the performance of an interested TFT-LCD pixel. The constructed RSMs are reduced using a step-wise regression. We verify the accuracy using the normal residual plots and their residual of squares. According to the models, we then analyze the sensitivity of the capacitances by considering the design parameters as changing factors (i.e., the size variation and the position shift) under an assumption of Gaussian distribution. Consequently, we further apply the models to optimize the designed circuit. The designing parameters of these models are selected and optimized to fit the designing target of the examined circuit by the genetic algorithm in our unified optimization framework. This computational statistics method predicts the capacitances' effects on the gate delay time and compares with full 3D simulation approaches, it shows the engineering practicability in display panel industry.

## 1 Introduction

Thin film transistors (TFTs) have found wide usage in active matrix liquid crystal displays [13]. The basic principle of operation of the liquid-crystal display (LCD) panel is to control the transparency of each pixel portion by bus lines to charge the pixel electrode. To obtain high display performance, the capacitance of each pixel plays very important role in display circuit design. However, the capacitance of a pixel is very hard to be analyzed in a computationally efficient way because of the three-dimensional (3D) complex geometry structure. In this paper, we complete a

Yiming Li, Hsuan-Ming Huang
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu 300, Taiwan, e-mail: ymli@faculty.nctu.edu.tw

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_30,

systematical method to analyze and optimize the capacitance of an interested TFT device, shown in Fig. 1, using a 3D technology computer aided design (TCAD) filed simulation [6], a computational statistic method and a genetic algorithm (GA). Figure 1(a) shows the equivalent circuit of TFT-LCD panel, which has 1280 × 1024

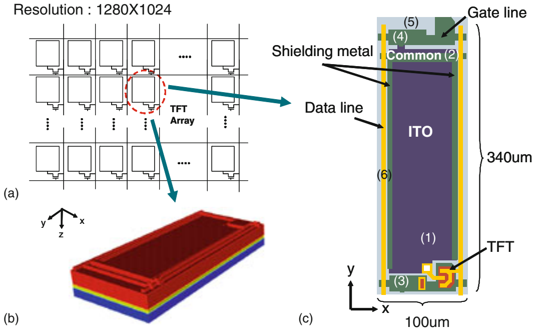


**Fig. 1: a** Equivalent circuit of TFT-LCD panel **b** 3D schematic plot of the TFT-LCD pixel **c** and perspective plot of the TFT-LCD pixel

**Table 1:** The upper and lower limits of the TFT-LCD designing parameters

| Variable | Parameters | Variation range ($\mu$m) |
|---|---|---|
| | Size variation | |
| A | Gate line | [0,10] in y |
| B | Shield metal (left) | [0,3] in x |
| C | Shield metal (right) | [0,3] in x |
| D | Data line | [0,3] in x |
| E | ITO electrode | [0,5] in x |
| F | ITO electrode | [0,5] in y |
| | Position shift | |
| G | Shield metal (left) | [0,5] in x |
| H | Shield metal (right) | [-5,0] in x |
| I | ITO electrode | [-2,2] in y |

resolution. The 3D schematic plot of the pixel in this panel, which has twelve layers, is shown in Fig.1(b) and the perspective plot is shown in Fig. 1(c). A computational statistics methodology is developed and implemented which consists of a Design of Experiment (DOE) setup and a second-order response surface model (RSM). By considering the designing parameters as changing factors (i.e., the size variation and the position shift), listed in Tab. 1, according to the DOE, we construct a RSM for the capacitances of TFT-LCD. Designing parameters such as the gate line, the shield metal, the data line and the ITO electrode are corresponding to the parts (4), (2), (6) and (1) as shown in Fig. 1(c), respectively. The RSM can explain the behavior of capacitances on the investigated TFT-LCD pixel. We simplify the RSMs using a step-wise regression, and verify their accuracy by the residual of squares. Under a Gaussian distribution, the model allows us to analyze the sensitivity of capacitances in a TFT-LCD pixel with respect to the aforementioned factors efficiently. These models also enable us to optimize the designing targets of the tested TFT-LCD pixel.

The paper is organized as follows. In Sec. 2, we briefly describe the computational statistics approach for the structural analysis and design optimization of TFT-LCDs. In Sec. 3, the simulation results are discussed. Finally, we draw conclusions and suggest future work.

## 2 Computational Methodology

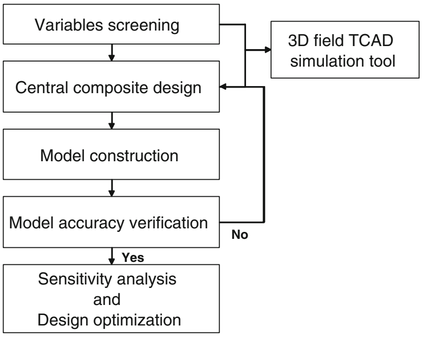

Variables screening
3D field TCAD simulation tool
Central composite design
Model construction
Model accuracy verification
No
Yes
Sensitivity analysis and Design optimization

**Fig. 2:** A flowchart of the computational method

The computational statistics methodology that can be used to account for the characteristic sensitivity and circuit design optimization is depicted in Fig. 2. Variables selection is a procedure to find the significant factors from a list of many potential candidates. Alternatively, we use a screening design or empirical check to identify significant main effects, rather than interaction effects, the latter being assumed an order of magnitude less important. A Plackett-Burman [5, 12] design was used to determine major contrasts and interactions. Based on the results of the screening experiment, parameters in need of further study were identified. With a Central Composite Design (CCD) DOE technique, a 3D field TCAD simulation [6] is performed to calculate the passive components of the studied TFT-LCD structure. From this we constructed the RSM [2,5,7,11]; mathematically, the response surface models can be represented as second-order polynomials:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k} \sum_{i \neq j}^{k} \beta_{ij} x_i x_j + \varepsilon, \tag{1}$$

where $k$ is the number of input factors, $x_i$ is the $i$th input factor, $\beta_i$ is the $i$th regression coefficient, and $\varepsilon$ represents model error. Several techniques, such as normality assumption and plot of residuals versus predicted value, to verify the adequacy of the RSM are then used [1–4, 11]. For the investigated structure, a $2^{nd}$ order model is established between capacitances (i.e., responses) and design parameters (i.e., factors). We notice that we didn't include scaling before the modelling. The details of the variables selection, the central composite design and the response surface model construction can be found in the reference [8].

Next, the sensitivity of the capacitances as approximately modeled by the RSM can be explored through a random procedure of statistics accordingly. Based upon our unified optimization framework (UOF) [9], we further develop a genetic algorithm (GA) [10] technique for the circuit design optimization. The flow of the GA evolutionary architecture is as follows. First, gene encoding is the way to encode the parameters into genes on the chromosome. Next step is to evaluate the fitness of each individual according to the stopping criteria. Then we select better chromosomes and breed a new generation through crossover and mutation. Finally, the fitness of the new generation is evaluated and the process is repeated for a specified number of generations or until achieving to desired targets. In the circuit design of TFT-LCD pixel, the capacitances in the SPICE netlist are encoded as optimization variables, and the fitness functions are constructed using an interested circuit performance. Here, the gate delay time of the studied whole display panel, shown in Fig. 1(a), is minimized. The designing parameters of RSMs for the capacitances that we have constructed are used to optimize the gate delay time of all TFT-LCD pixels in this study.

## 3 Results and Discussion

Among various designing parameters, variables screening has resulted in nine important factors as listed in Tab. 1. The table also shows the upper and lower limits of these designing parameters. We construct the $2^{\mathrm{nd}}$ order RSM using 149 runs with the CCD for the ten capacitances in a TFT-LCD pixel. We have constructed ten response surface models for different capacitances. Without loss of generality, here, we merely list two models for the most important capacitances $C_{12}$ and $C_{15}$, which are the capacitance between the part (1) and the part (2) and the capacitance between the part (1) and the part (5), as shown in Fig. 1(c), respectively.

$$\begin{aligned}\log \mathbf{C_{12}} = {} & +2.57101+0.021036\cdot B+0.025445\cdot C+0.051072\cdot E\\ & -0.10289\cdot F-0.021547\cdot G+0.022297\cdot H-0.025643\cdot I\\ & +0.019341\cdot E\cdot F+3.46849\cdot 10^{-3}\cdot E\cdot G-5.22133\cdot 10^{-3}\cdot E\cdot H\\ & +0.012046\cdot E\cdot I-4.20268\cdot 10^{-3}\cdot F\cdot G+3.78314\cdot 10^{-3}\cdot F\cdot H\\ & +9.85079\cdot 10^{-3}\cdot F\cdot I,\end{aligned} \tag{2}$$

and

$$\mathbf{1/\sqrt{C_{15}}} = +0.058861 + 1.16740 \cdot 10^{-4} \cdot B + 1.23106 \cdot 10^{-4} \cdot C \\ -6.42761 \cdot 10^{-4} \cdot E + 9.38689 \cdot 10^{-4} \cdot F + 5.39816 \cdot 10^{-5} \cdot G \\ -5.16325 \cdot 10^{-5} \cdot H - 2.69947 \cdot 10^{-5} \cdot I, \tag{3}$$

where $A$ to $I$ are designing parameters as listed in Tab. 1. The residual of squares for the formulated $C_{12}$ and $C_{15}$ are 0.9141 and 0.9887, the others are listed in Tab. 2. Figure 3 shows the residual normal probability plot and the residuals versus the predicted plot for the capacitance $\log \mathbf{C_{12}}$, and Fig. 4 shows the model adequacy checking for $\mathbf{1/\sqrt{C_{15}}}$. This examination highly reflects the modelling functionality for the RSM of these capacitances. The scatter plots of values calculated from the response surface models versus the simulated values obtained from the 3D field TCAD simulator for the models of these two capacitances, as shown in Fig. 5. The results show that there is a high linearity between the actual and predicted values. This confirms the accuracy of the constructed models. Figure 6(a) shows relationship between $C_{12}$ and ITO size variation along the y direction (i.e., the parameter $F$), given in Tab. 1, where the other parameters are set to the nominal values, and under Gaussian distribution (with more than 10000 trails and $3\sigma$, practically determined by the process variation, is about $0.25\,\mu$m). The standard deviation of the $C_{12}$ and $C_{15}$ due to the variation of the parameter $F$ of the tested pixel TFT-LCD is shown in Fig. 6(b). The sensitivity analysis between the $C_{15}$ and the parameter $F$ is depicted in Fig. 7. It is found that 1.7385 fF increase of $\sigma C_{12}$ and 0.2157 fF of $\sigma C_{15}$ when the ITO size varies from $4\,\mu$m to $1\,\mu$m. The increase of $\sigma C_{12}$ and $\sigma C_{15}$ are mainly due to relatively large variations of the circuit performance when the ITO size variance decreases in the y direction. Besides, to the other designing parameters in this TFT-LCD pixel sensitivity analysis can be performed by exploiting the RSMs. The gate delay is one of the most significantly limiting factors for the large-screen-size and high-resolution TFT-LCD design. We successfully reduce the gate delay time from 2877.1 ns to 8.2289 ns by the GA on the platform of UOF. The optimized designing parameters of the whole display panel are listed in Tab. 3. We notice that this estimation should be subject to further investigation by individually constructing RSMs with respect to each TFT-LCD pixel. Therefore, the individual behavior can be further examined for this TFT-LCD pixel.

**Table 2:** R-square of the constructed capacitance models

| Response | R-Square | Response | R-Square |
|---|---|---|---|
| $C_{12}$ | 0.9141 | $C_{14}$ | 0.9999 |
| $C_{15}$ | 0.9887 | $C_{13}$ | 0.9993 |
| $C_{56}$ | 0.9939 | $C_{26}$ | 0.9849 |
| $C_{35}$ | 0.9976 | $C_{46}$ | 0.9999 |
| $C_{16}$ | 0.9823 | $C_{24}$ | 0.9807 |

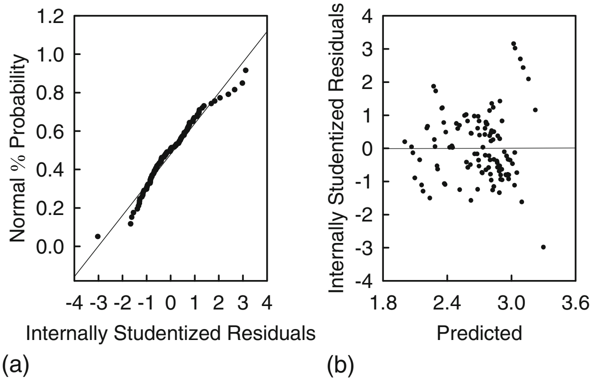


**Fig. 3:** **a** The residual normal probability plot and **b** the residuals vs. predicted plot for the response $\log \mathbf{C_{12}}$

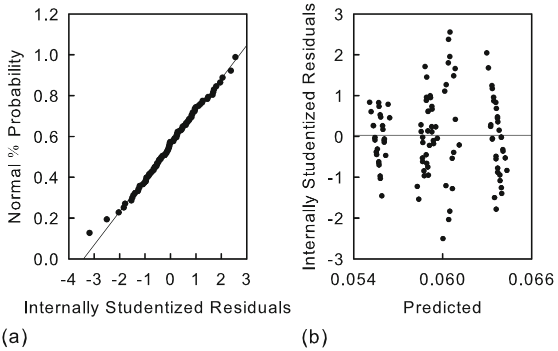


**Fig. 4:** **a** The residual normal probability plot and **b** the residuals versus the predicted plot for the response of $\mathbf{1/\sqrt{C_{15}}}$

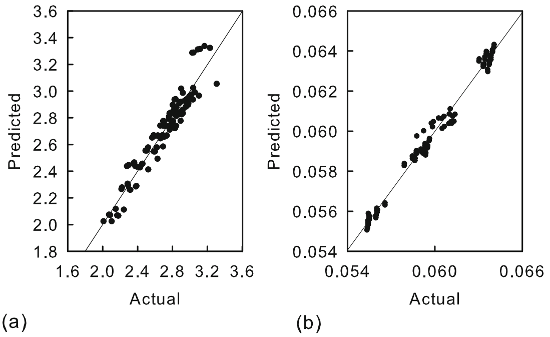


**Fig. 5:** Scatter plots calculated from the response surface model versus the simulated values obtained from the 3D field TCAD simulator for the **a** $\log \mathbf{C_{12}}$ and **b** $\mathbf{1/\sqrt{C_{15}}}$

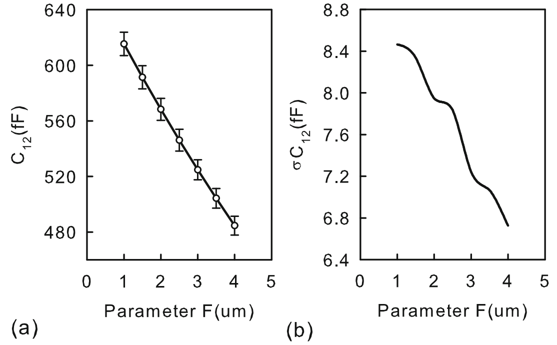


**Fig. 6:** **a** The relationship between the $C_{12}$ and the parameter $F$. **b** The standard deviation of the $C_{12}$ versus the parameter $F$

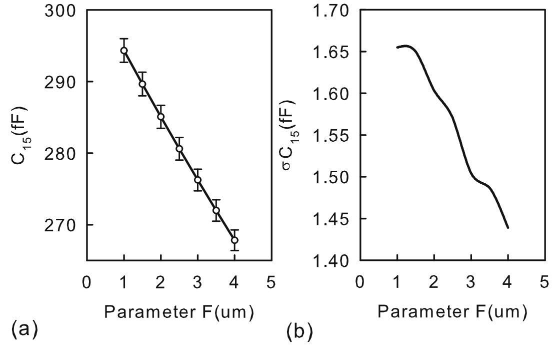


**Fig. 7:** **a** The relationship between the $C_{15}$ and the parameter $F$. **b** The standard deviation of the $C_{15}$ versus the parameter $F$

**Table 3:** A set of optimized designing parameters of the tested TFT-LCD pixel for the gate delay time minimization

| | A | B | C | D | E | F | G | H | I | Gate delay time |
|---|---|---|---|---|---|---|---|---|---|---|
| Original | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2887.1 ns |
| Optimized | 4.919 | 0.0525 | 0.7985 | 2.2996 | 4.527 | 4.535 | 3.2022 | -3.932 | -0.6898 | 8.229 ns |

# 4 Conclusions

In this work, we have successfully implemented a computational statistics technique for the capacitance sensitivity analysis of a TFT-LCD pixel and design optimization of the whole TFT-LCD circuit. Based on the 3D field solver and the central composite design method, the second-order response surface models have been constructed for the structural capacitances. Consequently, the constructed models were applied to study the sensitivity of capacitance with respect to the structural designing parameters and optimal design of the gate delay time of the tested TFT-LCD circuit in a computationally effective way, compared with a full 3D TCAD simulation. This approach can be incorporated into CAD tools for TFT-LCD design and can benefit the design automation of display panels.

**Acknowledgements** This work was supported by National Science Council (NSC),Taiwan under Contract NSC-97-2221-E-009-154-MY2 and Contract NSC-96-2221-E-009-210, and by the Innolux Display Corp., Chu-Nan, Taiwan under a 2006-2008 grant.

## References

1. Boning, D.S. and Mozumder, P.K.: DOE/Opt: A System for Design of Experiments, Response Surface Modeling, and Optimization Using Process and Device Simulation. IEEE Trans. Semiconductor Manufacturing, **7(2)**, 233–244 (1994)
2. Box, G.E.P. and Draper, N.R.: Empirical Model-Building and Response Surfaces, Wiley, New York (1987)
3. Daniel, C.: Use of Half-normal Plots in Interpreting Factorial Two-level Experiments. Technometrics, **1**, 311–341 (1959)
4. Dodgson, J.H.: A Graphical Method for Assessing Mean Squares in Saturated Fractional Designs. Journal of quality technology, **35**, 206–212 (2003)
5. Engineering Statistics Handbook [Online]. URL http://www.itl.nist.gov/div898/handbook/.
6. Li, Y., Chou, H.-M., Lee, J.-M. and Lee, B.-S.: A three-dimensional simulation of electrostatic characteristics for carbon nanotube array field effect transistors. Microelectronic Engineering, **81**, 434–440 (2005)
7. Li, Y. and Chou, Y.-S.: A Novel Statistical Methodology for Sub-100 nm MOSFET Fabrication Optimization and Sensitivity Analysis. In: Ext. Abs. 2005 Int. Conf. Solid State Devices and Materials, pp. 622–623 (2005)
8. Li, Y., Li, Y.-L. and Yu, S.-M.: Design optimization of a current mirror amplifier integrated circuit using a computational statistics technique. Mathematics and Computers in Simulation, **79**, 1165-1177 (2008)
9. Li, Y., Yu, S.-M. and Li, Y.-L.: Electronic design automation using a unified optimization framework. Mathematics and Computers in Simulation, doi:10.1016/j.matcom.2007.11.001 (2007)
10. Li, Y: An automatic parameter extraction technique for advanced CMOS device modeling using genetic algorithm. Microelectronic Engineering, **84(2)**, 260–272 (2007)
11. Myers, R.H. and Montgomery, D.C.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments, John Wiley Sons Inc., New York (2002)
12. Plackett, R.L. andBurman, J.P.: The design of optimum multifactorial experiments. Biometrika, **33**, 305–325 (1946)
13. Wu, I.-W.: Polycrystalline silicon thin film transistors for liquid crystal displays. Solid State Phenomena, **37-38**, 553–564 (1994)

# Lookup-Table Based Settling Error Modeling in SIMULINK

Marko Neitola and Timo Rahkonen

**Abstract** This work presents a data-based behavioral modeling scheme for switched-capacitor integrator settling error. In a typical SIMULINK behavioral model, settling behavior is implemented as a conditional, equation-based block. Here, the amplifier model is first characterized by a full range of amplifier's initial input and output voltages. The resulting settling errors are tabulated and finally, the settling error table is used directly as a lookup-table in behavioral simulations. One- or two-dimensional lookup-tables are standard library blocks in SIMULINK. This means that the actual settling error model is independent of the modeled amplifier topology, which is clearly a welcome feature in behavioral modeling.

## 1 Introduction

Continuous-time transient analyses for large mixed-signal circuits like switched-capacitor (SC) delta-sigma (DS) A/D-converters are known to be very slow. Therefore, a discrete-time behavioral model is usually necessary. The target simulator here is SIMULINK, a popular simulation platform integrated with MATLAB. This paper provides a new perspective on the modeling of the amplifier nonlinear settling behavior that avoids solving differential equations. The settling error model is a part of a larger SC DS-converter behavioral model. Being discrete-time, continuous-amplitude systems, SC-circuits suit very well into the SIMULINK modeling environment.

An accurate settling error model requires modeling of both slew-rate (SR) limited region and bandwidth-limited region. Since SC-amplifiers spend a lot of time in the slew-rate limited operation, any analytical settling analysis may become very complicated. The amplifier itself has properties like output-voltage dependent voltage gain, whose contribution to the settling error can be significant in case of

Marko Neitola, Timo Rahkonen
University of Oulu, Oulu, Finland, e-mail: marko.neitola@ee.oulu.fi, timo.rahkonen@ee.oulu.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_31,

low-voltage amplifiers. Moreover, analytical modeling of two- or multidimensional phenomena is usually obscure due to a large amount of non-dominant terms.

The method presented in this paper relies on the SC-integrator transient simulations followed by settling error tabulation. This part is called the characterization. After this, the resulting settling error table will be used in the DS-converter behavioral model as a lookup-table (LUT). The ease of use in the proposed method can be justified by the fact that a lookup-table is a standard library block in SIMULINK. The content of LUT is a re-definable variable in the MATLAB workspace. The settling error table can be originated from any type of simulator at any user-defined abstraction level. Obviously the most accurate results are obtained from a transistor level characterization. Here, we use a realistic two-pole amplifier model defined in MATLAB. The reason for this is that in our case study, we need a quick characterization platform for an exceptionally large circuit parameter sweep. The circuit parameters are swept to perform ca. thousand characterizations in order to visualize performance boundaries as a function of both slew-rate and phase margin. A transistor level characterization would be more feasible at the stage where all the circuit components are dimensioned.

There are several publications, e.g. [1–3], dedicated on modeling of SC DS-converters with nonlinearities in SIMULINK-environment. The proposed modeling method can be easily added to a behavioral model that also includes other significant non-idealities.

Section 2 discusses the characterization of the integrator settling error. The lookup-table range and size allocation are discussed in Section 3. Finally, Section 4 introduces a behavioral SIMULINK-model of a second order SC delta-sigma A/D converter along with the simulation and benchmarking results.

## 2 The Characterization

### 2.1 The Charge Distribution and Initial Voltages

The characterization is based on a group of amplifier settling simulations. Exciting the integrator with a group of input and output signals results to a group of initial voltage pairs, which are the prerequisite of settling simulations (Sect. 2.2). As an example model, we use a popular, parasitic insensitive SC-integrator, see Fig. 1. Capacitances are the sampling capacitor $C_S$, the feedback capacitor $C_I$ and the parasitic capacitance $C_p$. In the periodic transition between two integrator phases, the total charge is passively redistributed. This results in initial voltages $v_{i0}$ and $v_{o0}$ at the input and output nodes of the amplifier, respectively.

In our example, the SC-integrator is clocked so that the sampling capacitor $C_S$ is charged during one clock cycle (phase $\phi 1$). In the next cycle (phase $\phi 2$), due to closed feedback loop, the integrator forces $v_{i0}$ towards zero and the sampled charge is moved to load capacitor $C_L$. The load capacitor $C_L$ is typically the sampling capacitor of the next integrator in case of DS-converters. The SC-integrator in Fig. 1 is set to full-delaying and non-inverting mode.

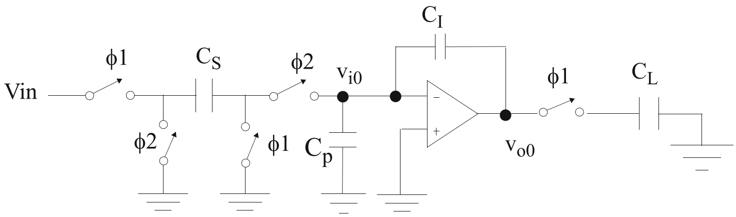


**Fig. 1:** Parasitic insensitive SC integrator phases: sampling phase $\phi 1$ and charge transfer phase $\phi 2$

The initial voltage at the input and the output of the amplifier is determined by studying the passive charge transfer in the capacitive network around the amplifier. In a generalized network of Fig. 2a, all capacitors are initially charged to voltages $V_{10}$ to $V_{40}$. When the capacitors are connected to voltages $V_1$ to $V_3$ in Fig. 2b, the total charge is passively re-distributed, resulting in initial voltages $v_{i0}$ and $v_{o0}$ at the input and output of the operational amplifier. The initial conditions can be solved from the charge-sharing equation:

$$\begin{bmatrix} C_S + C_I + C_p & -C_I \\ -C_I & C_I + C_L \end{bmatrix} \cdot \begin{bmatrix} v_{i0} \\ v_{o0} \end{bmatrix} = \begin{bmatrix} C_S V_{10} + C_I V_{20} + C_p V_{40} + C_S V_1 + C_p V_3 \\ -C_I V_{20} + C_L V_{30} + C_L V_2 \end{bmatrix} . (1)$$

At the right-hand side of (1), $V_{10}$ to $V_{40}$ are the voltages across capacitors charged at the previous switching phase. It is quite straightforward to apply charge-sharing equations for either switching phases of the circuit in Fig. 1.

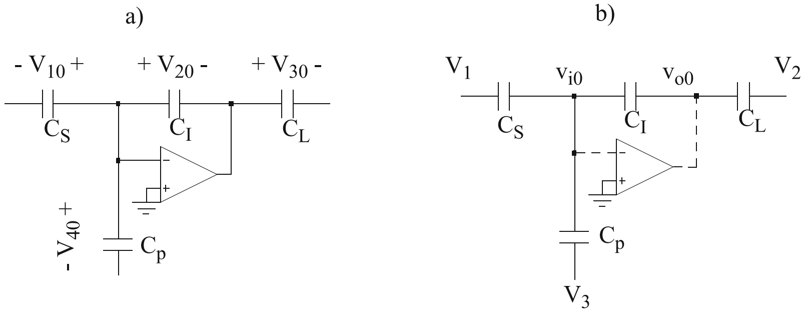


**Fig. 2:** The principle of calculating initial voltages: **a** Initial charges and **b** capacitors connected to source voltages

## 2.2 The Amplifier Model and the Settling Simulations

The settling simulations in the characterization stage can be performed by using a simulation platform of user's choice: Spice, Verilog-A, VHDL-AMS, MATLAB, etc. Here, a state-space model of the circuit in Fig. 3 was created using MATLAB. The model is a 2-pole folded-cascode amplifier complemented with nonlinearities that illustrate the usage of our method.

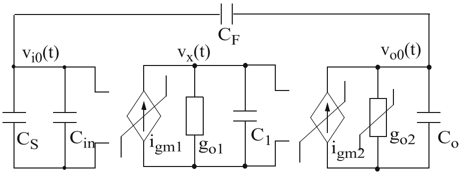


**Fig. 3:** The small-signal amplifier model

The effects of slewing are analyzed as in [4] where the slew-rate induced distortion of SC integrators is studied by assuming a piece-wise linear transconductance gm1, the current $i_{gm1}$ saturates to the value of bias current $i_{1MAX}$. The slewing is modeled by tanh-nonlinearity in the input transconductance:

$$i_{gm1}(t) = i_{1MAX} \cdot \tanh \frac{v_{i0}(t)}{V_{max}}, \tag{2}$$

where $V_{max}$ is the maximum input voltage level. The nonlinearity of the output stage conductance ($g_{o2}$ in Fig. 3) is realized by assuming a quadratic dependency on initial output voltage $v_{o0}$. The output stage current $i_{gm2}$ is

$$i_{gm2}(t) = -gm_2 \cdot v_x(t), \tag{3}$$

which constitutes the output voltage dependent gain. The voltage $v_x$ is the node-voltage at the previous stage of small-signal circuit in Fig. 3.

In our MATLAB-based model, the ideal integrator output voltages are the same (user-defined) output voltages used to calculate the initial voltages. All input-output voltage combinations result in different initial voltages and settling errors. The actual settling error (which will be tabulated) can be defined for an amplifier step response stepping from 0V to $V_{in}$ [1]:

$$\varepsilon_{settl} = (V_{in} - SR \cdot t_{sl}) \cdot e^{-t_{exp}/\tau}, \tag{4}$$

where $t_{sl}$ and $t_{exp}$ are the durations of nonlinear and linear settling, respectively. The linear part is defined by the amplifier bandwidth, or time constant $\tau$.

Our model is used in a parameter sweep. Sweeping the value of $I_{1MAX}$ in (2) affects the slew-rate (SR) defined for Fig. 3 in (5). Sweeping the capacitance $C_1$ moves the non-dominant pole in the transfer function shown in (6) i.e. it affects the gain bandwidth as well as the phase margin (PM) of the integrator.

$$SR = \frac{i_{1MAX}}{C_o + \frac{C_F \cdot C_S}{C_F + C_S}}, \tag{5}$$

$$H(\mathrm{j}\omega) = \frac{gm_1}{g_{01} + \mathrm{j}\omega C_1} \cdot \frac{gm_2}{g_{02} + \mathrm{j}\omega C_2}. \tag{6}$$

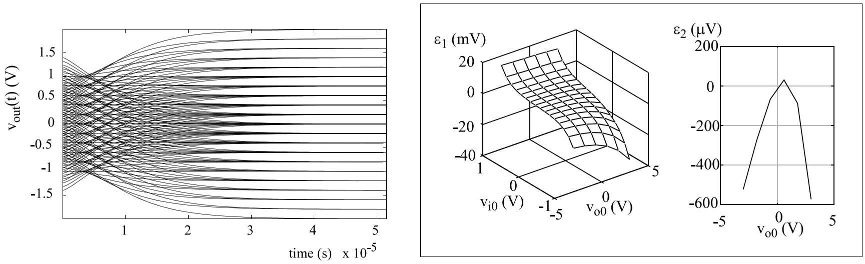
The switching between two capacitance networks requires two separate characterizations. The settling simulations are made for both capacitance networks (phases $\phi 1$ and $\phi 2$), and the simulation time-step is determined by the circuit's smallest time-constant.

The nonlinear settling simulation of a state-space function model can be presented using a general Laplace-domain nodal analysis in an iterative loop:

$$\begin{aligned}
&s \cdot \overline{C} \cdot \overline{V} + \overline{G} \cdot \overline{V} = \overline{I} \\
&\Leftrightarrow s \cdot \overline{V} = \overline{C}^{-1} \cdot \overline{I} - \overline{C}^{-1} \cdot \overline{G} \cdot \overline{V} \\
&\Leftrightarrow \overline{\Delta V} = \Delta t \cdot \overline{C}^{-1} \cdot \overline{I} - \Delta t \cdot \overline{C}^{-1} \cdot \overline{G} \cdot \overline{V} \\
&\overline{V}_{new} = \overline{\Delta V} + \overline{V}
\end{aligned}, \tag{7}$$

where $\Delta t$ is the time step and matrices $\overline{C}$, and $\overline{G}$ contain capacitances and conductances, respectively. At the beginning of a settling simulation, $\overline{V}$ contains the initial node voltages (Fig. 3). Nonlinear branch currents in $\overline{I}$ are solved by (2) and (3). The increment for the current voltage $\overline{\Delta V}$ is calculated an added to the original voltage vector, resulting the new voltage vector $\overline{V}_{new}$. $\overline{V}_{new}$ is calculated according to previous voltages and currents (forward Euler integration). At the next iteration, the new value is assigned as $\overline{V}$, and the next voltage vector is solved. The iteration is continued until the end of settling period.

In Fig. 4a, we have a set of amplifier output voltage settling curves for a group of initial input and output voltages. The output voltages at the end of settling period are then compared to ideal values and the settling errors are stored. For the settling error, being a function of one or two initial voltages is solely dictated by the capacitance network around the amplifier. Here, as seen in figure Fig. 4b, when the switching phase is $\phi 1$, the settling error is a function of both $v_{i0}$ and $v_{o0}$. At $\phi 2$, only $v_{o0}$ affects on the settling.



a) b)

**Fig. 4: a** The settling curves and **b** graphical presentation of the settling error tables

## 3 Lookup-Table Allocation

An important issue in the chosen modeling method is that only the characterized settling errors are tabulated, not the complete response. This alleviates the table allocation: much coarser presentation is adequate.

The lookup tables are used as SIMULINK's one- and two-dimensional lookup table blocks: "Lookup Table" and "Lookup Table (2-D)". Both blocks perform linear interpolation and extrapolation [5]. Constructing the error table requires some approximation for the upper and lower limits for the initial voltages. It is imperative that the tables are allocated so that they never extrapolate during simulation. This can be ensured by a group of simulations with default parameters and a proper safety margin. The range of initial voltages depends on the input signal type and the amplitude. In addition, the system topology contributes on the range, as there are feedback signals summed to the input of the integrating amplifier.

The LUT-model accuracy is naturally limited by the accuracy of the characterization model. Furthermore, the lookup-table needs sufficient amount settling error measurements in one characterization, but overdetermining the LUT slows down both characterization and simulation. One way to test and assure a proper amount of table data is to perform a polynomial LMSE-fit to the settling error table. Calculating the coefficient of determination [6] between true table and the fit reveals the required degree of a polynomial model. The degree is naturally proportional to the required table density.

Speaking of polynomial fitting, one might also consider polynomial modeling instead of using a LUT. Indeed, this is perfectly plausible and tested by the authors. A polynomial settling error model with alterable coefficients was created in SIMULINK using embedded m-file blocks. The simulations indicated that the results were highly identical, but LUT-based settling error model was ca. 4 times faster to simulate. Furthermore, a LUT-model is considerably easier to use as its contents is a matrix (or a vector) in MATLAB workspace.

## 4 The Behavioral Model

The objective for the behavioral model was to observe a delta-sigma A/D-converter's performance as a function of slew-rate and closed-loop phase margin. The model of a second order DS A/D-converter is presented in Fig. 5a. The order is defined by the number of consecutive integrators. The oversampling ratio is 32, sampling frequency is 10MHz and the integrator open-loop bandwidth is 65MHz. The DS-topology is a feed-forward type [7]. The modeled integrator has four inputs: the signal input *Vin*, the local resonator feedback *Vfb*, the D/A-feedback *Vref* and the integrator output *Vint*. The initial voltages are calculated in the behavioral model in every sampling interval by the numerical solution of (1). The capacitors were scaled according to the converter coefficients a, c and g in Fig. 5a. Only the first integrator (the left one in Fig. 5a) has the settling error lookup-table. This is because the distortion of the following integrator is divided by the gain of preceding integrator, i.e. the distortion is cancelled. The performance of the first integrator is crucial and it needs to be at least as linear as the overall DS converter's theoretical linearity.

The settling error $\varepsilon_2$(n) is one-dimensional and is summed to the output of the integrator. The error $\varepsilon_1$(n) is a two-dimensional error summed to the integrator input. The latter was found dominant settling error model throughout the parameter sweep of current $i_{1MAX}$ and $C_1$ ((2) and Fig. 3). A single example of such simulation with SNDR (signal to noise and distortion ratio) and SFDR (spurious-free dynamic range) results is shown in Fig. 5b. This is a "bad case scenario", where the phase margin and the slew rate are both very low ($50^o$ and $0.09V/ns$). The parameter sweep contained 1024 characterizations and simulations. For each simulation, both SNDR and SFDR performance were calculated. The object was to find the performance boundaries as a function of slew-rate (SR) and phase margin (PM). The maximum values for SNDR and SFDR were obtained from simulating the system with an ideal integrator. In Fig. 6a we have graph showing the boundary (and value), where the SFDR has deteriorated 6 dB from the maximum value. In Fig. 6b we have a boundary-graph of 3dB SNDR deterioration.

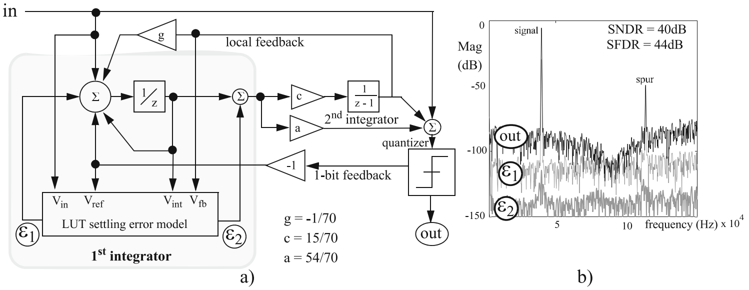


**Fig. 5: a** SIMULINK-model of a second order one-bit DS A/D converter and **b** an example of magnitude response with both settling error contributions

The black and gray graphs in Fig. 6 are the boundaries for one- and three-bit DS-converters, respectively. The three-bit converter had the same converter topology but had seven quantizer levels and different coefficients. With small slew-rates, the three-bit converter's SNDR and SFDR performance deteriorates less easily. This is explained by the fact that the three-bit feedback signal has significantly smaller step-size, making the amplifier less susceptible to slew-rate. In case of a low phase margin and a large slew-rate, the SNDR boundary of one-bit system is at lower phase margin values. This is simply due to lower SNDR performance.

The modeling methods presented in this work were applied using a laptop computer with 1,6GHz Pentium M processor and 1GB of RAM. MATLAB version was R2006b. The benchmarking times are based on the average of 1024 characterizations and simulations. One characterization for the lookup-table method also took ca 1,5 seconds. The SIMULINK model simulation times were very short, roughly one second in average for each $2^{16}$-point simulation.

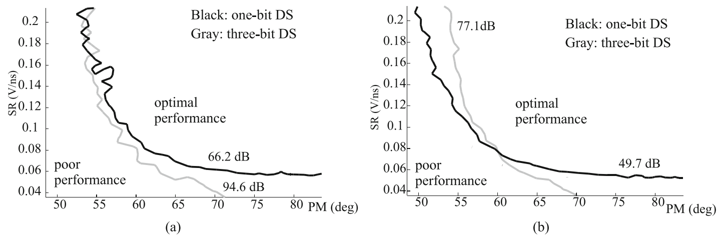


**Fig. 6:** Performance boundaries: **a** 6 dB SFDR and **b** 3 dB SNDR performance deterioration

# 5 Summary

Behavioral modeling of circuit nonlinearities is essential in case of large mixed-signal systems. A delta-sigma converter is a good example of such system, because it is inherently nonlinear and it needs long time-domain simulations to ensure stability. To this date, various SIMULINK-modeling methods for the most dominant non-idealities are reported. This work concentrated only on the settling error part and provided a very convenient model, which is based on a settling error lookup-table. The tabulation is made before the actual simulation in a characterization stage, which consists of a group of settling simulations.

The settling error lookup-table can be either one- or two-dimensional depending to the capacitor network. The settling error data can be originated from any simulator and with any abstraction level, or possibly even from device measurements. Once the amplifier model is constructed, a designer may re-characterize the amplifier and re-simulate without changing the behavioral model, because the simulation model is amplifier topology-independent. Moreover, this approach avoids constructing very complex analytical equations of strong nonlinearities like the slew-limitation.

# References

1. Koe W.M., Zhang, J.: Understanding the Effect of Circuit Non-idealities on Sigma-Delta Modulator. In: Proceedings of the 2002 IEEE International Workshop on Behavioral Modeling and Simulation, 94–101 (2002)
2. Fornasari, A., Malcovati, P., Maloberto, F.: Improved Modeling of Sigma-delta Modulator Non-idealities in Simulink. IEEE International Symposium on Circuits and Systems **6**, 5982–5985 (2005)
3. Zare-Hoseini, H., Kale I., Shoaei, O.: Modeling of Switched-Capacitor Delta-Sigma Modulators in Simulink. IEEE Tr. Instrumentation and Measurement **54**, 1646–1654 (2005)
4. Sansen, W., Qiuting, H., Halonen, K.: Transient Analysis of Charge Transfer in SC Filters - Gain Error and Distortion. IEEE J. Solid-State Circuits **22**, 268–276 (1991)
5. The Mathworks inc. *Simulink Release Notes, Version R14 and higher*.
6. Bock, R.K., Krischer, W.: The Data Analysis BriefBook. Springer. Series: Accelerator Physics. (1998)
7. Norsworthy, S.R., Schreier R., Temes, G.C.: Delta-Sigma Data Converters: Theory, Design and Simulation. IEEE Press Marketing. 476 p. (1997)

# Speed-Up Techniques for Time-Domain System Simulations

Timo Rahkonen

**Abstract** Many combined analog-digital (mixed-signal) systems involve quite a lot of Digital Signal Processing (DSP) functions. These circuits are simulated either by using sample-based behavioral models (often with fixed time-step), or by combining digital event-driven simulation with traditional transient analysis. The above approaches are expensive in some applications, and this paper presents ways of speeding up behavioural simulations of mixed-signal systems. As a system design example, linear state-space models are applied to study the effect of small timing errors in a time-interleaved digital-to-analog converter system.

## 1 Introduction

There are plenty of mixed-signal systems, that contain a lot of DSP, and some analog circuitry. Often the designer is interested in the spectral properties of modulated or pseudo-random data, and very long data sequences are needed to catch some statistical properties like spectral regrowth or average power efficiency. For example, designers would like to study quickly the effects of the following non-idealities:

- Settling errors of the switched-capacitor (SC) integrators in $\Sigma\Delta$ analog-to-digital converters.
- Small timing errors and slew rate errors (glitches) in the output of transmitter digital-to-analog converters [1].
- Losses and spectral response of modulated, linearly assisted switch-mode power supplies [3].

Such systems are usually simulated either by using sample-based (and often fixed time-step) Matlab or Simulink models, or by using hardware description

Timo Rahkonen
Department of Electrical and Information Engineering and Infotech Oulu, University of Oulu, Oulu, Finland, e-mail: timo.rahkonen@oulu.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_32,

languages like Verilog-A or VHDL-AMS. The latter employ normal non-linear transient analysis to solve the analog part. Traditional circuit simulators have been recently enhanced e.g. by employing hierarchical solvers and isomorphic mapping, where distinct blocks are solved separately, or similar types of circuits (like memory cells) are recognised and share the same solution (see e.g. [6]). These techniques speed up the analysis of very large digital circuits, but the basic problems of time-domain simulation of especially strongly resonant circuits remain: numerical integration algorithms warp the time constants, hence distorting the response of oscillatory circuits, and the spectral purity of the output signal is impaired due to excessive timestep and interpolation error problems [5].

This paper is focused on building easy and quick methods for modelling analogue effects without the need of solving differential equations numerically. This gives a non-warped response in highly oscillatory circuits, and fits nicely in behavioural Matlab and Simulink models.

## 2 Analysis Techniques

This paper presents two approaches, that offer quick modelling of analog behaviour in otherwise sample-driven systems, suited especially for Matlab and Simulink behavioural models. First, the spectral effects of small time-skew errors are estimated using the techniques presented in [1]. Second, state-space models of linear circuits are used to predict the response of the circuit, without the need of intermediate time steps. This is very handy in resonant circuits like the low-loss switch-mode power supplies [3] or switching amplifiers, but for consistency it is applied here to the analysis of time skew errors in a time-interleaved analog system.

### *2.1 Lumped Timing Error Modeling*

Small timing skew or glitches in a segmented D/A structure (illustrated in Fig. 1) cause very broadband spurious responses [7]. The time skew (in plot a and c) is often just some tens of picoseconds, and catching their spectral effect with FFT calls for very short time step. The effect of finite slew rate (SR, plots b and d) is even more expensive to model, as one needs several time points on the ramp to model its spectral response.

In [1] the following approach was used: standard Fourier integral was used instead of FFT to calculate the spectrum of the skew and slew-rate errors alone. The errors are small and occur only at sparsely distributed locations in time, and the short duration of these errors makes it possible to use piecewise constant or piecewise linear approximations when calculating their impact to the Fourier integral. This makes the simulation very efficient: we just need to determine the occurrence of the error, calculate the integral over the PWL error and add this increment in the

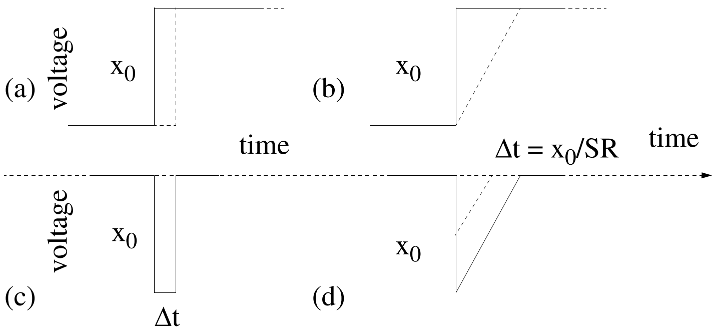


**Fig. 1:** Skew and slew rate errors in the output of a D/A converter

Fourier integrals. In the study it was also found out that below half the sampling rate $f_s/2$ the correct spectral response is actually achieved by using simple lumped error models, where the area of the errors are simply averaged over the sample duration $T$ and summed up to the ideal sample value. Below are the averaged models of a timing skew ($e_{skew}$) with amplitude $x_0$ and duration of $\Delta t$, and slew-rate error ($e_{SR}$) with amplitude $x_0$ and duration of $\Delta t$ (i.e., $SR = \frac{x_0}{\Delta t}$)

$$e_{skew} = x_0 \cdot \Delta t / T, \quad e_{SR} = \frac{x_0 \cdot \Delta t}{2 \cdot T}. \tag{1}$$

The upper approach is sufficient for finding tolerable amounts of timing errors in D/A converters. When calculating the Fourier integral, the ideal and distorted response are separated. This idea is somewhat tempting, as it may improve the dynamic range of frequency domain analysis (a somewhat similar approach has been used for transient noise analysis in [4]). However, this is complicated to implement in a general simulator, as the signal needs to be split into two parallel ones.

## 2.2 Time-Domain State-Space Modeling

Time domain solutions of state-space models have been used for quick design space exploration of e.g. class E power amplifiers [2] and linearly assisted switch-mode power supplies for envelope tracking RF power amplifiers [3]. The latter is a good example of a small continuous-time system, that is driven by a broadband modulated digital signal. Long sequences of modulated data combined with steeply driven switches and some critical timing issues result in excessive simulation times in normal transient analysis, but the state-modeling was a fast enough (roughly by a factor 1:10) to allow exhaustive sweeping of some critical design parameters, like the required bandwidth of the assisting amplifier.

Standard state space model of a linear system consists of state variables $X$, inputs $U$, and state matrices $A,B,C$, and $D$ :

$$\begin{aligned} sX &= A\cdot X + B\cdot U \\ Y &= C\cdot X + D\cdot U. \end{aligned} \tag{2}$$

Here $A$ is usually of form $C^{-1}\cdot G$ where $C$ contains the capacitive and inductive part, and $G$ the conductive part of the circuit. The state vector has the following general time-domain solution

$$x(t) = \phi(t)\cdot x(0) + \int_0^t \phi(t-\tau)Bu(\tau)\mathrm{d}\tau, \tag{3}$$

where $x(0)$ is the initial state vector and

$$\phi(t) = \exp(At) = I + At + \frac{A^2t^2}{2!} + \cdots . \tag{4}$$

Assuming that the input signals $U$ are stepwise constant (output of a D/A converter, for example), (3) can be simplified into the form

$$x(t) = \phi(t)\cdot x(0) + A^{-1}\cdot(\phi(t) - I)\cdot B\cdot U. \tag{5}$$

This is readily solvable at any time point with normal matrix operations, without iteration or intermediate time points. The analysis can also be cascaded so that the final solution at the end of one sample is used as an initial solution $x(0)$ to the next step. This approach is especially suited for strongly resonant circuits, that would require a small timestep in normal transient analysis. These include linear filters and linear but time-varying switching power-supply circuits.

One obvious disadvantage of the above approach is that it is limited to piece-wise constant input signals. This is sufficient for circuits controlled by D/A output, but one can also add a linear ramp into the state model at the cost of one additional state variable. Unfortunately, adding a lossless integrator makes the state transition matrix $A$ singular and prevents its inversion. Hence, small losses in the integrator are defined by $\delta$ in the below example that describes a 2nd order low-pass filter with quality factor $Q$ and corner frequency $\omega_0$

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -\omega_0^2 & -\omega_0/Q & 1 \\ 0 & 0 & -\delta \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, C = \begin{bmatrix} \omega_0^2 & 0 & 0 \end{bmatrix}, D = 0. \tag{6}$$

Here the driving input $U$ is two-dimensional, containing both the piecewise-constant value $V_{in}$ and its rate of change $\mathrm{d}V_{in}/\mathrm{d}t$:

$$U = \begin{bmatrix} V_{in} & \mathrm{d}V_{in}/\mathrm{d}t \end{bmatrix}^T. \tag{7}$$

## 3 State-Model of a D/A Converter Skew and Slew Rate Errors

The state modelling of section 2.2 is used here to solve the response of small skew and slew rate errors (shown in Fig. 1) in a D/A converter (DAC) followed by a 2nd order Butterworth low-pass filter. The sample rate of the DAC is $f_s$, and the low-pass corner frequency of the filter is $f_s/3$. Fig. 2a shows the sample-and-hold response (S/H) of a D/A converter, and the filtered response. The response of a single skew error of amplitude of 100 units and width of 0.01, 0.03 and 0.1 sample periods is shown in Fig. 2b. Similarly, the response of an SR error of the same width is shown in Fig. 2c (here the magnitude of the pulse is scaled down by multiplier 0.1 to make its triangular shape visible). In b the pulse energy is proportional to the pulse width $\Delta t$, while in c it is proportional to $\Delta t^2$, which is also seen in the amplitude of the ringing response.

In both Fig. 2b and c the narrow error pulse is seen to spread over several samples in the low-pass filter, making the calculation of Fourier integral more complicated than with short, unfiltered error pulses. However, the output responses are exact, and only two time points per sample is sufficient to calculate the response at the end of the sample: one at the beginning of the error, second at the end of it, and then one at the end of the sample, which is already the beginning of the next error. The responses in Fig. 2 were calculated by placing 20 time points per sample, but the response at time point t=1, for example, is exactly the same even if we calculate only two points per sample.

## 4 Test Case: Time-Interleaved D/A Converters

It is well known, that the normal sample-and-hold causes a $\frac{\sin f}{f}$ (sinc) shaped frequency response. Figure 3 illustrates a system [8], where the hold time $T_h$ of a digital-to-analog converter is extended to longer than one sample time $T_s$. This shifts the notch of the sinc response downwards, resulting in linear-phase filtering of the image frequencies. The presented setup is basically a time-interleaved system, and it is highly sensitive to timing and gain errors between the parallel paths. Hence, it is a good candidate to test the state modeling approach.

The D/A system is modelled as follows. The data is divided into even and odd samples, converted to voltage, multiplied by a 0/1 gating pulses $(g_o(t), g_e(t))$ that set the duration of the S/H pulses, summed together and fed to the state model of a 2nd-order low-pass filter. To catch the time-interleaved nature, the output is treated as pairs of even and odd samples, that may have different gain and offset. Timing errors are included so that the beginning time of the odd samples can be moved arbitrarily around time $(2 \cdot k + 1)T_s$, and the duration of even and odd samples $(T_{even}, T_{odd})$ can be set independently. Also skew and slew-rate errors can be created by additional points shown shortly after the beginning and end of the masking pulses. Fig. 4 shows the shape of the gating functions, and the circles present the breakpoints where the filter response needs to be evaluated.

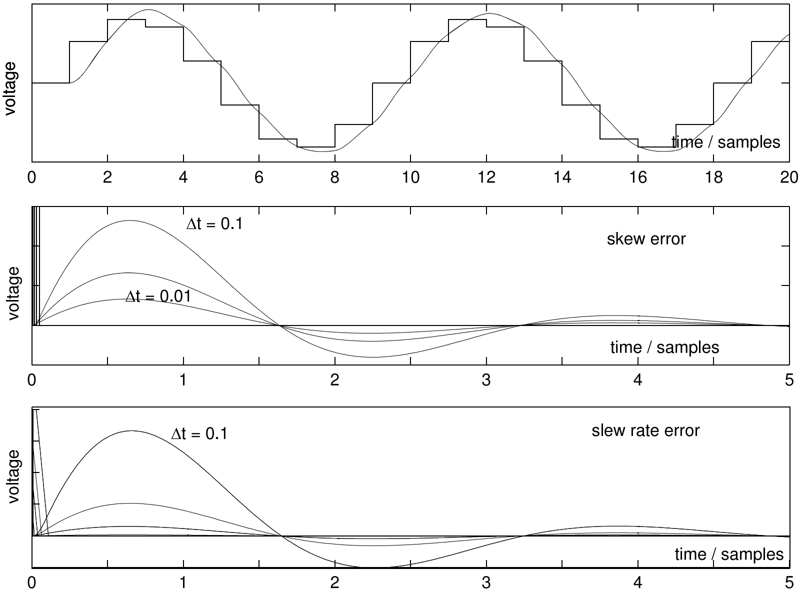


**Fig. 2:** Pulse responses in the output of a D/A converter **a** S/H and filtered response, **b** original and filtered rectangular skew error, **c** 1:10 attenuated triangular slew rate error and filtered response

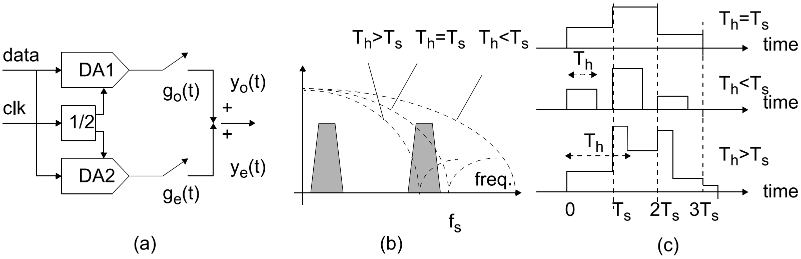


**Fig. 3:** **a** Structure of a time-interleaved D/A converter pair, **b** frequency response and **c** output waveforms as functions of varying hold time $T_h$

Figure 5a) shows an example of the intended operation and b) the effect of time-skew errors. In plot a), the hold time is 1.5 clock periods, no timing errors are present, and the spectrum is as expected. The four lower tones are the desired ones, and the 4 upper tones are sampling images, that need to be filtered away. Due to the $1.5T_s$ long hold-time, the $\frac{\sin f}{f}$ notch is placed to frequency bin 670 (bin 1024 corresponds to the sampling frequency) to attenuate the lowest image tone by more than 10 dB. To show the shape of the sinc response, some broadband noise is summed to the 4-tone test signal.

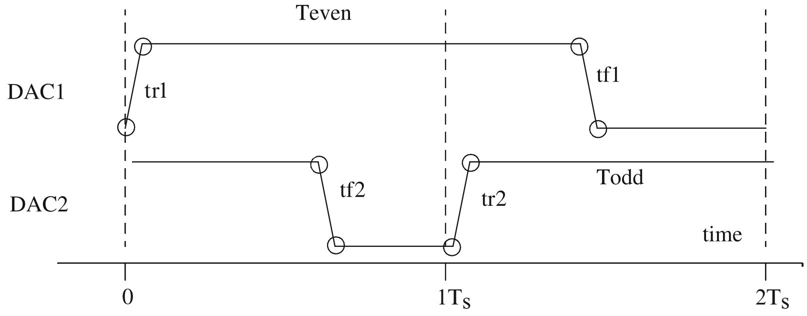


**Fig. 4:** Time masks for multiplying the D/A outputs. *Circles* show time instants, where the response needs to be calculated

Figure 5b), the output of the second D/A (DAC2) is delayed by 0.001 samples, and the spectral images caused by residual sampling at half the sampling frequency $(f_s/2)$ are clearly seen ca. 60 dB below the desired signals. Using normal fixed-spaced FFT, one would need 1000 points over the sample period $T_s$ to catch this effect, but here only 8 points are calculated per one sample. First, the values at every breakpoint are calculated, and then four points are evenly placed over the sample, and FFT is calculated. In a similar manner, one can include segmentation timing errors and slew rate errors into the model. If these are not needed (the idea is that they are masked by the gating signals $g_o(t), g_e(t)$), we can actually remove two of the break points, and repeat the above analysis by calculating only six points per sample.

# 5 Summary

This paper reviewed techniques for speeding up behavioural time-domain simulations and finding suitable design specifications of mixed-signal systems, especially in Matlab type modelling, where there is no real analog simulator kernel available. Fourier integrals were used instead of FFT to study the spectral effects of narrow error pulses. Closed-form linear state-space modeling was extended to operate with piecewise-linear input signals (opposed to piecewise-constant signals before), and it was applied to calculate the shape and magnitude of very small skew and slew-rate errors.

As a larger example, a time-interleaved D/A system was modelled and simulated, and the linear state model was shown to detect the effects of even very small timing errors with very modest oversampling, as the response needs to be calculated only at the breakpoints of the piecewise-linear signal and at a few fixed-spaced points needed for the FFT.

Linear state-space models are quite easy to apply to linear time-varying systems, too, as one only needs to know the time place where to switch the model. Modelling

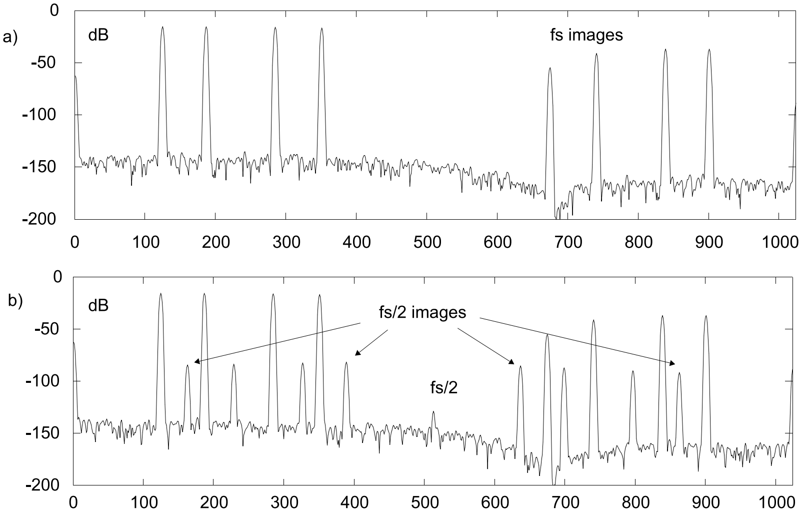


**Fig. 5:** Magnitude vs. frequency spectrum of the setup in Fig. 4 with 4-tone test. **a** Ideal response, **b** with time skew of 0.001 samples between the two DACs

nonlinear systems by piece-wise linear state models, however, already results in some iterative processing, as one needs to find the time where to switch the model.

## References

1. Rahkonen,T., Repo, H.: Efficient behavioral modelling of small timing errors in A/D and D/A converters. *Kluwer Academic journal on Analog Integrated Circuits and Signal Processing* **46**(1), 29–36 (2006)
2. Reynaert, P., Martens, K., Steyaert, M.: A state-space behavioral model for CMOS class E power amplifiers. *IEEE Trans on Computer-Aided Des. of Integrated Circ. Syst.*, **22**(2), 132–138 (2003)
3. Rahkonen, T., O-P Jokitalo, O-P: Design of a linearly assisted switcher for a supply modulated RF transmitter. *Springer journal on Analog integrated circ. and signal proc.* **54**(2), 113–119 (2008)
4. Demir, A., Sangiovanni-Vincentelli, A.: Analysis and Simulation of Noise in Nonlinear Electronic Circuits and Systems. Kluwer Academic Publishers (1998)
5. Kundert, K.: The Designer's Guide to Spice and Spectre. Kluwer Academic Publishers (1995)
6. Cadence Virtuoso Ultrasim. Online document (2008). URL http://www.cadence.com/rl/Resources/datasheets/virtuoso_mmsim.pdf p.10. Cited Dec 17 2008.
7. Doris, K., Leenaerts, D.M.W., van Roermund, A.H.M.: Time Non-linearities in D/A converters. *ECCTD01*, Espoo 2001, pp. III-353 – III-356 (2001)
8. Rahkonen, T., Aikkila, J.: Linear phase reconstruction filtering using hold time longer than one sample period. IEEE Trans. Circ. Syst. I **51**(1), 178–181 (2004)

# Part III
# Coupled Problems

# Introduction to Part III

Wil Schilders

This part addresses the challenging topic of solving *coupled problems*. The increasing necessity to solve complex problems in the science and engineering community, accounting for all the coupling occurring at the different scales of the problem, requires the development of new ideas and methods which can effectively provide accurate numerical solutions in affordable computation times. The state of the art is discussed here as well as mathematical, numerical, and computational methods for solving coupling problems of multidisciplinary character, with an emphasis on coupling with electromagnetic (EM) and/or circuit simulation. Special attention is paid to showing the potential of new computational methods for solving practical multidisciplinary problems of industrial interest.

The first three papers in this section are all aimed at thermal interactions in coupled device and circuit simulations. This is an important topic for the electronics industry and is also studied extensively in the European Research and Training Network CoMSON (see `http://www.comson.org`).

The invited paper by Brunk and Jüngel presents an overview on coupled simulations involving thermal effects in semiconductor devices and electronic circuits. A mathematical analysis of the coupling conditions of the two coupled models is carried out. Numerical results clearly show the significance of thermal effects in small semiconductor devices, leading to the conclusion that the inclusion of thermal models is indispensable in state-of-the-art simulations.

The paper by Ali et al. builds upon the analysis demonstrated in the paper by Brunk and Jüngel, and addresses the mathematical well-posedness of the steady-state and transient problems in coupled semiconductor–circuit systems. The paper shows the importance of mathematical analysis for coupled problems demonstrating that analysis is not at all straightforward, but requires extreme care.

One should also be careful, when using a reduced-order model for one part of the coupled problem and coupling this model to the full model. This is essentially

W.H.A. Schilders
NXP Semiconductors, Corp. I&T/DTF/A&M/Physical Design Methods, Mathematics, High Tech Campus 46, 5656 AE Eindhoven, The Netherlands, e-mail: wil.schilders@nxp.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_33,

the topic of the paper by Culpo et al., which concentrates on thermal issues, too. The paper also discusses how to cope with the multiscale nature of heat diffusion in VLSI circuits via a special meshing technique.

The contribution by Romano and Scordia departs from the purely thermal problem, and instead concentrates on energy-transport models based on the maximum-entropy principle. The coupled problem being investigated here is that of phonon–electron interaction in silicon. Such interactions lead to heating of the lattice. The paper concentrates on the numerical discretization scheme for the interaction equations, and some preliminary numerical results are shown.

The next two papers are again aimed at investigating, from a mathematical point of view, the coupling of different systems of equations, with an emphasis on electronic systems consisting of circuit and device equations. Baumans et al. discuss the problem of finding suitable and stable initial conditions and use the differential-algebraic structure as a vehicle for their analysis. The paper by Ali et al., in turn, addresses the problem of coupling circuit equations with a hydrodynamic device model, leading to a hyperbolic system of partial differential equations. The theoretical findings are confirmed by a numerical simulation of a unipolar device.

Continuing with the paper of Li and Hwang, we remain with coupled circuit and device equations, but now concentrate on the simulation of fluctuations caused by dopants. This is again an important topic in the electronics industry, known as variability, and has received a lot of attention recently. It entails the realistic point of view that the manufacturing process may lead to serious deviations from the original design. Hence, robust designs take into account potential deviations and fluctuations. The paper discusses the effects of such fluctuations on the coupled device–circuit model.

The industrial context is well represented in the invited paper by Schoenmaker at al., where the important problem of EM coupling between blocks in an integrated circuit is studied. This problem is also receiving much attention in the electronics industry, and known under names as chip peripheral co-design, parasitic electromagnetic coupling, or co-habitation. Undesirable EM coupling between various components is also an extremely difficult problem, as a full simulation is not feasible. Hence, techniques like domain decomposition must be used combined with intelligent reduced-order modeling strategies. The paper describes these techniques and also presents simulation results for a number of realistic industrial examples.

The next contribution by Plata et al. addresses similar issues, and presents in more detail a domain-decomposition method to address the co-habitation issue. New and rather revolutionary is the extraction of a reduced-order circuit that consists of both electrical and magnetic components, also referred to as EM hooks. The resulting algorithms based on this concept show very promising results.

Part III on coupled problems ends with a paper by Schöps et al. investigating the index of the differential-algebraic system consisting of coupling circuits with magnetoquasistatic conductor models. The contribution also discusses a convergence analysis and some numerical results are provided.

# Heating of Semiconductor Devices in Electric Circuits

Markus Brunk and Ansgar Jüngel*

*Invited speaker at the SCEE 2008 conference

**Abstract** Thermal effects in a coupled circuit-device system are modeled and numerically simulated. The circuit equations arise from modified nodal analysis. The transport in the semiconductor devices is modeled by the energy-transport equations for the electrons and the drift-diffusion equations for the holes, coupled to the Poisson equation for the electric potential. The lattice temperature is described by a heat equation with a heat source including energy relaxation heat, recombination heat, hole Joule heating, and radiation. The circuit-device model is coupled to a thermal network. The resulting system of nonlinear partial differential-algebraic equations is discretized in time using backward difference formulas and in space using (mixed) finite elements. Heating effects from numerical simulations in a *pn*-junction diode and a clipper circuit are presented.

## 1 Introduction

In modern ultra-integrated computer chips, secondary effects like self-heating strongly influence the switching behavior of the transistors and the performance of the circuit. In order to control the thermal effects, accurate circuit simulations are needed, which go beyond compact modeling and simplified temperature models. In this paper, we review a coupled circuit-device model taking into account the temperature of the electrons and the semiconductor lattice and the temperature of the circuit elements and present new numerical simulations illustrating the self-heating.

Markus Brunk
Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway, e-mail: markus.brunk@math.ntnu.no

Ansgar Jüngel
Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria, e-mail: juengel@anum.tuwien.ac.at

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_34,

First coupled circuit-device models were often based on a combination of device and circuit simulators [1]. More recently, electric network models were coupled to semiconductor transport equations, using drift-diffusion [2, 3] or energy-transport models [4]. Nonisothermal device modeling started in the 1970s, employing drift-diffusion-type equations and heat flow models for the lattice temperature [5]. A thermodynamic approach to extend the drift-diffusion equations to the nonisothermal case was presented in [6], later generalized in [7] using first principles of entropy maximization and partial local equilibrium. In [8], the energy-transport equations were coupled to a heat equation for the lattice temperature.

All these references are concerned with the modeling of certain subsystems. Here, based on our work [9], we present a complete coupled model, including (i) the device model consisting of the energy-transport equations for the electrons, the drift-diffusion equations for the holes, and a heat equation for the lattice temperature, (ii) the electric-network equations, and (iii) a thermal network model describing the heat evolution in the circuit elements, electric lines, and devices. The models are described in Section 2. The three subsystems are coupled by thermo-electric, electric circuit-device, and thermal network-device interfaces explained in Section 3. Finally, in Section 4, the heating behavior in a *pn*-junction diode and a clipper circuit is illustrated.

## 2 Model Equations

**Device modeling.** The electron transport in the semiconductor device is modeled by the energy-transport equations, whereas the hole transport is described by the drift-diffusion equations. The equations for the electron density $n$, the electron thermal energy $\frac{3}{2}k_BnT_n$ (with $k_B$ being the Boltzmann constant and $T_n$ the electron temperature), the hole density $p$, and the self-consistent electric potential $V$ read as [10]

$$\partial_t n - q^{-1}\mathrm{div}\, J_n = -R(n,p), \quad \partial_t p + q^{-1}\mathrm{div}\, J_p = -R(n,p), \tag{1}$$

$$\partial_t\left(\tfrac{3}{2}k_BnT_n\right) - \mathrm{div}\, J_w + J_n\cdot\nabla V = W(n,T_n) - \tfrac{3}{2}k_BT_nR(n,p), \tag{2}$$

$$\varepsilon_s \Delta V = q(n-p-C(x)), \tag{3}$$

where $q$ is the elementary charge, $\varepsilon_s$ the semiconductor permittivity, and $C(x)$ the doping profile. The function $R(n,p)$ models Shockley-Read-Hall recombination-generation processes and $W(n,T_n)$ the relaxation to the lattice temperature $T_L$,

$$R(n,p) = \frac{np-n_i^2}{\tau_p(n+n_i)+\tau_n(p+n_i)}, \quad W(n,T_n) = \frac{3}{2}\frac{nk_B(T_L-T_n)}{\tau_0}, \tag{4}$$

where $n_i$ is the intrinsic density, $\tau_n$ and $\tau_p$ the electron and hole lifetimes, respectively, and $\tau_0$ the energy relaxation time.

The constitutive relations for the electron current density $J_n$, the hole current density $J_p$, and the electron energy density $J_w$ are given by

$$J_n = q\Big(\nabla\Big(\mu_n \frac{k_B T_L}{q} n\Big) - \mu_n T_L \frac{n}{T_n} \nabla V\Big), \quad J_p = -q\Big(\nabla\Big(\mu_p \frac{k_B T_L}{q} p\Big) + \mu_p p \nabla V\Big), \tag{5}$$

$$J_w = \nabla\Big(\frac{3}{2}\mu_n T_n \frac{k_B^2 T_L}{q} n\Big) - \frac{3}{2}\mu_n k_B T_L n \nabla V, \tag{6}$$

where the mobilities for the electrons and holes, $\mu_n$ and $\mu_p$, respectively, are assumed to depend on the lattice temperature $T_L$ according to

$$\mu_j(T_L) = \mu_{j,0}\Big(\frac{T_0}{T_L}\Big)^{\alpha_j}, \quad j = n, p, \tag{7}$$

where $T_0 = 300\,\mathrm{K}$. The values $\mu_{j,0}$ and $\alpha_j$ $(j = n, p)$ are typically determined from measurements; see, e.g., [11, Table 4.1-1]. The current leaving the semiconductor device, which occupies the domain $\Omega \subset \mathbb{R}^d$ $(d \geq 1)$, at terminal $\Gamma_k$ is defined by

$$j_k = \int_{\Gamma_k} J_{\mathrm{tot}} \cdot \nu \, \mathrm{d}s \quad \text{with} \quad J_{\mathrm{tot}} = J_n + J_p + J_d, \tag{8}$$

where $\nu$ is the exterior normal unit vector to $\Gamma_k$ and $J_d = -\varepsilon_s \partial_t \nabla V$ the displacement current density. We choose one terminal as reference terminal. Due to charge conservation, the corresponding current can be computed by the negative sum of the other terminal currents collected in the vector $j_S$.

The model equation for the lattice temperature is derived from thermodynamic principles. Assuming that the thermal effects are due to the majority carriers (electrons), the free energy for the system of energy-transport and Poisson equations is the sum of the electric energy, the thermodynamic energy of the lattice subsystem, and the thermodynamic energy of the electron subsystem [7, 12],

$$f = \frac{\varepsilon_s}{2}|\nabla V|^2 + \rho_L c_L T_L (1 - \log T_L) + n\Big[k_B T_n \Big(\log \frac{n}{N_c} - 1\Big) + E_c\Big],$$

where $\rho_L$ denotes the material density, $c_L$ the heat capacity, $E_c$ the conduction-band energy, and $N_c$ the effective density of states depending on $T_n$ (see [13] for details). Then the internal total energy is given by

$$u = f - T_n \frac{\partial f}{\partial T_n} - T_L \frac{\partial f}{\partial T_L} = \frac{\varepsilon_s}{2}|\nabla V|^2 + \rho_L c_L T_L + n(E_c - T_L E_c') + \frac{3}{2} k_B n T_n,$$

where the prime denotes the derivative with respect to $T_L$. The associated total energy flux density $J_u$ is the sum of the energy flux in the electric field, the Fourier heat flux, and the electron energy flux:

$$J_u = V J_{\mathrm{tot}} - \kappa_L \nabla T_L - (E_c - T_L E_c') q^{-1} J_n - J_w,$$

where $\kappa_L$ is the heat conductivity of the lattice. Inserting the expressions for $u$ and $J_u$ into the energy balance equation $\partial_t u + \operatorname{div} J_u = -\gamma$ and employing the Poisson equation for $\partial_t V$, a straightforward computation leads to the heat equation for the lattice temperature (see [9] for details):

$$0 = \partial_t u + \operatorname{div} J_u + \gamma = \partial_t T_L(\rho_L c_L - nE_c') - \operatorname{div}(\kappa_L \nabla T_L) - H, \tag{9}$$

where $\gamma = S_L(T_L - T_{\text{env}})$ is the energy loss by radiation with the transmission constant $S_L$ and the environmental temperature $T_{\text{env}}$, and $H$ is the heat source term,

$$H = -W + R\big(E_c - T_L E_c' + \tfrac{3}{2}k_B T_n\big) + q^{-1} J_n \cdot \nabla(E_c - T_L E_c') - J_p \cdot \nabla V - S_L(T_L - T_{\text{env}}),$$

where the relaxation term $W$ is defined in (4). For related but different choices of the heat source term, we refer to the discussion in [6]. For nondegenerate homostructure devices, we can neglect the space dependency of the energy band. Furthermore, we neglect the dependency of the energy band on the lattice temperature since this dependency is rather small [11]. Thus, the heat source term becomes

$$H = -W + R\big(E_c + \tfrac{3}{2}k_B T_n\big) - J_p \cdot \nabla V - S_L(T_L - T_{\text{env}}),$$

The first term in $H$ represents the energy relaxation heat, the second term is the recombination heat, the third term expresses Joule heating from the holes, and the last term signifies the We notice that only the Joule heating from holes appears, as the Joule heating from electrons appears as source term in (2) and affects the lattice temperature indirectly via the relaxation term $W$.

The model equations (1)-(9) are complemented by initial and boundary conditions. The boundary $\partial\Omega$ of the semiconductor domain is assumed to consist of the union of Ohmic contacts $\Gamma_C = \cup_k \Gamma_k$ and the union of insulating boundary segments $\Gamma_I$ such that $\Gamma_C \cup \Gamma_I = \partial\Omega$ and $\Gamma_C \cap \Gamma_I = \emptyset$. We prescribe initial conditions for the electron density $n$, the electron temperature $T_n$, and the lattice temperature $T_L$ in $\Omega$.

On the insulating boundary parts, the normal components of the current densities, the electric field and the temperature flux are assumed to vanish,

$$J_n \cdot \nu = J_p \cdot \nu = J_w \cdot \nu = \nabla V \cdot \nu = \nabla T_L \cdot \nu = 0 \quad \text{on } \Gamma_I,\ t > 0. \tag{10}$$

The electric potential at the contacts is the sum of the applied voltage $V_{\text{app}}$ and the built-in potential $V_{\text{bi}}$,

$$V = V_{\text{app}} + V_{\text{bi}} \quad \text{on } \Gamma_C,\ t > 0, \quad \text{where} \quad V_{\text{bi}} = \operatorname{arsinh}(C(x)/2n_i). \tag{11}$$

According to the numerical results of [14], we may suppose that the normal component of the electron temperature vanishes on $\Gamma_C$. In order to model the temperature exchange between the semiconductor device and the surrounding network with the temperature $T_{\text{env}}$, we employ a Robin boundary condition for the lattice temperature:

$$\nabla T_n \cdot \nu = 0, \quad -\kappa_L \nabla T_L \cdot \nu = R_{\text{th}}^{-1}(T_L - T_{\text{env}}) \quad \text{on } \Gamma_C,\ t > 0, \tag{12}$$

where $R_{\text{th}}$ is the thermal resistivity of the contact. For the particle densities, we use, as motivated in [4], the Robin conditions

$$n + (\theta_n \mu_n)^{-1} J_n \cdot \nu = n_a, \quad p - (\theta_p \mu_p)^{-1} J_p \cdot \nu = p_a \quad \text{on } \Gamma_C,\ t > 0, \tag{13}$$

where $(n_a, p_a)$ is the solution of the charge-neutrality equation $n_a - p_a - C(x) = 0$ and the thermal equilibrium condition $n_a p_a = n_i^2$, and $\theta_n$, $\theta_p$ are some positive parameters ($\theta_n = \theta_p = 2500$ in the simulations; see [4]).

**Circuit modeling.** To simplify the presentation, the electric circuit is assumed to contain only one semiconductor device and (ideal) resistors, capacitors, inductors and voltage and current sources. The circuit is modeled by employing modified nodal analysis [3], whose basic tools are the Kirchhoff laws and the current-voltage curves of the basic elements. We replace the circuit by a directed graph with branches and nodes. Branch currents, branch voltages, and node potentials (without the mass node) are introduced as (time-dependent) variables. Then, the circuit can be characterized by the incidence matrix $A = (a_{ik})$ describing the node-to-branch relations,

$$a_{ik} = \begin{cases} 1 & \text{if the branch } k \text{ leaves the node } i, \\ -1 & \text{if the branch } k \text{ enters the node } i, \\ 0 & \text{else.} \end{cases}$$

The network is numbered in such a way that the incidence matrix consists of the block matrices $A_R$, $A_C$, $A_L$, $A_i$, and $A_v$, where the index indicates the resistive, capacitive, inductive, current source, and voltage source branches, respectively. The semiconductor device is included into the network model employing the semiconductor incidence matrix $A_S = (a_{ik}^S)$ defined by

$$a_{ik}^s = \begin{cases} 1 & \text{if the current } j_k \text{ enters the circuit node } i, \\ -1 & \text{if the reference terminal is connected to the node } i, \\ 0 & \text{else.} \end{cases}$$

The current-voltage characteristics for the basic elements are given by

$$i_R = g_R(v_R), \quad i_C = \frac{\mathrm{d}q_C}{\mathrm{d}t}(v_C), \quad v_L = \frac{\mathrm{d}\phi_L}{\mathrm{d}t}(i_L),$$

where $g_R$ denotes the conductivity of the resistor, $q_C$ the charge of the capacitor, and $\phi_L$ the flux of the inductor. Moreover, $i_\alpha$ and $v_\alpha$ with $\alpha = R$, $C$, $L$, are the branch current vectors and branch voltage vectors.

Denoting by $i_s = i_s(t)$, $v_s = v_s(t)$ the input functions for the current and voltage sources, respectively, the Kirchhoff laws lead to the following system of differential-algebraic equations in the charge-oriented modified nodal approach [3]:

$$A_C \frac{\mathrm{d}q_C}{\mathrm{d}t}(A_C^\top e) + A_R g_R(A_R^\top e) + A_L i_L + A_v i_v + A_S j_S = -A_i i_s, \tag{14}$$

$$\frac{\mathrm{d}\phi_L}{\mathrm{d}t}(i_L) - A_L^\top e = 0, \quad A_v^\top e = v_s, \tag{15}$$

for the unknowns $e(t)$, $i_L(t)$, and $i_v(t)$, where $e(t)$ denotes the vector containing the node potential. The circuit is coupled to the device through the semiconductor current $j_S$ (defined in (8)) in (14) and through the boundary conditions for the electric potential. At terminal $\Gamma_k$, it holds $V(t) = e_i(t) + V_{\rm bi}$ if the terminal $\Gamma_k$ is connected to the circuit node $i$.

Equations (14)-(15) represent a system of differential-algebraic equations. Under certain assumptions on the topology of the network, the (tractability) index of the system is at most two [3, 15]. Moreover, if the circuit does neither contain so-called LI-cutsets nor CV-loops with at least one voltage source, the index is at most one.

**Thermal network modeling.** Following [16], the thermal network consists of lumped thermal elements, i.e. zero-dimensionally modeled elements with temperature value $\widehat{T}^\ell(t)$; distributed thermal lines, i.e. spatially one-dimensional elements with temperature $T^d(x,t)$; and distributed semiconductor devices with the lattice temperature $T_L(x,t)$ as described above. Adjacent lumped elements are considered as a zero-dimensional unit with temperature $\widehat{T}$. We assign the temperature at the interface of connected distributed elements to an artificial zero-dimensional element (thermal node) with temperature $\widehat{T}$ and without thermal mass. This forms a network with lumped-distributed interfaces only, in which the nodes represent the zero-dimensional units and the branches represent the distributed elements.

The thermal network is characterized by the thermal incidence matrix $A_d^{\rm th} = (a_{ij}^{\rm th})$ and the thermal semiconductor incidence matrix $A_S^{\rm th} = (a_{S,ij}^{\rm th})$ defined by

$$a_{ij}^{\rm th} = \begin{cases} 1 & \text{if the contact at } x = 0 \text{ of branch } j \text{ is connected to node } i, \\ 1 & \text{if the contact at } x = L_{\rm th} \text{ of branch } j - m_d \text{ is connected to node } i, \\ 0 & \text{else,} \end{cases}$$

$$a_{S,ij}^{\rm th} = \begin{cases} 1 & \text{if the terminal } j \text{ is connected to thermal node } i, \\ 0 & \text{else,} \end{cases}$$

where $m_d$ is the number of thermal lines and $[0, L_{\rm th}]$ the interval of the distributed element. The embedding of the (possibly multi-dimensional) device model into the zero- and one-dimensional thermal network model is described in Section 3.

The temperature in the thermal nodes evolves according to the heat equation

$$\widehat{\mathbf{M}}\frac{\mathrm{d}\widehat{\mathbf{T}}}{\mathrm{d}t} = \widehat{\mathbf{F}}^d + \widehat{\mathbf{F}}^S - \widehat{\mathbf{S}}(\widehat{\mathbf{T}} - T_{\rm env}\mathbf{I}) + \widehat{\mathbf{P}}, \quad t > 0. \tag{16}$$

Here, $\widehat{\mathbf{M}}$ is a diagonal matrix containing the thermal masses of the thermal nodes, each of which is given as the sum of the thermal masses of the lumped elements contributing to the corresponding node. The thermal mass is the product of the heat capacity, the material density, and the physical volume of the corresponding element. Furthermore, $\widehat{\mathbf{T}}$ is the vector of all temperature values in the thermal nodes, and $\mathbf{I}$ is the identity matrix. The electro-thermal source vector for the thermal nodes $\widehat{\mathbf{P}}$ and the heat flux vectors from the distributed lines $\widehat{\mathbf{F}}^d$ and the device $\widehat{\mathbf{F}}^S$ are defined below in (20), (18), and (19), respectively. The temperature values in the lumped elements $\widehat{\mathbf{T}}^\ell$ can be computed from $\widehat{\mathbf{T}}$ by the formula $\widehat{\mathbf{T}} = M\widehat{\mathbf{T}}^\ell$, where the matrix $M = (m_{ij})$ relates the lumped elements to the thermal nodes, with $m_{ij} = 1$ if the lumped element $j$ belongs to the thermal node $i$ and $m_{ij} = 0$ else.

The vector $\mathbf{T}^d = (T_j^d)$ of all temperatures of the thermal lines satisfies

$$M_j \partial_t T_j^d = \partial_x(\kappa_j \partial_x T_j^d) - S_j(T_j^d - T_{\text{env}}) + P_j, \quad x \in (0, L_j),\ t > 0, \tag{17}$$

where $M_j$ denotes the thermal mass of the $j$-th element of length $L_j$, $\kappa_j$ is the thermal conductivity, $S_j$ the transmission function, and $\mathbf{P} = (P_j)$ the electro-thermal source vector defined in (20). The above equation is complemented by initial conditions and Dirichlet boundary conditions, collected in the vectors $\mathbf{T}_0^d$ and $\mathbf{T}_1^d$.

## 3 Coupling Conditions

The heat equations (16) and (17) are coupled through the boundary conditions, $(\mathbf{T}_0^d, \mathbf{T}_1^d)^\top = (A_d^{\text{th}})^\top \widehat{\mathbf{T}}$, and the following equation for the thermal flux:

$$\widehat{\mathbf{F}}^d = A_d^{\text{th}} \begin{pmatrix} \Lambda_0 \partial_x \mathbf{T}^d(0,t) \\ -\Lambda_1 \partial_x \mathbf{T}^d(L_{\text{th}},t) \end{pmatrix}, \tag{18}$$

where $L_{\text{th}}$ denotes the length of a thermal line and $\Lambda_0$, $\Lambda_1$ contain the products of thermal conductivities and the cross sections of the thermal lines at the contacts.

Next, we describe the coupling between the thermal network and the device. The influence of the network on the device is modeled by the last boundary condition in (12) on $\Gamma_k$, with $T_{\text{env}}$ replaced by the temperature of the connected elements, $\mathbf{T}_a = (A_S^{\text{th}})^\top \widehat{\mathbf{T}}$. The semiconductor heat flux at terminal $k$ is given by the integral

$$F_k^S = \int_{\Gamma_k} J_{\text{th}}^S \cdot \nu \, \mathrm{d}\sigma, \quad \text{such that} \quad \widehat{\mathbf{F}}^S(t) = A_S^{\text{th}}(F_j(t))_j, \quad t > 0. \tag{19}$$

The thermal flux density $J_{\text{th}}^S$ is derived by making the quasi-stationary assumption $\operatorname{div} J_u = 0$. Then, inserting the stationary balance equation for the electric energy, a computation shows that (see [9] for details)

$$\operatorname{div} J_{\text{th}}^S + \nabla V \cdot (J_n + J_p) = 0, \quad \text{where} \quad J_{\text{th}}^S = -\kappa_L \nabla T_L - q^{-1} E_c J_n - J_w.$$

This equation indicates that the flux $J_{\text{th}}^S$ is responsible for the heat production caused by the dissipated power and is therefore considered as a heat flux.

For the coupling between the electric and thermal network, we assume that only semiconductor devices and resistors are thermally relevant. (In the clipper simulations below, the thermal effects in the resistor are neglected.) Electric-to-thermal coupling occurs through the power dissipated by a resistor. We assume as in [16, Sec. 5.3] that the resistance is given by $R = 1 + \alpha_1 T_R + \alpha_2 T_R^2$, where $\alpha_1$ and $\alpha_2$ are some nonnegative parameters and $T_R$ is the temperature of the resistor. The vector $\mathbf{T}_R$ of all resistor temperature values can be determined from the temperature vectors of the thermal nodes $\widehat{\mathbf{T}}$ and of the distributed lines $\mathbf{T}^d$ by

$$\mathbf{T}_R = \widehat{K}^\top \widehat{\mathbf{T}} + K^\top \widetilde{\mathbf{T}}^d,$$

where the lumped values $\widetilde{\mathbf{T}}^d$ are computed from the distributed values $\mathbf{T}^d$ by taking the mean value, and the matrices $K = (k_{\ell j})$ and $\widehat{K} = (\widehat{k}_{\ell j})$ are defined by

$$k_{\ell j} = \begin{cases} 1 & \text{if the resistor } j \text{ corresponds to the thermal branch } \ell, \\ 0 & \text{else,} \end{cases}$$

$$\widehat{k}_{\ell j} = \begin{cases} 1 & \text{if the resistor } j \text{ corresponds to the thermal node } \ell, \\ 0 & \text{else.} \end{cases}$$

The electric-to-thermal coupling is realized by the source terms $\mathbf{P} = (P_j)$ and $\widehat{\mathbf{P}}$ in the heat equations (16) and (17):

$$\widehat{\mathbf{P}} = \widehat{K} P_R, \quad \mathbf{P} = L_R^{-1} K P_R, \quad \text{where } P_R = \operatorname{diag}(i_R) A_R^\top e, \tag{20}$$

$i_R$ contains the currents through all resistors, $A_R$ denotes the resistor incidence matrix, $e$ is the vector containing the node potentials, and $L_R$ is the resistor length. For a discussion about the proper choice of the local power distribution, we refer to [16].

## 4 Numerical Examples

The complete model is a system of nonlinear partial differential-algebraic equations. It consists of the partial differential equations (1)-(3), (9) for the device with current relations given in (5)-(6), the differential algebraic electric network equations (14)-(15), and the thermal network equations (16)-(18). The coupling conditions are given in (19)-(20). The unknowns of the system are the electron, energy and hole densities, $n, w, p$, the potential in the device $V$, the node potentials $e$, the currents through inductors, voltage sources, and semiconductor device, $i_L, i_V, j_S$, the displacement current $J_d$, the lattice temperature $T_L$, and the temperature values in the lumped and distributed elements of the thermal network, $\widehat{\mathbf{T}}, \mathbf{T}^d$.

It was shown in [2], that under certain conditions, the index of the system of semidiscretized drift-diffusion equations and electric network equations is not larger than two. To our knowledge, no index results for the coupled electro-thermal model equations are available.

In the following, we restrict ourselves to one-dimensional device models. The equations are discretized in time by backward difference formulas (BDF-1 or BDF-2) to pay tribute to the differential-algebraic character of the system. The heat equations and the Poisson equation are discretized in space by linear finite elements. The transport equations are discretized by an exponentially fitted mixed finite-element method using Marini-Pietra elements [17]. It is shown in [17] that, for the stationary model, this method guarantees current conservation and positivity of the discrete particle densities. These properties also hold for the BDF-1 time-discrete system and, under a step size restriction, for the BDF-2 time-discrete system. In the following simulations, the positivity of the discrete particle densities has always been obtained. The nonlinear discrete system is iterated by a combination of a fixed-point strategy and a variant of the Gummel method; see [9] for details.

**Bipolar junction diode.** We first illustrate the lattice heating in a 100 nm silicon *pn* diode consisting of a 50 nm *p*-doped part with doping $-C_0 = -5 \cdot 10^{23}\,\mathrm{m}^{-3}$ and a 50 nm *n*-doped part with doping $C_0$. Initially, the device is assumed to be in thermal equilibrium. The same physical parameters as in [9] are employed. We apply a forward bias of 1.5 V to the diode. The transient response of the electron and lattice temperature is illustrated in Figure 1. The electron temperature increases quickly in the entire device with a temperature maximum of about 3300 K in the *n*-region and then decreases slightly until the steady state is reached with a temperature minimum around the junction. The increase of the lattice temperature is significantly slower with a maximum of 325 K at steady state. Due to the high thermal conductivity, the lattice temperature is almost constant in the device.

The influence of the lattice heating on the electrical performance of the device is shown in Figure 2. In the left figure, we compare the results computed from the drift-diffusion (DD) model (using low-field mobilities) with those from the energy-transport (ET) equations with and without lattice heating. The current from the non-isothermal ET model is smaller than that from the ET model with constant lattice

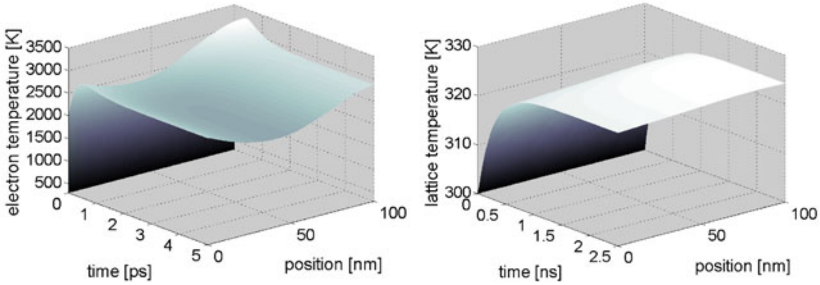


**Fig. 1:** Transient electron temperature (*left*) and lattice temperature (*right*) in a *pn* diode at 1.5 V

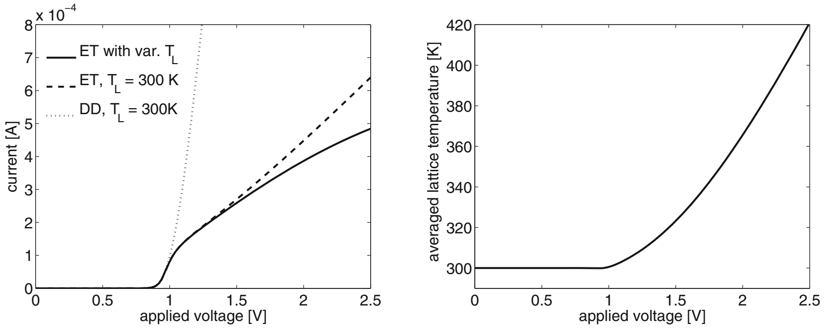


**Fig. 2:** *Left*: Current-voltage characteristics of a *pn* diode computed from different models. *Right*: Averaged lattice temperature in a *pn* diode (stationary computations)

temperature. The right figure shows the averaged lattice temperature as a function of the applied voltage. For high applied bias, the lattice temperature reaches up to 420 K. As for voltages below 1 V, the current from the DD and ET models almost coincide, the device heats up only for an applied bias larger than about 1 V.

**Clipper circuit.** A clipper is employed as an entrance protective circuit to avoid voltage peaks. It consists of two *pn* diodes (parameters as in the previous example), one resistor with resistivity $R = 5\,\mathrm{k}\Omega$, and three voltage sources (see Figure 3). We concentrate on the effect of lattice heating and neglect the thermal effects in the resistor. Here, $V_{\mathrm{in}}(t) = 5\sin(2\pi 10^{10}\,\mathrm{Hz}\,t)$ V represents the input signal. The remaining voltages are kept constant with $V_{\mathrm{min}}(t) = -U$ and $V_{\mathrm{max}}(t) = U$, where $U = 2$ V. A perfect clipper, with a much higher resistance, would clip the input signal between $\pm(U + V_{\mathrm{th}})$, where $V_{\mathrm{th}}$ is the threshold voltage of the diode. In the present case, it holds approximately $V_{\mathrm{th}} = 0.9$ V such that the signal is between ±2.9 V. However, we have chosen the resistance such that the output signal should stay below 4 V.

In Figure 4 we depict the input and output signals of the circuit. We observe that during the first oscillations the maximal output signal is below 4 V, with a slight increase of the maximal value (left figure). It increases during the first oscillations from 3.93 V to 3.96 V. This increase becomes more significant for larger time (right figure). In fact, after 30 oscillations the maximal output signal is 4.09 V, which corresponds to an increase of about 5 %. A simulation of the same circuit with constant lattice heating keeps the maximum output signal almost constant below 4 V. This

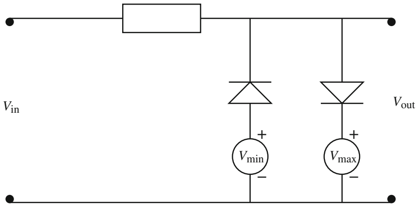


**Fig. 3:** Clipper circuit with two *pn* diodes, one resistor and three voltages sources

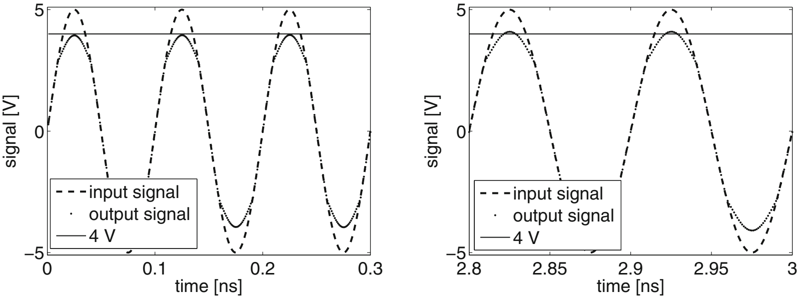


**Fig. 4:** Input and output signal of the clipper during the first oscillations (*left*) and after 30 oscillations of the input signal (*right*)

shows that the increasing maximal output voltage is caused by lattice heating, as the heated diode provides less current leading to a larger resistance.

The circuit is constructed in such a way that, at the maximal input signal of 5 V, we have a voltage drop of about 1 V at the resistor, 2 V at the forward-biased diode and 2 V at the additional voltage source. In the branch containing the backward-biased diode, the voltage drop is 1 V at the resistor, 6 V at the diode, and $-2$ V at the voltage source. This behavior is illustrated in Figure 5. According to Figure 2 (right), we expect a stationary lattice temperature of about 360 K in the diodes. This is confirmed by the results presented in Figure 6 showing the lattice temperature of one of the diodes in the circuit. We observe that the device heats up while being forward biased. As the backward bias period is to short to cool down the device, the lattice heating accumulates during the first oscillations up to about 360 K.

**Conclusions.** In this paper we have presented a coupled system for the thermal-electric modeling and simulation of semiconductor devices in electric circuits. The numerical results clarify the significance of the thermal effects in small semiconductor devices. In strongly biased devices, lattice heating occurs and influences the electrical performance considerably. This shows that for accurate simulations of (ultra) small semiconductor devices and integrated circuits, the inclusion of thermal models is indispensable.

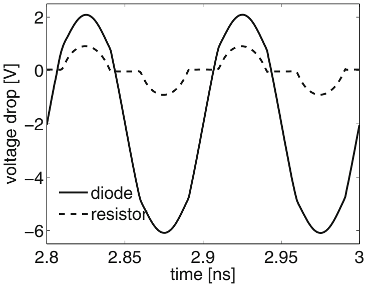


**Fig. 5:** Voltage drop at the second diode and the resistor during the 29th and 30th oscillation

**Acknowledgements** The authors acknowledge partial support from the German Federal Ministry of Education and Research (BMBF), grant 03JUNAVN. The first author was supported by the ERCIM "Alain Bensoussan" Fellowship Programme. The second author was supported by the Austrian Science Fund (FWF), grant P20214 and WK "Differential Equations", and from the German Science Foundation (DFG), grant JU 359/7. This research is part of the ESF program "Global and geometrical aspects of nonlinear partial differential equations (GLOBAL)".

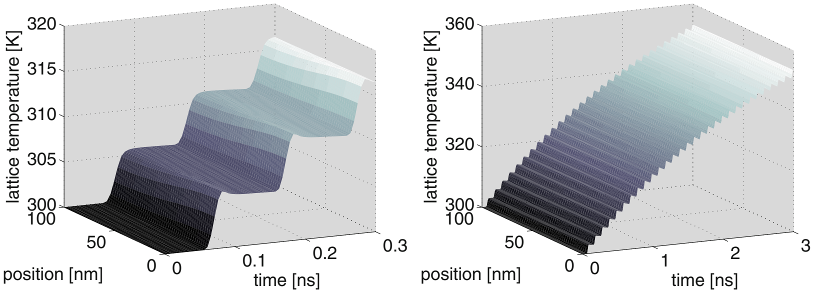


**Fig. 6:** Distribution of the lattice temperature in one of the *pn* diodes within the first 3 (*left*) and within the first 30 (*right*) oscillations of $V_{\mathrm{in}}$

# References

1. Einwich, K., Schwarz, P., Trappe, P., Zojer, H.: Simulatorkopplung für den Entwurf komplexer Schaltkreise der Nachrichtentechnik. In: 7. ITG-Fachtagung "Mikroelektronik für die Informationstechnik", Chemnitz, pp. 139–144 (1996)
2. Selva Soto, M., Tischendorf, C.: Numerical analysis of DAEs from coupled circuit and semiconductor simulation. Appl. Numer. Math., **53**, 471–488 (2005)
3. Tischendorf, C.: Coupled Systems of Differential Algebraic and Partial Differential Equations in Circuit and Device Simulations. Habilitation thesis, Humboldt-Universität zu Berlin, Germany (2003)
4. Brunk, M., Jüngel, A.: Numerical coupling of electric circuit equations and energy-transport models for semiconductors. SIAM J. Sci. Comput., **30**, 873–894 (2008)
5. Adler, M.: Accurate calculations of the forward drop and power dissipation in thyristors. IEEE Trans. Electr. Dev., **25**, 16–22 (1978)
6. Wachutka, G.: Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling. IEEE Trans. Comp. Aided Design, **9**, 1141–1149 (1990)
7. Albinus, G., Gajewski, H., Hünlich, R.: Thermodynamic design of energy models of semiconductor devices. Nonlinearity, **15**, 367–383 (2002)
8. Alì, G., Carini, M.: Energy-transport models for semiconductor devices and their coupling with electric networks. In: V. Cutello, G. Fotia, L. Puccio (eds) Applied and industrial mathematics in Italy II, pp. 13–24, World Sci. Publ., NJ (2007)
9. Brunk, M., Jüngel, A.: Self-heating in a coupled thermo-electric circuit-device model. Preprint, Vienna University of Technology, Austria (2008)
10. Ben Abdallah, N., Degond, P.: On a hierarchy of macroscopic models for semiconductors. J. Math. Phys., **37**, 3308–3333 (1996)
11. Selberherr, S.: Analysis and Simulation of Semiconductor Devices. Springer, Berlin (1984)
12. Bandelow, U., Gajewski, H., Hünlich, H.: Fabry-Perot lasers: thermodynamic-based modeling. In: J. Piprek (ed.), Optoelectronic Devices. Advanced Simulation and Analysis, pp. 63–85, Springer, Berlin (2005)
13. Jüngel, A.: Transport Equations for Semiconductors. Lecture Notes in Physics, Vol. 773. Springer, Berlin (2009)
14. Anile, A., Romano, V., Russo, G.: Extendend hydrodynamic model of carrier transport in semiconductors. SIAM J. Appl. Math., **61**, 74–101 (2000)
15. Tischendorf, C.: Topological index calculation of differential-algebraic equations in circuit simulation. Surv. Math. Industr., **8**, 187–199 (1999)
16. Bartel, A.: Partial Differential-Algebraic Models in Chip Design – Thermal and Semiconductor Problems. Ph.D. thesis, Universität Karlsruhe, Germany (2003)
17. Marini, L. D., Pietra, P.: New mixed finite element schemes for current continuity equations. COMPEL, **9**, 257–268 (1990)

# Analysis of a PDE Thermal Element Model for Electrothermal Circuit Simulation

Giuseppe Alì, Andreas Bartel, Massimiliano Culpo, and Carlo de Falco

**Abstract** In this work we address the well-posedness of the steady-state and transient problems stemming from the coupling of a network of lumped electric elements and a PDE model of heat diffusion in the chip substrate. In particular we consider the thermal element model presented in [1] and we prove that it can be controlled by any combination of voltage sources (imposing the average current in a region of the chip) and current sources (imposing the Joule power per unit area produced in a region) connected to its temperature nodes.

This result justifies the implementation of the element as a linear n-port conductance as carried out in [2].

## 1 Introduction

Due to downscaling, power densities become more important [3] and therefore thermal models to resolve the geometric layout, which fit seamless into the circuit design are necessary. A method for automatically deriving a thermal network model from the layout of an IC and substrate material properties was introduced in [1,4] and the numerical validation is reported in [2].

The novelty of this method compared to other existing approaches [5, 6] is that it does not work by fitting the parameters of a given network topology, but rather it

Giuseppe Alì
Università della Calabria, Ponte P.Bucci, 87036 Arcavacata di Rende (CS), Italy, and INFN-Gruppo c. Cosenza, e-mail: g.ali@mat.unical.it

Andreas Bartel, Massimiliano Culpo
Bergische Universität Wuppertal, Gaußstrasse 20, 42119 Wuppertal, Germany, and CoMSON RTN project, e-mail: bartel@math.uni-wuppertal.de, culpo@math.uni-wuppertal.de

Carlo de Falco
Dublin City University, Glasnevin, Dublin 9, Ireland, and MACSI consortium, e-mail: carlo.defalco@dcu.ie

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_35,

consists of a parabolic PDE which can be connected to a network of lumped (electrical) device models to perform coupled system-level electrothermal simulation with a standard spice-like circuit simulator. The coupling is performed by controlling the average temperature of some substrate regions via a set of (controlled) voltage sources and the total power dissipated in some other regions via a set of (controlled) current sources. A similar coupling, but based on 1-dimensional heat transport, was considered in [7].

To sketch the main idea, we can write symbolically the MNA equations for an IC as:

$$F\left(\dot{\mathbf{e}}, \mathbf{e}, \theta, t\right) = 0 \tag{1}$$

where $\mathbf{e}$ is a vector accounting for the electrical variables and $\theta$ is a vector comprising the local (lumped) temperatures of the, say, $n$ thermally active components. We denote by $w_k(\mathbf{e}, \theta)$ the thermal power produced by the $k$-th thermal component, and by $\Omega_k$ the region of the substrate where it is located ($k = 1, \dots, n$). Then thermal powers will act as localized source terms for a heat equation which describes a global (distributed) substrate temperature $T$:

$$\frac{\partial T}{\partial t} + \mathscr{L} T = \sum_{k=1}^{n} \frac{w_k(\mathbf{e}, \theta)}{|\Omega_k|} \mathbf{1}_{\Omega_k}, \tag{2}$$

where $\mathscr{L}$ is a linear diffusion-reaction operator and $\mathbf{1}_{\Omega_k}$ is the indicator function over $\Omega_k$. Finally, we identify the temperature $\theta_k$ with the average of $T$ over $\Omega_k$, i.e.:

$$\theta_k = \frac{1}{|\Omega_k|} \int_{\Omega_k} T \, \mathrm{d}\Omega \tag{3}$$

In this work we consider the equations obtained when the thermal element is controlled by a set of independent sources with finite internal resistance or conductance, neglecting the coupling with the electric part. The analysis of this simplified problem provides a sound theoretical basis to the approach to implementation of the thermal element followed in [1] and is an initial step towards the analysis of the coupled electro-thermal system which will be the subject of a forthcoming paper.

## 2 Statement of the Problem

Let the domain $\Omega \subset \mathbb{R}^d$, with $d = 1, 2, 3$, model the IC substrate and $\Omega_k$ be the thermally *active region* of the $k$-th circuit element. We assume that $\Omega$ be Lipschitz and that the family $\{\Omega_k, k = 1, \dots, n\}$ satisfies the requirements:

1. $\overset{\circ}{\Omega}_k \neq \emptyset$,
2. $\bar{\Omega}_k \subset \Omega \;\; \forall k = 1, \dots, n$
3. $\bar{\Omega}_k \cap \bar{\Omega}_j = \emptyset \;\; \forall j, k \in \{1, \dots, n\}, k \neq j.$

We denote by $u(\mathbf{x},t)$ the temperature at an instant $t$ at each point $\mathbf{x}$ in $\Omega$, we let $q_k(t)/|\Omega_k|$ be the average temperature in the region $\Omega_k$ at time $t$ and $p_k(t)$ the instantaneous Joule power per unit length, area or volume (in case $d = 1,2$ or 3, respectively) dissipated by element $k$. For simplicity, we consider constant thermal diffusivity, which in scaled variable can be assumed equal to 1. Then the heat diffusion in the substrate is governed by the linear heat equation

$$\frac{\partial u}{\partial t} - \Delta u + cu = \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}(\mathbf{x}), \quad \text{in } \Omega \times (0,T) \tag{4}$$

denoting by $\mathbf{1}_{\Omega_k}$ the indicator function of the set $\Omega_k$. The term $cu$ accounts for heat exchange with the environment (in a 2-dimensional model). Equation (4) is supplemented with initial-boundary conditions

$$u(\mathbf{x},0) = u_0(\mathbf{x}), \quad \text{in } \Omega, \tag{5}$$

$$u + \alpha \frac{\partial u}{\partial n} = g(t), \quad \text{on } \partial\Omega \times (0,T), \tag{6}$$

where the function $g$ represents the ambient temperature. Furthermore, the $k$-th average device temperature $q_k$ is connected to $u$ by the relation

$$\int_{\Omega_k} u(\mathbf{x},t)\,\mathrm{d}\Omega = q_k(t). \tag{7}$$

Finally, to close the system, we need to state constitutive relations for the set of average temperatures $q_k(t)$ and for the set of instantaneous powers $p_k(t)$ in terms of the electrical variables in the circuit. As anticipated in the introduction, in the present work we make the simplifying assumption that the thermal network be controlled via independent sources. Under this assumption the constitutive relations can be cast into the form

$$a_k p_k(t) \; + \; b_k q_k(t) = s_k(t), \quad k = 1,\dots,n \tag{8}$$

where the $s_k(t)$ are given functions and $a_k$ and $b_k$ denote constant coefficients. Notice that $a_k = 0$ for a given $k$ indicates that the $k$-th region is attached to a voltage source fixing the value of its average temperature, while $b_k = 0$ indicates that the Joule power dissipated in the $k$-th region has been assigned by attaching it to a current source. Summarizing, the problem we intend to investigate reads:

**Problem 1.** Given initial datum $u_0(\mathbf{x})$ and ambient temperature $g(t)$, find $u(\mathbf{x},t)$, $\mathbf{p}(t)$ and $\mathbf{q}(t)$ such that equations (4)–(8) are satisfied, where $c$, $\alpha$, $a_k$, $b_k$ are known quantities and $c$, $\alpha \geq 0$.

As our interest in this problem is mainly driven by the need to prove the suitability of the model (4)-(8) for implementation in a standard SPICE-like circuit simulator, it is convenient to restate Prob. 1 in a discrete-time form applying a semi-discretization approach based on Rothe's method. To this end, let us introduce a set of $N$ time steps $t_0 = 0 < t_1 < \ldots < t_N = T$, then, supposing for sake of simplicity that a BDF method of order $m$ is used for time discretization we can formulate the time-discrete problem as a sequence of problems of the form:

**Problem 2.** Let $\tau$ be an integer s.t. $m < \tau \leq N$. Given the $m$ functions $u_i(\mathbf{x}),\ i = \tau - 1 \dots \tau - m$ and the $2 \times m$ $n$-vectors $\mathbf{p}_i,\ \mathbf{q}_i,\ i = \tau - 1 \dots \tau - m$ satisfying the conditions

$$\int_{\Omega_k} u_i(\mathbf{x})\,\mathrm{d}\Omega = q_{i_k} \tag{7'}$$

and

$$a_k p_{i_k} \; + \; b_k q_{i_k} = s_{i_k}, \quad k = 1, \dots, n \tag{8'}$$

find $u_\tau(\mathbf{x})$, $\mathbf{p}_\tau$ and $\mathbf{q}_\tau$ satisfying

$$-\Delta u_\tau + \tilde{c}\, u_\tau = f_\tau + \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}(\mathbf{x}) \quad \text{in } \Omega \tag{4'}$$

and

$$u_\tau + \alpha \frac{\partial u_\tau}{\partial n} = g(t_\tau) \quad \text{on } \partial\Omega, \tag{6'}$$

where $f_\tau := -\sum_{i=1}^{m} \beta_i u_{\tau-i}$, $\tilde{c} := c + \beta_0$ and the coefficients $\beta_0, \dots, \beta_m$ depend on the BDF method chosen.

## 3 Linear Elliptic Kernel Problem

The above problem is related to the following kernel problem:

**Problem 3.**

$$\begin{cases} -\Delta u + cu = f + \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}, & \text{in } \Omega, \\ u + \alpha \dfrac{\partial u}{\partial n} = g, & \text{on } \partial\Omega, \\ \displaystyle\int_{\Omega_k} u\,\mathrm{d}\Omega = q_k, & k = 1, \dots, n. \end{cases} \tag{9}$$

where $c \in \mathbb{L}^\infty(\Omega)$, $f \in \mathbb{L}^2(\Omega)$, $\alpha \geq 0$ and:

$$g \in \begin{cases} \mathbb{L}^2(\partial\Omega), & \alpha > 0, \\ \mathbb{H}^{1/2}(\partial\Omega), & \alpha = 0. \end{cases} \tag{10}$$

We prove in the following existence and uniqueness of the solution for (9). To do this we cast the differential operator in a weak form and consider $\alpha > 0$ first. Defining:

$$a(u,v) := \int_\Omega \nabla u \nabla v \, \mathrm{d}\Omega + \int_\Omega cuv \, \mathrm{d}\Omega, \tag{11}$$

$$\mathscr{A}(u,v;g) := a(u,v) + \int_{\partial\Omega} \frac{u-g}{\alpha} v \, \mathrm{d}\gamma, \tag{12}$$

$$(u,v) := \int_\Omega uv \, \mathrm{d}\Omega, \tag{13}$$

$$\mathscr{B}_k(u) := \int_{\Omega_k} u \, \mathrm{d}\Omega. \tag{14}$$

we can state the following

**Theorem 1.** *Given $q_k > 0$ ($k = 1,\dots,n$) and $g$, there exist unique $u \in \mathbb{H}^1(\Omega)$ and $p_k \in \mathbb{R}$ ($k = 1,\dots,n$) such that:*

$$\mathscr{A}(u,v;g) = (f + \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}, v), \qquad \forall v \in \mathbb{H}^1(\Omega), \tag{15a}$$

$$\mathscr{B}_k(u) = q_k, \qquad k = 1,\dots,n. \tag{15b}$$

*Proof.* Since the differential operator (15a) is linear, we can represent the general solution, for any choice of $p_k$, as [8, 9]

$$u = u_* + \sum_{k=1}^{n} p_k u_k, \tag{16}$$

with $u_*(x)$ solution of the problem

$$\mathscr{A}(u_*,v;g) = f, \qquad \forall v \in \mathbb{H}^1(\Omega), \tag{17}$$

and $u_k(x)$, $k = 1,\dots,n$, solution of the problem

$$\mathscr{A}(u_k,v;0) = (\mathbf{1}_{\Omega_k}, v), \qquad \forall v \in \mathbb{H}^1(\Omega). \tag{18}$$

Substituting (16) into (15b) we get:

$$\mathscr{B}_k(u) = \mathscr{B}_k(u_* + \sum_{j=1}^{n} p_j u_j) = \mathscr{B}_k(u_*) + \sum_{j=1}^{n} p_j \mathscr{B}_k(u_j) = q_k, \tag{19}$$

with ($k = 1,\dots,n$). Then we can write the conditions for $p_k$ in (15b) as a linear algebraic system:

$$\sum_{j=1}^{n} \mathscr{B}_k(u_j) p_j = q_k - \mathscr{B}_k(u_*), \quad k = 1,\dots,n. \tag{20}$$

This system is uniquely solvable if and only if the matrix $\mathrm{B} = [\mathscr{B}_k(u_j)]$ is non singular, that is to say $\det(\mathrm{B}) \neq 0$. Noting that:

$$\mathscr{B}_k(u_j) = (\mathbf{1}_{\Omega_k}, u_j) = \mathscr{A}(u_k, u_j; 0), \tag{21}$$

we can derive an equivalent condition on the matrix $\mathrm{A} = [\mathscr{A}(u_k, u_j; 0)]$:

$$\det(\mathrm{A}) \neq 0. \tag{22}$$

By using the (extended) Cauchy-Schwartz inequality (see, for instance, [10], Chap. 5), the matrix A is positive semi-definite. In particular $\det(\mathrm{A}) \geq 0$, and the equality holds true if and only if there exist real numbers $\lambda_k$, $k = 1, \ldots, n$, not all equal to zero, such that

$$\sum_{k=1}^{n} \lambda_k u_k = 0. \tag{23}$$

In conclusion to prove existence and uniqueness it is sufficient to prove that $\sum_{k=1}^{n} \lambda_k u_k = 0$ implies $\lambda_k = 0$ for all $k = 1, \ldots, n$. This fact follows from the equality:

$$0 \equiv \sum_{k=1}^{n} \lambda_k \mathscr{A}(u_k, v; 0) = \sum_{k=1}^{n} \lambda_k (\mathbf{1}_{\Omega_k}, v) \quad \forall v \in \mathbb{H}^1(\Omega). \tag{24}$$

Then, for any $k = 1, \ldots, n$, we can choose $v$ such that $\mathrm{supp}(v) \subset \Omega_k$ and get $\lambda_k = 0$.

*Remark 1.* The matrix $A = [\mathscr{A}(u_k, u_j; 0)] = [\mathscr{B}_k(u_j)]$ appearing in the proof of Theorem 1 is thus positive definite.

*Remark 2 (Non-homogeneous Dirichlet boundary condition).* If $\alpha = 0$ the weak formulation has to be modified, as the boundary conditions change from Robin to non-homogeneous Dirichlet type. This case is standardly treated extending $g$ in the whole $\Omega$ domain, and denoting with $\tilde{g}$ this extension [8]. A solution of the problem:

$$a(\tilde{u}, v) = (f + \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}, v) - a(\tilde{g}, v), \qquad \forall v \in \mathbb{H}^1_0(\Omega), \tag{25a}$$

$$\mathscr{B}_k(\tilde{u}) = q_k - \mathscr{B}_k(\tilde{g}), \qquad k = 1, \ldots, n. \tag{25b}$$

is then searched. Modifications to Theorem 1 and subsequent proof are straightforward, and left to the reader.

## 4 Linear Elliptic Extended Problem

Using the results of the previous section we can now proceed to study the following problem which is equivalent to Prob. 2.

**Problem 4.** Given $g(x)$ with the same regularity as in the previous section, find $u(x)$, $p_k$ and $q_k$ $(k = 1, \ldots, n)$ such that:

$$\begin{cases} -\Delta u + cu = f + \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}, & \text{in } \Omega, \\ u + \alpha \dfrac{\partial u}{\partial n} = g, & \text{on } \partial\Omega, \\ \displaystyle\int_{\Omega_k} u \, \mathrm{d}\Omega = q_k, & k = 1, \dots, n, \\ a_k p_k + b_k q_k = s_k, & k = 1, \dots, n, \end{cases} \tag{26}$$

where $c > 0$, $\alpha > 0$, $a_k \geq 0$, $b_k > 0$, $s_k$ are data of the problem.

As it was the case for Problem (9) we can state the following:

**Theorem 2.** *Given $g(x) \in \mathbb{L}^2(\partial\Omega)$, $a_k \geq 0$, $b_k > 0$, there exist unique $u \in \mathbb{H}^1(\Omega)$ and $q_k, p_k \in \mathbb{R}$ $(k = 1, \dots, n)$ such that:*

$$\mathscr{A}(u, v; g) = (f + \sum_{k=1}^{n} p_k \mathbf{1}_{\Omega_k}, v), \qquad \forall v \in \mathbb{H}^1(\Omega), \tag{27a}$$

$$\mathscr{B}_k(u) = q_k, \qquad k = 1, \dots, n, \tag{27b}$$

$$a_k p_k + b_k q_k = s_k, \qquad k = 1, \dots, n. \tag{27c}$$

*Proof.* We can use the same decomposition as in the previous problem. The only difference from the previous problem is that in this case (27c) gives a system for $p_k$ and $q_k$, with matrix

$$\begin{pmatrix} \mathrm{B} & -\mathbf{Id} \\ \mathrm{diag}(a) & \mathrm{diag}(b) \end{pmatrix}, \tag{28}$$

where $a = (a_1, \dots, a_n)$, $b = (b_1, \dots, b_n)$. This is a block matrix, where the two lower blocks commute. Then, by using classical results [11], we have

$$\begin{aligned} \det \begin{pmatrix} \mathrm{B} & -\mathbf{Id} \\ \mathrm{diag}(a) & \mathrm{diag}(b) \end{pmatrix} &= \det\big(\mathrm{B}\ \mathrm{diag}(b) + \mathrm{diag}(a)\big) \\ &= \det\big(\mathrm{B} + \mathrm{diag}(a)\mathrm{diag}(b)^{-1}\big) \det\big(\mathrm{diag}(b)\big) > 0, \end{aligned}$$

since the matrix $\mathrm{B} + \mathrm{diag}(a)\mathrm{diag}(b)^{-1}$ is positive definite.

## Acknowledgments

The third author was supported by the European Commission in the framework of the CoMSON RTN project. The fourth author was supported by the Mathematics Applications Consortium for Science and Industry in Ireland (MACSI) under the Science Foundation Ireland (SFI) mathematics initiative.

## References

1. Culpo, M., de Falco, C.: A pde thermal model for chip-level simulation including substrate heating effects. Preprint MS-08-10, School of Mathematical Sciences (2008)
2. Culpo, M., de Falco, C., Denk, G., Voigtmann, S.: Automatic thermal network extraction and multiscale electro-thermal simulation. In: Proceedings of the SCEE 2008 Conference (Submitted) (2008)
3. roadmap commitee, I.: International tecnology roadmap for semiconductors 2007. Tech. rep., ITRS (2007)
4. Culpo, M., de Falco, C.: Dynamical iteration schemes for coupled simulation. In: Proceedings of the GAMM2008 meeting (Submitted) (2008)
5. Grasser, T., Selberherr, S.: Fully coupled electrothermal mixed-mode device simulation of sige hbt circuits. Electron Devices, IEEE Transactions on **48**(7), 1421–1427 (Jul 2001). DOI 10.1109/16.930661
6. Igic, P., Mawby, P., Towers, M., Batcup, S.: Dynamic electro-thermal physically based compact models of the power devices for device and circuit simulations. Semiconductor Thermal Measurement and Management, 2001. Seventeenth Annual IEEE Symposium pp. 35–42 (2001). DOI 10.1109/STHERM.2001.915142
7. Bartel, A.: Partial Differential-Algebraic Models in Chip Design Thermal and. Semiconductor Problems. VDI-Verlag (2004)
8. Quarteroni, A., Valli, A.: Numerical approximation of Partial Differential Equations. Computational Mathematics. Springer (1997)
9. Evans, L.: Partial Differential Equations. American Mathematichal Society (1998)
10. Deheuvels, R.: Formes quadratiques et groupes classiques. Presses Universitaires de France, Paris (1981)
11. Silvester, J.: Determinants of block matrices. The Mathematical Gazette **84**(501), 460–467 (2000)

# Automatic Thermal Network Extraction and Multiscale Electro-Thermal Simulation

Massimiliano Culpo, Carlo de Falco, Georg Denk, and Steffen Voigtmann

**Abstract** We present a new strategy to perform chip-level electro-thermal simulation. In our approach electrical behaviour of each circuit element is modeled by standard compact models with an added temperature node [1, 2]. Mutual heating is accounted for by a 2-D or 3-D diffusion reaction PDE, which is coupled to the electrical network by enforcing instantaneous energy conservation. To cope with the multiscale nature of heat diffusion in VLSI circuit a suitable spatial discretization scheme is adopted which allows for efficient meshing of large domains with details at a much smaller scale. Preliminary numerical results on a realistic test case are included as a validation of the model and of the numerical method.

## 1 Introduction

In this communication, we present a tool to automatically extract a thermal network model for a chip-level electro-thermal simulation. Starting from layout geometry and chip material data, it produces an *n*-port thermal device model, possibly non-linear, that can be coupled to the electrical circuit network via an extra temperature node in the electrical device compact models. Compared to other similar tools [3,4], the one we propose does not rely on fitting the parameters of a given network of lumped thermal resistors and capacitors; rather a full 2D or 3D discretization of the

Massimiliano Culpo
Bergische Universität Wuppertal, Gaußstrasse 20, 42119 Wuppertal, Germany, and CoMSON RTN project, e-mail: culpo@math.uni-wuppertal.de

Carlo de Falco
Dublin City University, Glasnevin, Dublin 9, Ireland, and MACSI consortium, e-mail: carlo.defalco@dcu.ie

Georg Denk, Steffen Voigtmann
Qimonda AG Munich, Am Campeon 1-12, 85579 Munich, Germany, e-mail: georg.denk@qimonda.com, steffen.voigtmann@qimonda.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_36,

heat equation on the whole chip is cast into a form which is analogous to that of a multi-port circuit element and simulation is performed within a spice-like circuit simulator directly. Although the ideas on which our method rely can be straightforwardly extended to the case of a non-linear heat equation, we focus here, to simplify the presentation, on the case where the material properties are independent of temperature so that the resulting circuit element is linear. Similarly, though our focus here is on transient analysis, the model is suitable for use in DC, AC and HB analyses as well.

## 2 Electro-Thermal Network Model

Below we briefly outline how the thermal and electrical subsystem are coupled, while we postpone the details about the model for the thermal element to Sec. 2.1.

The global system of equations describing our electro-thermal circuit is constructed following the well known MNA approach [5, 6]. Denoting by the superscript $\cdot^{(1)}$ the contributions stemming from the electrical part of the coupled network, and by $\cdot^{(2)}$ the ones stemming from the thermal part, the system of equations modeling the coupled electro-thermal system read

$$A_{\mathbf{e}}^{(1)}\dot{Q}^{(1)}(\mathbf{e},\theta)+F_{\mathbf{e}}^{(1)}(\mathbf{e},\theta)=0 \tag{1a}$$

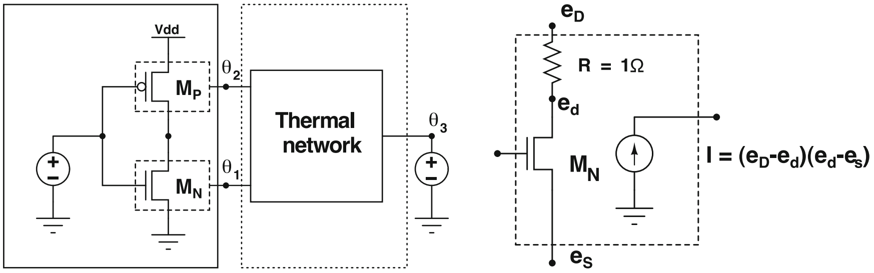
$$A_{\theta}^{(1)}\dot{Q}^{(1)}(\mathbf{e},\theta)+F_{\theta}^{(1)}(\mathbf{e},\theta)+A_{\theta}^{(2)}\dot{Q}^{(2)}(\theta,\mathbf{r})+F_{\theta}^{(2)}(\theta,\mathbf{r})=0 \tag{1b}$$

$$A_{\mathbf{r}}^{(2)}\dot{Q}^{(2)}(\theta,\mathbf{r})+F_{\mathbf{r}}^{(2)}(\theta,\mathbf{r})=0\ , \tag{1c}$$

Here $\mathbf{e}(t)$ and $\mathbf{r}(t)$ represent the state variables of the electrical and thermal subsystems, respectively, $\theta(t)$ is the vector of lumped temperatures which are the *interface variables* shared by the two components of the system and $A_{\mathbf{e}}^{(1)}$, $A_{\theta}^{(1)}$, $A_{\theta}^{(2)}$, $A_{\mathbf{r}}^{(2)}$ are incidence matrices. If we let $L$ be the number of temperature dependent circuit elements, then $\theta$ will be of the form

$$\theta=[\theta_1(t),\dots,\theta_L(t),\theta_{L+1}(t)]^T, \tag{2}$$

where the first $L$ components are device temperatures, while $\theta_{L+1}(t)$ is the ambient temperature. To better separate the two components of the system, it is convenient to rewrite (1) in the following form



**(a)** Coupling of an electrical circuit (*solid frame*) with a thermal 3-port element (*dotted frame*) via the temperature pins $\theta_1$ and $\theta_2$. Note the additional pin $\theta_3$ used to set the ambient temperature

**(b)** Macro-model for a temperature dependent $n$-MOSFET. The controlled source represents the Joule power while the device temperature is equal to the voltage at the pin on the right

**Fig. 1:** Assembly of a coupled electro-thermal system

$$A_{\mathbf{e}}^{(1)}\dot{Q}^{(1)}(\mathbf{e},\theta)+F_{\mathbf{e}}^{(1)}(\mathbf{e},\theta)=0 \tag{3a}$$

$$A_{\theta}^{(1)}\dot{Q}^{(1)}(\mathbf{e},\theta)+F_{\theta}^{(1)}(\mathbf{e},\theta)=J_1 \tag{3b}$$

$$A_{\theta}^{(2)}\dot{Q}^{(2)}(\theta,\mathbf{r})+F_{\theta}^{(2)}(\theta,\mathbf{r})=J_2 \tag{3c}$$

$$A_{\mathbf{r}}^{(2)}\dot{Q}^{(2)}(\theta,\mathbf{r})+F_{\mathbf{r}}^{(2)}(\theta,\mathbf{r})=0 \tag{3d}$$

$$J_1+J_2=0\,. \tag{3e}$$

In this form the contributions of the electrical (3a, 3b) and thermal (3c, 3d) subsystem to the conservation law (1b) have been separated by introducing the quantities $J_1$ and $J_2$ which represent the *Joule power* produced by the lumped devices and the power dissipated in the substrate, respectively. The equation (3e) is equivalent to (1b), and is a statement of *instantaneous energy conservation*. The coupling could in principle be non-local in time as *e.g.* in [7]. Enforcing condition (3e) is analogous to the procedure used by standard circuit simulators to assemble system equations by enforcing Kirchhoff Current Law at each circuit node. Therefore, by casting the equations stemming from the discretization of the heat equation in the chip substrate into the form (3c)-(3d) we allow the thermal subsystem to be interpreted as a standard $n$-port device. As a consequence, to implement coupled electro-thermal simulation within a spice like simulator we need to:

1. add a temperature node to each temperature-dependent circuit element
2. implement an *element evaluator* for the thermal $n$-port device

Item 2. is addressed in Sec. 2.1 while Fig. 1b depicts an approach to achieve 1. via a macromodel without implementing new evaluators for the circuit elements; the power $J_1$ is represented by the current of a controlled source with one pin connected to ground, the temperature of the device will be represented by the voltage at the other pin. Fig. 1a shows the coupling of an electric circuit to a thermal 3-port element; the temperature pins $\theta_1$ and $\theta_2$ allow for flux of thermal energy between the two subsystems while the pin $\theta_3$ allows to set the temperature of the external environment by means of a voltage source.

## 2.1 Thermal Element Model

In this section we briefly describe the model for our distributed thermal element which has been discussed in more detail in [8]. In addition we introduce a simple technique which allows to avoid the large increase in the size of the state vector for the coupled system by locally eliminating the internal variables of the thermal element. We model the chip substrate by a bounded open domain $\Omega \subset \mathbb{R}^d$ $(d = 2,3)$ and the active region of the $q$-th circuit element is represented by a subdomain $\Omega_q$, $q = 1\dots L$. We assume that the family $\{\Omega_q\}$ satisfy $\bar{\Omega}_q \subset \Omega$, $|\Omega_q| \neq 0$ and $\bar{\Omega}_i \cap \bar{\Omega}_j = \emptyset \forall i \neq j$. If we denote with $p_q$, $q = 1\dots L$ the Joule power per unit area/volume on $\Omega_q$ and by $T(x,t)$ the temperature at a point $x \in \Omega$ at the time $t$, then the heat equation describing the evolution of $T$ in $\Omega$ reads

$$\begin{cases} \dot{Q}(T) + \mathrm{div} S(T) = \sum_{q=1}^{L} p_q \chi_{\Omega_q} & \text{in } \Omega \\ RS(T) \cdot \nu|_{\partial\Omega} = T|_{\partial\Omega} - \theta_{L+1}, \end{cases} \tag{4}$$

where $Q(T)$ denotes the *internal energy*, $S(T)$ the *heat flux vector*, $\chi_{\Omega_q}(x)$ is the *indicator function* for the set $\Omega_q$, $R$ is the *thermal resistance to the external environment* and $\nu$ is the outward unit normal to $\partial\Omega$. In a linear material $Q(T) = c_v T$ and $S(T) = \kappa \nabla T$ where the heat capacity $c_v$ and thermal conductivity $\kappa$ are constants independent of $T$. The lumped temperatures $\theta_q$, $q = 1\dots L$ are defined via $\dfrac{\int_{\Omega_q} T\,\mathrm{d}\omega}{\int_{\Omega_q} \mathrm{d}\omega}$ and the *vector of internal variables* is $\mathbf{r} = [p_1(t), \dots, p_L(t), T(x,t)]^T$. The spatial discretization of (4) by the method of Patches of Finite Elements (PFE) [9] is addressed in [8] here we wish to point out that the choice of this particular method is suggested by two main features:

1. it allows to efficiently mesh large domains with geometrical features of much smaller scale
2. it allows for even more convenient meshing when a large number of sub-domains of identical geometry is present

In Sec. 3 we present an example where both such features produce a noticeable advantage. Using PFE for spatial discretization and a $p$-step BDF formula of the form

$$\dot{Q}^{(2)}(t_k) = \sum_{i=0}^{p} \alpha_i Q^{(2)}(t_{k-i}) = \alpha_0 Q^{(2)}(t_k) + \beta(t_k)\ , \tag{5}$$

we can express the discrete counterpart of (3c) and (3d) as

$$\left[A_\theta^{(2)} \sum_{i=0}^{p} \alpha_i Q^{(2)}(t_{k-i})\right] + F_\theta^{(2)}(t_k) = J_2(t_k)\ , \tag{6a}$$

$$\left[A_\mathbf{r}^{(2)} \sum_{i=0}^{p} \alpha_i Q^{(2)}(t_{k-i})\right] + F_\mathbf{r}^{(2)}(t_k) = 0\ . \tag{6b}$$

As we have assumed linearity of the substrate material we have

$$\begin{bmatrix} A_\theta^{(2)} Q^{(2)}(t_k) \\ A_\mathbf{r}^{(2)} Q^{(2)}(t_k) \end{bmatrix} = \begin{bmatrix} C_{\theta\theta} & C_{\theta\mathbf{r}} \\ C_{\mathbf{r}\theta} & C_{\mathbf{rr}} \end{bmatrix} \begin{bmatrix} \theta(t_k) \\ \mathbf{r}(t_k) \end{bmatrix}\ , \quad \begin{bmatrix} F_\theta^{(2)}(t_k) \\ F_\mathbf{r}^{(2)}(t_k) \end{bmatrix} = \begin{bmatrix} G_{\theta\theta} & G_{\theta\mathbf{r}} \\ G_{\mathbf{r}\theta} & G_{\mathbf{rr}} \end{bmatrix} \begin{bmatrix} \theta(t_k) \\ \mathbf{r}(t_k) \end{bmatrix}\ , \tag{7}$$

which allows us to write

$$\begin{aligned} \begin{bmatrix} \alpha_0 C_{\theta\theta} + G_{\theta\theta} & \alpha_0 C_{\theta\mathbf{r}} + G_{\theta\mathbf{r}} \\ \alpha_0 C_{\mathbf{r}\theta} + G_{\mathbf{r}\theta} & \alpha_0 C_{\mathbf{rr}} + G_{\mathbf{rr}} \end{bmatrix} \begin{bmatrix} \theta(t_k) \\ \mathbf{r}(t_k) \end{bmatrix} + \begin{bmatrix} \beta_\theta(t_k) \\ \beta_\mathbf{r}(t_k) \end{bmatrix} = \\ \begin{bmatrix} B_{\theta\theta} & B_{\theta\mathbf{r}} \\ B_{\mathbf{r}\theta} & B_{\mathbf{rr}} \end{bmatrix} \begin{bmatrix} \theta(t_k) \\ \mathbf{r}(t_k) \end{bmatrix} + \begin{bmatrix} \beta_\theta(t_k) \\ \beta_\mathbf{r}(t_k) \end{bmatrix} = \begin{bmatrix} J_2(t_k) \\ 0 \end{bmatrix}\ . \end{aligned} \tag{8}$$

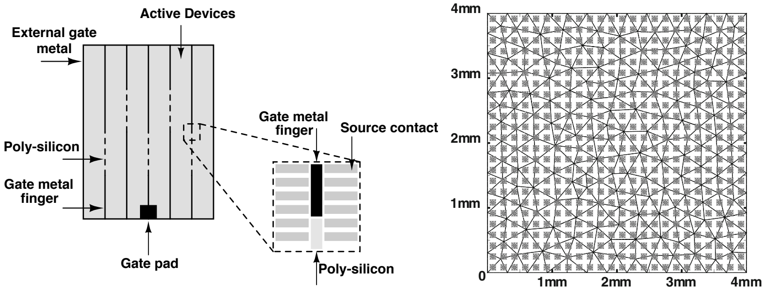
By the arguments used in [10] $B_{\mathbf{rr}}$ can be shown to be invertible, so that $\mathbf{r} = B_{\mathbf{rr}}^{-1}(B_{\mathbf{r}\theta}\theta + \beta_\mathbf{r})$ from which we get

$$G^{(2)}\theta = \mathbf{J}_2 + \widetilde{\mathbf{J}}_2, \tag{9}$$

where the conductance matrix $G^{(2)} = (B_{\theta\theta} - B_{\theta\mathbf{r}} B_{\mathbf{rr}}^{-1} B_{\mathbf{r}\theta})$ is of size $(L+1)\times(L+1)$ and the dynamical current $\widetilde{\mathbf{J}}_2 = -(\beta_\theta - B_{\mathbf{rr}}^{-1} B_{\mathbf{r}\theta}\beta_\mathbf{r})$ is of size $(L+1)\times 1$ regardless of the number of mesh nodes.

## 3 Numerical Results

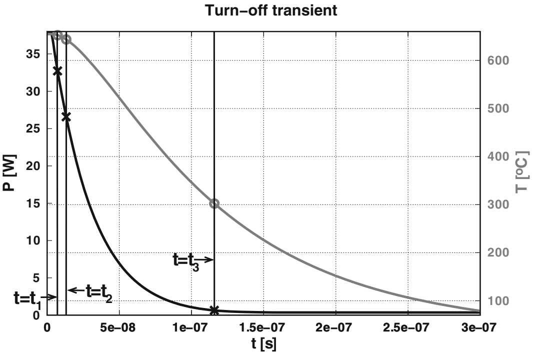
As a preliminary validation of the proposed method we present results obtained by its application to a simplified version of the power device discussed in [11]. In Fig. 2a we show a sketch of the complete device layout, composed of a large number of identical cells arranged in a regular grid. Notice that the gate metal fingers do not cover the entire layout from the upper to the lower part, but they leave some empty space in which the gate signal is propagated through Poly-Silicon. This lack of a direct metal connection produces hot-spot phenomena during high frequency switching. In [11] a distributed electrical network was introduced which allowed to observe local maxima in the current density distribution, in [12] the model is extended to account for the self-heating of the cells, still the dependence of the electrical characteristics of each cell on the dissipated power remains purely local. The non locality introduced in our model by the distributed thermal element provides a



**(a)** Sketch of the device layout **(b)** Thermal network mesh

**Fig. 2:** Simplified power MOS-FET structure

**Fig. 3** Turn-off transient for the device of Fig. 2a: total dissipated power vs. mean temperature on chip



further extension as it allows to account for mutual-heating effects previously neglected. In Fig. 2b we show a picture of the mesh underlying the distributed thermal network: a coarse grid covers the 4mm × 4mm die, while a fine one (approximately 80 $\mu$m × 80 $\mu$m) is replicated at each active region position. The parameters $c_v$ and $\kappa$ in this particular example where chosen by fitting the turn-off time of our simplified model to the results presented in [11]; the resulting values were $c_v = 10^{-4}$ J $\times$ m$^{-2}$ $\times$ K$^{-1}$ and $\kappa = 2 \times 10^{-2}$J $\times$ sec$^{-1}$ $\times$ K$^{-1}$. In Fig. 3 the total dissipated power and the mean temperature of the device are plotted against time during a turn-off transient. As expected to a lowering of the power corresponds a cooling of the device; however these two effects exhibit different relaxation times. Finally in Fig. 4 we present the power densities and lumped temperatures of the cells for three different time-points defined in Fig. 3. We can see clearly a delay in the propagation of the signal from the gate-pad in the lower part of the die to the single cells, and the presence of an hot-spot in the central upper part of the die for $t = t_2$. Moreover the presence of a non negligible temperature gradient over the device area is detected at times $t = t_1$ and $t = t_2$. Furthermore the different spatial distribution of heat density and temperature are an indication that non-local effects may not be negligible in estimating the device performance.

## 4 Conclusions

In Fig. 5 we summarize the expected work-flow using the tool we presented for a coupled electro-thermal simulation. In designing the IC the thermally active regions are defined by adding an extra mask layer to the layout. A 2D or 3D mesh is formed automatically and from that a passive thermal element will be assembled. This element is then attached to the devices temperature nodes; ambient temperature is set via additional temperature sources. The full system can be simulated by a standard circuit simulator, producing as extra output the average temperature in each device as well as the full multidimensional temperature field in the IC.

## Acknowledgments

The first author was supported by the European Commission in the framework of the CoMSON RTN project. The second author was supported by the Mathematics Applications Consortium for Science and Industry in Ireland (MACSI) under the Science Foundation Ireland (SFI) mathematics initiative.

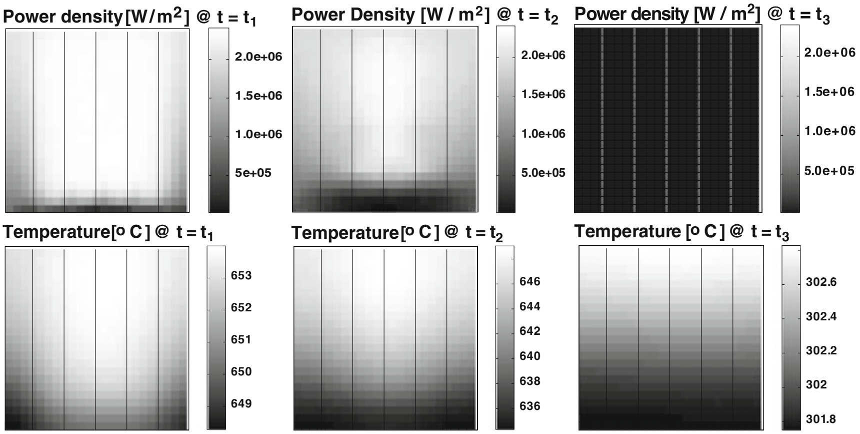


**Fig. 4:** Power densities and lumped temperatures at times $t_1$, $t_2$ and $t_3$ defined in Fig. 3

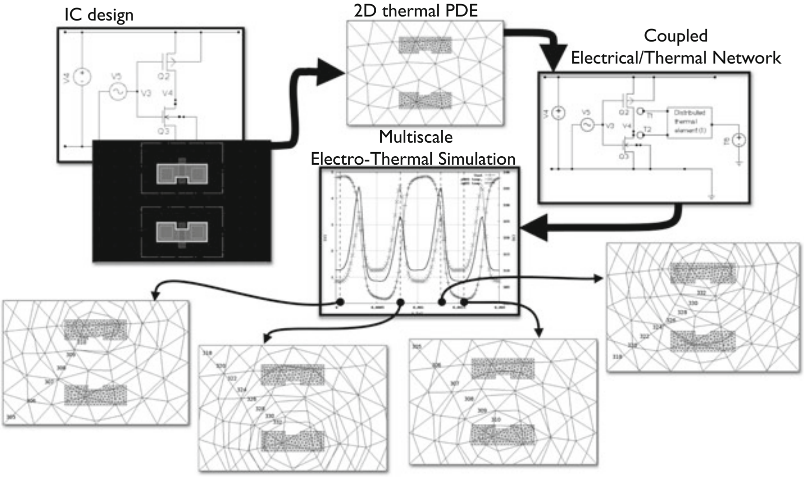

IC design
2D thermal PDE
Coupled
Electrical/Thermal Network
Multiscale
Electro-Thermal Simulation
Distributed thermal element

**Fig. 5:** Electro-thermal simulation: expected work-flow in an industrial environment

## References

1. Osman, A., Osman, M., Dogan, N., Imam, M.: An extended tanh law mosfet model for high temperature circuit simulation. Solid-State Circuits, IEEE Journal of **30**(2), 147–150 (1995)
2. Ku, J.C., Ismail, Y.: On the scaling of temperature-dependent effects. Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on **26**(10), 1882–1888 (2007)
3. Szekely, V., Poppe, A., Pahi, A., Csendes, A., Hajas, G., Rencz, M.: Electro-thermal and logi-thermal simulation of vlsi designs. VLSI Systems, IEEE Transactions on **5**(3), 258–269 (1997)
4. Igic, P., Mawby, P., Towers, M., Batcup, S.: Dynamic electro-thermal physically based compact models of the power devices for device and circuit simulations. Semiconductor Thermal Measurement and Management, 2001. XVII Annual IEEE Symposium pp. 35–42 (2001)
5. Ho, C., Ruehli, A., Brennan, P.: The modified nodal approach to network analysis. Circuits and Systems, IEEE Transactions on **22**(6), 504–509 (1975)
6. Günther, M., Feldmann, U.: CAD-based electric-circuit modeling in industry. I. Mathematical structure and index of network equations. Surveys Math. Indust. **8**(2), 97–129 (1999)
7. Bartel, A., Günther, M.: From SOI to abstract electric-thermal-1d multiscale modeling for first order thermal effects. Math. Comput. Model. Dyn. Syst. **9**, 25–44 (2003)
8. de Falco, C., Culpo, M.: A PDE thermal model for CHIP-level simulation including substrate heating effects. Tech. rep., DCU School of Math. Sciences (2008)
9. Rezzonico, V., Lozinski, A., Picasso, M., Rappaz, J., Wagner, J.: Multiscale algorithm with patches of finite elements. Math. Comput. Simulation **76**(1-3), 181–187 (2007)
10. Alí, G., Bartel, A., Culpo, M., de Falco, C.: Analysis of a PDE thermal element model for electrothermal circuit simulation (2008). Submitted to SCEE08 prooceedings
11. Biondi, T., Greco, G., Allia, M., Liotta, S., Bazzano, G., Rinaudo, S.: Distributed modeling of layout parasitics in large-area high-speed silicon power devices. Power Electronics, IEEE Transactions on **22**(5), 1847–1856 (2007)
12. Greco, G., Rallo, C.: Xa integration in custom power mosfet analysis flow. In: SNUG 2008 proceedings (2008)

# Simulations of an Electron-Phonon Hydrodynamical Model Based on the Maximum Entropy Principle

V. Romano and C. Scordia

**Abstract** Recently an energy-transport models has been formulated based on the maximum entropy principle for the coupled phonon-electron system in silicon in order to cope with the effects of heating of the crystal lattice. Here the numerical simulations of some benchmark devices are presented in order to assess the validity of the model.

## 1 The Model

Thermal effects in the crystal lattice influence the electrical behaviour, in particular in nanoscale devices. At macroscopic level, several heuristic models of lattice heating have been proposed. They are represented by the lattice energy balance equation and differ for the proposed form of thermal conductivity and energy production, e.g. [1–5]. A critical review can be also found in [6].

Recently, a consistent hydrodynamical model for charge carriers has been developed starting from the moment system associated with the transport equations, obtaining the closure relations with the maximum entropy principle (hereafter MEP) [7–10]. The same approach has been adopted in [11] for the electron-phonon system, obtaining also an energy-transport and a drift-diffusion model under appropriate scalings. The electrons are described with the 8-moment system as in [7–9]. The phonons are considered as two populations: acoustic and non polar optical. The non-polar optical phonons are described with a Bose-Einstein distribution while the acoustic ones are described by the MEP distribution function in the 9-moment approximation already introduced in [12]. For the acoustic phonons the linear

V. Romano
Department of Mathematics and Computer Science, University of Catania, Catania, Italy, e-mail: romano@dmi.unict.it

C. Scordia
University of Wuppertal, Wuppertal, Germany, e-mail: scordia@dmi.unict.it

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_37,

dispersion relation has been adopted while the non-polar optical phonons energy has been taken as constant. Moreover the non polar optical phonons are assumed to be described by the Bose-Einstein distribution.

The direct solution of the Boltzmann equations describing the electron-phonon system is a daunting computational task and requires long CPU times, not yet suitable for CAD purposes. Macroscopic models are therefore warranted. Starting from the transport equations, in [11] the 8-moment model for electron and the 9-moment model for phonons have been considered.

Since the number of unknowns exceeds the number of equations, closure relations must be introduced. To this aim MEP has been adopted. It gives a systematic way for obtaining constitutive relations on the basis of information theory. Explicit constitutive relations have been obtained with coefficients depending on the electron energy $W$ and crystal temperature $T_L$ and related to the scattering parameters (see [11] for more details).

Under the diffusion scaling, from the moment system closed with MEP, the following electron energy transport and lattice heating model has been deduced

$$\frac{\partial n}{\partial t} + \operatorname{div}(n\mathbf{V}) = 0 \tag{1}$$

$$\frac{\partial (nW)}{\partial t} + \operatorname{div}(n\mathbf{S}) + nq\mathbf{V}\cdot\nabla\phi = nC_W \tag{2}$$

$$\rho c_V \frac{\partial T_L}{\partial t} - \operatorname{div}[k(T_L)\nabla T_L] = H \tag{3}$$

$$\mathbf{E} = -\nabla\phi, \quad \varepsilon\Delta\phi = -q(N_D - N_A - n). \tag{4}$$

with $n$ the electron density, $W$ the electron energy, $\phi$ the electrostatic potential and $\mathbf{E} = -\nabla\phi$ the electric field. $N_D$ and $N_A$ are the donors and acceptors density respectively (assumed as known function of the position). $q$ is the elementary charge, $\rho$ the silicon density, $c_V$ the specific heat, $C_W$ the energy production term, which is in a relaxation form $C_W = -\frac{W-W_0}{\tau_W}$ with $W = 3/2k_BT_L$ and $\tau_W(W)$ the energy relaxation time.

The thermal conductivity $k(T_L)$ and heat source $H$ are given by [11]

$$k(T_L) = \frac{\rho c_V \tau_R(T_L)c^2}{3}, \quad H = -nC_W - c^2\operatorname{div}\left(\tau_R n c_{11}^{(p)}\mathbf{V} + \tau_R n c_{12}^{(p)}\mathbf{S}\right). \tag{5}$$

$c_{11}^{(p)}$ and $c_{12}^{(p)}$ arise from the phonon momentum production term and depend on $W$ and $T_L$ while $\tau_R$ is the phonon relaxation time in resistive processes. The electron velocity $V$ and energy flux $S$ are given by

$$\mathbf{V} = D_{11}(W,T_L)\nabla\log n + D_{12}(W,T_L)\nabla W + D_{13}(W,T_L)\nabla\phi,$$
$$\mathbf{S} = D_{21}(W,T_L)\nabla\log n + D_{22}(W,T_L)\nabla W + D_{23}(W,T_L)\nabla\phi$$

(see [11] for the explicit expressions of the coefficients $D_{ij}$). Equations 5 generalize the models proposed in [1–5] with an explicit form of the coefficients.

The argument of the divergence operator in $H$ can be rewritten as

$$-P_n\mathbf{J} - P_S n\mathbf{S}, \tag{6}$$

with $\mathbf{J} = -qn\mathbf{V}$ the current density, $P_n = \frac{c^2\,\tau_R\,c_{11}^{(p)}}{q}$ and $P_S = -c^2\,\tau_R\,c_{12}^{(p)}$ thermoelectric power coefficients (note that $c_{12}^{(p)}$ is negative).

If we consider $P_n$ and $P_S$ as constant, in the stationary case one can use the electron energy balance equation (2) to eliminate div $(n\mathbf{S})$ obtaining the simplified model for the phonon energy production

$$H = -(1+P_S)\,nC_W + P_S\,\mathbf{J}\cdot\mathbf{E} \tag{7}$$

This indeed alters the results of the transient but leads to the same stationary solutions.

Since the electron production terms are slowly changing with respect to $K_B T_L$, we will use the further simplification that they are evaluated with $T_L = 300$ K.

A phenomenological radiation term $S_L(T_L - T_{en})$ for the exchange of energy with the environment is added to $H$, $T_{en}$ being the environment temperature and $S_L$ the transmission coefficient.

The following boundary conditions for the 1D $n^+ - n - n^+$ silicon diode are assumed (the device is represented by the interval $[0,L]$):

$$\begin{aligned}
n(0,t) = n_D(0), \quad n(L,t) = n_D(L) &\quad t \geq 0\\
\frac{\partial W}{\partial x}(0,t) = \frac{\partial W}{\partial x}(L,t) = 0 &\quad t \geq 0\\
\phi(0,t) = 0, \quad \phi(L,t) = V_b &\quad t \geq 0\\
-k(T_L)\frac{\partial T_L}{\partial x} = R_{th}^{-1}(T_L - T_{en}) \quad x = 0, L &\quad t \geq 0
\end{aligned}$$

where $R_{th}$ is the thermal resistivity of the contact. With a good approximation the latter relation can be replaced with the homogeneous Neumann condition $\frac{\partial T_L}{\partial x} = 0$ due to the high value of $R_{th}$.

## 2 Numerical Scheme

We discretize the balance equations by adopting the following coupling strategy:

- first we integrate the balance equations for electrons, with the crystal lattice frozen at the previous time step, obtaining the electron density and energy at the next time step.
- then we integrate the lattice energy balance equation in a semi-implicit way with $n$ and $W$ given by the step 1.

### 2.1 Step 1

In this step the numerical approach is similar to that in [13]. First we rewrite the current density $\mathbf{J} = n\,\mathbf{V}$ and the energy-flux density $\mathbf{H} = n\,\mathbf{S}$ as

$$\mathbf{J} = \mathbf{J}^{(1)} - \mathbf{J}^{(2)}, \quad \mathbf{H} = \mathbf{H}^{(1)} - \mathbf{H}^{(2)},$$

where

$$\mathbf{J}^{(1)} = \frac{c_{22}}{D}\left[\nabla(nU) - qn\lambda^W U \nabla\phi\right], \quad \mathbf{J}^{(2)} = \frac{c_{12}}{D}\left[\nabla(nF) - qn\lambda^W F \nabla\phi\right],$$
$$\mathbf{H}^{(1)} = \frac{c_{11}}{D}\left[\nabla(nF) - qn\lambda^W F \nabla\phi\right], \quad \mathbf{H}^{(2)} = \frac{c_{12}}{D}\left[\nabla(nU) - qn\lambda^W U \nabla\phi\right],$$

with $D = c_{11}c_{22} - c_{12}c_{21}$. The $c_{ij}(W)$'s arise from the electron momentum and energy-flux production, $\lambda_W(W)$ is the Lagrangian multiplier relative to the electron energy and $U$ and $F$ depend on $W$. Their expressions are given also in [11].

The basic idea is to introduce suitable average mobilities that are constant in each cell so that $\mathbf{J}^{(i)}$ and $\mathbf{H}^{(i)}$, i = 1, 2, can be expressed by means of *local* Slotboom variables and a Scharfetter-Gummel finite difference scheme can be used. The details can be found in [13].

Let us introduce the grid point $0 = x_0 < x_1 < \ldots x_i < \ldots x_{N-1} < x_N = L$, with $N$ a positive integer. For simplicity we assume a uniform grid so $x_i = ih$ with $h = L/N$, and uniform time steps. Moreover we set $I_{i+1/2} = [x_i, x_{i+1}]$ and $x_{i\pm1/2} = x_i \pm h/2$. In the sequel the notation $u_i^l$ will indicate the value of the variable $u(x,t)$ for $x = x_i$ and $t = l\,\Delta t$, $l$ being a positive integer.

By replacing the partial derivatives with finite differences, the balance equations (1)-(2) can be discretized as

$$\frac{n_i^{l+1} - n_i^l}{\Delta t} + \frac{J_{i+1/2} - J_{i-1/2}}{h} + O(h^2, \Delta t) = 0, \tag{8}$$

$$\frac{(nW)_i^{l+1} - (nW)_i^l}{\Delta t} + \frac{H_{i+1/2} - H_{i-1/2}}{h} - q\frac{J_{i+1/2} + J_{i-1/2}}{2}\frac{\phi_{i+1} - \phi_{i-1}}{2h} + $$
$$+ \frac{3}{2}\, n_i \frac{W_i - W_0}{(\tau_W)_i} + O(h^2, \Delta t) = 0, \tag{9}$$

$$\frac{1}{h^2}\left(\phi_{i+1} - 2\phi_i + \phi_{i-1}\right) + \frac{q}{\varepsilon}(C_i - n_i) + O(h^2) = 0 \tag{10}$$

where $C_i = N_D(x_i) - N_A(x_i)$. The variables with no temporal index, in particular the lattice temperature, must be considered evaluated at the time step $t = l\,\Delta t$. We approximate the electric potential $\phi$ by piece-wise linear function in each $I_{i+1/2}$

$$\phi(x) \simeq \phi_i + \frac{(x - x_i)}{h}(\phi_{i+1} - \phi_i), \quad x \in I_{i+1/2}$$

and $c_{ij}(W)$ by functions that are constant on each interval $I_{i+1/2}$. This enable us to introduce the *local* mobilities

$$g_{11} = -\frac{c_{22}}{D}, \quad g_{12} = -\frac{c_{12}}{D}, \quad g_{21} = -\frac{c_{11}}{D}, \quad g_{21} = -\frac{c_{12}}{D} \tag{11}$$

and write the significant components of $\mathbf{J}^{(i)}$ and $\mathbf{H}^{(i)}$ as

$$J^{(i)} \simeq -\frac{\partial g_{1i}}{\partial x} + q\overline{\lambda}^{W} g_{1i}\frac{\partial \phi}{\partial x}, \quad H^{(i)} \simeq -\frac{\partial g_{2i}}{\partial x} + q\overline{\lambda}^{W} g_{2i}\frac{\partial \phi}{\partial x} \tag{12}$$

where $\overline{\lambda}^{W}$ is the cell mean value of $\lambda^W$, we approximate as

$$\overline{\lambda}^{W} \simeq \frac{1}{2}\left[\lambda^W(W_i) + \lambda^W(W_{i+1})\right]. \tag{13}$$

After introducing $U_T = 1/q\lambda^W$, which plays the role of a thermal potential, and indicating by $\overline{U}_T$ its constant approximation in each cell $I_{i+1/2}$, it is possible to define the *local* Slotboom variables $s_{kr} = \exp\left(-\phi/\overline{U}_T\right) g_{kr}$ that satisfy

$$\frac{\partial s_{1r}}{\partial x} \simeq -\exp\left(-\phi/\overline{U}_T\right) J^{(r)}, \quad \frac{\partial s_{2r}}{\partial x} \simeq -\exp\left(-\phi/\overline{U}_T\right) H^{(r)}. \tag{14}$$

In each cell $I_{i+1/2}$ we can express $J^{(r)}$ as a Taylor expansion

$$J^{(r)}(x) = J^{(r)}_{i+1/2} + (x - x_{i+1/2})\left(\frac{\partial J^{(r)}}{\partial x}\right)_{x_{i+1/2}} + O(h^2). \tag{15}$$

By integrating the relations (14) over $I_{i+1/2}$, we find up to $O(h^2)$

$$(s_{1r})_{i+1} - (s_{1r})_i = -\int_{x_i}^{x_{i+1}} \exp\left(-\phi/\overline{U}_T\right)\, J^{(r)}_{i+1/2}\, \mathrm{d}x.$$

By taking into account that $\phi(x)$ is linear in $I^{(r)}_{i+1/2}$, the last integral can be explicitly evaluated, obtaining, after some elementary algebra

$$J^{(r)}_{i+1/2} = -z\coth z\,\frac{(g_{1r})_{i+1} - (g_{1r})_i}{h} + z\frac{(g_{1r})_{i+1} + (g_{1r})_i}{h}, \quad r = 1,2 \tag{16}$$

with $z = \frac{\phi_{i+1}-\phi_i}{2\overline{U}_T}$. If $z = 0$, that is if $\phi_{i+1} = \phi_i$, the previous expression remains valid provided that $z\coth z$ is replaced with the limit as $z \mapsto 0$ which is equal to one.

Similarly for the energy density current one finds

$$H^{(r)}_{i+1/2} = -z\coth z\,\frac{(g_{2r})_{i+1} - (g_{2r})_i}{h} + z\frac{(g_{2r})_{i+1} + (g_{2r})_i}{h}, \quad r = 1,2. \tag{17}$$

The complete numerical scheme of this first step is summarized below

$$n_i^{l+1} = n_i^l - \Delta t\, \frac{J_{i+1/2} - J_{i-1/2}}{h} = 0, \tag{18}$$

$$(nW)_i^{l+1} = (nW)_i^l - \Delta t \left[ \frac{H_{i+1/2} - H_{i-1/2}}{h} - q \frac{J_{i+1/2} + J_{i-1/2}}{2} \frac{V_{i+1} - V_{i-1}}{2h} + \right. \\ \left. + \frac{3}{2}\, n_i \frac{W_i - W_0}{(\tau_W)_i} \right], \tag{19}$$

$$\frac{1}{h^2}\left(\phi_{i+1} - 2\phi_i + \phi_{i-1}\right) + \frac{q}{\varepsilon}(C_i - n_i) = 0 \tag{20}$$

$$J_{i+1/2} = J^{(1)}_{i+1/2} - J^{(2)}_{i+1/2}, \quad H_{i+1/2} = H^{(1)}_{i+1/2} - H^{(2)}_{i+1/2} \tag{21}$$

$$J^{(r)}_{i+1/2} = -z \coth z\, \frac{(g_{1r})_{i+1} - (g_{1r})_i}{h} + z \frac{(g_{1r})_{i+1} + (g_{1r})_i}{h}, \quad r = 1,2 \tag{22}$$

$$H^{(r)}_{i+1/2} = -z \coth z\, \frac{(g_{2r})_{i+1} - (g_{2r})_i}{h} + z \frac{(g_{2r})_{i+1} + (g_{2r})_i}{h}, \quad r = 1,2 \tag{23}$$

supplemented with a CFL condition $\Delta t/(\Delta x)^2 < c$, where $c$ is a suitable positive constant.

### 2.2 Step 2

Regarding the discretization of the lattice energy equation, an explicit scheme is used. Setting $u = K_B T_L$, one has for the internal nodes

$$u_i^{n+1} = u_i^n + \frac{\Delta t_L}{\Delta x^2}\left[ \frac{\tilde{K}_i + \tilde{K}_{i+1}}{2}(u^n_{i+1} - u^n_i) - \frac{\tilde{K}_i + \tilde{K}_{i-1}}{2}(u^n_i - u^n_{i-1}) \right] \\ + \frac{a \Delta t_L}{(\tau_W)_i}(1 + P_S) n_i^{n+1} \left( W_i^{n+1} - \frac{3}{2} u_i^n \right) + b \Delta t_L J_i^{n+1} E_i^{n+1} + \frac{\Delta t_L S_L}{\rho c_V}(u_i^n - k_B T_{en}). \tag{24}$$

where $a = \dfrac{k_B}{\rho c_V}$, $b = a P_S$, $\tilde{K} = \dfrac{k}{\rho c_V}$. The time step $\Delta t_L$ in (24) is related to the time step $\Delta t$ in the discretization of (1),(2) by $\Delta t_L \approx 10^{-2} \Delta t$, which implies about 100 iterations of (24) for each time step in the numerical integration of (1)-(2).

## 3 Numerical Simulations

Concerning the physical parameters, we have modeled the thermal conductivity with the fitting formula $k(T_L) = 1.5486\,(T_L/300K)^{-4/3}$ V A/cm K, assumed $c_V = 703$ m$^2$/sec$^2$ K (see [6]) and set $T_{en}$ = 300 K.

Two nanodevices under the bias voltage of 1 Volt are considered: an $n^+ - n - n^+$ silicon diode with a channel of 50 nanometers and another with a channel of 25 nanometers. In submicron diodes with longer channels the effects related to the crystal heating are not relevant. The overall stationary state is reached in about 100 picoseconds, even if the electron parts is practically in the steady state after about 5 ps. The lattice temperature depends strongly on the thermal power coefficient and the radiation coefficient $S_L$. If $P_S$ is set equal to zero, that is only the relaxation term is considered in the source term of (3), a negligible crystal heating is obtained and this implies that the major role is played by the Joule effects, casting doubts on the models including only the relaxation part in $H$.

In Fig. 1 the lattice temperature is plotted along the device for $S_L = 4.0 \times 10^{12}, 2.0 \times 10^{13}, 4.0 \times 10^{13}$ W/m$^3$ K, considering the device with the channel 50 nanometers long. For the lowest values of $S_L$ there is a dramatic heating, which appears as unphysical. For the higher values of $S_L$ the maximum raise of the temperature is more realistically less than 10 K.

Similar results are found for the diode having a 25 nanometers channel, but with a more pronounced rise in temperature, implying an increase of the thermal effects as shrinking the dimension of the device. This is qualitatively in agreement with the results obtained by using other models, e.g. in [14].

The electron variables, density and energy, are only slightly influenced by the thermal effects. In particular the current changes less than two percent, see Fig. 2. Note that higher the crystal temperature and smaller is the current. However, even a small deviation in the temperature of each single device implies a relevant cumulative effects in integrated circuits with a very high number of components.

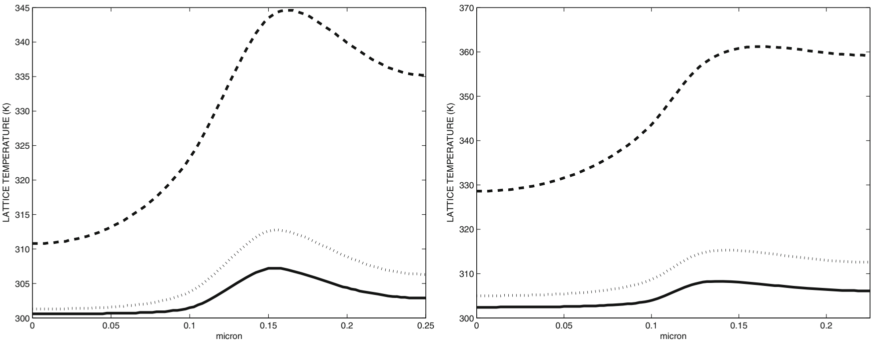


**Fig. 1:** Lattice temperature versus the position in the device with the channel of 50 nanometers (*left*) and 25 nanometers (*right*), for different values of the transmission coefficient: $S_L = 4.0 \times 10^{12}$ (*dashed line*), $2.0 \times 10^{13}$ (*dotted line*), $4.0 \times 10^{13}$ (*continuous line*) W/m$^3$ K

**Acknowledgements** The authors acknowledge the financial support by M.I.U.R. (PRIN 2007 *Equazioni cinetiche e idrodinamiche di sistemi collisionali e complessi*), P.R.A. University of Catania (ex 60 %), the EU Marie Curie RTN project COMSON grant n. MRTN-CT-2005-019417.

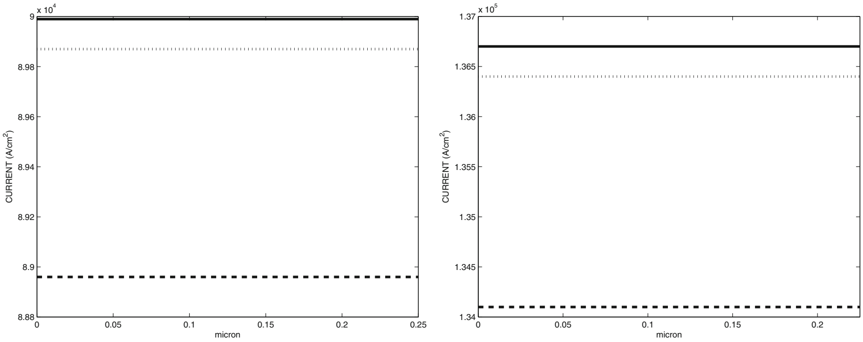


**Fig. 2:** Current versus the position in the device with the channel of 50 nanometers (*left*) and 25 nanometers (*right*), for different values of the transmission coefficient: $S_L = 4.0 \times 10^{12}$ (*dashed line*), $2.0 \times 10^{13}$ (*dotted line*), $4.0 \times 10^{13}$ (*continuous line*) W/m$^3$ K

# References

1. Gaur, S.P., Navon, D.H.: Two-dimensional carrier flow in a transistor structure under non-isothermal conditions. IEEE trans. Electron. Devices **ED-23**, 50–57 (1976)
2. Sharma, D.K., Ramanthan, K.V.: Modeling thermal effetcs on MOS I-V characteristics. IEEE Electron. Device Lett. **EDL-4**, 362–364 (1983)
3. Adler, M.S.: Accurate calculations of the forward drop and power dissipation in thyristors. IEEE Trans. Electron. Devices, **ED-25**, 16–22 (1979)
4. Chryssafis, A., Love, W.: A computer-aided analysis of one dimensional thermal transient in n-p-n power transistors. Solid-State-Electron. **22**, 249-256 (1978)
5. Wachutka, G.: Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling. IEEE Trans. on Computer-Aided Design **9**, 1141–1149 (1990)
6. Selberherr, S.: Analysis and simulation of semiconductor devices. Wien - New York, Springer-Verlag (1984)
7. Anile, A.M., Romano, V.: Non parabolic band transport in semiconductors: closure of the moment equations. Continuum Mech. Thermodyn. **11**, 307–325 (1999)
8. Romano, V.: Non parabolic band transport in semiconductors: closure of the production terms in the moment equations. Continuum Mech. Thermodyn. **12**, 31–51 (2000)
9. Romano, V.: Non parabolic band hydrodynamical model of silicon semiconductors and simulation of electron devices. Mathematical Methods in the Applied Sciences **24** 439 (2001)
10. Jaynes, E.T.: Information theory and statistical mechanics. Physical Review **106**(4), 620 (1957).
11. Romano, V., Zwierz, M.: *Electron-phonon hydrodynamical model for semiconductors*, preprint (2008)
12. Dreyer, W., Struchtrup, H.: Heat pulse experiment revisited. Continuum Mech. Thermodyn. **5**, 3–50 (1993)
13. Romano, V.: 2D numerical simulation of the MEP energy-transport model with a finite difference scheme. J. Comp. Physics **221** 439–468 (2007)
14. Brunk, M., Jüngel, A.: Numerical coupling of electric circuit equations and energy-transport models for semiconducotrs. SIAM J. Sci. Comput. **30**, 873–894 (2008)

# Consistent Initialization for Coupled Circuit-Device Simulation

Sascha Baumanns, Monica Selva Soto, and Caren Tischendorf

**Abstract** For a coupled circuit device simulation in the time domain, consistent initial values have to be calculated. We study the structure and properties of the differential-algebraic equations (DAEs) that arise after space discretization of the partial differential equation part coming from the device modelling. Exploiting the special DAE structure, we show that a consistent initial value can be computed within two steps. Firstly, one determines an operation point. Secondly, a linear system is solved for correcting the operation point such that the hidden constraints are also satisfied. Finally, an algorithm for the calculation of such values is proposed.

## 1 Introduction

Nowadays semiconductor devices in an electrical circuit are modeled via equivalent circuits containing only basic elements that can be described by algebraic and ordinary differential equations. With the rapid development of chip technology these equivalent circuits have become more and more complex. This has motivated the idea of using distributed device models, represented by a system of Partial Differential Equations (PDE), to describe the behavior of the semiconductor devices in the circuit [1, 2]. The resulting mathematical model couples the differential algebraic equations (DAEs) describing the circuit and the partial differential equations (PDEs) modeling the semiconductor devices.

In order to numerically simulate electrical circuits described by such a model, we discretize the partial differential equations in space first. This results in a DAE for the coupled simulation problem. The numerical simulation of this DAE involves the problems of finding consistent initial values for the integration. DAEs are known for the fact that solutions have to fulfill certain constraints. Correspondingly, initial values have to be found that satisfy these constraints.

The main objective of this article is the determination of appropriate initial values for the DAE arising after space discretization. We study which conditions initial values should satisfy in order to be consistent and present an algorithm for their calculation. This algorithm is based on the ideas presented in [3, 4]. Due to the

S. Baumanns, M. Selva Soto, C. Tischendorf
Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Cologne, Germany, e-mail: sbaumann@math.uni-koeln.de, mselva@math.uni-koeln.de, tischendorf@math.uni-koeln.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_38,

special properties of this DAE, discussed later on in this paper, a consistent initial value for it can be calculated in two steps: in the first one an operating point is computed and in the second one this point is corrected by solving a linear system of equations.

This paper is organized as follows. First we describe briefly the coupled DAE-PDE model of the coupled circuit device system as well as the DAE system that is obtained after spatial discretization. Section 3 is devoted to the properties of this DAE. In section 4, the conditions for consistent initial values are studied and an algorithm for the calculation of such values is proposed. This algorithm has been implemented in MATLAB.

## 2 Coupled System for the Circuit and Device Simulation

For simplicity and shorter description, we restrict to the case of coupling only one semiconductor device to an electrical circuit. Assume this semiconductor device to have $n_S$ metal semiconductor contacts and let $n_N$ be the number of nodes in the graph associated to the circuit. Each contact of the semiconductor device is joined to a node of the electrical circuit. The contacts of the semiconductor joined to the same node of the electrical circuit define a terminal. Let $n_T$ be the number of terminals of the semiconductor device. We define the following incidence matrix $A_S \in \mathbb{R}^{(n_N-1)\times(n_T-1)}$ by

$$A_S(i,j) = \begin{cases} 1, & \text{if terminal } j \text{ is joined to the node } i \\ -1, & \text{if the reference terminal is attached to node } i \\ 0, & \text{else} \end{cases}$$

The system proposed in [2] couples the modified nodal analysis (MNA) equations for electrical circuits to the drift diffusion (DD) equations for semiconductor devices. The MNA equations have the form

$$A_C \frac{\mathrm{d}q_C(A_C^T e,t)}{\mathrm{d}t} + A_R g_R(A_R^T e,t) + A_L j_L + A_V j_V + A_S j_S + A_I i_S(t) = 0, \tag{1a}$$

$$\frac{\mathrm{d}\phi(j_L,t)}{\mathrm{d}t} - A_L^T e = 0, \tag{1b}$$

$$A_V^T e - v_S(t) = 0 \tag{1c}$$

with $t \in [t_0, t_F]$. The matrices $A_C, A_R, A_L, A_V, A_S$ and $A_I$ describe the element related reduced incidence matrices. The functions $v_S(t)$, $i_S(t)$, $q_C(u,t)$, $g(u,t)$ and $\phi(j,t)$ describe the constitutive relations for the circuit elements. As unknowns we have the node potentials $e(t) : \mathbb{R} \to \mathbb{R}^{n_N-1}$, except of the mass node, as well as the currents $j_L(t) : \mathbb{R} \to \mathbb{R}^{n_L}$ through inductors, the currents $j_V(t) : \mathbb{R} \to \mathbb{R}^{n_V}$ through voltage sources and the currents $j_S : \mathbb{R} \to R^{n_T-1}$through semiconductor devices. Note that the term $A_S j_S$ within the Kirchhoff's current law equation (1a) involves a coupling to the DD model since the current $j_S$ at the semiconductor's contacts depends on the DD variables.

Suppose $\Omega$ to be a bounded domain in $\mathbb{R}^d$, $d \in \{1,2,3\}$ and let $x \in \Omega$ represent the space variable. The DD equations are given by the following set of PDEs for the electrostatic potential $\psi(x,t)$ and the electrons and holes densities, $n(x,t)$ and $p(x,t)$ respectively.

$$\nabla \cdot (-\varepsilon \nabla \psi) - q(C - n + p) = 0, \tag{1d}$$

$$-\frac{\partial n}{\partial t} + \frac{1}{q}\mathrm{div} J_n - R(n,p) = 0, \qquad J_n - q\mu_n(U_T \nabla n - n \nabla \psi) = 0, \tag{1e}$$

$$\frac{\partial p}{\partial t} + \frac{1}{q}\mathrm{div} J_p + R(n,p) = 0, \qquad J_p + q\mu_p(U_T \nabla p + p \nabla \psi) = 0. \tag{1f}$$

For simplicity, we consider the mobilities $\mu_n$ and $\mu_p$ as well as the material quantities $\varepsilon$ and $U_T$ as constants. The elementary charge $q$ is always constant.

The boundary of the semiconductor device is here divided into two disjoint parts $\Gamma = \Gamma_D \cup \Gamma_N$. The first one includes the metal semiconductor contacts (Ohmic contacts) where the external potentials are applied and can divided into $n_T$ disjoints contacts. The contact $n_T$ is chosen as reference contact. The corresponding boundary conditions have the form

$$n = n_D(x), \quad p = p_D(x), \quad \psi = \psi_{bi}(x) + \psi_{ext}(x, A_S^{\mathrm{T}} e) \tag{1g}$$

for all $x \in \Gamma_D$ and $t \in [t_0, t_F]$. On $\Gamma_N$ homogeneous Neumann boundary conditions are imposed, i.e.

$$\nabla \psi \cdot \nu = 0, \quad J_n \cdot \nu = 0, \quad J_p \cdot \nu = 0 \tag{1h}$$

for all $x \in \Gamma_N$ and $t \in [t_0, t_F]$ with $\nu$ being the unit vector pointing in the outer normal direction of $\Omega$. In (1g), $\psi_{ext}(x, A_S^{\mathrm{T}} e)$ denotes the externally applied voltage[1] and $\psi_{bi}(x)$, $n_D(x)$ as well as $p_D(x)$ are given functions that do not depend on time.

The currents $j_{S_1}, j_{S_2}, \ldots, j_{S_{n_T-1}}$ at the semiconductor terminals can be calculated as

$$j_{S_i} = -\int_\Omega (J_n + J_p) \cdot \nabla w_i \,\mathrm{d}x - \frac{\mathrm{d}}{\mathrm{d}t} q_{S_i}, \quad q_{S_i} = -\int_\Omega \varepsilon \nabla \psi \cdot \nabla w_i \,\mathrm{d}x, \tag{1i}$$

where the functions $w_i$, $i = 1, 2, \ldots, n_T - 1$ are chosen such that

$$\nabla \cdot (-\varepsilon \nabla w_i) = 0, \quad \text{in } \Omega, \tag{2a}$$

$$w_i|_{\Gamma_j \subset \Gamma_D} = \delta_{ij},\ j = 1, 2, \ldots, n_T \quad \text{and} \quad \nabla w_i \cdot \nu = 0, \ \text{on } \Gamma_N. \tag{2b}$$

This way, we may express $\psi_{ext}(x, A_S^{\mathrm{T}} e)$ as

$$\psi_{ext}(x, A_S^{\mathrm{T}} e) = (w_1\ w_2\ \cdots\ w_{n_T-1}) \cdot A_S^{\mathrm{T}} e.$$

[1] Since we consider the semiconductor device as part of an electrical circuit, $\psi_{ext}$ is not an independent function to be assigned, but it is to be determined by the electrical network

The discretization of (1d)-(1f) in space results in a DAE system [5] for

$$y(t) = (e, j_L, j_V, j_S, q_S, \Psi, N, P)^{\mathrm{T}} : \mathbb{R} \to \mathbb{R}^m$$

with $m = n_N - 1 + n_L + n_V + 2(n_T - 1) + 3M$. Here $M$ denotes the number of interior and Neumann nodes in the spatial mesh used to discretize the PDEs in the system. For each $t \in [t_0, t_F]$, the vector $\Psi(t)$ contains the approximations to the values of $\psi$ at the mesh points or mesh elements. The same holds for $N(t)$ and $P(t)$. The resulting DAE has the form

$$A \frac{\mathrm{d}}{\mathrm{d}t} d(y,t) + b(y,t) = 0 \tag{3a}$$

with

$$A = \begin{pmatrix} A_C & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & M_h & 0 \\ 0 & 0 & 0 & 0 & M_h \end{pmatrix}, \qquad d = \begin{pmatrix} A_C^{+} A_C q_C(A_C^{\mathrm{T}} e(t), t) \\ \phi(j_L(t), t) \\ q_S(t) \\ N(t) \\ P(t) \end{pmatrix} \tag{3b}$$

and

$$b = \begin{pmatrix} A_R g_R(A_R^{\mathrm{T}} e, t) + A_L j_L + A_V j_V + A_S j_S + A_I i_S(t) \\ -A_L^{\mathrm{T}} e \\ A_V^{\mathrm{T}} e - v_S(t) \\ q_S + f(A_S^{\mathrm{T}} e, \Psi) \\ j_S + g(A_S^{\mathrm{T}} e, \Psi, N, P) \\ T_h \Psi + h(A_S^{\mathrm{T}} e, N, P) \\ r_1(A_S^{\mathrm{T}} e, \Psi, N) \\ r_2(A_S^{\mathrm{T}} e, \Psi, P) \end{pmatrix}. \tag{3c}$$

The matrix $A_C^{+}$ above denotes the Moore-Penrose inverse of $A_C$. All functions of the system above are assumed to be continuously differentiable with respect to all their components. Suppose further the matrices $T_h$ and $M_h$ to be symmetric and positive definite.[2] The partial derivatives

$$C(u,t) = \frac{\partial q_C(u,t)}{\partial u}, \quad L(j,t) = \frac{\partial \phi(j,t)}{\partial j}, \quad G(u,t) = \frac{\partial g_R(u,t)}{\partial u}$$

are also assumed to be positive definite, that means, we consider all capacitors, inductors and resistors to be passive.[3] Finally, suppose that

$$J_h = \frac{\partial f(u,\Psi)}{\partial u} - \frac{\partial h(u,N,P)}{\partial u} T_h^{-1} \frac{\partial f(u,\Psi)}{\partial \Psi}$$

[2] This is always true for Galerkin approximations with basis functions $\varphi_i(x)$ that provide independent functions $\frac{\mathrm{d}}{\mathrm{d}x}\varphi_i(x)$.

[3] We need passivity of resistors only if they are not connected by a capacitive path.

is a symmetric and positive definite matrix[4].

In what follows we assume that the circuit contains neither loops of voltage sources only nor cut sets of current sources only. It is a natural assumption since a violation would lead to a short circuit in reality.

## 3 Properties of the DAE Obtained After Spatial Discretization

Let $D(y,t) = \frac{\partial d(y,t)}{\partial y}$. Note that it is such that $\operatorname{im} D(y,t)$ is constant. The DAE (3) also has a properly stated leading term, that means,

$$\operatorname{im} D(y,t) \oplus \ker A = \mathbb{R}^k \tag{4}$$

with $k = n_C + n_L + n_T - 1 + 2M$ is satisfied and there exists a projector $R \in \mathbb{R}^{k \times k}$ that realizes the decomposition (4), i.e. $R^2 = R$, $\operatorname{im} D(y,t) = \operatorname{im} R$ and $\ker A = \ker R$. One possible choice is $R = A^+ A$, where $A^+$ denotes the Moore-Penrose inverse of $A$.

Since the network equations usually do not fulfill high smoothness conditions, we use the tractability index concept[6] for the index determination. Additionally, this concept leads us easily to network topological conditions characterizing the index of the discretized coupled DAE system. The DAE-index depends on the regularity of certain matrices $G_i, i = 0, 1, 2, \dots$ that are recursively constructed.

1. Since the matrix $G_0(y,t) = AD(y,t)$ is singular for all $(y,t) \in \mathbb{R}^m \times \mathbb{R}$, the DAE has always an index greater than zero [5, 7].
2. The DAE index is one if and only if the matrix $G_1(y,t) = G_0(y,t) + \frac{\partial b(y,t)}{\partial y} Q_0$ is nonsingular with $Q_0$ being a projector onto $\ker G_0$. It is shown that this is the case if the circuit contains neither loops of capacitors, voltage sources and semiconductor devices with at least one semiconductor device or one voltage source (CVS-loops) nor cut sets of inductors and currents sources (LI-cut sets) [5, 7]. Since $\operatorname{im} D(y,t)$ is constant, the DAE is also numerically qualified in this case.
3. In all other cases the matrix $G_2 = G_1 + \frac{\partial b(y,t)}{\partial y} P_0 Q_1$ with $Q_1(y,t)$ being a projector onto $\ker G_1(y,t)$ is nonsingular, i.e. the DAE index equals to two [5, 7].

[4] Let $w_{i,h}$ denote the approximations to the functions $w_i(x)$ defined in (2) with a Galerkin method. If they are written as linear combination of the same functions as the approximations to $\psi$ and $\frac{\mathrm{d}}{\mathrm{d}x} w_{i,h}(x)$ are linearly independent (this is e.g. the case if the spatial mesh is sufficiently fine), then it holds that $J_h$ is symmetric and positive definite, since

$$J_h(i,j) = \int_\Omega \nabla w_{i,h} \cdot \nabla w_{j,h} \, \mathrm{d}x, \quad i,j = 1, 2, \dots, n_T - 1.$$

## 4 Consistent Initial Values for the DAE Associated to the Coupled System

One of the difficult parts in solving DAEs numerically is to determine a consistent set of initial conditions in order to start the integration. In order to calculate consistent initial values for the DAE system (3), we exploit its special structure.

If (3) has index one, its flow is restricted to

$$\mathscr{M}_0(t) = \left\{ y \in \mathbb{R}^m | \exists z \in \mathbb{R}^k : Az + b(y,t) = 0 \right\}.$$

and $\mathscr{M}_0(t)$ is completely filled by this flow [6].

**Theorem 1.** *If the DAE (3) satisfies the conditions in section 2 and the circuit contains neither CVS-loops nor LI-cut sets, then the system*

$$\begin{aligned} Az_0 + b(y_0,t_0) &= 0, \\ (I-R)z_0 + d(y_0,t_0) - Ry^0 &= 0 \end{aligned}$$

*is locally uniquely solvable for $z_0$, $y_0$ and provides a consistent initial value $y_0$ for (3). The vector $y^0$ can be arbitrarily chosen. [6, 7]*

Speaking in terms of electrical variables, if the circuit contains neither CVS-loops nor LI-cut sets, initial values for the inductive currents, the capacitive branch voltages $A_C^{\mathrm{T}}e$, the charges at the semiconductor contacts $q_S$ and the concentrations of electrons and holes on the mesh nodes can be arbitrarily chosen.

The flow of index-two DAEs is additionally restricted by so-called hidden constraints and the set of consistent values at $t_0$ is a proper subset of $\mathscr{M}_0(t_0)$. In this case, we can compute a consistent initial value for (3) as follows [3, 7]

- Describe the hidden constraints.
- Compute a value $y_0$ that satisfies the explicit equation of the DAE system with $y_0 \in \mathscr{M}_0(t_0)$.
- Correct this value in order to fulfill the hidden constraints, i.e. calculate a value $y_* \in \mathscr{M}_1(t_0) \subset \mathscr{M}_0(t_0)$.

To describe the hidden constraints we follow the idea in [3, 4, 7] and reduce the index of the DAE system. For this reason, we introduce the DAE

$$(A \quad W_1) \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} d(y,t) \\ W_1 b(y,t) \end{pmatrix} + (I - W_1) b(y,t) = 0 \tag{5}$$

where $W_1$ is a projector along $\operatorname{im} G_1(y,t)$. It holds $W_1 W_0 = W_1$ for any projector $W_0$ along $\operatorname{im} G_0(y,t) = \operatorname{im} A$. The DAE (5) has been obtained by replacing $W_1 b(y,t)$ in the original DAE (3) by its differentiated form. The DAE (5) has also a properly stated leading term and index one. It is clear that every solution of (3) is also a solution of (5). Conversely, every solution $y$ of (5) that satisfies $W_1 b(y(t),t) = 0$ at least at one point $t \in [t_0, t_F]$, is also a solution of original DAE (3).

This approach suggests that the solution $y(t)$ of the index-two DAE should satisfy $y(t) \in \mathscr{M}_1(t)$, $\forall t \in [t_0, t_F]$ with

$$\mathscr{M}_1(t) = \left\{ y \in \mathscr{M}_0(t) \mid \exists z \in \mathbb{R}^m : W_1 \left( \frac{\partial b(y,t)}{\partial y} z + \frac{\partial b(y,t)}{\partial t} \right) = 0 \right\}.$$

The next step is to compute a value $y_0 \in \mathscr{M}_0(t_0)$. This can be done by solving the system $b(y_0, t_0) = 0$. For general DAEs, one can not expect that this system is always solvable. However, if the circuit-device system is well posed then the system has a unique equilibrium solution for physical reasons.

For DAE systems of the form (3) it has been shown [7] that the index-two components[5] of the solution can be described by $Ty$ with

$$T = \begin{pmatrix} Q_{CRVS} & 0 & 0 & 0 \\ 0 & 0_{n_L} & 0 & 0 \\ 0 & 0 & Q_{C-VS} & 0 \\ 0 & 0 & 0 & 0_{n_T - 1 + 3M} \end{pmatrix},$$

if $Q_{CRVS}$ denotes a projector onto $\ker (A_C\, A_R\, A_V\, A_S)^{\mathrm{T}}$, $Q_{C-VS}$ a projector onto $\ker Q_C^{\mathrm{T}} (A_V\, A_S)$ and $Q_C$ is a projector onto $\ker A_C^{\mathrm{T}}$. In (3) these components occur only linearly, i.e. with $U = I - T$ it can be written as

$$A \frac{\mathrm{d}}{\mathrm{d}t} d(y,t) + \tilde{b}(Uy,t) + \mathscr{B}Ty = 0.$$

**Theorem 2.** *If the DAE (3) satisfies the conditions in section 2 and the circuit contains either CVS-loops or LI-cut sets, then the following linear system provides a consistent initialization* $(z_*, y_*)$*:*

$$\begin{pmatrix} A & \mathscr{B}T \\ 0 & U \\ W_1 B_0 D_0^- & 0 \\ (I - R) & 0 \end{pmatrix} \begin{pmatrix} z^* \\ y^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ W_1 B_0 D_0^- d_0 - W_1 b_0 \\ 0 \end{pmatrix}$$

*with* $z_* := z^*$, $y_* := y^* + y_0$, $y_0$ *being any value belonging to* $\mathscr{M}_0(t_0)$ *and*

$$B_0 := \frac{\partial b(Uy_0, t_0)}{\partial Uy}, \quad D_0 := \frac{\partial d(Uy_0, t_0)}{\partial Uy}, \quad b_0 := \frac{\partial b(Uy_0, t_0)}{\partial t}, \quad d_0 := \frac{\partial d(Uy_0, t_0)}{\partial t}.$$

*The matrix* $D_0^-$ *denotes a generalized inverse of* $D_0$ *satisfying* $D_0 D_0^- = R$.

*Proof.* It is shown [7] that this linear system is uniquely solvable. In order to prove that $(z_*, y_*)$ is a consistent initialization for (3), we have to check that the hidden constraints are fulfilled. With $z = D_0^- z_* - D_0^- d_0$ it holds that

[5] By index-two components we mean those components that depend on derivatives of the input functions.

$$W_1\left(\frac{\partial b(y_0,t_0)}{\partial y}z+\frac{\partial b(y_0,t_0)}{\partial t}\right)=0$$

The explicit equations are fulfilled with $Uy^* = 0, Az_* + \mathscr{B}Ty_* = 0$ and $b(y_0,t) = 0$.

## 5 Conclusions

It has been shown that the differential-algebraic equations arising from a monolithic coupled circuit device simulation have the special structure that the higher index components, i.e. here the index-two components, appear only linearly in the systems. This result extends the knowledge from circuit simulation [3] to the coupled circuit device simulation.

Uniquely solvable equation systems have been presented that allow a computation of consistent initial values. Starting from an initial solution that satisfies the index-1 constraint, e.g. from an operating point, only a linear system has to be solved in order to get a consistent initialization.

As known already for circuits [3], the special structure implies that two Euler integration steps yield always a consistent value. However, this value is a consistent one at the timepoint $t_0 + 2h$ supposed the system is integrated by a stepsize $h$. The systems presented here provide a consistent initialization at the initial time point $t_0$.

Following our approach, a non-linear system of equations must be solved in order to obtain an initial solution that satisfies the index-1 constraints, e.g. an operating point. By solving then a linear system of equations, a consistent initialization at the time point $t_0$ is obtained. However, in order to construct the linear system of equations constant projectors must be computed.

## References

1. Alì, G., Bartel, A., Günther, M., Tischendorf, C.: Elliptic partial differential-algebraic multiphysics models in electrical network design. Math. Models Meth. Appl. Sci. **13**(9), 1261–1278 (2003)
2. Tischendorf, C.: Coupled systems of differential algebraic and partial differential equations in circuit and device simulation. Modeling and numerical analysis (2004). Habilitation thesis at Humboldt Univ. of Berlin
3. Estévez Schwarz, D.: Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation. Ph.D. thesis, Humboldt-Univ. Berlin (2000)
4. Estévez Schwarz, D., Lamour, R.: The computation of consistent initial values for nonlinear index 2 daes. Numerical Algorithms **26**(1), 49–75 (2001)
5. Soto, M.S., Tischendorf, C.: Numerical analysis of daes from coupled circuit and semiconductor simulation. Appl. Numer. Math. **53**(2-4), 471–488 (2005)
6. März, R.: Differential-algebraic systems with properly stated leading term and MNA equations. In: K. Anstreich, R. Bulirsch, A. Gilg, P. Rentrop (eds.) Modelling, Simulation and Optimization of Integrated Circuits, pp. 135–151. Birkhäuser (2003)
7. Baumanns, S.: Consitent initialization of differential algebraic equations from coupled circuit and device simulation (in german). Master's thesis, University of Cologne, Math. Institute (2008)

# Hyperbolic PDAEs for Semiconductor Devices Coupled with Circuits

Giuseppe Alì, Giovanni Mascali, and Roland Pulch

**Abstract** We address the problem of coupling a system of network equations corresponding to an electric circuit with a detailed model for a device connected to the circuit. The device is modeled by an hydrodynamic model based on the maximum entropy principle, which results in a hyperbolic system of partial differential equations. We perform a numerical simulation with an oscillator coupled with an $n^+$-$n$-$n^+$ channel.

## 1 Introduction

Traditionally, in microelectronics, an integrated circuit is described as a network of lumped components, neglecting any secondary effect between the components and with the substrate. If some secondary effect becomes important under given operating conditions, it may be included by simply modifying the network. This scheme is severely undermined by the increasing miniaturization of integrated circuits, and by the transition of microelectronics to nanoelectronics.

For this reason, the development of new models for the nonlinear components of an integrated circuit, and for their coupling to an electric network is a mandatory task for semiconductor industry. A lot of theoretical and applied work has been done in this direction. In particular, the coupling between electric networks and semiconductor devices, modeled by means of partial differential equations, has been addressed in a series of paper. This coupling leads to systems of partial-differential-algebraic

Giuseppe Alì, Giovanni Mascali
Dipartimento di Matematica, Università della Calabria, and INFN-Gruppo c. Cosenza, via P. Bucci 30/B, 87036 Cosenza, Italy, e-mail: giuseppe.ali@unical.it, g.mascali@unical.it

Roland Pulch
Chair of Applied Mathematics and Numerical Analysis, Department of Mathematics and Sciences, Bergische Universität Wuppertal, Wuppertal, Germany, e-mail: pulch@math.uni-wuppertal.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_39,

equations (or PDAEs, for short). So far, the elliptic case [1] and the parabolic-elliptic case [2] have been addressed.

In this paper we present a first study on hyperbolic PDAEs occurring in network-device coupling. A similar coupling had been studied in [4], between two circuits connected by a lossy transmission line, described by a (linear) telegrapher equation. Here we consider a nonlinear hyperbolic model for semiconductor devices, based on the maximum entropy principle.

In the following section, we set up the coupled system of the network equations corresponding to an electric circuit and the partial differential equations modelling a device connected to the circuit.

## 2 Modelling of Electric Networks with Devices

### *2.1 MNA Equations for Electric Networks*

Modified nodal analysis (MNA) represents a common technique for achieving a mathematical model of integrated circuits [3]. In this approach, an integrated circuit is modeled by an RLC network, that is by a directed graph with $n_v$ nodes and $n_a$ branches, whose branches contain resistances, capacitors and inductors, labelled with the letters $R$, $C$, $L$, respectively. The network will also contain branches with current sources ($I$) and branches with voltage sources ($V$). Here, we assume that the network contains branches with 2-contact semiconductor devices, labelled by $D$. To each node, is attached a potential, $\mathbf{u} \in \mathbb{R}^{n_v}$, and to each branch a current, $\mathbf{i} \in \mathbb{R}^{n_a}$.

By using constitutive relations for the $R$, $C$, $L$ components, and Kirchhoff current law, this network approach yields a system of differential algebraic equations (DAEs) of the form

$$\begin{aligned}
\mathbf{0} &= \mathbf{A}_C \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}_C(\mathbf{A}_C^\top \mathbf{u}(t),t) + \mathbf{A}_R \phi_R(\mathbf{A}_R^\top \mathbf{u}(t),t) + \mathbf{A}_L \mathbf{i}_L(t) \\
&\quad + \mathbf{A}_V \mathbf{i}_V(t) + \mathbf{A}_I \mathbf{I}(t) + \mathbf{A}_D \mathbf{i}_D(t), \\
\mathbf{0} &= \frac{\mathrm{d}}{\mathrm{d}t}\phi_L(\mathbf{i}_L(t),t) - \mathbf{A}_L^\top \mathbf{u}(t), \\
\mathbf{0} &= \mathbf{A}_V^\top \mathbf{u}(t) - \mathbf{V}(t).
\end{aligned} \tag{1}$$

Table 1 illustrates the meaning of the involved variables. Here, nonlinear constitutive relations for the device currents $\mathbf{i}_D$ will be provided by the distributed device model described later. The unknown functions are the state variables $\mathbf{y} := (\mathbf{u}, \mathbf{i}_L, \mathbf{i}_V)^\top$. The currents $\mathbf{i}_D$ represent coupling variables, since they leave devices and enter the electric network (or vice versa).

Considering system (1), consistent initial values

$$\mathbf{y}(t_0) = (\mathbf{u}(t_0), \mathbf{i}_L(t_0), \mathbf{i}_V(t_0))^\top = (\mathbf{u}_0, \mathbf{i}_{L,0}, \mathbf{i}_{V,0})^\top \tag{2}$$

have to be specified at some initial time $t_0$.

**Table 1:** Variables in circuit equations

| | | | | | |
|---|---|---|---|---|---|
| $\mathbf{u}$ | node voltages | $\mathbf{q}_C$ | charge term (capacitances) | $\mathbf{I}$ | current sources |
| $\mathbf{i}_L$ | currents through inductances | $\phi_R$ | functions for resistances | $\mathbf{V}$ | voltage sources |
| $\mathbf{i}_V$ | currents through voltage sources | $\phi_L$ | flux term (inductances) | $\mathbf{A}_X$ | incidence matrices |
| $\mathbf{i}_D$ | currents through devices | | | | |

## 2.2 MEP-Based Hydrodynamical Models for Devices

We describe the semiconductor devices contained in the $D$-branches by means of a hydrodynamical model for semiconductors, based on the maximum entropy principle (MEP) [5]. For simplicity, we consider unipolar, one-dimensional devices. Each of the $n_D$ devices will be modeled by an interval $[0,\ell_i]$, $i = 1,\ldots,n_D$, in which a family of charge carriers lives. In a compact way (neglecting the index $i$ labeling the device), the system of equations which we will consider can be written in the form:

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = \mathbf{G}(\mathbf{U},E). \tag{3}$$

Equation (3) can be formally obtained from the semiclassical Boltzmann Transport Equation (BTE) for semiconductors. The vector $\mathbf{U}(x,t)$ collects a group of macroscopic variables corresponding to the moments of the carrier distribution function. Several choices are possible for the moments, depending on the choice of the weight functions. In the following section we will make a definite choice. The vectors $\mathbf{F}$, $\mathbf{G}$, respectively, correspond to the moment fluxes, and to the moments, with respect to the same weight functions, of the carrier collision term and driving term due to the electric field $E$, which appear in the BTE. According to this generic definition, the fluxes and production terms are not defined as functions of the moments $\mathbf{U}$. A systematic way of obtaining closure relations is provided by the maximum entropy principle. According to the MEP, the variables corresponding to quantities not directly related to the chosen moments $\mathbf{U}$, can be evaluated by using the maximum entropy distribution function, that is, the distribution function which: 1) preserves the chosen moments; 2) maximizes the physical entropy of the system. Application of MEP leads to a definite form of the functions $\mathbf{F}(\mathbf{U})$, $\mathbf{G}(\mathbf{U},E)$, once the moments and the desired order of accuracy have been fixed.

Equation (3) is coupled with a Poisson equation for the electric potential $\phi$,

$$\frac{\partial}{\partial x}(\varepsilon E) = \rho_{\mathrm{bi}} + \rho(\mathbf{U}), \tag{4}$$

where $E = -\partial\phi/\partial x$ is the electric field, $\varepsilon$ is the dielectric constant, $\rho_{\mathrm{bi}}$ the built-in charge density, $\rho(\mathbf{U}) = qn$, with $q$ charge of a carrier, and $n$ carrier number density — i.e., the moment with respect to the weight function 1.

System (3)-(4) has to be supplemented with appropriate initial-boundary data.

## 2.3 Coupling Conditions

The coupling between the electric network and the device is done by corresponding node voltages and branch currents. For simplicity we consider only one device.

The electric network affects the device by means of the applied potential $u_D$, which is related to some components of $\mathbf{u}$. The exact relation can be expressed by means of the incidence matrix $\mathbf{A}_D$:

$$u_D(0,t) = \mathbf{A}_D^\top \mathbf{u}, \quad u_D(\ell,t) = 0. \tag{5}$$

Note that we have fixed the external potential of one of the contacts (at $x = \ell$) as zero potential for the device, which means that the potential at the other contact is measured with respect to that at the first contact. This choice does not affect the expression of the current because, as we will see, the latter depends only on the electric field, which is the space derivative of the potential and does not sense a time-dependent additive function.

The device affects the electric network by means of the current. We denote by $j$ the carrier number density flux. Then, the charge density conservation law, which is compatible with (3), holds:

$$\frac{\partial n}{\partial t} + \frac{\partial j}{\partial x} = 0. \tag{6}$$

Combining this equation with Poisson equation, we get

$$\frac{\partial J}{\partial x} := \frac{\partial}{\partial x}\left(A\varepsilon\frac{\partial E}{\partial t} + Aqj\right) = 0, \tag{7}$$

where we have introduced the cross-sectional area $A$. Thus, we can identify the total current transmitted to the electric network as $J$, which is the sum of the displacement current and the carrier current. Thanks to (7), $J$ is conserved through the device and can be identified as the current through the branch with the device,

$$i_D(t) = J(0,t) \equiv J(x,t). \tag{8}$$

This scalar current is replaced by a vector comprising the currents through all branches with devices. The coupling term $\mathbf{A}_D\mathbf{i}_D$, appearing in (1), can be written in a more convenient way by using the following decomposition for the potential:

$$\phi = \hat{\phi}_{\mathrm{bi}} + \psi\mathbf{A}_D^\top\mathbf{u} + \hat{\phi}(n),$$

with

$$\begin{cases} -\frac{\partial}{\partial x}\left(\varepsilon\frac{\partial}{\partial x}\hat{\phi}_{\mathrm{bi}}\right) = \rho_{\mathrm{bi}}, \\ \hat{\phi}_{\mathrm{bi}}(0) = \phi_{\mathrm{bi}}(0),\ \hat{\phi}_{\mathrm{bi}}(\ell) = \phi_{\mathrm{bi}}(\ell), \end{cases} \quad \begin{cases} -\frac{\partial}{\partial x}\left(\varepsilon\frac{\partial}{\partial x}\psi\right) = 0, \\ \psi(0) = 1,\ \psi(\ell) = 0, \end{cases} \quad \begin{cases} -\frac{\partial}{\partial x}\left(\varepsilon\frac{\partial}{\partial x}\hat{\phi}\right) = \rho, \\ \hat{\phi}(0) = \hat{\phi}(\ell) = 0, \end{cases}$$

where $\phi_{\mathrm{bi}}$ is the built-in potential. Noting that $\hat{\phi}_{\mathrm{bi}}$ is time independent, solving the last two Poisson equation and substituting the results into the expression of $i_D(t)$, one finds

$$\mathbf{A}_D\mathbf{i}_D = \mathbf{A}_D C_D \mathbf{A}_D^\top \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} + \mathbf{A}_D C_D \int_0^\ell \varepsilon^{-1} q j \,\mathrm{d}x, \tag{9}$$

with $C_D = -A\varepsilon\frac{\partial\psi}{\partial x} = A\left(\int_0^\ell \varepsilon^{-1}\,\mathrm{d}x\right)^{-1}$. Thus, the contribution of the device to the network can be split into an additional capacitance term, in parallel to the device, plus a term proportional to weighted integral average of the carrier current. If the dielectric constant does not depend on $x$, the expression for the device capacitance simplifies as $C_D = A\varepsilon/\ell$, and the weighted average reduces to the average.

## 3 Test Case

To test the presented model, we consider a one-dimensional device of type $n^+$-$n$-$n^+$, coupled with a simple electric network, as shown in Figure 1. The device that we have in mind is a MOSFET, whose gate channel is well modelled by a one-dimensional MEP-based hydrodynamical model. The electric network is a simple oscillator, containing: a resistance, with resistivity $R = 25\,\mathrm{ohm}$, a capacitor, with

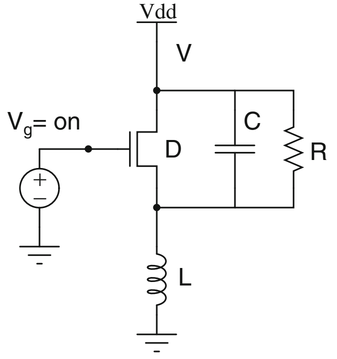


**Fig. 1:** Schematics of the coupled system

capacitance $C = 0.82 \times 10^{-12}\,\mathrm{F}$, an inductor, with inductance $L = 3.5 \times 10^{-12}$ henry, a voltage source (battery), with bias $V = 1\,\mathrm{V}$.

The equations of the oscillator can be rearranged and split in a differential part and an algebraic part:

$$\begin{cases} C\dfrac{\mathrm{d}v_D}{\mathrm{d}t} - i_L + i_D - \dfrac{1}{R}v_D = 0, \\ L\dfrac{\mathrm{d}i_L}{\mathrm{d}t} + v_D = V, \end{cases} \qquad \begin{cases} v_L = V - v_D, \\ i_V = -i_L. \end{cases} \tag{10}$$

where $v_D$ is the voltage drop through the device. For the circuit variables, we assign the initial conditions

$$v_D(0) = V \quad i_L(0) = 0.$$

The oscillator is connected to a 1-dimensional device of $n^+$-$n$-$n^+$ type, with length $\ell = 0.4\,\mu\mathrm{m}$, cross-sectional area $A = 1.0 \times 10^{-5}\,\mathrm{cm}^2$, and doping profile

$$N_+(x) = \begin{cases} 10^{18} & \text{for} \quad x < 0.1\,\mu\mathrm{m}, \\ 10^{16} & \text{for} \quad 0.1\,\mu\mathrm{m} < x < 0.3\,\mu\mathrm{m}, \\ 10^{18} & \text{for} \quad x > 0.3\,\mu\mathrm{m}, \end{cases} \quad (\text{donors/cm}^3).$$

We model the device by means of a MEP-based hydrodynamical model, for the variables $(n, u, W, S)$, where $\mathbf{U} = (n, nu, nW, nS)$ correspond to the moments relative to the weight functions $(1, v_c, \mathscr{E}_c, \mathscr{E}_c v_c)$, with $\mathscr{E}_c$, $v_c$ dispersion relation and group velocity of the carriers. More precisely, the equations we solve numerically, can be formally derived by using the weight functions $(1, \hbar k, \mathscr{E}_c, \mathscr{E}_c v_c)$, with $k$ wavenumber vector, and adopting Kane's approximation for the dispersion relation, which implies $\hbar k = m^*(1 + 2\alpha\mathscr{E}_c)v_c$. In this way, we obtain system (3), with

$$\mathbf{U} = \begin{pmatrix} n \\ nu \\ nW \\ nS \end{pmatrix}, \qquad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} nu \\ n(\frac{1}{m^*}G_1 - 2\alpha G_2) \\ nS \\ nG_2 \end{pmatrix}, \tag{11}$$

$$\mathbf{G}(\mathbf{U}, E) = \begin{pmatrix} 0 \\ d_{11}(W)nu + d_{12}(W)nS + q\left(2\alpha G_3 - \frac{1}{m^*}\right)nE \\ C_W(W) - qnuE \\ d_{21}(W)nu + d_{22}(W)nS - qnG_3E \end{pmatrix}.$$

The meaning of the main variables is explained in Table 2. The involved fluxes and production terms can be expressed as functions of the unknowns by exploiting the maximum entropy principle [5]. The functions $d_{ij}$ depend on the energy and appear in the expressions of the production terms for the average velocity and the average energy flux. The evaluations of $d_{ij}$ and of the energy production term $C_W$ are performed by tabulated values [5].

The hyperbolic system (3), (11) is coupled with the Poisson equation for the electric potential (4), which we write in the form:

$$-\varepsilon \frac{\partial^2 \phi}{\partial x^2} = q_e(N_+(x) - n), \quad E = -\frac{\partial \phi}{\partial x}. \tag{12}$$

Hence the complete system of PDEs consists of hyperbolic equations coupled with an elliptic equation.

We consider the following equilibrium initial value data:

$$n(x,0) = N_+(x), \qquad W(x,0) = W^{\text{eq}}, \qquad u(x,0) = S(x,0) = 0, \tag{13}$$

and boundary conditions [5]:

$$\begin{aligned}
&n(0,t) = N_+(0), \quad n(\ell,t) = N_+(\ell), \quad W(0,t) = W(\ell,t) = W^{\text{eq}}, \\
&\phi(0,t) = \phi_{\text{bi}}(0) - v_D(t), \quad \phi(\ell,t) = \phi_{\text{bi}}(\ell), \\
&\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(\ell,t) = 0, \quad \frac{\partial S}{\partial x}(0,t) = \frac{\partial S}{\partial x}(\ell,t) = 0.
\end{aligned}$$

Finally, the current through the device can be computed as

$$i_D = \frac{\varepsilon A}{\ell}\frac{\mathrm{d}v_D}{\mathrm{d}t} - \frac{q_e A}{\ell}\int_0^\ell nu\,\mathrm{d}x,$$

establishing the final coupling with (10).

**Table 2:** Variables related to PDE system

| | | | | | |
|---|---|---|---|---|---|
| $n$ | number density | $G_i$ | fluxes | $\hbar$ | reduced Planck constant |
| $u$ | average velocity | $W^{\text{eq}}$ | equilibrium energy | $\alpha$ | non-parabolicity factor |
| $W$ | average energy | $\varepsilon$ | dielectric constant | $m^*$ | effective electron mass |
| $S$ | energy flux | $q_e$ | unit charge (in absolute value) | $\ell$ | length of the device |
| $\phi$ | electric potential | $N_+$ | donor concentration | $A$ | device cross-sectional area |
| $E$ | electric field | | | | |

No analytical solution to the Riemann problem for the device model under investigation is available at the present time, therefore an approach based on the full numerical evaluation of the Roe matrix is not practical. We have to resort to an extension [6] of the traditional central differencing schemes to one-dimensional balance laws with (possibly *stiff*) source terms, which has been developed on the basis of the Nessyhau and Tadmor scheme [7] for homogeneous hyperbolic systems.

The complete method is based on a second-order splitting technique which separately solves the system with the source put equal to zero (convection step) and the system with the flux put equal to zero (relaxation step). At each time step the bias applied to the device is determined by solving the circuit equations, with $i_D$ fixed at the previous time step, by using a 4-th order Runge-Kutta method. The results for the device voltage and current are represented in Fig. 2.

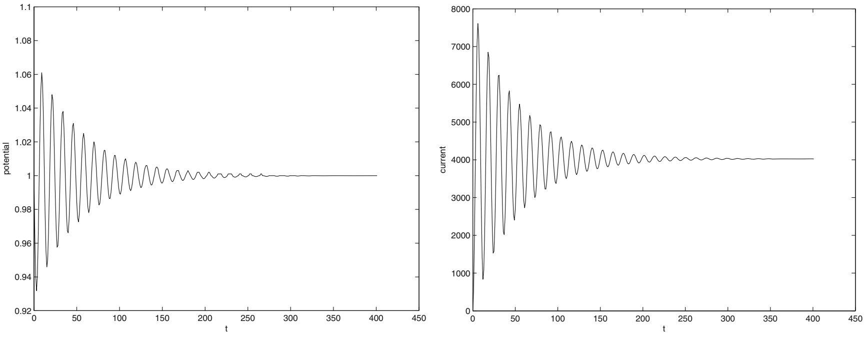


**Fig. 2:** *On the left*: potential (V) vs time (ps). *On the right*: current (A) vs time (ps)

**Acknowledgements** The authors acknowledge the financial support by P.R.A. University of Calabria (ex 60 %) and by the EU Marie Curie RTN project COMSON grant n. MRTN-CT-2005-019417.

## References

1. Alì, G., Bartel, A., Günther, M., Tischendorf, C.: Elliptic partial differential-algebraic multiphysics models in electrical network design, M3AS, **13**, 1261–1278 (2003)
2. Alì, G., Bartel, A., Günther, M.: Parabolic differential-algebraic models in electrical network design, SIAM Multiscale Model. Simul., **4**, 813-838 (2005)
3. Günther, M.; Feldmann, U.; ter Maten, E.J.W.: Modeling and discretization of circuit problems. in: Schilders, W.H.A.; ter Maten, E.J.W. (eds.): Handbook of Numerical Analysis. Special Volume Numerical Analysis of Electromagnetism. Elsevier North Holland, Amsterdam, 2005, 523–659.
4. Günther, M.: A PDAE model for interconnected linear RLC networks, Mathematical and Computer Modelling of Dynamical Systems, **7**, 189–203 (2001).
5. Anile, A.M., Mascali, G., Romano, V.: Recent developments in Hydrodynamical Modeling of semiconductors. In : Anile A. M. , Allegretto, W. Ringhofer, C. (ed.) Mathematical Problems in Semiconductor Physics, Lecture Notes in Mathematics n. 1823, pp. 1-56, Springer, Berlin (2003)
6. F. Liotta, V. Romano and G. Russo, *Central schemes for balance laws of relaxation type*, SIAM J. Num. Analysis 38 (2000), pp. 1337–1356.
7. H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation law*, J. Comp. Physics 87 (1990), pp. 408–463.

# Large-Scale Atomistic Circuit-Device Coupled Simulation of Discrete-Dopant-Induced Characteristic Fluctuation in Nano-CMOS Digital Circuits

Yiming Li and Chih-Hong Hwang

**Abstract** The increasing characteristics variability in nano-CMOS devices becomes a major challenge to scaling and integration. In this work, a large-scale statistically sound "atomistic" circuit-device coupled simulation methodology is presented to explore the discrete-dopant-induced characteristic fluctuations in nano-CMOS digital circuits. According to the simulation scenario, the discrete-dopant-induced characteristic fluctuations are examined for a 16-nm-gate MOSFET and inverter circuit. The fluctuations of the intrinsic current-voltage and capacitance-voltage characteristics, and timing behaviors for the explored device and circuit are estimated. The timing fluctuation may result in a significant signal delay in the digital circuit. Consequently, links should be established between circuit design and fundamental device technology to allow circuits and systems to accommodate the individual behavior of every transistor on a silicon chip. The proposed simulation approach could be extended to outlook the fluctuations in various digital and analog circuits.

## 1 Introduction

Yield analysis and optimization, which take the manufacturing tolerances, model uncertainties, variations in the process parameters, etc, into account, have been known as indispensable components of the circuit design methodology[1]. Various randomness effects resulting from the random nature of manufacturing process, such as ion implantation, diffusion, and thermal annealing, have induced significant fluctuations of electrical characteristics in nano-MOSFETs. The number of dopants is of the order of tens in the depletion region of a MOSFET, whose influence on device characteristic is large enough to be distinct[2]. Diverse approaches have recently been reported to study fluctuation-related issues in semiconductor devices and circuits [2–7]. However, the attention is less drawn on the existence of timing characteristic fluctuations of an active device due to random dopant placement.

Yiming Li, Chih-Hong Hwang
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu 300, Taiwan, e-mail: ymli@faculty.nctu.edu.tw

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_40,

Moreover, due to the randomness of the dopant position in the device, the fluctuation of the device's gate capacitance is hard to be modeled in the current compact models [7]. Therefore, in this study, we propose a large-scale statistically sound circuit-device coupled simulation approach to analyze the random dopant effect in nano-CMOS circuit, concurrently capturing the discrete-dopant-number- and discrete-dopant-position-induced fluctuations. Based on the statistically generated large-scale doping profiles, the device simulation is performed by solving a set of three-dimensional (3D) drift-diffusion equations with quantum corrections by the density gradient method [8, 9] on a parallel computing system [10, 11]. In the estimation of the circuit-level characteristics fluctuations, to capture the nonlinearity of gate capacitance fluctuation, the aforementioned device equations are coupled with the circuit nodal equations of the studied circuit and solve simultaneously. The proposed simulation approach can outlook the fluctuations in circuit characteristics and benefit the development of next generation nanoelectronic circuits and systems.

The paper is organized as follows. In Sec. 2, we introduce the large-scale statistically sound "atomistic" simulation approach and simulation techniques for studying the random dopant effect in nanoscale device and circuit. In Sec. 3, we investigate the discrete-dopant-induced device and circuit characteristic fluctuations in the 16-nm-gate CMOS circuit. Finally, we draw conclusions and suggest future work.

## 2 Simulation Technique

Figure 1 shows the simulation flow for the proposed approach. To consider the effect of random fluctuation of the number and location of discrete dopants in the channel region, 758 dopants are randomly generated in a (80 nm)$^3$ cube, in which the equivalent doping concentration is $1.48\times10^{18}$ cm$^{-3}$ (the nominal channel doping concentration), as shown in Fig. 1(a). The cube of (80 nm)$^3$ is then partitioned into sub-cubes of (16 nm)$^3$. The number of dopants in the sub-cubes may vary from zero to 14, and the average number is six, as shown in Figs. 1(b), 1(c), and 1(f). Then the coordinates of discrete dopants in these sub-cubes are equivalently mapped into the corresponding coordinates in device channel region for the 3D device simulation, as shown in Fig. 1(d). The device simulation is performed by solving a set of 3D drift-diffusion equations with density gradient quantum correction [8, 9]. The step function is used to include the discrete dopant effect into the source of the Poisson equation. The step function $H(x,y,z) = 1$, for $x \geq 0$, $y \geq 0$, and $z \geq 0$; $H(x,y,z) = 0$, otherwise. The effect of the discrete dopant is considered by including the following term into the source doping concentration of the Poisson equation

$$N_A = \sum_{i=0}^{k} N_A^{dopant} \left( H(x-x_l, y-y_l, z-z_l) - H(x-x_u, y-y_u, z-z_u) \right), \qquad (1)$$

$k$ is numbers of dopant in the device channel. $N_A^{dopant}$ is the associated doping concentration for a dopant within a box. The volume of the box is defined by two coordinates, the lower point ($x_l$, $y_l$, $z_l$) and the upper point ($x_u$, $y_u$, and $z_u$). To calculate the numerical solution of the 3D device transport equations, we first decouple

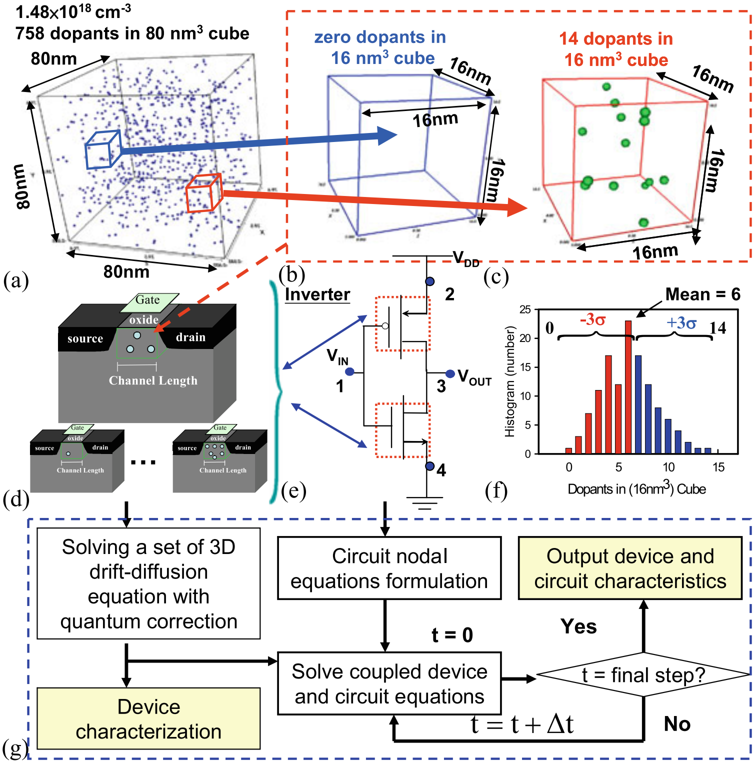

1.48×10^18 cm^-3
758 dopants in 80 nm^3 cube
80nm
80nm
80nm
zero dopants in 16 nm^3 cube
16nm
16nm
16nm
14 dopants in 16 nm^3 cube
16nm
16nm
16nm
(a)
(b)
(c)
Gate
oxide
source
drain
Channel Length
(d)
Inverter
V_DD
2
V_IN
1
V_OUT
3
4
(e)
Mean = 6
0
-3σ
+3σ
14
Histogram (number)
Dopants in (16nm^3) Cube
(f)
Solving a set of 3D drift-diffusion equation with quantum correction
Circuit nodal equations formulation
Output device and circuit characteristics
Device characterization
t = 0
Yes
Solve coupled device and circuit equations
t = final step?
t = t + Δt
No
(g)

**Fig. 1: a** Discrete dopants randomly distributed in the (80 nm)$^3$ cube with the average concentration $1.48\times10^{18}$ cm$^{-3}$. The dopants in sub-cubes of (16 nm)$^3$ may vary from zero to 14 (the average number is six), [(**b**), (**c**), and (**f**)]. **d** The sub-cubes are then mapped into 3D device channel region for device characterization. **e** A CMOS inverter for the analysis of circuit characteristic fluctuations, where the upper device is the P-MOSFET and the lower one is the N-MOSFET. **g** The simulation flow for device and circuit-device coupled simulations

the coupled partial differential equations by the Gummels decoupling method. The device transport equations are approximated by the finite volume method over a non-uniform mesh. Then the nonlinear algebraic equations are solved with the monotone iteration method [12] on our parallel computing system [10, 11]. An inverter circuit is then adopted as an example for estimating the circuit characteristics fluctuations, as shown in Fig. 1(e). The circuit nodal equations are then formulated (node1: $V_1 = V_G$, node2: $V_2 = V_{DD}$, node3: $I_{d,P-MOSFET} = I_{d,N-MOSFET}$, node4: $V_4 = 0$, for example). Currently, there is no well-established compact model available for describing the discrete-dopant-induced nonlinear device characteristic fluctuations, instead of using a compact modeling approach, the circuit nodal equations are coupled with device transport equations and solved simultaneously to examine the circuit characteristic fluctuations [13, 14]. The simulation flow for the proposed device and circuit-device coupled simulations is shown in Fig. 1(g). The large-scale simulation technique is statistically sound for random dopant fluctuation characterization.

## 3 Results and Discussion

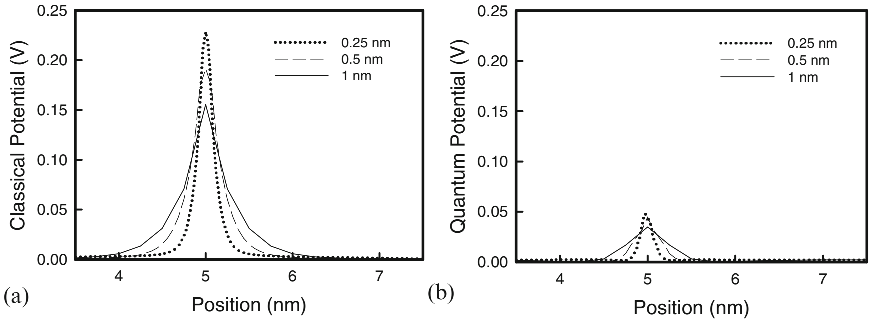


**Fig. 2:** Potential profiles for **a** classical and **b** quantum potential with different mesh size

Figure 2(a) and 2(b) illustrate the mesh size dependence of the classical and quantum mechanical potentials for a single discrete dopant within the silicon channel. In the “atomistic” simulation, the key point to study random impurities induced fluctuation relies on how to introduce the microscopic non-uniformity of localized impurity distributions inside the device. In conventional drift-diffusion approach for a large device size, the number of impurities included in each mesh exceeds one and the equivalent doping concentration does not change abruptly at every mesh node. Also, the dopant density at each mesh node changes gradually and the non-uniformity of impurity arrangement is averaged.

However, for the nanoscale transistor, the corresponding number of impurities is significantly reduced. Most meshes contain no dopant or, at most, one dopant. The dopant density at each mesh node changes its order of magnitude and behaves like a $\delta$-function. The resolution of individual impurities for the conventional drift-diffusion simulation using a fine mesh creates problems of singularities in the Coulomb potential, as shown in Fig. 2(a). The sharp Coulomb potential wells may un-physically trap majority carriers, reduce the mobile electron concentration, modify the depletion region, and alter the threshold voltage ($V_{th}$). Therefore, the density gradient quantum correction [8, 9] is used to handle the discrete dopant effect by properly introducing the related quantum mechanical effects, as plotted in Fig. 2(b). The quantum mechanical potential shows less sensitivity to the mesh size and is quite similar for mesh spacing below 0.5 nm. We notice that the potential barrier of the Coulomb well is about 45 mV, which roughly corresponds to the ground state of a hydrogenic model of an impurity in silicon.

Figure 3(a) shows the $I_D$-$V_G$ characteristics of the discrete-dopant-fluctuated 16-nm-gate planar MOSFETs, where the solid line shows the result of the nominal case ($1.48\times10^{18}$ cm$^{-3}$ continuously doping concentration), and the dashed lines are

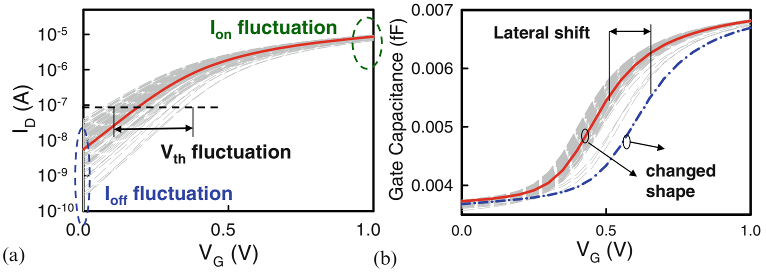


**Fig. 3:** Fluctuations of **a** $I_D$-$V_G$ **b** and C-V curves for the studied 16-nm-gate planar MOSFETs

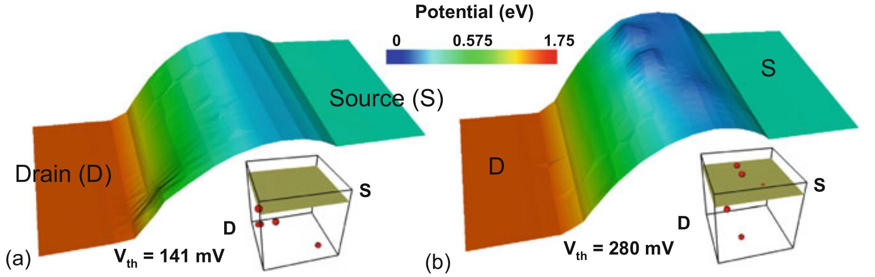


**Fig. 4:** Cutting-plane plots of the off-state potential profiles ($V_G = 0$ V; $V_D = 1$ V) for the simulated device with the same dopant number (six dopants) in channel region but with different $V_{th}$

discrete-dopant-fluctuated devices. The fluctuations of the on- and off-state currents ($I_{on}$ and $I_{off}$) and $V_{th}$ characteristics are observed. The detailed physical mechanism was described somewhere else [3–5]. Figure 3(b) shows the capacitance-voltage (C-V) characteristics of the discrete-dopant-fluctuated 16-nm-gate planar MOSFETs. The lateral shift and the changed shape for the C-V curves are observed. The lateral shift of gate capacitance results from the variation of $V_{th}$ and may be described by the correspond parameters in a compact model. However, the changed shape of the C-V curves result from the position of discrete dopants in the channel, and it is hard to be described by any compact modeling approach [8, 9]. Figure 4 compares the off-state potential profiles for two devices with the same dopant number (six dopants) in the channel region but with different $V_{th}$. The potential barriers in Fig. 4 are induced by the corresponding dopants within the device channel. The different distribution of discrete dopants may induce different potential profiles and thus alter the device's transport characteristics. The $V_{th}$ difference between Figs. 4(a) and 4(b) is about 139 mV, which is 99% of the nominal $V_{th}$, 140 mV. Therefore, instead of using a compact modeling approach, we have to use device simulation, and a circuit and device coupled simulation approach to capture the nonlinear variations induced by the discrete-dopant-position effect.

Figure 5(a) shows the voltage transfer curves for the discrete-dopant-fluctuated 16-nm-gate CMOS inverters. Two points on the voltage transfer curve determine the noise margins of the inverter. These are the maximum permitted logic "0" at the input, $V_{IL}$, and the minimum permitted logic "1" at the input, $V_{IH}$. The two

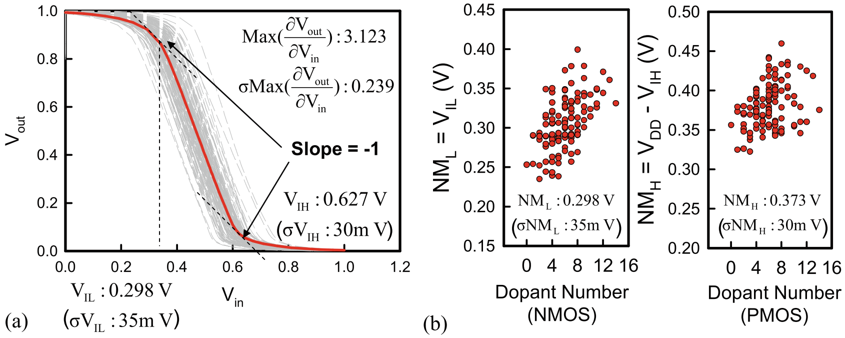
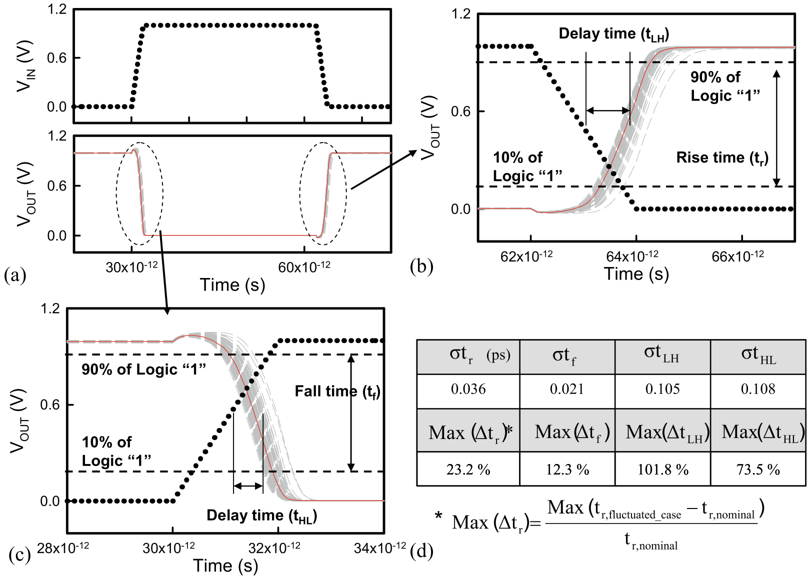


**Fig. 5: a** Voltage transfer curves for the studied 16-nm-gate planar MOSFET circuit. **b** Noise margins, $NM_L$ and $NM_H$, as a function of the dopant number in the N-MOSFET and P-MOSFET

points on the voltage transfer curve are defined as those values of $V_{\mathrm{in}}$ where the incremental gain is unity; the slope -1 V/V. The nominal value and fluctuations of the $V_{IL}$ and $V_{IH}$ are shown in the insets of Fig. 5(a). The $V_{IL}$ fluctuation is larger than the $V_{IH}$ due to the larger $V_{th}$ fluctuation of the N-MOSFET than the P-MOSFET. In the inverter circuit, the maximum slope of the voltage transfer curve implies the maximum voltage gain of the inverter. Therefore, the voltage gain fluctuation of the inverter is estimated, which is about 7% of the nominal value, as shown in the inset of Fig. 5(a). Noise margins for the logic "0" and "1", $NM_H$ and $NM_L$, as a function of the dopant number are plotted in Fig. 5(b), where the $NM_H$ and $NM_L$ are defined. The $NM_L$ is increased with the increasing dopant number in the N-MOSFET due to the increased $V_{th}$ of device. For the $NM_H$, as numbers of dopant in the P-MOSFET increases, the increased $V_{th}$ of device may decrease the $V_{IH}$ of voltage transfer curve and thus increase the $NM_H$. We notice that even for cases with the same number of dopants within device channel, their noise margins are still quite different due to the different distribution of random dopants. The noise margins of the inverter circuit may be increased as dopant number increases; however, the fluctuations of the noise margins are also increased due to the more sources of fluctuation in the device channel region.

Figure 6 shows the timing characteristics of the CMOS inverter. The input and output signals are shown in Fig. 6(a). Figures. 6(b) and 6(c) are the zoom-in plots for the fall and rise transition characteristics of the output signal, where the rise time ($t_r$), the fall time ($t_f$), low-to-high delay time($t_{LH}$) and high-to-low delay time ($t_{HL}$) are defined in the insets. The timing fluctuations are consequently summarized in Fig. 6(d). For the studied inverter circuits, the $t_r$ fluctuation is larger than the $t_f$ fluctuation because of the smaller driving capability of the P-MOSFET than that of the N-MOSFET. The device with the larger driving capability may require less time to charge and discharge the load capacitance and thus exhibits less timing fluctuations. The $t_r$ and $t_f$ fluctuations may not play an important role in timing characteristics; however, their maximum difference are about 23% and 12%, which bring a significant impact on timing. The delay time is dependent on the starting point of the signal transition, for example the time of 90% of the logic "1" for $t_{HL}$ and the time of 10%



| $\sigma t_r$ (ps) | $\sigma t_f$ | $\sigma t_{LH}$ | $\sigma t_{HL}$ |
|---|---|---|---|
| 0.036 | 0.021 | 0.105 | 0.108 |
| Max $(\Delta t_r)$* | Max $(\Delta t_f)$ | Max$(\Delta t_{LH})$ | Max$(\Delta t_{HL})$ |
| 23.2 % | 12.3 % | 101.8 % | 73.5 % |

$$* \ \text{Max}\,(\Delta t_r)=\frac{\text{Max}\,(t_{r,\text{fluctuated\_case}}-t_{r,\text{nominal}})}{t_{r,\text{nominal}}}$$

**Fig. 6: a** Input and output signals for the discrete-dopant-fluctuated 16-nm-gate inverter circuits. **b**, **c** Zoom-in plots for the fall and rise transitions, where the insets define the rise, fall, high-to-low, and low-to-high delay times. **d** Summarized timing characteristic fluctuations

of the logic "1" for $t_{LH}$. Since the time of 90% and 10% of the logic "1" are related to the $V_{th}$ of the N-MOSFET and P-MOSFET, respectively, the $t_{HL}$ fluctuation is larger than the $t_{LH}$ fluctuation due to the larger $V_{th}$ fluctuation of the N-MOSFET. For the fall transition characteristics, the signal falls as the N-MOSFET is turned on. Therefore, as the $V_{th}$ of the N-MOSFET is increased, the starting point of the fall transition is delayed. The $t_{HL}$ is increased as the dopant number of N-MOSFET increases. Moreover, the $t_{HL}$ fluctuation is increased as numbers of dopant increases due to the more sources of fluctuation inside device channel. Similarly, we can infer that the $t_{HL}$ and $t_{HL}$ fluctuation are increased as the dopant number of P-MOSFET increases.

## 4 Conclusions

Statistical variability introduced by discreteness of charge and granularity of matter could not be completely eliminated by advanced process control and already critically affects timing issues in digital logic circuits. In this paper, a large-scale 3D "atomistic" circuit-device coupled simulation approach has been implemented to investigate the discrete-dopant-induced characteristic fluctuations in nano-CMOS digital circuits. The quantum mechanical potential is less sensitive to the mesh size and quite similar for mesh spacing below 0.5 nm. The quantum mechanical potential barrier is about 45 mV, which roughly corresponds to the ground state of a hydrogenic model of an impurity in silicon. According to the proposed simulation scenario, the nonlinearity of device characteristic fluctuations including discrete-dopant-number and discrete-dopant-position effects have been estimated in terms of surface potential, I-V, and C-V curves. For the discrete-dopant fluctuated 16-nm-gate inverter circuit, The maximum difference of $t_f$, $t_r$, $t_{HL}$ and $t_{LH}$ are about 23%, 12%, 101.8% and 73.5%, respectively. The significant timing variation may result in significant timing violation and delay in state-of-art nano-CMOS circuits and systems. The study may benefit the development of next generation nanoscale circuits and systems, where the design paradigms have to change to acclimatize the even increasing variability. Besides the inverter circuits, the simulation approach could be further applied for various digital and analog circuits characteristic fluctuations. Also, the fluctuation suppression techniques could be verified and developed.

**Acknowledgements** This work was supported by Taiwan National Science Council under Contract NSC-97-2221-E-009-154-MY2 and Contract NSC-96-2221-E-009- 210, and by the Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan under a 2007-2008 grant..

## References

1. Li, Q., Zhang, J., Li, W., Yuan J. S., Chen, Y., and Oates, A.S.: RF Circuit Performance Degradation Due to Soft Breakdown and Hot-Carrier Effect in Deep-Submicrometer CMOS Technology. IEEE Trans. Microwave Theory Tech., **49**, 1546–1551 (2001)
2. Wong, H.-S., Taur, Y., and Frank D.J.: Discrete Random Dopant Distribution Effects in Nanometer-Scale MOSFETs. Microelectronics Reliability, **38**, 1447–1456 ( 1999)
3. Li, Y., and Hwang, C.-H.: Discrete-dopant-induced characteristic fluctuations in 16 nm multiple-gate silicon-on-insulator devices. J. Appl. Phy., **8**, 084509 (2007)
4. Li, Y., Yu, S.-M., Hwang, J.-R., and Yang, F.-L.: Discrete Dopant Fluctuated 20nm/15nm-Gate Planar CMOS. IEEE Trans. Electron Device, **55**, 1449–1455 (2008)
5. Li, Y., and Yu, S.-M.: Coupled-Simulation-and-Optimization Approach to Nanodevice Fabrication With Minimization of Electrical Characteristics Fluctuation. IEEE Trans. Semi. Manufacturing, **20**, 432–438 (2007)
6. Mahmoodi, H., Mukhopadhyay, S., and Roy K.: Estimation of Delay Variations Due to Random-Dopant Fluctuations in Nanoscale CMOS Circuits. IEEE Journal of Solid-State Circuits, **40**, 1787–1796 (2005)
7. Brown, A. and Asenov, A.: Capacitance Fluctuations in Bulk MOSFETs Due to Random Discrete Dopants. J. Comp. Elect., **7**, 115–118 (2008)
8. Shimada, T. and Odanaka, S.: A Numerical Method for a Transient Quantum Drift-Diffusion Model Arising in Semiconductor Devices. J. Comp. Elect., **7**, 485–493 (2008)
9. Tang, T.-W., Wang, X.,and Li, Y.: Discretization Scheme for the Density-Gradient Equation and Effect of Boundary Conditions. J. Comp. Elect., **1**, 389–393 (2002)
10. Li, Y., Lu, H.-M., Tang, T.-W., and Sze, S. M.: A Novel Parallel Adaptive Monte Carlo Method for Nonlinear Poisson Equation in Semiconductor Devices. Math. Comp. Simulation, **62**, 413–420 (2003)
11. Li, Y., Sze, S.-M., and Chao, T.-S.: A Practical Implementation of Parallel Dynamic Load Balancing for Adaptive Computing in VLSI Device Simulation. Engineering with Computers, **18**, 124–137 (2002)
12. Li, Y. : A Parallel Monotone Iterative Method for the Numerical Solution of Multidimensional Semiconductor Poisson Equation. Computer Physics Communications, **153**, 359–372 ( 2003)
13. Grasser, T. and Selberherr, S.: Mixed-mode device simulation. Microelectronics Journal, **31**, 873–881 (2000)
14. Li, Y., Huang, J.-Y., and Lee, B.-S.: Effect of Single Grain Boundary Position on Surrounding-Gate Polysilicon Thin Film Transistors. Semiconductor Science and Technology, **23**, 015019 (2008)

# Evaluation of Electromagnetic Coupling Between Microelectronic Device Structures Using Computational Electrodynamics

Wim Schoenmaker*, Peter Meuris, Walter Pflanzl, and Alexander Steinmair

*Invited speaker at the SCEE 2008 conference*

**Abstract** Electromagnetic coupling between devices in an microelectronic layout can become a serious design concern. In this paper, the problem of electromagnetic coupling is addressed from field computational point of view. Approximation schemes are justified by evaluating dimensionless parameters in the set up of the field equations and scale considerations of devices. The discretization scheme is reviewed and a simulation method is presented to compute the S-matrix directly by imposing boundary conditions that map directly to the experimental set up. An example demonstrates the validity of the scheme.

## 1 Introduction

With the use of increasing frequency ranges, electromagnetic coupling becomes a more pronounced design concern because induced electric fields are proportional to the rate of change of the magnetic induction. However, not only the pace of time variations are determining for including electromagnetic coupling but also the problem scale and the intensity of the currents that are responsible for the induced fields must be considered. An overall picture of the scaling arguments is presented in Section 2 which helps to identify the needed steps and inclusion of non-negligible effects. Once we note from scale considerations that electromagnetic coupling terms represent a non-negligible contribution to the full system of equations, we move to the the solution of these equations. In section 3, we review and update the approach

Wim Schoenmaker, Peter Meuris
MAGWEL NV, Martelarenplein 13, 3000 Leuven, Belgium, e-mail: wim.schoenmaker@magwel.com, peter.meuris@magwel.com

Walter Pflanzl, Alexander Steinmair
austriamicrosystems AG, Schloss Premstätten, 8141 Unterpremstätten, Austria, e-mail: walter.pflanzl@austriamicrosystems.com, alexander.steinmair@austriamicrosystems.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_41,

that was proposed some years ago by the first author and co-workers [1–3]. We will refer to this approach as 'computational electrodynamics'.

At several occasions we were inquired if this method is equivalent to the method based on Nedelec's edge elements [4,5]. The main difference is that we do not refer to test functions at all. Our method is more related to finite-integration techniques (FIT) [6,7].

Scale considerations are not the only an issue for deciding if some terms in the full system of Maxwell equations and constitutive laws can be neglected. When discussing the coupling of devices, it is also important to realize that different devices can have intrinsic or geometrical scales that differ orders of magnitude. In such scenarios the coupled problem is most easily split in computational domains. Computational electrodynamics gives, rather straightforwardly, a series of prescriptions for matching the interface conditions of the various domains.

Electromagnetic coupling of microelectronic devices is an RF issue and is most conveniently measured using s-parameters. In section 5, we present our method to compute these matrix elements. In fact, s-parameter extraction is straightforwardly achieved as a post-processing of the results of a computational electrodynamics problem with the appropriate setting of the boundary conditions.

In Section 6 we will present an example of a coupled problem, that we have addressed recently.

## 2 Scaling Rules for the Maxwell Equations

The use of scaling arguments is definitely not new to the field of computing in electromagnetic modeling. Well-known approximations are the so-called EQS (electro-quasi-static) and MQS (magneto- quasi-static) approximations. Approximations can be put in a different perspective by considering the scaling step that is necessary when converting the full set of equations to dimensionless equations before the actual computing can start. For our present argument it suffices to consider insulators and metals only. Diffusive currents in semiconductors can easily be added to the equations. Therefore, we start from the Maxwell equations in which $\mathbf{J}_c$ is the conductive current :

$$\mathbf{J}_{\mathrm{c}} = \sigma\mathbf{E}, \qquad \mathbf{D} = \varepsilon_0\varepsilon_r\mathbf{E}, \qquad \mathbf{H} = \frac{1}{\mu_0\mu_r}\mathbf{B}\,, \tag{1}$$

$$\mathbf{E} = -\nabla V - \frac{\partial\mathbf{A}}{\partial t}, \qquad \mathbf{B} = \nabla\times\mathbf{A}\,. \tag{2}$$

We consider the Maxwell equations in the potential formulation. The Poisson equation is used to solve the scalar field in insulators and semiconducting regions and the current-continuity equation is used in metals to find the scalar potential. The electric system is :

$$\nabla.\left[\varepsilon\left(\nabla V + \mathrm{i}\omega\mathbf{A}\right)\right] + \rho = 0\,, \qquad \nabla.\left[\left(\sigma + \mathrm{i}\omega\varepsilon\right)\left(\nabla V + \mathrm{i}\omega\mathbf{A}\right)\right] = 0\,. \tag{3}$$

The Maxwell-Ampere equation is :

$$\nabla \times \left( \frac{1}{\mu} \nabla \times \mathbf{A} \right) - \; (\sigma + \mathrm{i}\omega\varepsilon)\,(-\nabla V - \mathrm{i}\omega\mathbf{A}) = 0 \,. \tag{4}$$

This system must be completed with a gauge condition

$$\nabla \cdot \mathbf{A} + \mathrm{i}\omega\xi\varepsilon\mu V = 0 \,, \tag{5}$$

where $\xi$ is a parameter that allows us to slide over different gauge conditions. Now let $L$ be the 'natural' length scale of the problem that is considered. For example $L = 1\mu$m. Furthermore, let $T$ be the natural time scale, for example $T = 10^{-9}$ sec. It is possible to reformulate the equations ( 3) and ( 4) in *dimensionless* variables $V$ and $\mathbf{A}$ and the set of equations is controlled by two dimensionless variables, $K$ and $\nu$

$$\nabla . \left[ \varepsilon_r \left( \nabla V + \mathrm{i}\omega\mathbf{A} \right) \right] + \rho = 0 \,, \qquad \nabla . \left[ \left( \sigma + \mathrm{i}\omega\varepsilon_r \right) \left( \nabla V + \mathrm{i}\omega\mathbf{A} \right) \right] = 0 \,, \tag{6}$$

and

$$\nabla \times \left( \frac{1}{\mu_r} \nabla \times \mathbf{A} \right) - \; K\omega^2 \left( \varepsilon_r \; - \mathrm{i} \; \nu \right) \mathbf{A} - \mathrm{i} \; \omega K \left( \varepsilon_r - \mathrm{i}\nu \right) \nabla V = 0 \,. \tag{7}$$

The constants $K = \varepsilon_0\mu_0 L^2/T^2$ and $\nu = \sigma T/\varepsilon_0$. Note that for $\sigma = 10^4$ S/m we obtain $K\nu = 10^{-5}$. This value corresponds to the conductance of an inversion layer in the on-state of a transistor. This number enters into the Maxwell-Ampere equation and suggests that in this scenario the magnetic sector is negligible. For a single transistor finger this is a valid conclusion, but one should be aware that in actual designs many fingers may operate in a parallel mode therefore the value of $K$ could increase since $L$ must be adapted to this situation. Taking into account the presence of the back-end processing, one encounters metallic conductance of $10^7$ S/m, such that magnetic effects are important.

## 3 Discretization

In our earlier work, we presented a discretization method that decided for each variable where on the grid it belongs. It was concluded that the geometrical and physical meaning of variables plays a key role. For instance, a scalar variable, e.g. the Poisson potential, $V$, is a number assigned to each space location and for a computational purpose, its discretized value should be assigned to the nodes of the grid. On the other hand the vector potential $\mathbf{A}$ is a variable of the same character as $\nabla V$ and should therefore be assigned to the links of the computational grid. Geometrical considerations have been an important guide for correctly discretizing Maxwell's equations, as was also elaborated by Bossavit [8, 9].

The conversion of continuous variables to discrete variables on the computation grid also has consequences for the particular discretization route that is followed when implementing discrete versions of the Maxwell equations. Gauss' law is discretized by considering elementary volumes around the nodes of the grid and one next perform an integration of Gauss' law over these volume cells. The flux assigned to each segment of the enclosing surface is assumed to be constant which allows for expressing this (constant) flux in terms of the node variables and link variables. This scheme has been the key to the success of the simulation of the semiconductor devices. The Scharfetter-Gummel formulation of the discretized currents can be set up following the above approach [10]. Since links variables are fundamentally different from node variables, we expect that the discretization of the Maxwell-Ampere equation has to be done taking this geometrical aspect into account. Whereas it was quite 'natural' to regard node variables as a representative of some volume element, in the same way we consider a link variable representing some area element. Thus to each link is associated an area element and in order to discretize the Maxwell-Ampere equation on a grid we now apply Stokes' law to arrive at the discretized equations.

After having obtained a scheme to discretize the Maxwell equations, we proceed with expanding them into a small signal analysis. This means that each variable is written as a time-independent part and an harmonic part

$$X = X_0 + X_1 \mathrm{e}^{\mathrm{i}\omega t} \,. \tag{8}$$

If we apply boundary conditions of a similar form and collect terms independent of $\omega$ and terms proportional to $\mathrm{e}^{\mathrm{i}\omega t}$ and omit terms proportional to $X_1^2$ then we obtain a system of equations for the phasors $X_1$. Of particular interest is the treatment of the spurious modes in the fields. These modes can be eliminated by selecting a 'gauge tree' in the mesh, adding a ghost field to the equation system or apply a projection method while iterating towards the solution. We can also apply a gauge condition and construct discrete operators that resemble the continuous operators as close as possible including having a semi-definite spectrum. Using a two-fold application of Stokes' law, the term $\nabla \times \left( \frac{1}{\mu_r} \nabla \times \mathbf{A} \right)$ appears in the discretized formulation as a collection of closed-loop circulations. By subtracting a discretized version of $\nabla\left(\nabla \cdot \mathbf{A}\right)$ we arrive at an operator that resembles $-\nabla^2 \mathbf{A}$. However, since $\mathbf{A}$ is a vector field, the latter can only have meaning in terms of the foregoing expressions. The discretization of the first term in (4) can be illustrated as shown in Fig. 1. The primary link $PQ$ has a dual area assigned to it. This area is denoted with the links $a$, $b$, $c$ and $d$. The curl-curl operator is realized as a sum of circulations around all primary surfaces that contain this link. The most-left picture of Fig. 1 illustrates this aspect. The subtraction of the grad-div operator is done in two steps : The grad means that both at $P$ and at $Q$ a divergence is evaluated. The center- and right drawing show these divergences. Next, these terms are added with opposite sign.

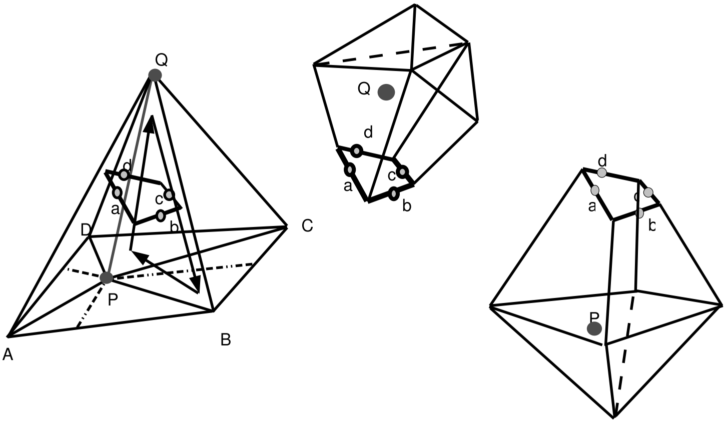


**Fig. 1:** Discretized version of the regularized curl-curl operator acting on a vector field

## 4 The E*V* Solver

Besides scaling and geometrical considerations, another important ingredient for a successful discretization is to avoid unnecessary matrix fill when selecting dynamical variables. In this section, we present a method to reduce the cross coupling between the $V$ and $\mathbf{A}$ system. Let us consider the Ampere-Maxwell equation. For notational convenience we will introduce the notation: $\phi = \sigma + \mathrm{i}\omega\varepsilon_{\mathrm{r}}$. Then we can write (7) as

$$\nabla \times \left( \frac{1}{\mu_{\mathrm{r}}} \nabla \times \mathbf{A} \right) + K\,\phi\;\left(\nabla V + \mathrm{i}\omega\mathbf{A}\right) - K\,\mathbf{J}_{\mathrm{diff}} = 0\,, \tag{9}$$

where $\mathbf{J}_{\mathrm{diff}}$ is the diffusive part of the current. Furthermore, we will need the gauge condition

$$\nabla \cdot \mathbf{A} + \mathrm{i}\,\omega\xi K\,\varepsilon_{\mathrm{r}} V = 0\,, \tag{10}$$

where $\xi$ is the slider between 0 (Coulomb gauge) and 1 (Lorentz gauge).
The crucial observation now is that for any scalar field, the equation $\nabla \times \nabla V = 0$ is valid. This leads to

$$\frac{1}{\mathrm{i}\omega}\,\nabla \times \left( \frac{1}{\mu_{\mathrm{r}}}\,\nabla \times\,\left[\mathrm{i}\omega\,\mathbf{A} + \nabla V\right] \right) + K\,\phi\;\left(\nabla V + \mathrm{i}\omega\mathbf{A}\right) - K\,\mathbf{J}_{\mathrm{diff}} = 0\,. \tag{11}$$

We recognize $\mathrm{i}\omega\mathbf{A} + \nabla V = -\mathbf{E}$ and therefore we find that

$$\nabla \times \left( \frac{1}{\mu_{\mathrm{r}}}\,\nabla \times\,\mathbf{E} \right) + K\,\mathrm{i}\,\omega\,\phi\,\mathbf{E} + K\,\mathrm{i}\,\omega\,\mathbf{J}_{\mathrm{diff}} = 0\,. \tag{12}$$

Of course, this equation could have been straightforwardly obtained from the Maxwell equations by noting that $\mathbf{B} = -1/(\mathrm{i}\omega)\nabla \times \mathbf{E}$. However, here we consider $\mathbf{E}$ as a variable transformation of $\mathbf{A}$. Just as for the $\mathbf{A}$ system, we must regularize the operator $\nabla \times \nabla \times \mathbf{E}$. This is achieved by subtracting the gauge condition. Using

$$\mathbf{A} = \frac{\mathrm{i}}{\omega}\left( \mathbf{E} + \nabla V \right) , \tag{13}$$

we obtain

$$\nabla \cdot \left\{ \frac{\mathrm{i}}{\omega}\left[ \mathbf{E} + \nabla V \right] \right\} + \mathrm{i}\,\omega\, K\xi\ \varepsilon_{\mathrm{r}}\, V = 0\,. \tag{14}$$

This is equivalent to the following expression :

$$\nabla \cdot \mathbf{E} + \nabla^2 V + \omega^2 K\xi\ \varepsilon_{\mathrm{r}}\, V = 0\,. \tag{15}$$

The regularization is now achieved by subtraction of the gradient of equation ( 15) from equation (12).

$$\begin{aligned} &\nabla \times \left( \frac{1}{\mu_{\mathrm{r}}} \nabla \times \mathbf{E} \right) - \nabla \left( \nabla \cdot \mathbf{E} \right) + K\,\mathrm{i}\,\omega\,\phi\,\mathbf{E} \\ &- \nabla \left( \nabla^2 V \right) - \omega^2 K\xi\ \nabla(\varepsilon_{\mathrm{r}}\, V) + K\,\mathrm{i}\,\omega\,\mathbf{J}_{\mathrm{diff}} = 0\,. \end{aligned} \tag{16}$$

As is seen from this equation the coupling to the variables $V$ has a strength of order one and is not growing with $\sigma$. Furthermore it should be noticed that the Poisson equation is not part of the set of equations that must be solved. It is an implicit consequence of the Ampere-Maxwell system. Therefore, the equation to be used for determining $V$, is the gauge condition :

$$\nabla^2 V + \nabla \cdot \mathbf{E} + K\,\xi\ \omega^2\ \varepsilon_{\mathrm{r}}\, V = 0\,. \tag{17}$$

With equations (16) for the solution of $\mathbf{E}$ and (17) for the solution of $V$, we can compute the full $\mathbf{E}V$ system. The cross couplings will not explode for large $\sigma$ in the bulk of the material. Thus we expect that this set-up of equations would have lead to linear systems that will solve faster at high high-frequencies in comparison with the system of equations based on the $\mathbf{A}V$ formulation. However, it should be noted that a third-order derivative term is present. As a consequence the matrix fill increases substantially. We were able to solve (16) and (17) self-consistently for a series of applications at the cost of using *direct* solvers. Finally we note that a full-wave solution needs again*four* fields, i.e. $E_x, E_y, E_z$ and $V$, to be solved.

### 4.1 Boundary Conditions

Although no strong coupling exists in the bulk of the material, the boundary conditions introduce again this coupling in some circumstances.

The boundary conditions for the vector equation (16) can be deduced from the boundary conditions for the vector potential $\mathbf{A}$. Since for each link in the surface of the simulation domain we have put the boundary condition $\mathbf{A} \cdot \hat{\mathbf{t}} = 0$, and $\hat{\mathbf{t}}$ is a tangential unit vector, we obtain

$$\mathbf{E} \cdot \hat{\mathbf{t}} = -\hat{\mathbf{t}} \cdot \nabla V \,. \tag{18}$$

The boundary conditions for the scalar equation (17) can be deduced from the condition that for surface regions outside the contacts, the outward pointing electric field component is taken equal to zero, i.e. $\mathbf{E} \cdot \hat{\mathbf{n}} = 0$ where $\hat{\mathbf{n}}$ is a normal unit vector. However, this will not be sufficient to determine the boundary condition for $V$, since an additional unknown, $\partial V/\partial n$ needs to be given outside the contact regions. Fortunately, there is still room for further restriction. The boundary condition for $\mathbf{A}$ was only provided for the tangential components of $\mathbf{A}$. We will now include also a boundary condition for the normal component of $\mathbf{A}$ that consists of stating that the normal component of $\mathbf{A}$ will have be continuous when crossing the simulation surface

$$\hat{\mathbf{n}} \cdot \mathbf{A}_{\text{inside}} = \hat{\mathbf{n}} \cdot \mathbf{A}_{\text{outside}} \,. \tag{19}$$

This can also be written as $\partial A_{\perp}/\partial\, n = 0$, or in other words: a Neumann boundary condition is used for the perpendicular component of $\mathbf{A}$. However, the surface nodes of the simulation domain can also be determined by applying the Poisson equation and/or current continuity equation for these nodes.

$$\nabla \cdot (\phi\, \mathbf{E}\,) = 0 \,. \tag{20}$$

For internal nodes, this equation is a consequence of the Maxwell-Ampere system. However, at the surface it must explicitly be enforced by the boundary condition. Thus for the boundary nodes, we apply the Poisson and current-continuity equations, using the inwards pointing link variables $E_{ij}$. This enables one to get boundary conditions for the $V$ variables on the simulation boundary.

## 5 Scattering Parameters

In order to determine the S matrix, a rather straightforward procedure is followed. For that purpose a collection of ports is needed and each port consists of two contacts. A contact is defined as a collection of nodes that are electrically identified. A rather evident appearance of a contact is a surface segment on the boundary of the simulation domain. A slightly less trivial contact consists of two or more of these surfaces on the boundary of the simulation domain. The nodes that are found on these surfaces are all at equal potential. Therefore, although there may be many nodes assigned to a single contact, all these nodes together generate only one potential variable to the system of unknowns. Of course, when evaluating the current entering or leaving the contact, each node in the contact contributes to the total contact current. Assigning prescribed values for the contact potential can be seen as applying Dirichlet's boundary conditions to these contacts. This is a familiar technique in technology CAD. Outside the contact regions, Neumann boundary conditions are applied. Unfortunately, since we are now dealing with the full system of Maxwell equations, providing boundary conditions for the scalar potential will not suffice. We also need to provide boundary conditions for the vector potential. Last but not least, since the set of variable $V$ and $\mathbf{A}$ are not independent, setting a boundary condition for one variable has an impact on the other. Moreover, the choice of the gauge condition also participates in the appearance of the variables and their relations. A convenient set of boundary conditions is given by the following set of rules :

- Contact surface $V = V|_c^i$. To each contact area a prescribed potential value is assigned.
- Outside the contact area on the simulation domain $\mathbf{D}_n = 0$. There is no electric field component in the direction perpendicular to the surface of the simulation domain.
- For the complete surface of the simulation domain, we set $\mathbf{B}_n = 0$. There is no magnetic induction perpendicular to the surface of the simulation domain.

We must next translate these boundary condition to restrictions on $\mathbf{A}$. We start with the last one. Since there is no normal component $\mathbf{B}$, we may assume that the vector potential is perpendicular to the surface of the simulation domain. That means that the links at the surface of the simulation domain do not generate a degree of freedom. It should be noted that more general options exist. Nevertheless, the above set of boundary conditions provide the minimal extension of the TCAD boundary conditions if vector potentials are present.

In order to evaluate the scattering matrix, say of an N-port system, we iterate over all ports and put a voltage difference over one port and put an impedance load over all other ports. Thus the potential variables of the contacts belonging to all but one port, become degrees of freedom that need to be evaluated. The following variables are required to understand the scattering matrices, where $Z_0$ is a real impedance that is usual taken to be 50 Ohms

$$a_i = \frac{V_i + Z_0 I_i}{2\sqrt{Z_0}} \tag{21}$$

$$b_i = \frac{V_i - Z_0 I_i}{2\sqrt{Z_0}} \,. \tag{22}$$

The variables $a_i$ represent the voltage waves incident on the ports labeled with index $i$. The variables $b_i$ represent the reflected voltages at ports $i$. The scattering parameters $s_{ij}$ describe the relationship between the incident and reflected waves

$$b_i = \sum_{j=1}^{N} s_{ij} a_j \,. \tag{23}$$

The scattering matrix element $s_{ij}$ can be found by putting a voltage signal at port $i$ and place an impedance of $Z_0$ over all other ports. Then $a_j$ is zero by construction, since for those ports we have that $V_j = -Z_0 I_j$. Note that $I_j$ is defined positive if the current is ingoing. In this configuration $s_{ij} = b_i / a_j$. In a simulation setup, we may put the input signal directly over the contacts that correspond to the input port. This would imply that the input load is equal to zero. The $s$-parameter evaluation set up is illustrated in Fig. 2.

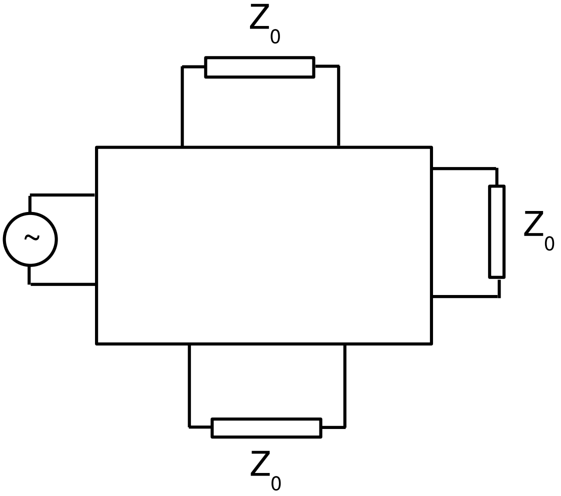


**Fig. 2:** Set up of the $s$-parameter evaluation: 1 port is excited and all others are floating

# 6 Applications

Using the solver based on computational electrodynamics, we are able to compute the s-parameters by setting up a field simulation of the full structure. This allows us to study in detail the physical coupling mechanisms. As an illustration, we consider two inductors which are positioned on a substrate layer separated by a distance of 14 micron. This structure was processed and characterized and the $s$-parameters were obtained. It is quite convenient when studying a compact model parameters to obtain a quick picture of the behavior of the structure. For this device a convenient variable is the 'gain', which corresponds to the ratio of the injected power and the delivered power over an output impedance [11]

$$G = \frac{P_{in}}{P_{out}} . \tag{24}$$

The structure is shown in Fig. 3.

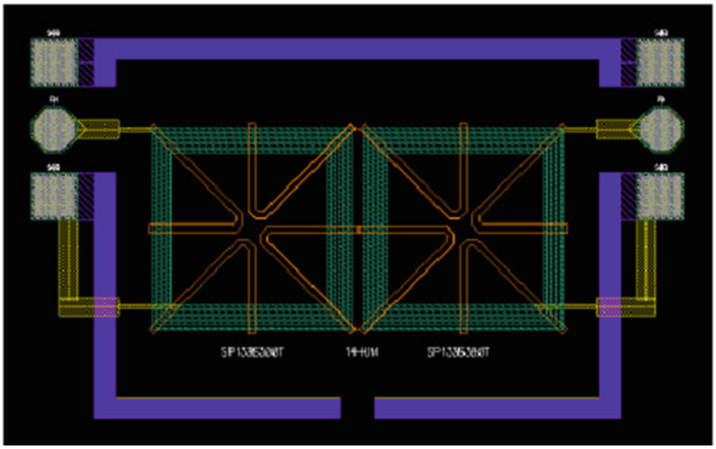


**Fig. 3:** View on the coupled spiral inductor using the Virtuosa design environment

When computing the *s*-parameters, we put the signal source on one spiral (port 1) and place 50 Ohm impedance over the contacts of the second spiral (port 2).The $s_{11}$-parameter is shown in Fig. 4 and the $s_{12}$-parameter is shown in Fig. 5. Finally,

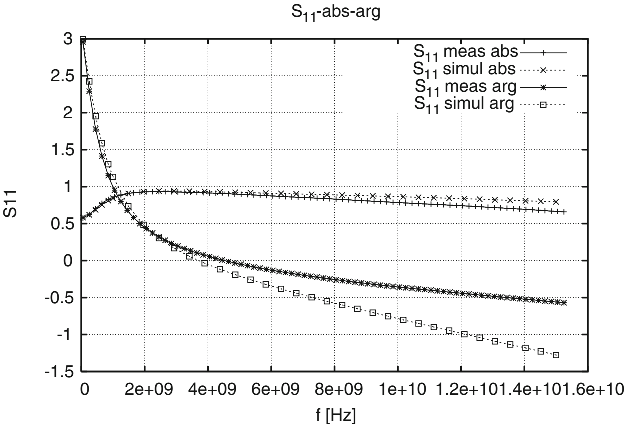


**Fig. 4:** Comparison of the experiment and simulation results for $s_{11}$

the gain plot is shown in Fig. 6. This results shown here have been obtained without any calibration of the material parameters. The silicon is treated ‘as-is’. This means

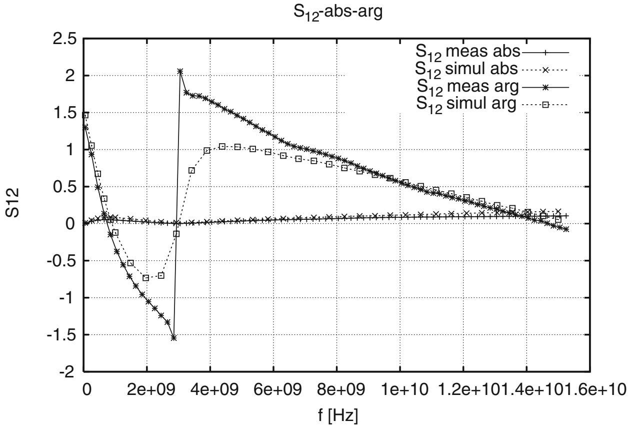


**Fig. 5:** Comparison of the experiment and simulation results for $s_{12}$

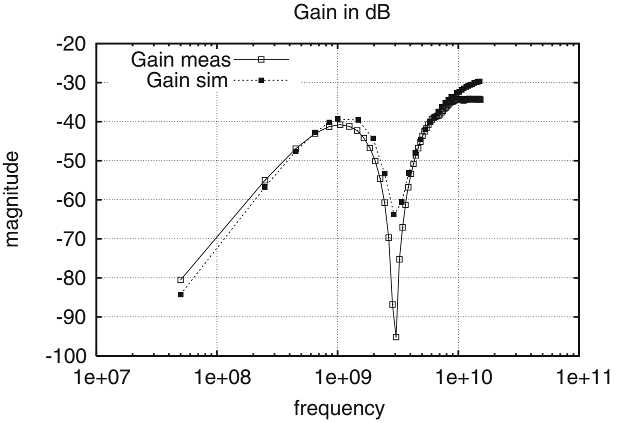


**Fig. 6:** Comparison of the experimental and simulation results for the gain

that the substrate and the eddy current suppressing n-wells are dealt with as doped silicon.

## 7 Conclusions

In this paper we presented a version of computational electrodynamics which is based on the scalar and vector potential formulation. Whereas the finite-integration technique directly deals with the field intensity quantities **E** and **B**, our formulation deals with the more fundamental gauge fields. It should be emphasized that the field quantities are derived variables and once that the potentials have been computed, whereas all other variables are obtained by 'post-processing'. Our approach is a discrete implementation of the geometrical interpretation of electrodynamics [12]. According to this interpretation, the field intensities correspond to the curvature and the potentials are connections in the geometrical sense. The practical capabilities of our method are comparable to other field solvers that focus directly on the fields **E** and **B**, with one exception: if the potentials are needed in the evaluation of the constitutive relations then our method has a clear advantage. This happens if semiconductor modeling is needed and one can not mimic the semiconductor with moderately conductive material. Another area of application is the unified solving of quantum problems and magnetic induction problems where the potential approach is definitely the most natural choice. We have shown with a realistic application that the method is capable of producing fairly good results. The deviations at higher frequency are an indication that adaptive meshing methods are mandatory.

**Acknowledgements** This work was performed in the EU funded projects CODESTAR (IST-2001-34058) CHAMELEON-RF (FP6-2004-IST-4-027378) ICESTARS (FP7-214911) and the IP project PULLNANO (IST 026828). The authors acknowledge the numerous stimulating discussions with the partners in this project.

## References

1. Meuris, P., Schoenmaker, W., W, M.: Strategy in Electromagnetic Interconnect Modeling. IEEE Trans. on CAD of Integr. Syst. **20**, 739–752 (2001)
2. Schoenmaker, W., Meuris, P.: Electromagnetic interconnects and passives modeling: Software implementation issues software implementation issues. IEEE Trans. on CAD of Integr. Syst. **21**, 534–543 (2002)
3. Schoenmaker, W., Magnus, W., Meuris, P.: Ghost fields in classical gauge theories. Phys. Rev. Lett. **88**, 181,602–1 – 181,602–4 (2002)
4. Nedelec, J.C.: Mixed finite elements in $r^3$. Numer. Math. **35**, 315–341 (1980)
5. Lee, J.F.S.D.K., J, C.Z.: Tangential vector finite elements for electromagnetic field computation. IEEE Transactions on Magnetics **27**, 4032–4035 (1991)
6. Weiland, T.: A discretization method for the solution of maxwells equations for six-component fieldss. Electronics and Communications AEU **31 No. 3**, 116120 (1977)
7. Schuhmann, R., Weiland, T.: A stable interpolation technique for fdtd on nonorthogonal grids. International Journal on Numerical Modelling **11**, 299306 (May 1998)
8. Bossavit, A.: Discretization of electromagnetic problems: The "generalized finite differences" approach. In: W.H.A. Schilders, E.J.W. ter Maten (Eds), Handbook of numerical analysis , Elsevier North-Holland **XIII**, 105–197 (2005)
9. Bossavit, A.: The sommerville mesh in yee-like schemes. In: W.H.A. Schilders, E.J.W. ter Maten, S.H.M.J. Houben, Scientific computing in electrical engineering, Series Maths. Series Mathematics in Industry , Springer **4**, 128–136 (2003)
10. Scharfetter, D., Gummel, H.: Large signal analysis of a silicon read diode oscillator. IEEE Trans. Electron Devices **ED-16**, 66–77 (1969)
11. Niknejad, A.M., Meyer, R.G.: Analysis, design and optimization of spiral inductors and transformers for si rf ics. IEEE Journ. of Solid-State Circuits **33**, 1470–1481 (1998)
12. Frankel, T.: The Geometry of Physics. University Press, Cambridge (1997)

# Evaluation of Domain Decomposition Approach for Compact Simulation of On-Chip Coupled Problems

Jagoda Plata, Michal Dobrzynski, and Sebastián Gim

**Abstract** Continued device scaling into the nanometer region has given rise to new effects that previously had a negligible impact but now present greater challenges to successful design of mixed-signal silicon. This paper evaluates Domain Decomposition (DD) strategies for compact simulation of on-chip coupled problems from a computational perspective, using the recently completed CHAMELEON-RF software prototype on several standard benchmark structures.

## 1 Introduction

Incessant miniaturization of the transistor according to Moore's Law has lead to generational improvements in microprocessor technology [1]. However, continued scaling of devices into the nanometer region has given rise to new effects that previously had a negligible impact, but now present challenges to a continued scaling. This has resulted in an increased complexity in engineering resources essential for a successful design. The International Technology Roadmap for Semiconductors (ITRS) suggests extreme scaling of CMOS technology until the 10 nm region and operating frequencies of up to 60 GHz in future generation devices [2]. At such short dimensions, fabrication process variations, substrate noise and electromagnetic (EM) coupling between circuit components make mixed-signal RF silicon designs extremely challenging.

Because of this, the CHAMELEON-RF project was conceived as part of an initiative to address these issues [3]. The project is a research platform for the development of prototype tools and methodologies for comprehensive high accuracy

Jagoda Plata, Michal Dobrzynski, Sebastian Gim
Numerical Methods Laboratory, Electrical Engineering Faculty, Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania, e-mail: plata@lmn.pub.ro, d_michal@lmn.pub.ro, seb@lmn.pub.ro

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_42,

modeling of on-chip electromagnetic effects using the Domain Decomposition (DD) approach and the concept of electromagnetic interconnectors or ‘hooks’ [4].

The CHAMELEON-RF nano-EDA research platform incorporates a novel dual [5] Finite Integral Technique [6] (dFIT) EM field simulator with systematic All Level Reduced Order Modeling (ALROM) [7] to keep manageable the complexity of Integrated Circuit (IC) [8]. The method allows tractable multi-scale parameterized model extraction of coupled structures with the possibility of sensitivity analysis [9]. This approach has advantages over alternative approaches in full 3D field simulators - such as FEM, BEM etc. - which, although enable greater accuracy, are intractable for most practical real world designs.

In this paper, we evaluate the DD strategy for compact simulation of on-chip coupled problems from a computational perspective. By decomposing the modeling domain into distinct domains connected by hooks, computational saving can be obtained. Non essential domains, such as the substrate or air layers, can be simulated just in the simplified field regime, instead of a Full Wave analysis. This approach results in state space matrices with a reduced number of Degrees Of Freedom (DOFs).

## 2 Domain Decomposition approach

An efficient approach to manage the complexity of the IC structures is a decomposition of the computational domain into sub-domains. In this case, each sub-domain generates a simpler field problem that can be simulated independently.

In the proposed approach the defined sub-domains are interconnected by the means of hooks and have to comply with the boundary conditions. The definition of the IC component terminals (intentional interconnections) and connectors - hooks - is based on the correct formulated EM field problem, associated with the concept of ElectroMagnetic Circuit Element (EMCE) [10] and related to the boundary of the domain (Fig. 1). By definition, an EMCE is a simply connected domain $D$ bounded by a fixed surface $\Sigma$ comprising $n'$ disjoint parts $S'_1$, $S'_2$, …, $S'_{n'}$ called electric terminals and $n''$ disjoint parts $S''_1$, $S''_2$, …, $S''_{n''}$ called magnetic terminals on which equations (1-4) apply [4].

$$\mathbf{n}\cdot\mathrm{curl}\mathbf{E}(P,t)=0,\forall P\in\sum-S_k^{'} \tag{1}$$

$$\mathbf{n}\cdot\mathrm{curl}\mathbf{H}(P,t)=0,\forall P\in\sum-S_k^{''} \tag{2}$$

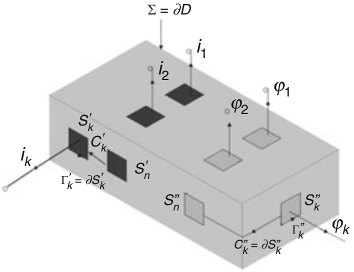
$$\mathbf{n}\times\mathbf{E}(P,t)=0,\forall P\in\bigcup S_k^{'} \tag{3}$$

$$\mathbf{n}\times\mathbf{H}(P,t)=0,\forall P\in\bigcup S_k^{''} \tag{4}$$

Here, we denote with $\mathbf{n}$ the normal unitary vector of the $\Sigma$ surface in the point $P$.

The condition (1) prevents inductive couplings with environment through the element boundary, with an exception of magnetic terminals. Second, the condition (2) implies absence of both conductive and capacitive couplings through the element boundary, except at the electric terminals. The variation of electric potential over

**Fig. 1** Concept of EMCE



every electric terminal is excluded with the condition (3). Therefore it allows a connection of the electric terminal to a node of an external electric circuit. The last (4) condition excludes the variation of magnetic potential over every magnetic terminal. Hence its connection to an external magnetic circuit node is allowed.

With the boundary conditions (1-4) the interaction between EMCE and the environment is described by four scalar variables: terminal current and voltage for electric terminals, and flux and magnetic voltage for the magnetic terminals.

The coupling of the IC with its surrounding is realized in three basic ways: electric interconnect terminals (intentional), electric connectors (virtual) and magnetic connectors (virtual), however from the theoretical point of view there is no difference between terminals and connectors. Assuming, that the electric environment is represented by an electric circuit, while the magnetic environment by a magnetic circuit, these two circuits can be coupled together by the appropriate formulation of (1-4) by means of virtual magnetic and electric connectors. The electric terminals allow the electric interaction while the magnetic terminals allow the inductive interaction, thus the component can be coupled with its electromagnetic environment.

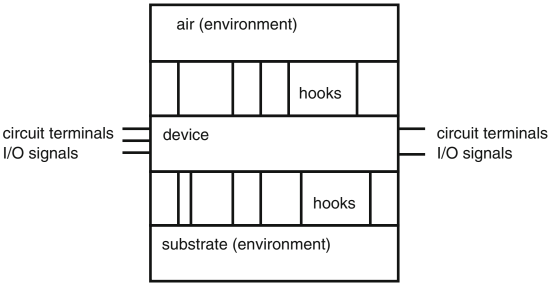


**Fig. 2:** Typical domain decomposition of the IC RF block

The numerical approach we propose is based on the DD of the RF block into its environmental components - namely the silicon substrate lower sub-domain and the upper sub-domain representing the air (Fig. 2). Consequently, the computational effort of such divided domains is based on state space matrices with a reduced size when applying the discretization mesh. Furthermore, computational saving result from an appropriate field regime description and simulation in each of the domains; although the analysis of metal components placed in silicon dioxide layers has to be determined with appropriate Full Wave (FW) equations, the upper air and lower substrate sub-domains can be simulated just in the Magneto Static (MS) and/or Electro Quasi Static (EQS) regime.

The models have been extracted solving the Maxwell equations with FIT [11] combined with ALROM and simulated on the prototype software, developed in the Numerical Methods Laboratory (LMN), implementing all the mentioned methods.

Several benchmark structures have been simulated - passive microstructures designed and fabricated on a 0.35$\mu$m BiCMOS process, measured at the industrial partner site austriamicrosystems AG. In this paper an exemplary benchmark test structure is used to evaluate and present the main ideas of our modeling approach.

## 3 Simulation Results

The benchmark test structure presented in this paper is a CHRF203 metal stack placed over thick n-well layer that acts as shielding. It is a widely used structure in the semiconductor industry for RF filtering purposes.

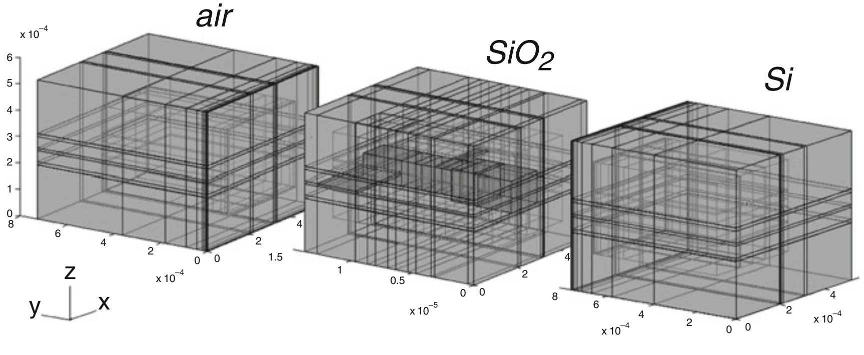


**Fig. 3:** Computational domain split into air, silicon dioxide and substrate sub-domains; with discretization meshes applied: electric (*black*) and magnetic grid (*grey*)

In our approach, the computational domain of the device was decomposed into three parts containing the following layers: 725 $\mu$m of silicon, 10 $\mu$m of silicon dioxide and 725 $\mu$m of air (Fig. 3). Each of the sub-domains was analyzed independently and a compact, reduced model was extracted. On the interface of the sub-domains both electric and magnetic hooks were placed in order to enable the reconnection of the split model.

The simulation of the reconnected sub-domains was first validated against the full simulation of the entire device. For this purpose we applied a discretization mesh for each sub-domain of 9x10 in the XZ direction. The electric and magnetic hooks used were only one-dimensional (points), enabling the node-by-node interconnection of the model. Hence a total number of 160 hooks was set in each node of XZ surface interconnecting the split sub-domains. On the electric grid 89 electric hooks and one reference grounding node were placed. Likewise, we positioned on magnetic grid surface one reference magnetic ground node and 71 magnetic hooks in all the other nodes. Next, the simulation of the sub-domain test cases was carried out in the FW (device sub-domain), MS and EQS regime (environment sub-domains). The results and comparison between the DD test cases are presented in Table 1 and Fig. 4. In the comparison we use scattering parameters (S-parameters), which describe how the energy couples between each pair of device ports.

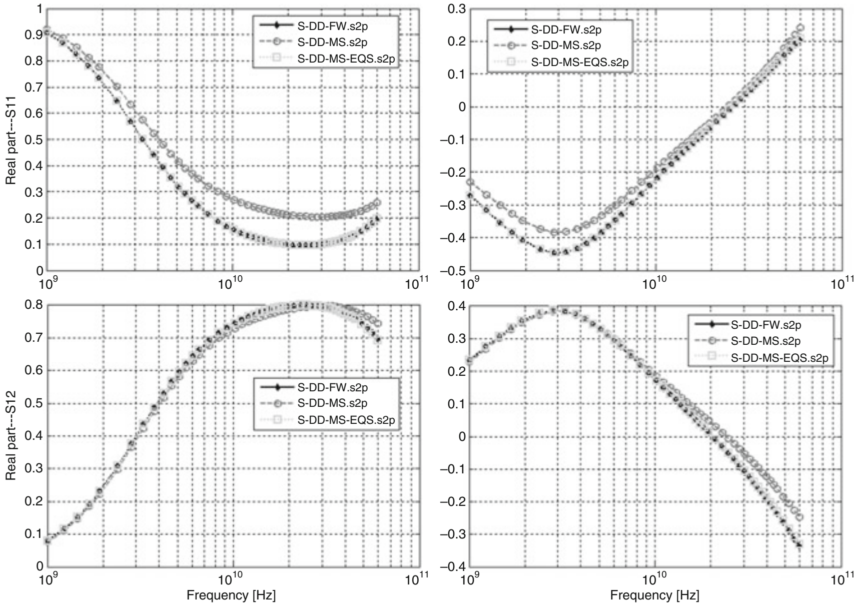


**Fig. 4:** S-parameters FW+FW+FW (black) vs MS+FW+MS (*dark grey circles*) vs MS-EQS+FW+MS-EQS (*light grey squares*). Minimal grid 9 × 17 × 10

Note that simplifying the description of the field problem in the component, the *m*-number of DOF decreases. As a result, the solution of a liner system of *m*-equations and therefore the frequency response of the analyzed device is found faster.

The computational effort to carry out the simulation in the MS regime for substrate and air sub-domains was the most beneficial; however the results obtained carry a too large error compared with the FW simulation. Nevertheless simulations

**Table 1:** Simulation comparison for various sub-domains and field description approach

| *n* | Description | DOF | Time$^{a}$(s) | Error$^{b}$(%) | |
|---|---|---|---|---|---|
| *1* | Simulation in the FW regime for all three sub-domains | 7942 | 2 | - | Black$^{c}$ case |
| *2* | MS regime simulation for the two environment sub-domains; the device sub-domain was analysed in FW | 6198 | 1.5 | 14.27 | Dark grey$^{c}$ case |
| *3* | MS+EQS simulation of the upper and lower sub-domains; the device sub-domain was analyzed in FW | 7002 | 1.6 | **0.74** | Light grey$^{c}$ case |

[a] Time to compute one frequency

[b] Error obtained between the *n*-case approach and the 'black' case (1)

[c] Curves denoted in Fig. 4

of the sub-domains in MS+EQS regime reveal both reasonable accuracy and computational saving (error below 1% in Table 1).

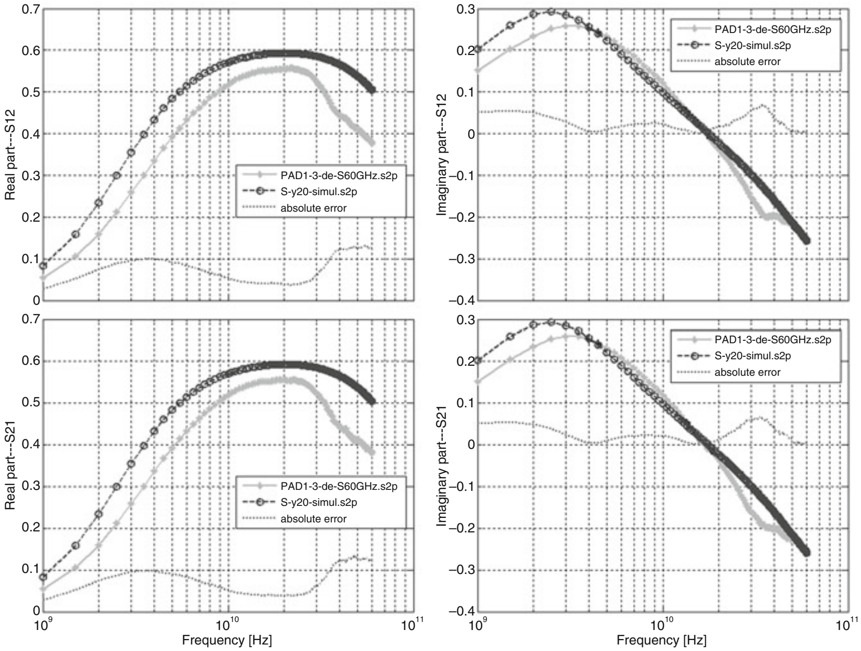


**Fig. 5:** S-parameters simulation (*black*) vs measurement (*grey*). DOF of interconnected model = 23.242; refined grid 9 × 51 × 10; maximum error (simulation vs measurement) = 27.15%

Finally, the sub-domains were discretized with refined, exponentially partitioned orthogonal grids in order to compare simulation results with the measurement data provided by the industrial partner (Fig. 5). The successful comparison validated the extracted model. Figure 5 illustrates the average agreement with measurement (grey) using the MS+EQS (air and silicon sub-domains) and FW (middle silicon dioxide) configuration for simulation (black). These are the initial results of the prototype software running on desktop computers. The denser discretization grid would provide better accuracy results. A high performance Beowulf cluster is currently being installed at the authors' institution and will enable simulations of larger DOFs in the near future.

## 4 Conclusions

The use of DD and electromagnetic hooks has been shown to be an accurate and computationally efficient method to model compact IC structures by dividing it into independent sub-domains. The decomposition of the modeled device into environmental and structural component domains is a beneficial approach especially for large challenging structures. The reduction in the complexity for the model extraction process was clearly illustrated with a reduced number of DOF equations in a sparse linear system for the partitioned domain compared with a full simulation. Computational effort to solve these systems was also lower compared with the full simulation. Furthermore, the division into sub-domains enables time saving also due to the parallelization of the modeling process and enables reuse of previously simulated domains.

Another advantage of the approach is the possibility of using only MS and/or MS+EQS regime analysis in the simple sub-domains consisting exclusively of air and silicon (without interconnect structures). This allows minimizing the computational effort which is currently the most time and memory consuming for the FW simulation. Simulating the air and silicon substrate sections in either MS or EQS regime results in a 1/6 computational saving compared with the FW simulation. Of the various scenarios considered for decomposition, the MS+EQS (air) / FW (structure) / MS+EQS (substrate) provided the best computational efficiency with the lowest error with respect to a full FW simulation.

The Chamy prototype software based on the presented theory and developed in LMN was validated for realistic, challenging engineering problems in microprocessor design. Chamy is an accurate tool to manage the problems of complexity in the integrated component structures working at high frequencies.

**Acknowledgements** The authors would like to gratefully acknowledge support and insightful discussion with Prof. Daniel Ioan, Asc. Prof. Gabriela Ciuprina and the LMN Chameleon project team. The authors would also like to acknowledge the financial support offered by the European Commission under the Marie Curie Fellowship program (FP6/EST3, FP6/ToK4nEDA) and thank austriamicrosystems AG for the experimental data.

The views stated herein reflect only the authors' views and the European Commission is not liable for any use that may be made of the information contained.

## References

1. Mollick, E.: Establishing Moore's Law. IEEE Annals of the History of Computing, **31**, 62–75 (2006)
2. ITRS Roadmap. URL http://www.itrs.net.CitedSep112008
3. Janssen, H.H.J.M., Niehof, J., Schilders, W.H.A.: Accurate Modeling of Coupled Functional RF Blocks: CHAMELEON RF. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 81–87. Springer, Berlin Heidelberg New York (2007)
4. Schilders, W.H.A., Ioan, D., Ciuprina, G., Van der Meijs, N., Schoenmaker, W.: Models for Integrated Components Coupled with their EM Environment. COMPEL, **27**, no.4, 820–829 (2008)
5. Ioan, D., Ciuprina, G., Mihalache, D.: Reduced Order Electromagnetic Models for On-Chip Passives Based on Dual Finite Integrals Technique. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 287–294. Springer, Berlin Heidelberg New York (2007)
6. Munteanu, I., Weiland, T.: RF & Microwave Simulation with the Finite Integration Technique - From Component to System Design. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 247–260. Springer, Berlin Heidelberg New York (2007)
7. Ioan, D., Ciuprina, G., Kula, S.: Reduced Order Models for HF interconnect over lossy semiconductor substrate. In: Proceedings of the 11th IEEE Workshop on Signal Propagation on Interconnects, SPI 2007, pp. 233-236. Ruta di Camogli, May 13–16 (2007)
8. Ioan, D., Ciuprina, G., Radulescu, M., Seebacher, E.: Compact modeling and fast simulation of on-chip interconnect lines. IEEE Transactions on Magnetics, **42**, 547–550 (2006)
9. Ioan, D., Ciuprina, G., Radulescu, M.: Theorems of parameter variations applied for the extraction of compact models of on-chip passive structures. In: Proceedings of the 2005 International Symposium on Signals, Circuits and Systems, ISSCS 2007, pp. 147–150. Iasi, July 14–15 (2005)
10. Munteanu, I., Ioan, D.: Missing Link Rediscovered: The Electromagnetic Circuit Element Concept. JSAEM Studies in Applied Electromagnetics and Mechanics, **8**, 302–320 (1999)
11. Weiland, T., Clemens, M.: Discrete electromagnetism with the Finite Integration Technique. Progress in Electromagnetics Research (PIER), **32**, 65–87 (2001)

# DAE-Index and Convergence Analysis of Lumped Electric Circuits Refined by 3-D Magnetoquasistatic Conductor Models

Sebastian Schöps, Andreas Bartel, Herbert De Gersem, and Michael Günther

**Abstract** In this paper the field/circuit coupling is reconsidered for (non-linear) lumped electric circuits refined by 3-D magnetoquasistatic conductor models, where the circuit is described by modified nodal analysis and the field is discretized in terms of the finite integration technique. This leads to the coupling of systems of differential-algebraic equations, for which two numerical approaches are proposed, the weak coupling (co-simulation) and strong coupling (monolithic). The DAE-index of the subproblems and of the full problem are analyzed, then convergence properties of the co-simulation are studied. Finally computational results of a simple half rectifier circuit are exemplarily given to prove the applicability of the concepts.

## 1 Introduction

Basic elements in circuit analysis are described by (non-)linear relations, disregarding distributed field effects. Sometimes complex *companion models* are employed to meet reality. These give, however, only a partial insight into field effects. In contrast, *refined models* directly rely upon Maxwell's equations and are coupled here with electric network equations. We analyze this coupling with two distributed conductor types, which exhibit proximity and skin effects related to eddy currents.

The coupled problem is a system of differential-algebraic equations (DAEs) originating from Kirchhoff's laws and the discrete Maxwell equations. It can be directly addressed by solving one *monolithic* system using a field- or circuit-oriented approach. In the field approach, commonly the circuit is described using loop/branch techniques and is solved within the field simulator. This approach is quite successful and well understood [1], but it is neither efficient for coupling with very large

Sebastian Schöps, Andreas Bartel, Michael Günther
BU Wuppertal, Wuppertal, Germany, e-mail: schoeps@math.uni-wuppertal.de, bartel@math.uni-wuppertal.de, guenther@math.uni-wuppertal.de

Herbert De Gersem
KU Leuven, Leuven, Belgium, e-mail: herbert.degersem@kuleuven-kortrijk.be

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_43,

circuits nor usable within modern circuit simulators that are based on modified nodal analysis (MNA). The circuit-oriented approach relies on MNA and although intensive research has been carried out [2], companion models are still widespread.

Obviously, the strongly coupled approaches do not have the advantages of problem-specific simulators. In this context *co-simulation* can becomes beneficial [3]. It allows to use different simulators for each subproblem, and thus provides a natural support for diversifying integration methods and time-stepping (*multirate*) with respect to the subproblems. Here, the coupling is given mathematically by a waveform relaxation scheme.

The paper is organized as follows: In Sections 2 and 3 the circuit and field settings are recalled; in Section 4 we analyze the index of the field-system; in Section 5 we introduce the weak and strong coupling and provide an index and convergence analysis; in Sections 6 and 7 we give an illustrative example and conclusions.

## 2 Lumped Electric Circuit

Electric circuits are described by basic element relations and Kirchhoff's laws. Using standard MNA, this yields a DAE system since the variables are redundant. In the charge-flux oriented formulation [4], the system reads

$$\begin{aligned} A_{\mathrm{C}}\tfrac{\mathrm{d}}{\mathrm{d}t}q + A_{\mathrm{R}}r(A_{\mathrm{R}}^{T}e,t) + A_{\mathrm{L}}i_{\mathrm{L}} + A_{\mathrm{V}}i_{\mathrm{V}} + A_{\mathrm{I}}i(t) + A_{\lambda}i_{\lambda}(A_{\lambda}^{T}e,t) &= 0,\\ \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi - A_{\mathrm{L}}^{T}e = 0, \qquad A_{\mathrm{V}}^{T}e - v(t) &= 0, \\ q - q_{\mathrm{C}}(A_{\mathrm{C}}^{T}e,t) = 0, \quad \Phi - \Phi_{\mathrm{L}}(i_{\mathrm{L}},t) &= 0, \end{aligned} \tag{1}$$

with incidence matrices $A$, node potentials $e$, independent and controlled current and voltage sources $i$, $i_\lambda$ and $v$, currents through voltage and flux controlled elements $i_V$ and $i_L$, charges $q$ and fluxes $\Phi$, functions of charges, fluxes and resistances $q_C$, $\Phi_L$ and $r$ (with positive definite derivatives), respectively.

Several index concepts were introduced to classify DAEs. Since these notations are equivalent for linear systems, we state here only the (differential) index, [5]: For the given system $F\left(t,\tfrac{\mathrm{d}}{\mathrm{d}t}x,x\right)=0$, the index $\nu \in \mathbb{N}_0$ is the smallest number, such that the enlarged set of equations

$$F\left(t,\tfrac{\mathrm{d}}{\mathrm{d}t}x,x\right)=0, \quad \tfrac{\mathrm{d}}{\mathrm{d}t}F\left(t,\tfrac{\mathrm{d}}{\mathrm{d}t}x,x\right)=0, \quad \ldots, \quad \tfrac{\mathrm{d}^\nu}{\mathrm{d}t^\nu}F\left(t,\tfrac{\mathrm{d}}{\mathrm{d}t}x,x\right)=0$$

allows to deduce a system of ordinary differential equations (ODEs) by algebraic manipulations. In this way, $\nu$ denotes the inherent number of derivatives and measures the expected numerical difficulties.

In this respect, the numerical properties of (1) are well known, the DAE-index has been discussed by decomposing the unknown $(e,i_{\mathrm{V}},i_{\mathrm{L}},q,\Phi)$ into algebraic and differential parts using a projector $Q_C$ onto the kernel of $A_C^T$, i.e.,

$$Q_C \ker A_C^T = \ker A_C^T \text{ and } A_C^T Q_C = 0$$

and its complement $P_C = I - Q_C$. We assume in the above terms:

**C1** No loops of capacitors and voltage sources, i.e., $\ker Q_C^T A_V = \{0\}$.
**C2** No cutsets of inductors and current sources[1], i.e., $\ker(A_C, A_R, A_V)^T = \{0\}$.
**C3** Voltage controlled current sources parallel to capacitors, i.e., $Q_C^T A_\lambda i_\lambda = 0$.

This splits the unknown into a differential part $y := (P_C e, j_L)^T$ and an algebraic part $z := (Q_C e, j_V, q, \Phi)^T$, such that

$$\tfrac{\mathrm{d}}{\mathrm{d}t} y = f_1(y, z, i_\lambda), \qquad 0 = g_1(y, z), \tag{2}$$

is an index-1 description of (1) since the derivative $\frac{\partial}{\partial z} g_1$ can be shown to be nonsingular assuming **C1-C3**. It is possible to prove [6]:

**Theorem 1.** *Let us consider a lumped electric circuit in form* (1) *that respects **C3**, then the flux/charged oriented MNA leads to an index-1 DAE iff **C1-C2** hold, it leads otherwise to an index-2 DAE.*

## 3 Electromagnetic Field

The electromagnetic field is described by Maxwell's equations. We assume a spatial discretization based on staggered grids (e.g. the finite integration technique) [7, 8]. In magnetoquasistatics with linear materials one can deduce the curl-curl equation

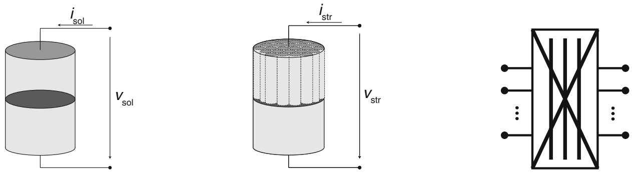
$$M_\sigma \tfrac{\mathrm{d}}{\mathrm{d}t} \widehat{a}(t) + K_\nu \widehat{a}(t) = \widehat{\widehat{j}}_{\mathrm{src}}(t)\ , \tag{3}$$

where $\widehat{a}$ denotes the discrete magnetic vector potential (MVP), $M_\sigma$ the diagonal positive semi-definite conductivity matrix, $\widehat{\widehat{j}}_{\mathrm{src}}$ the source current density and $K_\nu := \tilde{C} M_\nu C$ is the curl-curl matrix composed of the curl-operators for the primary and dual grid $C$ and $\tilde{C}$, respectively and the diagonal positive definite reluctivity matrix $M_\nu$. Due to the non-trivial nullspace of $M_\sigma$ this is a DAE, which is generally not uniquely solvable because of the additional nullspace of the curl-operators. Thus a gauge is needed to select one solution within the equivalent class $\widehat{\widehat{b}} = C\widehat{a}$, [10].

## 4 Field Models as Refined Network Elements

Conductor models for connecting field and circuit parts are well-known. Most common are solid and stranded conductors (Fig. 1). We use the given symbol for a (multiport) device that consists of (multiple) conductors of both types which are tightly coupled by the field. The field is described by the curl-curl equation and excited by $\widehat{\widehat{j}}_{\mathrm{src}}$ due to the connected circuit [9]. Typically voltage drops of solid conductors ($v_{\mathrm{sol}}$) and the currents through stranded conductors ($i_{\mathrm{str}}$) are considered to be given and thus the excitation reads

[1] neither independent nor voltage controlled current sources, i.e., solid/stranded conductors



**(a)** Solid **(b)** Stranded **(c)** Symbol

**Fig. 1:** Conductor models (**a**), (**b**) and device symbol (**c**) that embeds both into the circuit

$$\widehat{\widehat{j}}_{\mathrm{src}} = M_\sigma Q_{\mathrm{sol}} v_{\mathrm{sol}} + Q_{\mathrm{str}} i_{\mathrm{str}} \ . \tag{4}$$

Here $Q = [Q_{\mathrm{sol}}, Q_{\mathrm{str}}]$ denotes the coupling matrix. Each column corresponds to a conductor model and imposes currents/voltages onto edges of the grid. The unknown currents $i_{\mathrm{sol}}$ and voltages $v_{\mathrm{str}}$ are obtained by the additional equations

$$i_{sol} = G_{\mathrm{sol}} v_{\mathrm{sol}} - Q_{\mathrm{sol}}^T M_\sigma \tfrac{\mathrm{d}}{\mathrm{d}t}\widehat{a} \ , \qquad v_{\mathrm{str}} = R_{\mathrm{str}} i_{\mathrm{str}} + Q_{\mathrm{str}}^T \tfrac{\mathrm{d}}{\mathrm{d}t}\widehat{a} \ , \tag{5}$$

with the diagonal conductance matrices $G_{\mathrm{sol}}$ for solid and the diagonal resistance matrix $R_{\mathrm{str}} = G_{\mathrm{str}}^{-1}$ for stranded conductors. Let us assume the following:

**F1** The matrix pencil is regular, i.e., $[M_\sigma, K_\nu] := \det(\lambda M_\sigma + K_\nu) \neq 0$ for a $\lambda$.
**F2** The models are non-overlapping, i.e., $Q_{(i)}^T \, Q_{(j)} = 0$, for all $i \neq j$.
**F3** The excitation is consistent, i.e., $\ker(CQ_{\mathrm{sol}}) = \{0\}$, $\ker(CM_{\sigma,\mathrm{aniso}}^+ Q_{\mathrm{str}}) = \{0\}$.

where $M_{\sigma,\mathrm{aniso}}^+$ is the pseudoinverse of the anisotropic conductivity matrix for stranded conductors. **F1** is equivalent to a gauging of (3) and **F2** prohibits the smearing of spatially separate models into each other, this allows to obtain (6) from (3-5),

$$M_{\sigma,\mathrm{fillin}} \tfrac{\mathrm{d}}{\mathrm{d}t}\widehat{a} + K_\nu \widehat{a} = M_\sigma Q_{\mathrm{sol}} v_{\mathrm{sol}} + Q_{\mathrm{str}} G_{\mathrm{str}} v_{\mathrm{str}} := \widehat{\widehat{j}}_{\mathrm{src}}^{\,*} \ , \tag{6a}$$

$$Q_{\mathrm{sol}}^T K_\nu \widehat{a} = i_{\mathrm{sol}} \ , \tag{6b}$$

$$G_{\mathrm{str}} Q_{\mathrm{str}}^T M_{\sigma,\mathrm{aniso}}^+ K_\nu \widehat{a} = i_{\mathrm{str}} \ , \tag{6c}$$

where $M_{\sigma,\mathrm{fillin}} := M_\sigma + Q_{\mathrm{str}} G_{\mathrm{str}} Q_{\mathrm{str}}^T$ is the (dense) conductivity matrix for both types.

**Lemma 1.** *Let the field problem consist of solid and stranded conductors which fulfill* ***F1-F2****, then the curl-curl equation* (6a) *is index-1 for given voltages and the algebraic part of the MVP is zero.*

*Proof.* By **F1**, the symmetric positive semi-definiteness of $M_{\sigma,\mathrm{fillin}}$ implies that (6a) is index-1 and the Kronecker Normal Form [5] for this system reads

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\widehat{a}_1(t) + U_1 K_\nu V_1 \ \widehat{a}_1(t) = U_1 \ \widehat{\widehat{j}}_{\mathrm{src}}^{\,*} \ , \tag{7a}$$

$$\widehat{a}_2(t) = U_2 \ \widehat{\widehat{j}}_{\mathrm{src}}^{\,*} \ , \tag{7b}$$

and this splits the MVP $\widehat{a} = V_1 \widehat{a}_1 + V_2 \widehat{a}_2$ into differential and algebraic parts by using the regular matrices $U^T = \left(U_1^T \ , U_2^T\right)$ and $V = (V_1 \ , \ V_2)$. From

$$U_2 M_{\sigma,\mathrm{fillin}} = U_2 \left( M_\sigma + Q_{\mathrm{str}} G_{\mathrm{str}} Q_{\mathrm{str}}^T \right) = 0$$

follows that both $U_2 M_\sigma$ and $U_2 Q_{\mathrm{str}} G_{\mathrm{str}}$ vanish because the images of $M_\sigma$ and $Q_{\mathrm{str}}$ are distinct, since **F2** is assumed. Hence we finally conclude that the algebraic part of the MVP is zero: $\widehat{a}_2 = U_2 \,\widehat{\widehat{j}}{}^{*}_{\mathrm{src}} = 0$. □

Let us now study the full system (6) in the abstract form

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\widehat{a} = f_{2a}(\widehat{a}, v_\lambda), \qquad 0 = f_{2b}(\widehat{a}, v_\lambda), \qquad 0 = g_2(\widehat{a}, i_\lambda), \tag{8}$$

where the voltages $v_\lambda = (v_{\mathrm{sol}}, v_{\mathrm{str}})^T$ and the currents $i_\lambda = (i_{\mathrm{sol}}, i_{\mathrm{str}})^T$ are combined in vectors. The algebraic evaluation $f_{2b}$ is trivial in our case because of Lemma 1 and the algebraic function $g_2$ consists of (6b), (6c), which can be written in the form

$$0 = g_{\mathrm{sol}}(\widehat{a}, i_{\mathrm{sol}}), \qquad 0 = g_{\mathrm{str}}(\widehat{a}, i_{\mathrm{str}}).$$

System (8) establishes a relation between currents ($i_{\mathrm{sol}}$, $i_{\mathrm{str}}$) and voltages ($v_{\mathrm{sol}}$, $v_{\mathrm{str}}$) and we can choose which quantity is treated as unknown for each conductor type in the field system, since then the other quantity is defined by the coupled electric circuit. Therefore we will distinguish between the possible sets in the following.

**Theorem 2.** *Let the field problem consist of solid and stranded conductors which fulfill **F1-F3**. Iff all the voltages (*$v_{\mathrm{sol}}$, $v_{\mathrm{str}}$*) are given, then system* (6) *is index-1 and in all other cases it is index-2.*

*Proof.* In the case of given voltages the currents $i_\lambda$ are obtained by evaluations of the algebraic equation $g_2$. Thus one differentiation with respect to time yields an ODE, hence we have index-1. In all other cases the arguments are analogue to the case of given $i_{\mathrm{sol}}$ and $v_{\mathrm{str}}$. Now the function $f_{2a}$ in (8) depends on the unknown $v_{\mathrm{sol}}$ and one time derivative yields the additional *hidden constraint*:

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} g_{\mathrm{sol}}(\widehat{a}, i_{\mathrm{sol}}) = \frac{\partial}{\partial \widehat{a}} g_{\mathrm{sol}} \cdot f_2(\widehat{a}, v_{\mathrm{sol}}) + \frac{\mathrm{d}}{\mathrm{d}t} i_{\mathrm{sol}} =: h_{\mathrm{sol}}(\widehat{a}, v_{\mathrm{sol}}, \frac{\mathrm{d}}{\mathrm{d}t} i_{\mathrm{sol}}),$$

and since the conductivity matrices $M_\sigma$ and $M_{\sigma,\mathrm{aniso}}$ reflect **F2** ($M_{\sigma,\mathrm{aniso}} Q_{\mathrm{sol}} = 0$), another differentiation of this constraint gives

$$\frac{\partial}{\partial v_{\mathrm{sol}}} h_{\mathrm{sol}} = Q_{\mathrm{sol}}^T K_\nu M_{\sigma,\mathrm{fillin}}^{+} M_\sigma Q_{\mathrm{sol}} = Q_{\mathrm{sol}}^T K_\nu Q_{\mathrm{sol}} = Q_{\mathrm{sol}}^T C^T M_\nu C Q_{\mathrm{sol}} \,,$$

which is non-singular due to **F3**; thus it is index-2. □

If voltages are considered unknown, then (6) is an index-2 Hessenberg system (with index-1 evaluations), [11]. Since the index-2 variables enter only linearly and without time-dependence, the differential variables are not affected by the derivatives of perturbations and thus the numerical difficulties still correspond to index-1 [12].

# 5 Coupling

We assign $v_\lambda$ to the differences of applied node potentials $e$ for elements with topology $A_\lambda$ and assign the $i_\lambda$ to the currents through the conductors

$$v_\lambda = A_\lambda^T e\,, \qquad\qquad i_\lambda = (i_{\mathrm{sol}}, i_{\mathrm{str}})^T. \tag{9}$$

Now, the *monolithic* system is composed of (1), (6) and (9).

**Theorem 3.** *Let us consider an electric circuit in the form* (1) *with* ***C1-C2****, which is monolithically coupled via* (9) *to a field model* (6) *of solid and stranded conductors fulfilling* ***F1-F3****, then the full system is index-1.*

*Proof.* The algebraic components of the MVP are insignificant for solid and stranded conductors according to Lemma 1. Hence after embedding the field into the circuit system the separated unknowns of the full system read

$$y := (P_C e, j_L, \widehat{a}_1)^T, \qquad\qquad z := (Q_C e, j_V, q, \phi, i_\lambda)^T. \tag{10}$$

The critical partial derivative of the algebraic equation $\frac{\partial}{\partial z} g$ consisting of $g_1$ and $g_2$ is non-singular, since the first is regular due to **C1-C2** and the second is just an evaluation of a differential variable ($\widehat{a}_1$). Thus we have index-1. □

Assumption **C3** is not required in the monolithic coupling because the algebraic part of the MVP was shown to vanish for any excitement of solid and stranded conductors.

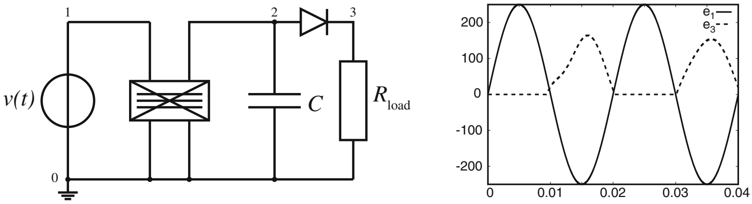
Alternatively, the subproblems could be treated separately by a *waveform relaxation scheme* (of Jacobi or Gau-Seidel type). When applying these schemes to DAEs one has to pay attention to algebraic constraints to avoid numerical instabilities, [13]. We suggest the Gau-Seidel scheme (11) that computes the functions $a^{(1)}$, $y^{(1)}$ and $z^{(1)}$ on a time frame $T = [t_0, t_0 + H]$ for given initial values at time $t_0$ and previous iterates $y^{(0)}$ and $z^{(0)}$.

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\widehat{a}^{(1)} &= f_2(\widehat{a}^{(1)}, v^{(0)}), \quad v^{(0)} := A_\lambda^T(y^{(0)} + z^{(0)}), \quad & \tfrac{\mathrm{d}}{\mathrm{d}t} y^{(1)} &= f_1(y^{(1)}, z^{(1)}, i_\lambda^{(1)}), \\ 0 &= g_2(\widehat{a}^{(1)}, i_\lambda^{(1)}), & 0 &= g_1(y^{(1)}, z^{(1)}, i_\lambda^{(1)}). \end{aligned} \tag{11}$$

The convergence is guaranteed since there is no dependence in algebraic constraints ($g_1$, $g_2$) on previous algebraic iterates ($i_\lambda^{(0)}$, $z^{(0)}$), [14]. Hence we obtain:

**Lemma 2.** *Let us consider an electric circuit* (1) *fulfilling* ***C1-C2*** *and a field model* (6) *respecting* ***F1-F3*** *and employing the interface* (9)*. Then the waveform-relaxation* (11) *will converge.*

The additional assumption **C3** can eliminate the $i_\lambda$-dependence of the algebraic equation $g_1$ and allows us to exchanges the computational order of the subproblems (we may compute the circuit first) without losing the convergence guarantee.



(**a**) Half rectifier: $v_{\text{eff}} = 250\,\text{V}$, $f = 50\,\text{Hz}$, $R_{\text{load}} = 100\,\Omega$ and Shockley diode $I_s = 1\,\mu\text{A}$

(**b**) Voltage in nodes 1 and 3, obtained by `mono`, $H = 5\mu\text{s}$

**Fig. 2:** Refined half rectifier circuit and its input and computed output voltages

# 6 Numerical Experiments

The simulations were obtained with code that is implemented within the *COMSON DP* using field models constructed by *EM Studio* from *CST* (`www.comson.org` and `www.cst.com`). The code is capable of both, the monolithic (`mono`) and the co-simulation of non-linear circuits refined by conductor models. The co-simulation uses scheme (11) with no (`cosim1`) and two iterations (`cosim3`) of each time frame $T$. The integration was kept simple by applying backward Euler.

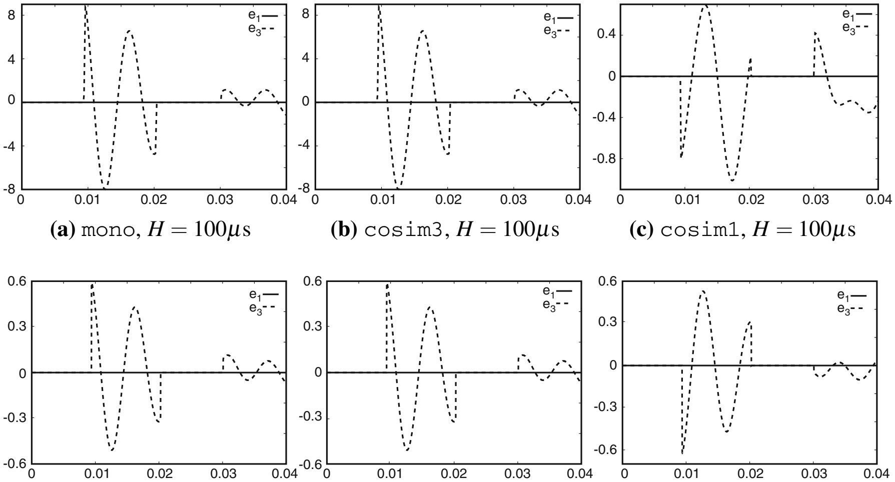
The example of Fig. 2 is a refined half rectifier with a transformer consisting of two stranded conductors and a solid core. `cosim1` performs slightly faster than `mono` using the step size $H$ and it yields better results if the accuracy requirement is quite low. For decreasing step sizes `cosim1` does not linearly improve its accuracy as `mono` and `cosim3` do (Fig. 3), but `cosim3` suffers from an increased computational effort due to the additional iterations.

Adaptive time-integrators in the co-simulation apply the same step size to both subproblems, as long as they do not have multirate potential itself. This is in line with the fact that the field reflects the dynamics of the coupled circuit nodes.

# 7 Conclusions

The field problem is essentially an index-1 DAE, the monolithic coupled system is still index-1 and the convergence of the proposed co-simulation is guaranteed, as illustrated by the computation of a refined rectifier circuit. The co-simulation may use problem-specific software packages and exploits multirate potentials if available in the circuit, but its efficiency can be improved, for example by applying a time frame and iteration control, and the use of more complex equivalent circuits (e.g. additional inductivities) might require fewer field updates [15, 16].

**Acknowledgements** This work was partially supported by the European Commission within the framework of the COMSON RTN project and the post-doc program of the “FG Mathematik and Informatik” of the Bergische Universitt Wuppertal.



**(a)** mono, $H = 100\mu$s **(b)** cosim3, $H = 100\mu$s **(c)** cosim1, $H = 100\mu$s

**(d)** mono, $H = 10\mu$s **(e)** cosim3, $H = 10\mu$s **(f)** cosim1, $H = 10\mu$s

**Fig. 3:** Errors in the voltages compared to the results of mono, $H = 5 \cdot 10^{-6}$ from Fig. 2b

## References

1. Benderskaya, G. et al.: Transient Field-Circuit Coupled Formulation Based on the FIT and a Mixed Circuit Formulation. COMPEL **23**(4), 968 – 976 (2004)
2. Dreher, T., Meunier, G.: 3D Line Current Model of Coils and External Circuits. IEEE Trans. Mag. **31**(3), 1853 – 1856 (1995)
3. Bedrosian, G.: A New Method for Coupling Finite Element Field Solutions with External Circuits and Kinematics. IEEE Trans. Mag. **29**(2), 1664 – 1668 (1993)
4. Günther, M., Rentrop, P.: Numerical Simulation of Elec. Circuits. GAMM **1/2**, 51 – 77 (2000)
5. Hairer, E., Wanner, G.: Solving ODEs II. Springer, Berlin, 1991
6. Estévez Schwarz, D., Tischendorf, C.: Structural Analysis of Electric Circuits and Consequences for MNA. Int. J. Circ. Theor. Appl. **28**(2), 131 – 162 (2000)
7. Bossavit, A.: Computational Electromagnetism. Academic Press, San Diego (1998)
8. Clemens, M.: Large Systems of Equations in a Discrete Electromagnetism: Formulations and Numerical Algorithms. IEE Proc Sci Meas Tech **152**(2), 50 – 72 (2005)
9. De Gersem, H. et al.: Field-circuit Coupled Models in Electromagnetic Simulation. JCAM **168**(1-2), 125 – 133 (2004)
10. Kettunen, L. et al.: Gauging in Whitney Spaces. IEEE Trans. Mag. **35**(3), 1466 – 1469 (1999)
11. Brenan, K.E. et al.: Numerical Solution of IVPs in DAEs. SIAM, Philadelphia, 1996
12. Arnold, M. et al.: Errors in the Numerical Solution of Nonlinear Differential-Algebraic Systems of Index 2. Martin-Luther-University Halle (1995)
13. Arnold, M., Günther, M.: Preconditioned Dynamic Iteration for Coupled Differential-Algebraic Systems. BIT **41**(1), 1 – 25 (2001)
14. Bartel, A.: Partial Differential-Algebraic Models in Chip Design - Thermal and Semiconductor Problems. Ph.D. thesis, TU Karlsruhe, VDI Verlag (2004)
15. Lange, E. et al.: A Circuit Coupling Method Based on a Temporary Linearization of the Energy Balance of the Finite Element Model. IEEE Trans. Mag. **44**(6), 838 – 841 (2008)
16. Zhou, P. et al.: A General Co-Simulation Approach for Coupled Field-Circuit Problems. IEEE Trans. Mag. **42**(4), 1051 – 1054 (2006)

# Part IV
# Mathematical and Computational Methods

# Introduction to Part IV

Vittorio Romano

This part is concerned with *mathematical and computational methods* in electrical engineering, including also multiobjective optimization and space-mapping methods. Theoretical results, novel approaches, and simulations cover some important issues mainly arising in the field of computational electromagnetics, including finite-element/volume discretization and the differential/integral formulation of Maxwell's equations, large interconnect structures, uncertainty quantification, and electron devices. Both the mathematical aspects and the applicative importance are outlined in this part, and as such may appeal to both engineers and theoretically-oriented readers.

The invited paper by Benderskaya et al. first gives a compact review of the electromagnetic (EM) fundamental relations, their classification in the static and quasi-static regimes, and the most general form of the computational model. Then, the authors restrict the investigation to the quasi-static case. The numerical methods arising from a spatial discretization with the finite-element method (FEM) or finite-integration technique (FIT) are critically discussed. For the resulting ordinary differential-equation system, one-step time-integration methods are considered and collected into three general classes: $\theta$-type, Runge–Kutta, and Rosenbrock. Comparisons between explicit and implicit methods are also made. Applications to conducting hollow spheres in a uniform transiently varying magnetic field complete the analysis.

The contribution by Lau et al. offers a novel staggered finite-volume time-domain method for Cartesian grids in order to solve Maxwell's equations in the integral form. The volume of control are brick-shaped and coordinates parallel, with the constraint that the field components are constant within each control volume and the stencil is as small as possible. After a heuristic overview, a matrix formulation is given and the Verlet leap-frog scheme is used for the time discretization. Assuming a homogeneous grid and a homogeneous material, the stability conditions are

Vittorio Romano
Department of Mathematics and Computer Science, University of Catania, viale A. Doria 6, 95125 Catania, Italy, e-mail: romano@dmi.unict.it

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_44,

investigated with a dispersion analysis based on plane-wave solutions. A numerical validation has been performed by simulating a rectangular resonator, homogeneously filled with a dielectric.

The paper by Herberthson addresses the problem of calculating the radar cross section of a perfect electric conducting surface, solving the field integral equations. In the case of surfaces homeomorphic to a sphere, the author applies the Hodge decomposition theorem on one-form in a compact set, reformulating the problem with the introduction of two scalar potentials. This procedure has the advantage of reducing the numerical effort with respect to the standard approaches known in the literature, because it leads to a smaller system of equations. As a counterpart the moment matrices are more costly to compute, but the problem can be overcome because the approach allows easy parallelization. Numerical examples in the case of a sphere are presented in order to give a preliminary assessment of the approach.

The article of the invited speaker Levadoux is concerned with a new family of source integral equations for the time-harmonic Maxwell scattering problems. Regardless of the composition of the obstacle — metallic, full dielectric, or coated with an impedance layer — a general methodology leading to the construction of some special equations, whose main feature are that they are well-conditioned, is presented. These equations do not contain spurious modes and can be viewed as compact perturbations of positive operators to which fast iterative schemes can be applied without any preconditioner. The proposed equations depend on the choice of an operator which is an approximation of the admittance of the diffracting body. The topic can be a challenge for future development.

The contribution by Harutyunyan et al. considers the problem of fast simulations of interconnect structures which consists in solving Maxwell's equations in the potential formulation. One of the main related difficulties is that the discretized equations result in large ill-conditioned matrices, thus making the use of efficient preconditioners necessary. The authors use the dual threshold incomplete factorization for improving the convergence rate of the BICGSTAB iterative solution algorithm. The efficiency of the approach is confirmed with simulations of an interconnect structure of micrometer size and an on-chip inductor with dimension of about one thousand microns.

The invited paper by Hesthaven et al. discusses the basic discontinuous Galerkin methods for computational electromagnetics. The benefits of such a method with respect to the widely used classical finite-difference time-domain method are highlighted: geometric flexibility, high-order accuracy, explicit time advancement, and very high parallel performance for large-scale applications. To validate the above considerations, a TM plane-wave scattering of a metallic cylinder and scattering from a metallic plate have been simulated. As an additional topic, the authors explore efficient probabilistic ways of dealing with uncertainty in EM problems, with application to the scattering of a plane wave from a perfect electric conducting sphere, having a random radius, and a rocket when the direction of the incident field is a random variable, uniformly distributed over a suitable interval.

The next contribution by van Belzen and Weiland presents the computation of empirical projection spaces by decomposing tensors that can be associated with the measured data. The notion of singular vales of a tensor is recalled along with some approximation properties and used for model-order reduction in the simulation of heat diffusion on a rectangular plate and two-dimensional incompressible fluid flow. The proposed method seems to reduce considerably the computing time with respect to the standard FEM approach.

The paper by Pechstein and Scheichl investigates the robustness of the finite-element tearing and interconnecting (FETI) methods which are efficient parallel domain-decomposition solvers for large-scale finite-element equations. Typically, in the problems using this approach, the degrees of freedom are very large, which means that direct solvers of the resulting systems are out of question and efficient preconditioners are necessary. The authors investigate the case of highly heterogeneous coefficients, giving theoretical condition-number bounds. The analytical estimates are confirmed by numerical tests, computing magnetic fields in cases where both large jumps and large variation in the reluctivity coefficient may arise.

The contribution by La Rosa et al. presents exact closure relations for the 8-moment and 9-moment models for charge transport in semiconductors obtained by using the maximum-entropy principle. These models improve the standard drift-diffusion and energy-transport ones that become inaccurate in the high-field regime, in particular when shrinking the typical dimension of the devices at nanoscale. The validity of proposed models is assessed by numerical simulations in the case of an $n^+$-n-$n^+$ silicon diode, comparing the results with those obtained solving the electron-transport equation by Monte Carlo and directly by a finite-difference scheme.

In the article by Jakobsson et al., multiobjective optimization is applied to antenna design. The used optimization algorithm is a novel response-surface method based on approximations with radial-basis functions, combined with CAD and mesh-generation software along with EM solvers. A key property of the algorithm is that the result is both a set of approximately Pareto-optimal solutions and an approximation of all objective functions as expansions in radial-basis functions. As an example, the optimization objective to transmit as much EM energy as possible through an antenna in a given frequency band minimizing, at the same time, the footprint of the antenna is studied.

In the next contribution, Lahaye and Drago solve an optimal doping-profile-control problem for semiconductors using the manifold-mapping technique. As coarse and fine approximations they employ the standard drift-diffusion and energy-transport models, respectively. One of the novelties is that the manifold-mapping technique is applied for the first time to a problem where the number of design variables depends of the finite-element-mesh points. One advantage is the possibility to optimize the energy-transport model without implementing an adjoint code and preserving computational efficiency at the same time. As an application, the problem of achieving current amplification in a ballistic diode by changing the doping profile is given.

In the last paper of Part IV, Simsek and Sengör propose a space-mapping-based surrogate method for solving inverse problems. Although the mapping between the coarse and the fine model is defined similarly to the linear inverse-mapping algorithm, parametric extraction is no longer necessary and the inverse coarse model, generated as a multilayer perceptron, is used instead of the coarse model. The efficiency of the method is addressed by considering the inverse problem to reconstruct a conducting cylinder and comparing the results with those obtained using conventional artificial neural networks, aggressive space-mapping methods, and the linear inverse-mapping algorithm.

# Numerical Time Integration in Quasistatic Computational Electromagnetics

Galina Benderskaya*, Wolfgang Ackermann, Oliver Sterz, and Thomas Weiland

*Invited speaker at the SCEE 2008 conference

**Abstract** Under certain conditions, electromagnetic time-domain modeling can be performed using the regimes of quasistatic approximations. The corresponding mathematical models represent then systems of first order ordinary differential equations or index 1 differential-algebraic equations. To resolve the time dependencies of the transient processes described by these equations, numerous time integration schemes can be employed. In this work, we give an overview of the mostly used time integration algorithms and discuss the main features, peculiarities and typical numerical difficulties associated with them. The materials presented in the paper are illustrated with corresponding numerical examples.

## 1 Introduction

Electromagnetic low frequency time-domain modeling can be performed using the regimes of quasistatic approximations or employing the full set of Maxwell's equations. Naturally, the usage of a quasistatic approximation introduces a modeling error. However, under certain conditions (see Sect. 2), this modeling error is very small or even negligible compared to the numerical discretization error. An advantage of quasistatic approximations is the possibility to reduce the corresponding computational costs considerably while obtaining at the same time the simulation results of comparable accuracy.

Galina Benderskaya, Oliver Sterz
CST - Computer Simulation Technology AG, Bad Nauheimer Straße 19, 64289 Darmstadt, Germany, e-mail: galina.benderskaya@cst.com, oliver.sterz@cst.com

Wolfgang Ackermann, Thomas Weiland
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, Schlossgartenstraße 8, 64289 Darmstadt, Germany, e-mail: ackermann@temf.tu-darmstadt.de, thomas.weiland@temf.tu-darmstadt.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_45,

The paper is organized as follows: in Sect. 2 we revisit Maxwell's equations as well as static and quasistatic approximations and state briefly the conditions under which these approximations can be employed. Quasistatic continuous electromagnetic field formulations represent (non)linear parabolic partial differential equations (PDEs) with time- and space-dependent operators. To resolve these dependencies, a number of different discretization techniques can be utilized. A first possibility is to discretize simultaneously in space and time using for example a Galerkin method. The classical method of lines (MOL) requires first a discretization in space, thus transforming a time-dependent PDE into first order ordinary differential equations (ODEs) or index 1 differential-algebraic equations (DAEs) which are further solved by an appropriate numerical time integration scheme. Finally, one can use Rothe's method where at first the time dependencies are resolved [1]. In this work, we follow the MOL approach where the indispensable spatial discretization is performed with e.g. the Finite Integration Technique (FIT) or the Finite Element Method (FEM) (see Sect. 3). The numerical integration of the obtained semi-discrete quasistatic electromagnetic formulations is a subject of the discussion in Sect. 4. Here, positive as well as negative aspects of different time integration methods are considered. Finally, illustrative examples are shown in Sect. 5.

## 2 Numerical Modeling

To construct an appropriate mathematical model for a quasistatic electromagnetic phenomenon, we start with a revision of Maxwell's equations and present a general classification of electromagnetic problems.

### 2.1 Fundamental Relations

The macroscopic behavior of electromagnetic fields is governed by Maxwell's equations. They represent a system of coupled partial differential equations and describe the relation between five vector fields and one scalar field:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1a}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \tag{1b}$$

$$\nabla \cdot \mathbf{D} = \rho \tag{1c}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{1d}$$

Equation (1a) known as *Faraday's law* establishes a correspondence between the electric field strength **E** and the magnetic flux density **B**. Equation (1b) is referred to as *Ampere's law* and sets up a link between the magnetic field strength **H**, the electric flux density **D** and the electric current density **J**. To complete the description of the electromagnetic fields, (1c) referred to as *Gauss law* and (1d) defining the inexistence of isolated magnetic charges are stated.

Maxwell's equations are completed by material relations which can be approximated using the following relations

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \approx \varepsilon(\mathbf{E})\,\mathbf{E} \tag{2a}$$

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \approx \mu(\mathbf{H})\,\mathbf{H} \tag{2b}$$

$$\mathbf{J} = \mathbf{J}_{\text{cond}} + \mathbf{J}_{\text{s}}\,. \tag{2c}$$

The vector fields **P** and **M** denote the polarization and the magnetization of the medium, respectively. Here, the parameters $\varepsilon_0$ and $\mu_0$ represent the permittivity and the permeability of free space, whereas $\varepsilon$ and $\mu$ specify the corresponding quantities of the medium. For linear homogeneous isotropic materials, the simplifications $\varepsilon(\mathbf{E}) = \varepsilon_{\text{const}}$ and $\mu(\mathbf{H}) = \mu_{\text{const}}$ hold. Equation (2c) expresses the superposition of different kind of currents: the conduction current density $\mathbf{J}_{\text{cond}}$ is defined according to the generalized Ohm's law as $\mathbf{J}_{\text{cond}} = \sigma\mathbf{E}$ with $\sigma$ being the conductivity of the medium and $\mathbf{J}_{\text{s}}$ denotes the imposed source current density.

The full electromagnetic spectrum ranging from statics to high frequency can such be modeled on a macroscopic scale with the help of Maxwell's equations (1) and approximated material relations (2). Furthermore, in case of a bounded computational domain, proper boundary conditions and initial data have to be specified in order to define a well-posed problem.

## 2.2 *Classification of Electromagnetic Problems*

The appearing time dependencies in Maxwell's equations can be handled in different ways which allows to establish a suitable classification of electromagnetic simulation regimes:

- *Static simulations* are performed if the time dependences in (1) can be completely omitted. The corresponding numerical formulations are then decoupled into electrostatic, magnetostatic and stationary current ones.
- *Quasistatic regimes* are justified if the rate of the dynamic changes in a model are so slow that the time delays stemming from the electromagnetic wave propagation can be neglected and if either the electric or the magnetic energy is dominant.
- *The full set of Maxwell's equations* has to be taken into account if neither $\partial\mathbf{B}/\partial t$ nor $\partial\mathbf{D}/\partial t$ can be neglected. This is the most general form of the computational model.

Within each simulation regime all time-dependent formulations can be stated in frequency or time domain. Here, we restrict ourselves to the quasistatic time-domain formulations.

The electroquasistatic (EQS) approximation is employed whenever the influence of the magnetic induction in the full set of Maxwell's equations can be omitted which simplifies (1a) to

$$\nabla \times \mathbf{E} = 0\,. \tag{3}$$

Equations (1) with Faradays law replaced by (3) and the appropriately applied boundary conditions allow to determine the electric field uniquely and express in this form the fundamental laws governing the EQS approximation regime.

The magnetoquasistatic (MQS) approximation is useful when the influence of the displacement current in (1) is negligible with respect to the conductive currents. This assumption results in the following simplified formulation:

$$\nabla \times \mathbf{H} = \mathbf{J}\,. \tag{4}$$

Equations (1) with Ampere's law replaced by (4) and the appropriately applied boundary conditions are sufficient to determine the magnetic field uniquely and represent the fundamental laws governing the MQS approximation regime.

It is well known (see [2] and [3] for heuristic arguments or [4] for a rigorous mathematical proof for the MQS approximation) that a necessary condition for the validity of any quasistatic approximation is $\mu\varepsilon\ell^2/\tau^2 \ll 1$, where the symbol $\ell$ represents a spatial model length and $\tau$ is equal to a characteristic time constant of an excitation signal. This means that $\tau_{\mathrm{em}} \ll \tau$, where $\tau_{\mathrm{em}}$ is the time of an electromagnetic wave propagating at the velocity $c = 1/\sqrt{\mu\varepsilon}$ over the distance $\ell$. (This corresponds to the condition $\lambda \gg \ell$ for a wave length $\lambda$ in frequency domain).

To determine the suitable regime uniquely, the influence of the conductivity $\sigma$ has to be taken into account additionally. This can be uniformly achieved if the electroquasistatic charge relaxation time $\tau_{\mathrm{e}} = \varepsilon/\sigma$ and the magnetoquasistatic current diffusion time $\tau_{\mathrm{m}} = \mu\sigma\ell^2$ are normalized to the characteristic wave propagation time $\tau_{\mathrm{em}}$ and the spatial length is normalized to the characteristic length $\ell^*$, where $\ell^* = 1/\sigma\eta$ and $\eta = \sqrt{\mu/\varepsilon}$. Then a graphical interpretation of the conditions that are necessary and sufficient for an appropriate classification of any electromagnetic problem can be established (Fig. 1) [2], [3].

## 3 Continuous and Discrete Quasistatic Formulations

A proper introduction of auxiliary scalar and vector fields allows to simplify the process of obtaining analytical or numerical solutions of Maxwell's equations.

Since for the EQS simulation regime (3) holds in the whole calculational domain, we can set

$$\mathbf{E} = -\nabla\varphi \tag{5}$$

where $\varphi$ is a scalar electric potential. Applying a divergence operator to (1b) and substituting (2a), (2c) and (5) into it, leads to the following PDE representing the continuous EQS formulation:

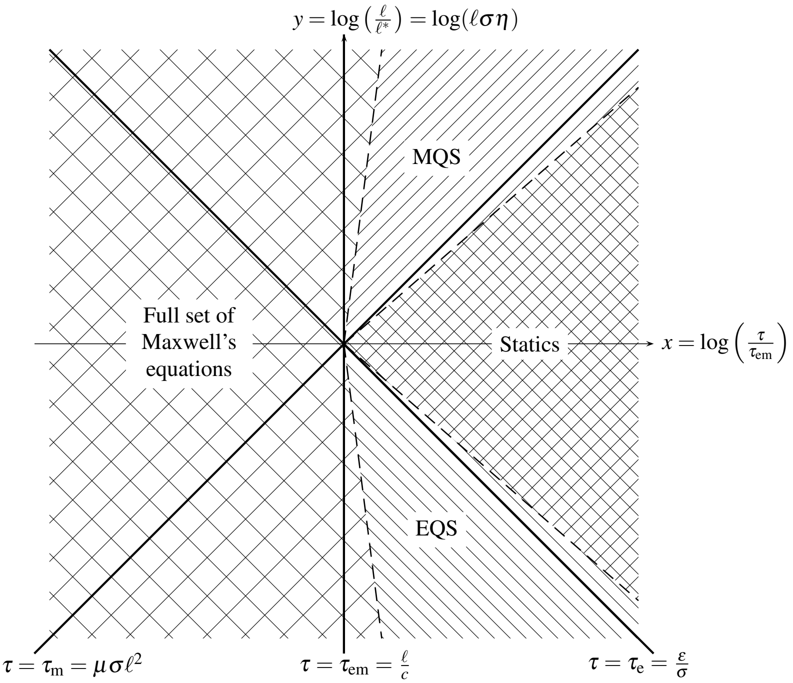

$y = \log\left(\frac{\ell}{\ell^*}\right) = \log(\ell\sigma\eta)$
MQS
Full set of Maxwell's equations
Statics
$x = \log\left(\frac{\tau}{\tau_{\mathrm{em}}}\right)$
EQS
$\tau = \tau_{\mathrm{m}} = \mu\sigma\ell^2$
$\tau = \tau_{\mathrm{em}} = \frac{\ell}{c}$
$\tau = \tau_{\mathrm{e}} = \frac{\varepsilon}{\sigma}$

**Fig. 1:** Graphical representation of the typical electromagnetic problem classification. Important characteristics are given by the model length $\ell$ with respect to the characteristic length $\ell^*$ and the excitation time constant $\tau$ with respect to the characteristic wave propagation time $\tau_{\mathrm{em}}$

$$\nabla\cdot\left(\frac{\partial}{\partial t}(\varepsilon\nabla\varphi)+\sigma\nabla\varphi\right)=0\,. \tag{6}$$

The main advantage of (6) is that it is expressed in terms of a scalar field $\varphi$ instead of the original vector field description.

For the MQS simulation regime various formulations are possible. Due to the solenoidal nature of field $\mathbf{B}$ we can set $\mathbf{B}=\nabla\times\mathbf{A}$, where $\mathbf{A}$ is a magnetic vector potential. Substituting this relation into (1a) results in

$$\mathbf{E}=-\frac{\partial\mathbf{A}}{\partial t}-\nabla\varphi \tag{7}$$

where $\varphi$ is an additional electric scalar potential. Here, we consider so-called temporal gauge setting $\varphi=0$ which simplifies (7) to $\mathbf{E}=-\partial\mathbf{A}/\partial t$. Plugging this into (1b) and using (2b) and (2c) yield

$$\sigma\frac{\partial\mathbf{A}}{\partial t}+\nabla\times(\frac{1}{\mu}\nabla\times\mathbf{A})=\mathbf{J}_s \tag{8}$$

for the MQS simulation regime.

Further on, (6) and (8) are discretized following the MOL approach where the indispensable spatial discretization can be performed with e.g., the Finite Integration Technique (FIT) [5] or the Finite Element Method (FEM) [6].

After a spatial discretization, the semi-discrete electro- and magnetoquasistatic formulations can be written in the following general form

$$\mathbf{M}\,\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t)+\mathbf{K}(t,\mathbf{x}(t))\,\mathbf{x}(t)=\mathbf{r}(t)\,. \tag{9}$$

For the EQS simulation regime, matrix $\mathbf{M}$ in (9) represents the discrete analogue of the continuous operator $\nabla\cdot(\varepsilon\nabla)$, matrix $\mathbf{K}$ stands for the continuous term $\nabla\cdot(\sigma\nabla)$ and the vector of unknowns $\mathbf{x}$ contains the introduced degrees of freedom.

For the MQS simulation regime, $\mathbf{M}$ identifies the discrete conductivity matrix, $\mathbf{K}$ represents the discrete counterpart of the continuous operator $\nabla\times(\frac{1}{\mu}\nabla\times)$ while the vector $\mathbf{r}(t)$ denotes the components of discrete current sources.

Independent of the chosen formulation, (9) can be in general integrated in time only by means of numerical techniques.

## 4 Numerical Time Integration

In case of a nontrivial invertible matrix $\mathbf{M}$, (9) describes a system of first order implicit differential equations [7]. Such systems can be integrated in time using so called *one-step* or *multi-step* numerical integration techniques. To update a solution at a new time point, a one-step time integration method uses the information about the solution only from the previous time instant [8]. In contrast to this, a solution update in a multi-step time integration method is based on several values of the solutions calculated at the previous time instants. In this work, we restrict ourselves to the one-step time integration schemes.

Conceptually, most of the one-step time integration methods can be classified using three main groups: $\theta$-type methods, Runge-Kutta (RK) methods and Rosenbrock-type methods which represent a special extension of RK-type methods.

### *4.1 $\theta$-Type Time Integration Schemes*

The $\theta$-time discretization scheme applied to (9) reads:

$$[\mathbf{M}+\Delta t\theta\mathbf{K}(\mathbf{x}_{n+1})]\mathbf{x}_{n+1}=\mathbf{M}\mathbf{x}_n-\Delta t\left[(1-\theta)\Big(\mathbf{K}(\mathbf{x}_n)\mathbf{x}_n-\mathbf{r}(t_n)\Big)-\theta\mathbf{r}(t_{n+1})\right]. \tag{10}$$

Different choices of the parameter $\theta$ lead to the following classical methods of numerical integration: $\theta=0$, specifies the forward Euler method, $\theta=1$ yields the backward Euler method, $\theta=1/2$ corresponds to the Crank-Nicolson method, and

$\theta = 2/3$ defines the Galerkin method. $\theta$-time discretization schemes are not originally equipped with a built-in error-controlled mechanism. Computations with the variable time step lengths can however, be performed by means of additionally constructed step size controllers. From (10) it is evident that in the presence of the nontrivial invertible matrix $\mathbf{M}$ every $\theta$-method requires the solution of a (non)linear system of equations.

## 4.2 Runge-Kutta Time Integration Methods

Despite the huge variety of the RK time integration methods, each of them can be compactly defined with the help of the so-called Butcher table [8]:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \\ \hline & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_s \end{array} . \tag{11}$$

In this table, the components of vector $\mathbf{c}$ are called the *abscissae*. Vector $\mathbf{b}$ represents a weight vector and $\mathbf{A}$ is a matrix specifying the method. The introduction of the second weight vector $\hat{\mathbf{b}}$ is necessary to construct an *embedded* RK method.[1] The integer value of $s$ defines the number of stages in the RK method.

The RK methods may be sorted according to the structure of the matrix $\mathbf{A} = [a_{ij}]$ in the Butcher table (11). The classical *explicit* methods are the methods where $\mathbf{A}$ is a lower triangular matrix with vanishing coefficients on the main diagonal. If this condition is not satisfied, the methods are called *implicit* RK methods (IRK). *Diagonally implicit* RK (DIRK) methods (if $a_{ij} = 0$ for $j > i$) represent a special case of IRK methods. Additionally, if all diagonal coefficients of the DIRK method are the same, the method is referred to as the *singly diagonally implicit* RK (SDIRK) method.

The application of the $s$-stage RK method defined by (11) to (9) with a possibly singular matrix $\mathbf{M}$ and possibly nonlinear matrix $\mathbf{K}$ leads to the following numerical scheme:

$$\mathbf{M}\mathbf{k}_i = \Delta t\Big(\mathbf{r}(t_n + c_i\Delta t) - \mathbf{K}(\mathbf{x}_n + \sum_{j=1}^{s} a_{ij}\mathbf{k}_j)(\mathbf{x}_n + \sum_{j=1}^{s} a_{ij}\mathbf{k}_j)\Big), \tag{12a}$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \sum_{i=1}^{s} b_i\mathbf{k}_i, \quad i = 1, \dots, s. \tag{12b}$$

[1] Embedded RK methods are discussed below in a separate subsection.

The values $\mathbf{k}_i, i = 1,\dots,s$ are called *stage derivatives*. From (12) it follows that for each time step of IRK method, a (non)linear system of dimension $ms$, where $m$ is the number of degrees of freedom in (9) has to be solved. In contrast to IRK methods, explicit RK methods are computationally much cheaper since for each stage in (12) they do not require a solution of the (non)linear system of equations.

### 4.2.1 Diagonally Implicit RK Methods

In the case of DIRK methods, all stage derivatives can be found successively requiring for each stage the solution of the (non)linear system with only $m$ unknowns [9]:

$$\Big(\frac{\mathbf{M}}{\Delta t a_{ii}} + \mathbf{K}(\mathbf{Y}_i)\Big)\mathbf{Y}_i = \mathbf{r}(t_n + c_i\Delta t) + \frac{\mathbf{M}}{\Delta t a_{ii}}\big(\mathbf{x}_n + \sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j\big)\,, \tag{13a}$$

$$\mathbf{k}_i = \frac{1}{a_{ii}}\,\big(\mathbf{Y}_i - \mathbf{x}_n - \sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j\big)\,, \tag{13b}$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \sum_{i=1}^{s} b_i\mathbf{k}_i\,. \tag{13c}$$

In case of linear quasistatic formulation (9), the SDIRK method does not need any reassembling of the system matrix in (13a) for the computation of stage values within one integration step. This allows considerable redaction of computational time.

### 4.2.2 Rosenbrock-Type Methods

In spite of the fact that DIRK methods allow the process of numerical solution of the nonlinear systems arising during the process of the numerical integration of (9) to be accelerated significantly, they still do not make it possible to avoid the solution of the nonlinear equations completely. Rosenbrock methods belong to the group of numerical schemes that circumvent the solution of the nonlinear systems of equations. This step is replaced here by a solution of a sequence of linear systems [7], [10].

An $s$-stage Rosenbrock method can be derived from (13) by application of one Newton iteration to each stage of DIRK method using the start values $\mathbf{k}_i^{(0)} = 0$. The upper index specifies the number of the nonlinear iteration and the additional set of coefficients $\mathbf{L} = \{l_{ij}\}, i = 1,\dots,s, j = 1,\dots,i$:

$$\mathbf{M}\mathbf{k}_i = \Delta t\Big(\mathbf{r}(t_n{+}c_i\Delta t) - \mathbf{K}(\mathbf{x}_n{+}\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j)(\mathbf{x}_n{+}\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j) - \mathbf{J}\sum_{j=1}^{i} l_{ij}\mathbf{k}_j\Big), \tag{14a}$$

$$\mathbf{x}_{n+1} = x_n + \sum_{i=1}^{s} b_i\mathbf{k}_i, \quad i = 1,\dots,s. \tag{14b}$$

where $\mathbf{J} = \partial\big(\mathbf{r}(t_n) - \mathbf{K}(\mathbf{x}_n, t_n)\mathbf{x}_n\big)/\partial\mathbf{x}_n$ represents the Jacobian matrix for (9).

According to (14), for each stage a linear system of equations with matrix $(\mathbf{M} + \Delta t\, \mathbf{J}\, l_{ii})$ for a vector of unknowns $\mathbf{k}_i$ has to be solved. Additional computational costs arise due to the necessity to calculate the matrix-vector multiplication $\mathbf{J}\sum l_{ij}\mathbf{k}_j$. This can be avoided, however, according to the technique presented in [7].

### 4.2.3 Adaptive Stepsize Control

RK methods can be naturally equipped with a technique allowing an adaptive time stepping control. In addition to a main solution $\mathbf{x}^{(p)}$ of a given order $p$ obtained with a weight vector $\mathbf{b}$, a so-called *embedded* solution $\mathbf{x}^{(\hat{p})}$ of a lower order $\hat{p}$ can be calculated. The main and the embedded methods share the same coefficient matrix $\mathbf{A}$. Consequently, the computation of the embedded solution does not require additionally the solution of a (non)linear system, but just a new superposition of the already calculated stage derivatives using merely different weight factors.

The difference between the main and the embedded solution defines the error vector $\mathbf{y} = \mathbf{x}^{(p)} - \mathbf{x}^{(\hat{p})} = \sum_{i=1}^{s}(b_i - \hat{b}_i)\mathbf{k}_i$. The norm of this vector $\|\mathbf{y}\|_{\text{err}}$ is then employed to estimate the local error for a given time step. In the literature, one can find different norms allowing the estimation of the error $\|\mathbf{y}\|_{\text{err}}$ [3, 8, 11, 12].

The first task of the *step-size controller* is to make a decision whether the last integration step has to be repeated with a smaller time step length or the simulation can be further advanced in time. The solution is rejected if $\|\mathbf{y}\|_{\text{err}} > \mu\,\varepsilon_{\text{tol}}$ holds true and a new attempt is made with a smaller step size; otherwise the time step is accepted. In this scheme, $\mu$ is an accelerating factor usually taken as 1.2 [12].

Secondly, the step-size controller calculates the length of a next time step using the formula

$$\Delta t_{n+1} = \rho \left( \frac{\varepsilon_{\text{tol}}}{\|\mathbf{y}\|_{\text{err}}} \right)^{1/(\hat{p}+1)} \Delta t_n \tag{15}$$

where $\rho$ denotes a safety factor that is usually set to 0.9 [10]. According to (15), the newly calculated time step length is decreased if the step before is rejected, otherwise it is increased.

## *4.3 Implicit Versus Explicit Methods*

We have listed so far a variety of the currently available time integration schemes. To decide which of them is the most appropriate for quasistatic electromagnetic simulations, let us review a number of issues associated with the differential equations and one-step time integration methods.

Differential equations can be classified as *stiff* and *non-stiff* ones. In the literature one can find a lot of attempts to come out with a satisfactory definition of the term "stiffness" [7, 13, 14]. Here, we adopt the following pragmatic definition for stiffness: a given dynamic problem describing some physical process is called stiff when it is more efficient to use an implicit time integration method than an explicit one for the time interval of interest. The reason for this is a boundness of a stability domain of any explicit time integration method [7]. On the contrary, stability domains of the implicit methods are unbounded and, consequently, the stability requirement does not put any limitation on the choice of the length of the time step. In other words, in general stiffness means that when using an explicit method a small time step is necessary due to stability reasons and not due to approximation reasons.

One reason for stiffness is that the components of the dynamic system may have incomparable characteristic time constants [14]. Since in the EQS formulation (6) the permittivity $\varepsilon$ never vanishes, the matrix $\mathbf{M}$ corresponding to $\nabla \cdot (\varepsilon \nabla)$ is always regular. Consequently, the discrete EQS formulation represents a (non)linear ODE system which is in practice not stiff, since permittivity values for different materials do not diverse too much. Under condition that obtaining an inverse to matrix $\mathbf{M}$ is computationally cheap, such systems can be integrated using any explicit time integration method.

Stiffness also appears in combination with the solution of DAEs. The DAE system incorporates two type of equations - differential and algebraic equations and are considered to be extremely stiff [7]. In MQS formulation (8), the conductivity $\sigma$ obviously equals zero exactly for those parts of the model which are filled with non-conductive materials. In this case, the matrix $\mathbf{M}$ is singular and consequently the discrete MQS formulation represents a (non)linear DAE system of index 1 [3], [15]. In this case, only implicit time integrators have to be employed.

Finally, for parabolic partial differential equations the stability limit $\Delta t \leq \mathscr{O}(\Delta x^2)$ on the size of the time step in the explicit time integration methods, implies that an enormous amount of time steps is necessary to follow the evolution of the time process if one reduces the spatial resolution $\Delta x$ to improve the overall accuracy of the numerical solution [16].

## 5 Numerical Simulations

Our first application example is the TEAM 11 workshop problem that demonstrates the calculation of a conducting hollow sphere in a uniform transiently varying magnetic field (Fig. 2(a)). The external field is instantaneously switched on to a uniformly distributed magnetic flux density $\mathbf{B}_0 = 1\ \mathrm{T}\ \mathbf{e}_z$ disturbed locally by the transient compensation due to induced eddy currents. The detailed description of the specified geometry together with the definition of the transiently varying field excitation as well as the analytical solution data for selected physical quantities can be found in [17].

For various spatial mesh resolutions, this problem is integrated in time employing the SDIRK method equipped with adaptive time stepping control. Here, the accuracy of the linear solver is set to $10^{-4}$ while the relative error tolerance for the adaptive time stepping control is chosen to be $10^{-5}$. These settings translates into 19 accepted steps versus 3 rejected ones for the finest spatial mesh resolution. For the coarsest spatial resolution, all 6 performed time steps are accepted. Figure 3 illustrates the simulation results for the monitored magnetic flux density component $B_z$ at the center of a conducting sphere for different simulations runs.

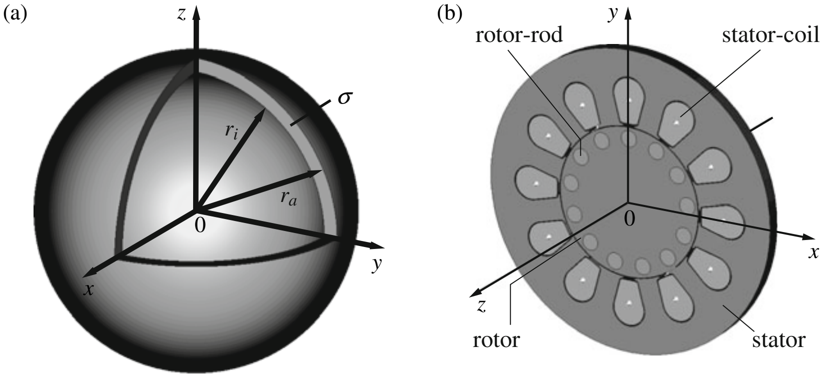


**Fig. 2: a** Geometrical model of a hollow conducting sphere ($\sigma = 5 \times 10^8$ S/m) exposed to a spatially homogeneous but transiently varying magnetic field. The inner radius is given by $r_i = 5$ cm whereas the outer radius is set to $r_a = 5.5$ cm. **b** Sliced model of an asynchronous motor including the three-phase excitation coils of the stator together with the cylindrical cage inductor elements of the rotor

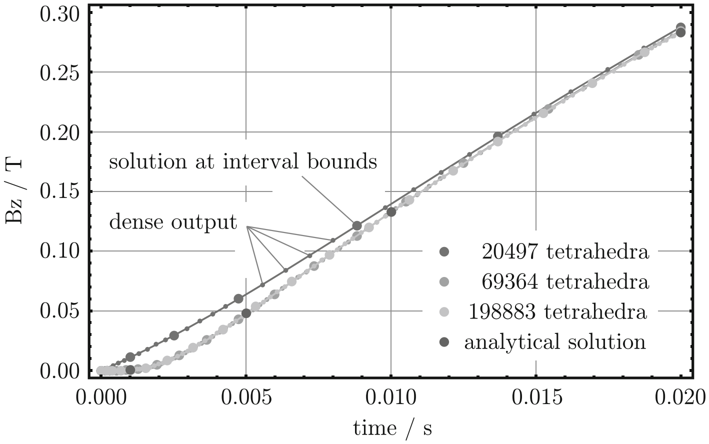


**Fig. 3:** Simulated time dependency of the magnetic flux density component $B_z$ in the center of the hollow sphere for various spatial resolutions. For comparison reasons, only the time interval where analytical data [17] are available is considered. In addition to the calculated solution at the interval bounds, also intermediate values obtained via the dense output interpolation are shown

Analogous simulations with the SDIRK method are also performed for a sliced model of an asynchronous motor (Fig. 2(b)). For a computational model consisting of approximately 100000 unknowns, the same solver settings result in 405 accepted and 122 rejected time steps. However, for the asynchronous machine model no analytical solution is available so that the reliability of the method can only be checked via classical convergence studies.

**Acknowledgements** The authors would like to thank Franz Hirtenfelder (CST - Computer Simulation Technology AG) for providing a detailed computational model of the applied asynchronous induction machine.

## References

1. Lang, J.: Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Springer, Berlin (2001)
2. Haus, H.A., Melcher, J.R.: Electromagnetic Fields and Energy. Prentice Hall, Englewood Cliffs (1989)
3. Benderskaya, G.: Numerical methods for transient field-circuit coupled simulations based on the Finite Integration Technique and a mixed circuit formulation. Ph.D. thesis, Darmstadt University of Technology, Darmstadt (2007). URL http://tuprints.ulb.tu-darmstadt.de/epda/000807/
4. Schmidt, K., Sterz, O., Hiptmair, R.: Estimating the eddy-current modeling error. IEEE Trans. Magn., **44(6)**, 686–689, (2008)
5. Weiland, T.: A discretization method for the solution of Maxwell's equations for six-component fields. AEÜ, **31**, 116–120 (1977)
6. Nédélec, J. C.: Mixed finite elements in $R^3$. Numer. Math. **35**, 315–341, (1980)
7. Hairer, E., Wanner., G.: Solving Ordinary Differential Equations. Stiff and Differential-Algebraic Problems. Springer-Verlag, Berlin (2002)
8. Hairer, E., Nørsett, S. P., Wanner, G.: Solving Ordinary Differential Equations I. Nonstiff Problems. Springer-Verlag, Berlin, (2000)
9. Nicolet, A., Delincé, F.: Implicit Runge-Kutta methods for transient magnetic field computation. IEEE Trans. Magn., **32(3)**, 1405–1408, (1996)
10. Lang, J., Two-dimensional fully-adaptive solutions of reaction-diffusion equations. Appl. Numer. Math., **18**, 223–240, (1995)
11. Wang, H., Taylor, S., Simkin, J., Biddlecombe, C., Trowbridge B.: An adaptive-step time integration method applied to transient magnetic field problems. IEEE Trans. Magn., **37(5)**, 3478–3481, (2001)
12. Gustafsson, K.: Control-theoretic techniques for stepsize selection in implicit Runge-Kutta methods. ACM Trans. Math. Soft., **20(4)**, 496–517, (1994)
13. Lambert, J.: Numerical Methods for Ordinary Differential Systems. John Wiley and Sons. Chichester (1991)
14. Kahaner. D., Moler, C., Nash, S.: Numerical Methods and Software. Prentice Hall, (1989)
15. Clemens, M., Weiland, T.: Numerical algorithms for the FDiTD and FDFD simulation of slowly varying electromagnetic fields. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, **12**, 3–22 (1999)
16. Morton K. W., Mayers, D. F.: Numerical Solutions of Partial Differential Equations. Cambridge University Press, Cambridge, (2005)
17. Emson, C.R.I.: Summary of results for hollow conducting sphere in uniform transiently varying magnetic field (Problem 11). COMPEL, **9(3)**, 191–203, (1990).

# A Novel Staggered Finite Volume Time Domain Method

Thomas Lau, Erion Gjonaj, and Thomas Weiland

**Abstract** In this work a novel, staggered finite volume time domain method for Cartesian grids is presented, analyzed and validated. An important characteristic of the method is the use of a rather unorthodox staggering of the degrees of freedom.

## 1 Introduction

In the majority of finite volume time-domain (FVTD) methods the electric degrees of freedom (DOF) and the magnetic DOF are co-located on the same spatial position [1]. Unfortunately, compared to the classical Yee scheme [2], this co-location reduces the accuracy of the FVTD method. Hence, one approach to increase the accuracy of the FVTD method, motivated by the Yee scheme, is to stagger the DOFs. However, the staggering can be performed in three different ways: first, staggering the electric and magnetic fields, only. Second, staggering the vectorial component of the DOFs, only. Third, combining both approaches and staggering the electric and magnetic DOFs and each of their vectorial components. The last approach corresponds to the staggering applied in the Yee scheme.

In this work the authors restrict their investigation to the second strategy, and only consider a specific staggering (fig. 1) of the vectorial components of the DOFs while keeping each vectorial components of the electric and magnetic field in the same place, respectively. The FVTD obtained through this staggering is denoted as SFVTD in the rest of the work.

Thomas Lau, Erion Gjonaj, Thomas Weiland
Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, 64289 Darmstadt, Germany,
e-mail: lau@temf.tu-darmstadt.de, gjonaj@temf.tu-darmstadt.de, weiland@temf.tu-darmstadt.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_46,

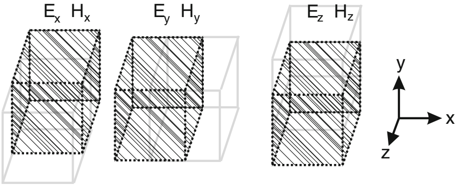


**Fig. 1:** Staggering of the field DOFs. The DOFs are defined, according to the vectorial component they represent, on three different dual volumes (*shaded*) with respect to the primary grid cell (*gray*)

## 2 Integral Formulation of Maxwell's Equations

Starting point for constructing the SFVTD method is the volume integral formulation of Maxwell's equations in terms of the electric field, **E**, and the magnetic field, **H**, neglecting currents and charges, on a finite domain $\Omega$. For an arbitrary volume $V$ with normal vector **n** in $\Omega$ the following two integral relationships must be fulfilled:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \varepsilon \mathbf{E}\,\mathrm{d}V = \int_{\partial V} \mathbf{n}\times\mathbf{H}\,\mathrm{d}A = \int_{\partial V} \mathrm{d}\mathbf{A}\times\mathbf{H}, \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \mu \mathbf{H}\,\mathrm{d}V = -\int_{\partial V} \mathbf{n}\times\mathbf{E}\,\mathrm{d}A = -\int_{\partial V} \mathrm{d}\mathbf{A}\times\mathbf{E}. \tag{2}$$

This formulation of Maxwell's equations is mathematically equivalent to the differential formulation (see [3], pp. 54). The different materials in $\Omega$ are characterized by their permeability, $\mu$, and permittivity, $\varepsilon$, respectively. It is assumed that the materials are linear, nondispersive and isotropic and therefore $\varepsilon$ and $\mu$ are characterized by scalar functions on $\Omega$.

## 3 Spatial Discretization

The basic idea of the SFVTD method is to construct an approximate solution of (1) and (2) in terms of volume averages of the electromagnetic fields over a finite set of so called control volumes. This is achieved by enforcing a discrete version of (1) and (2) on the control volumes. In the following, the discretization of (1), applying this idea, is sketched.

The first step in discretizing (1) consists of restricting the arbitrary volumes $V$ in (1) to a set of brick shaped and coordinate parallel control volumes $V_x, V_y$ and $V_z$, which are different for each vectorial component[1] (see fig. 1). In the following, the

[1] At this point it should be noted that for the SFVTD only the relative position of the staggered volumes to each other is relevant. Therefore the choice of the volumes $V_x, V_y$ and $V_z$ is not unique.

construction of the SFVTD scheme is based on the demands that the field components are constant inside their control volumes, respectively, and that the stencil of the resulting scheme should be as small as possible.

For example, the x-component of (1) reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{V_x} \varepsilon E_x \,\mathrm{d}V = \int\limits_{\partial V_x} \mathrm{dA}_y H_z - \int\limits_{\partial V_x} \mathrm{dA}_z H_y. \tag{3}$$

The physical interpretation of (3) is that the time evolution of the volume integral of $\varepsilon E_x$ over the control volume $V_x$ is solely determined by the fluxes generated by $H_z$ and $H_y$ through the $y$- and $z$-face of $V_x$. Analog expressions are derived for the $y$- and $z$-components of (1). However, for brevity's sake the following discretization steps are illustrated for the $x$-component, only.

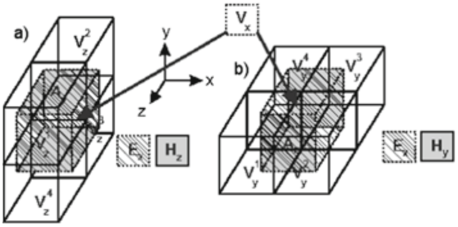


**Fig. 2:** The time evolution of the integral of $E_x$ over its control volumes (*dashed cell*) solely depends on the fluxes through the $y$- and $z$-faces, respectively. Figure (**a**) shows the flux through the $y$-face (*shaded*) which is uniquely defined by the values of $H_z$ on its control volumes (*solid cells*). Figure (**b**) shows the flux through the $z$-face (*shaded*) of the same cell which is uniquely defined by the values of $H_y$ on its control volumes (*solid cells*)

An inspection of (3) shows that neither the volume integral on the left hand side nor the flux integrals on the right hand side can be expressed exactly by volume averages of the electric or magnetic field. Therefore, approximations are inevitably for the discretization of (3).

First, the volume integral is approximated by the volume average of $E_x$, taking an inhomogeneous permittivity $\varepsilon$ inside $V_x$ into account, by

$$\int\limits_{V_x} \varepsilon E_x \,\mathrm{d}V \approx \int\limits_{V_x} \varepsilon \,\mathrm{d}V \frac{1}{V_x}\int\limits_{V_x} E_x \,\mathrm{d}V. \tag{4}$$

Second, the flux integrals on the right hand side of (3) are approximated by volume averages of $H_y$ and $H_z$. In order to link volume averages with fluxes, which are surface integrals, a constant field inside each control volume is assumed. For the further approximation of the fluxes, the positions of the different control volumes $V_x$, $V_y$ and $V_z$, with respect to each other, have to be specified. A formal definition of the control volumes is postponed to section 4, a schematic drawing of the staggering is shown in fig. 2. The staggering is chosen in such a way, that each flux face is immersed in exactly two control volumes. On the left hand side, the shaded $y$-surfaces are part of the control volumes $V_z^1$, $V_z^2$ and $V_z^3$, $V_z^4$ of $H_z$. On the right hand side, the shaded $z$-surfaces are part of the control volumes $V_y^1, V_y^2$ and $V_y^3, V_y^4$ of $H_y$. Thus, the left surface integrals in (3) is approximated by

$$\int\limits_{\partial V_x} \mathrm{dA}_y H_z \approx \frac{A_y}{2V_z^1}\int\limits_{V_z^1} H_z\,\mathrm{d}V + \frac{A_y}{2V_z^2}\int\limits_{V_z^2} H_z\,\mathrm{d}V - \frac{A_y}{2V_z^3}\int\limits_{V_z^3} H_z\,\mathrm{d}V - \frac{A_y}{2V_z^4}\int\limits_{V_z^4} H_z\,\mathrm{d}V$$

and an analogue expression is obtained for the right surface integral. This concludes the discretization of (3).

The preceding discretization steps fix the location of the control volumes for **E**, **H** and renders the fields to be piecewise constant. Thus, the discrete version of Faraday's and Ampere's law is uniquely obtained by steps analogue to the presented discretization steps. Perfect electric conducting (PEC) boundary conditions are considered by setting the flux generated by **E** of a PEC face for a control volume to zero.

## 4 Matrix Formulation of the SFVTD Method

In the following, the discretization approach sketched in the previous section is established formally. All quantities appearing with an index $\alpha$ are meant to be defined for $\alpha \in \{x,y,z\}$.

First, a primary cartesian grid, $G$, and the grid translation operators $\mathbf{T}_x$, $\mathbf{T}_y$ and $\mathbf{T}_z$ are defined. Then, the $x$-, $y$- and $z$-edges in $G$ are associated with the diagonal matrices $\mathbf{L}_x$, $\mathbf{L}_y$ and $\mathbf{L}_z$. In order to account for the different control volumes, also matrices for the dual edge length $\tilde{\mathbf{L}}_x$, $\tilde{\mathbf{L}}_y$ and $\tilde{\mathbf{L}}_z$ are defined

$$\tilde{\mathbf{L}}_\alpha = 0.5\left(1+\mathbf{T}_\alpha^{-1}\right)\mathbf{L}_\alpha.$$

Hereafter, diagonal matrices $\mathbf{A}_x$, $\mathbf{A}_y$ and $\mathbf{A}_z$ for the areas of the $x$-, $y$- and $z$-faces in $G$ are defined. Finally, the matrices for the control volumes $\mathbf{V}_x$, $\mathbf{V}_y$,$\mathbf{V}_z$ and $\tilde{\mathbf{V}}$, are defined by

$$\mathbf{V}_x = \mathbf{L}_x \otimes \mathbf{L}_y \otimes \tilde{\mathbf{L}}_z, \qquad \mathbf{V}_y = \tilde{\mathbf{L}}_x \otimes \mathbf{L}_y \otimes \mathbf{L}_z, \mathbf{V}_z = \mathbf{L}_x \otimes \tilde{\mathbf{L}}_y \otimes \mathbf{L}_z.$$

In the following, the components of the previously defined matrices will be used in two different contexts. First, as measure for length, areas and volumes and second, indicating the different integration domains.

The discrete electric field, $\mathbf{e}_\alpha$, and magnetic field, $\mathbf{h}_\alpha$ are defined by control volume averages according to

$$[V\alpha]_{i,j,k}\,[\mathbf{e}\alpha]_{i,j,k} = \int\limits_{[V\alpha]_{i,j,k}} \mathbf{E}_\alpha\,\mathrm{d}V, \quad [V\alpha]_{i,j,k}\,[\mathbf{h}\alpha]_{i,j,k} = \int\limits_{[V\alpha]_{i,j,k}} \mathbf{H}_\alpha\,\mathrm{d}V.$$

and are grouped into vectors of the form $\mathbf{e} = \left(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\right)^T$ and $\mathbf{h} = \left(\mathbf{h}_x, \mathbf{h}_y, \mathbf{h}_z\right)^T$.

Hereafter, diagonal material matrices $\mathbf{M}_\varepsilon$ and $\mathbf{M}_\mu$ are defined by

$$[\mathbf{M}_{\varepsilon,\alpha}]_{ijk} = \mathrm{diag}( \int\limits_{[V\alpha]_{ijk}} \varepsilon \,\mathrm{d}V), \quad \left[\mathbf{M}_{\mu,\alpha}\right]_{ijk} = \mathrm{diag}( \int\limits_{[V\alpha]_{ijk}} \mu \,\mathrm{d}V),$$

and are arranged into matrices of the form $\mathbf{M}_\varepsilon = \mathrm{diag}\left(\mathbf{M}_{\varepsilon,x}, \mathbf{M}_{\varepsilon,y}, \mathbf{M}_{\varepsilon,z}\right)$ and $\mathbf{M}_\mu = \mathrm{diag}\left(\mathbf{M}_{\mu,x}, \mathbf{M}_{\mu,y}, \mathbf{M}_{\mu,z}\right)$.

The component wise fluxes for the discretization of the transient equations (1) and (2) are expressed with the help of the matrices

$$\mathbf{P}_x = 0.5\mathbf{A}_x(\mathbf{1} - \mathbf{T}_x^{-1})(\mathbf{1} + \mathbf{T}_y), \qquad \mathbf{P}_y = 0.5\mathbf{A}_y(\mathbf{1} - \mathbf{T}_y^{-1})(\mathbf{1} + \mathbf{T}_z),$$
$$\mathbf{P}_z = 0.5\mathbf{A}_z(\mathbf{1} - \mathbf{T}_z^{-1})(\mathbf{1} + \mathbf{T}_x).$$

The fluxes are arranged into the discrete curl matrix $\mathbf{C}$

$$\mathbf{C} = \begin{pmatrix} \mathbf{0} & \mathbf{P}_z & \mathbf{P}_y^T \\ \mathbf{P}_z^T & \mathbf{0} & \mathbf{P}_x \\ \mathbf{P}_y & \mathbf{P}_x^T & \mathbf{0} \end{pmatrix}, \quad \mathbf{C} = \mathbf{C}^T.$$

Finally, the SFVTD discretized version of Ampere's and Faraday's law is established

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{M}_\varepsilon \mathbf{e} = \mathbf{C}\mathbf{h}, \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{M}_\mu \mathbf{h} = -\mathbf{C}\mathbf{e} \tag{5}$$

# 5 Time Discretization

The discrete Faraday's and Ampere's law form a system of ordinary differential equations (ODEs). For their numerical integration the time is discretized with time step $\Delta t$ by $t^{(n)} = t^{(0)} + n\Delta t$. Denoting with $\mathbf{e}^{(n)}$ and $\mathbf{h}^{(n)}$ the discrete values of the electric- and magnetic field strength sampled at the time instance, $t^{(n)}$, (5) is discretized by a Verlet-Leap-Frog (VLF) time integrator [4]

$$\mathbf{h}^{(*)} = \mathbf{h}^{(n)} - \frac{\Delta t}{2}\mathbf{M}_\mu^{-1}\mathbf{C}\mathbf{e}^{(n)}, \qquad \mathbf{e}^{(n+1)} = \mathbf{e}^{(n)} + \Delta t \mathbf{M}_\varepsilon^{-1}\mathbf{C}\mathbf{n}^{(*)},$$
$$\mathbf{h}^{(n+1)} = \mathbf{h}^{(*)} - \frac{\Delta t}{2}\mathbf{M}_\mu^{-1}\mathbf{C}\mathbf{e}^{(n+1)}.$$

## 6 Dispersion Analysis

Assuming a homogeneous grid $(x_i = i\Delta, y_j = j\Delta, z_k = k\Delta)$ with grid spacing $\Delta$ and a homogeneous material with the local velocity of light, $c$, a plane wave ansatz

$$[(\mathbf{e},\mathbf{h})]_{ijk}^{(n)} = (\mathbf{e}_0,\mathbf{h}_0)\exp(\mathrm{j}\,(\tilde{\omega}n - \beta_x i - \beta_y j - \beta_z k)),$$
$$\tilde{\omega} = \Delta t\omega, \beta_x = k_x\Delta, \beta_y = k_y\Delta, \beta_z = k_z\Delta,$$

is made [5]. The numerical dispersion relation for the propagating modes is

$$\sin^2(\frac{\tilde{\omega}}{2}) = \sigma^2(K_{\text{Yee}} - \delta K), \tag{6}$$
$$K_{\text{Yee}} = \sum_{\gamma}\sin^2(\frac{\beta_\gamma}{2}), \quad \delta K = \frac{1}{2}\sum_{\gamma\neq\delta}\sin^2(\frac{\beta_\gamma}{2})\sin^2(\frac{\beta_\delta}{2}),$$

with the Courant number $\sigma = c\Delta t/\Delta$. For waves propagating along the $\alpha$-coordinate axis, the dispersion relation (6) reduces to

$$\sin^2(\frac{\tilde{\omega}}{2}) = \sigma^2\sin^2(\frac{\beta_\alpha}{2}), \tag{7}$$

and therefore the SFVTD method has no numerical dispersion along the coordinate axes for $\sigma = 1$. Especially, it can be shown from the condition $\left|\sin^2(\frac{\tilde{\omega}}{2})\right| \leq 1$ and (6) that the stability limit, $\sigma_{\max}$, of the SFVTD method is equal to one.

For an analysis of the wave propagation properties [5], the numerical phase velocity, $v_p$, is calculated from (6). For the following discussion, it is convenient to analyze the relative error in the numerical phase velocity, $\mathrm{Er}_p$, with respect to their physically exact value $c$. Additionally, in order to have an error measure independent of the direction $\mathrm{Er}_p$ is averaged over all propagation directions.

$$\mathrm{Er}_p = \left|\frac{v_p - c}{c}\right|, \quad \mathrm{Er}_p^{\mathrm{av}} = \frac{1}{4\pi}\int \mathrm{d}\Omega\,\mathrm{Er}_p. \tag{8}$$

Defining the points per wavelength (PPW) by PPW $= 2\pi/\beta$, the error in the phase velocity of the SFVTD scheme is compared to that of Yee's scheme at the stability limit of both schemes, respectively.

Figure 3 (a) shows the anisotropic behavior of $\mathrm{Er}_p$ for 3 PPW. Yee's scheme approaches the maximal error along the x-, y- and z-axes. In contrast to this, the SFVTD scheme approaches its maximal error along the space diagonals. Especially, the magnitude of the maximal error $\mathrm{Er}_p$ is significant smaller in the SFVTD scheme in comparison to Yee's scheme.

The behavior for other PPW is similar to the behavior presented in fig. 3 (a). Therefore, the phase velocity error averaged over all angles, $\mathrm{Er}_p^{\mathrm{av}}$ for the SFVTD scheme, shown in 3 (b), is for all PPW smaller than that of Yee's scheme. For $\mathrm{Er}_p^{\mathrm{av}}$

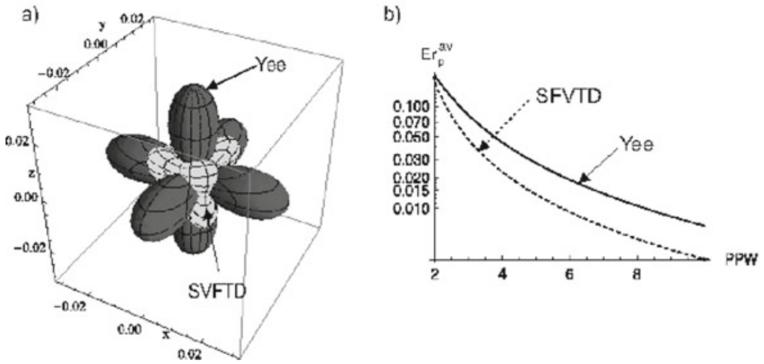


**Fig. 3:** Analysis of the relative error in the phase velocity. In (**a**) the directional dependency of the relative error in the phase velocity for 3 PPW for Yee's and the SFVTD method is shown. Figure (**b**) shows the dependence of the direction averaged relative error in the phase velocity for Yee's (*solid curve*) and the SFVTD scheme (*dashed curve*)

less than 1% Yee's scheme needs roughly 10 PPW in contrast to 6 PPW for the SFVTD method.

# 7 Numerical Validation

For the numerical validation of the SFVTD method a rectangular resonator with an edge length of 1m, homogeneously filled with a dielectric is investigated. The simulations are performed with a homogeneous grid and an inhomogeneous grid. The inhomogeneous grid is only inhomogeneous in $z$-direction. In this direction the mesh step is abruptly changed from $\Delta$ to $3\Delta$ in the middle of the resonator. The Courant number is set to the stability limit of the SFVTD method for each simulation.

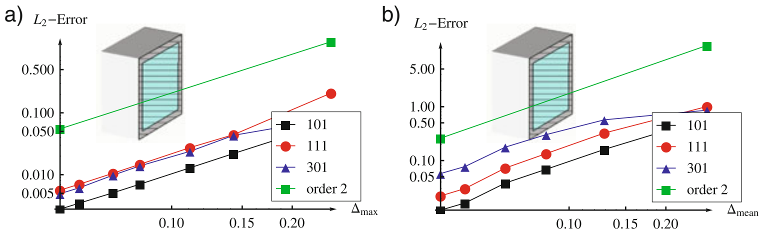


**Fig. 4:** Maximal $L_2$ error in the fields over one period for different eigenmodes of the homogeneous rectangular resonator for a homogeneous grid (**a**) and an inhomogeneous grid (**b**)

Figure 4 shows the maximal relative error in the eigenmodes $(m,n,p)$ over one period for the homogeneously filled resonator with a homogeneous and an inhomogeneous mesh. For both grids the SFVTD converges to the exact result with an order of 2.

## 8 Conclusions

A novel, staggered FVTD method has been presented. In comparison to Yee's scheme, the SFVTD method has better wave propagation properties and an approximately 1.7 larger stability limit on a homogeneous mesh for the same number of DOFs. At the stability limit the SFVTD method has no numerical dispersion along the axes. However, the computational costs of the scheme are higher than those of Yee's scheme. A simple validation problem has been presented, which shows second order convergence of the method.

## References

1. Leveque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge University Press (2002)
2. Yee, K.S.: Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**, 302–307 (1966)
3. Rothwell, E.J., Cloud, M.J.: Electromagnetics. CRC PRESS (2001)
4. Lau, T., Gjonaj, E., Weiland, T.: Time integration methods for particle beam simulations with the finite integration theory. FREQUENZ **59**, 210–219 (2005)
5. Strikwerda, J.C.: Finite Difference Schemes and Partial Differential Equations. SIAM (2004)

# EM Scattering Calculations Using Potentials

Magnus Herberthson

**Abstract** EM scattering from PEC surfaces are mostly calculated through the induced surface current **J**. In this paper, we consider PEC surfaces homeomorphic to the sphere, apply Hodge decomposition theorem to a slightly rewritten surface current, and show how this enables us to replace the unknown current with two scalar functions which serve as potentials for the current. Implications of this decomposition are pointed out, and numerical results are demonstrated.

## 1 Introduction

There are numerous ways to address the problem of calculating the radar cross section of PEC surfaces [1,2]. One method is to solve the electric field integral equation, (EFIE), where a standard reference is [3]. In frequency domain, taking the incoming field $\mathbf{E}_i$ to be a plane wave, we have $\mathbf{E}_i(\mathbf{r}) = \mathbf{E}_0 e^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$. By choosing an adapted orthonormal basis, where $\mathbf{E}_0 = E_0\hat{\mathbf{x}}, \mathbf{k} = k\hat{\mathbf{z}}$, the EFIE becomes [1,3,4],

$$\forall \mathbf{r} \in S : E_0 e^{-\mathrm{i}kz}\hat{\mathbf{x}} \hat{=} \mathrm{i}kc\mu_0(\mathbf{I} + \frac{1}{k^2}\nabla\nabla\cdot)\int_S g(\mathbf{r},\mathbf{r}')\mathbf{J}(\mathbf{r}')\mathrm{d}S' \tag{1}$$

where $\mathbf{J}$ is the surface current on $S$, $g$ is the Greens function $g(\mathbf{r},\mathbf{r}') = \frac{e^{\mathrm{i}k|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}$, and where $\hat{=}$ means tangential equality (on $S$).

Since the electric field is a covariant vector field (i.e. a one-form) rather than a contravariant vector field the equality in (1) is, for each $\mathbf{r} \in S$, evaluated in $T_p^*S$, the cotangent space at $p$. It is therefore natural to use the theory of forms, and in particular the Hodge decomposition theorem when addressing (1). For simplicity, we will assume that the surface S is closed and homeomorphic to a sphere.

Magnus Herberthson
Department of Sensor Informatics, Swedish Defence Research Agency, Box 1165, SE-581 11 Linköping, Sweden, e-mail: magnus.herberthson@foi.se
Department of Mathematics, Linköpings Universitet, SE-581 83 Linköping, Sweden, e-mail: maher@mai.liu.se

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_47,

## 2 Reformulation of EFIE

One first notes that when viewed as a one-form on $S$ the LHS of (1), i.e., $E_0 e^{-\mathrm{i}kz}\mathrm{d}x$ is not exact. However, by multiplying both sides of (1) with $e^{\mathrm{i}kz}$, the LHS is just the tangential part of $\hat{\mathbf{x}}$, i.e., the one-form $\mathrm{d}x$, which is exact, i.e., a gradient of a scalar function. For this reason, we introduce the following functions

$$h(\mathbf{r},\mathbf{r}') = g(\mathbf{r},\mathbf{r}')e^{\mathrm{i}k(z-z')}, \quad \mathbf{K}(\mathbf{r}') = e^{\mathrm{i}kz'}\mathbf{J}(\mathbf{r}') \tag{2}$$

After multiplication with $e^{\mathrm{i}kz}$ and some manipulation, (1) becomes

$$e_0\nabla x - \nabla\left[\int_S h\hat{z}\cdot \mathbf{K}\mathrm{d}S' + \frac{\mathrm{i}}{k}\int_S \mathbf{K}\cdot\nabla h\mathrm{d}S'\right] \hat{=} \tag{3}$$
$$\mathrm{i}k\int_S h\mathbf{K}\mathrm{d}S' - \hat{z}\left[\mathrm{i}k\int_S h\hat{z}\cdot\mathbf{K}\mathrm{d}S' - \int_S \mathbf{K}\cdot\nabla\, h\mathrm{d}S'\right]$$

This formulation is still in vector calculus notation, and theoretical advantage is that the LHS of (3) is now an exact one-form. Note, however, that the fictive current $\mathbf{K}$ is a 'down sampled' version of $\mathbf{J}$, i.e., the oscillatory part $e^{-\mathrm{i}kz}$ is factored out. In practice, this may therefore allow for a much sparser sampling of the 'current' $\mathbf{K}$, and therefore reduced numerics. Also, note that complicated objects may require dense sampling over areas of resonance. Before we apply the Hodge decomposition to $\mathbf{K}$, we make the following remark on the notation.

### 2.1 Notation

As usual, $\nabla$ and $\nabla'$ refer to the nabla operator with respect to $\mathbf{r}$ and $\mathbf{r}'$, respectively. On $S$ we denote the corresponding exterior derivative operators with d and d′; in particular they coincide with the covariant derivatives on $S$ when acting on scalar functions. These will be denoted $\nabla_S$ and $\nabla'_S$. We will use a symmetric scalar product.

There are numerous conventions for p-forms. We will only need the following facts/conventions (see for instance [5]) on $S$, which is two-dimensional. The exterior derivative d takes scalar functions (0-forms) into one-forms, one-forms into two-forms and annihilates two-forms. The Hodge star $*$ takes functions into volume (area) forms and vice versa. It also takes one-forms into one-forms through 'a rotation $\pi/2$'. In terms of vector calculus, $*\mathbf{v} = \widehat{\mathbf{n}}\times\mathbf{v}$, where $\mathbf{v}$ is tangent to $S$ and $\widehat{\mathbf{n}}$ is normal to $S$. The coderivative $\delta$ is $\delta = -*\mathrm{d}*$ and the Laplace-Beltrami operator $\Delta_S$ on $S$ is $-(\mathrm{d}\delta+\delta\mathrm{d})$. The Hodge decomposition theorem asserts that, when $S$ is compact, any one-form $\alpha$ can be written uniquely as $\alpha = \mathrm{d}\Phi + \beta + \delta\psi$ where $\Phi$ is a scalar, $\beta$ is a harmonic one-form and $\psi$ is a two-form. However, $\beta \in \mathrm{Harm}^1(S)$ is zero since there are no nontrivial harmonic one-forms on $S$ (assuming that $S$ is homeomorphic to a sphere). Thus, with $\Psi = *\psi$, so that $\Psi$ is a scalar function, we have that $\alpha = \mathrm{d}\Phi + \delta*\Psi \cong \nabla_S\Phi + \widehat{\mathbf{n}}\times\nabla_S\Psi$ on $S$, which means that $\alpha$ is expressed through the two potentials $\varPhi$ and $\varPsi$. We will also use the facts that $\int_S < \omega, \mathrm{d}f > \mathrm{d}S = \int_S < \delta\omega, f > \mathrm{d}S$ and $\int_S < \mathrm{d}\omega, \psi > \mathrm{d}S = \int_S < \omega, \delta\psi > \mathrm{d}S$ where $f$ is a function, $\omega$ is a one-form and $\psi$ is a two-form. In particular, we will use

$$\forall f, \psi : \int_S < \mathrm{d}f, \delta\psi > \mathrm{d}S = \int_S < \mathrm{d}^2 f, \psi > \mathrm{d}S = 0, \tag{4}$$

since $\mathrm{d}^2 f = 0$ is always true for any scalar function $f$.

## 2.2 Reformulation Through Scalar Potentials

To proceed we now use the Hodge decomposition [5] applied to $\mathbf{K}$. Thus,

$$\mathbf{K} = \mathrm{d}\varPhi + \delta * \varPsi \cong \nabla_S \varPhi + \widehat{\mathbf{n}} \times \nabla_S \varPsi \tag{5}$$

on $S$, which means that $\mathbf{K}$ is expressed through the two scalar potentials $\varPhi$ and $\varPsi$. Here, $\hat{\mathbf{n}}$ is a unit normal to $S$ and $\nabla_S$ stands for the intrinsic gradient operator on $S$. Equation (3) can now be written as

$$\nabla \left[ e_0 x - \int_S (< \mathrm{d}' z', \mathbf{K} > + \frac{\mathrm{i}}{k} \Delta_S' \varPhi) h \, \mathrm{d}S' \right] \hat{=} \\ \mathrm{i} \int_S (k\mathbf{K} - \hat{z} \left[ k < \mathrm{d}' z', \mathbf{K} > + \mathrm{i} \Delta_S' \varPhi \right]) h \, \mathrm{d}S' \tag{6}$$

where $\mathbf{K}$ is decomposed as in (5) and $\Delta_S$ is the intrinsic Laplace operator on S.

Depending on approach, equation (6) can be addressed in several ways. In the next section, we will consider a few of these.

# 3 Calculational Benefits

The factorization $\mathbf{K}(\mathbf{r}') = e^{\mathrm{i}kz'} \mathbf{J}(\mathbf{r}')$ may lead to sparser sampling. Since the external applied field is $-E_0 \hat{\mathbf{x}} e^{-\mathrm{i}kz}$ we can expect that the induced current $\mathbf{J}$ largely contains the oscillatory part $e^{-\mathrm{i}kz}$. Therefore, in non-resonant areas, the 'current' $\mathbf{K}$ resembles an envelope, which can be sampled much sparser than $\mathbf{J}$. For high frequencies, this can reduce the numerical problem substantially.

Another way of using (6) is to use the fact that the left hand side is an exact one-form on $S$. Namely, by a discretization of $\mathbf{K} = \mathrm{d}\varPhi + \delta * \varPsi$ which gives $\varPhi$ and $\varPsi$ $n$ degrees of freedom each, we must in principle produce and solve a $2n \times 2n$ system of equations

$$\begin{pmatrix} X & X \\ X & X \end{pmatrix} \begin{pmatrix} [\varPhi]_{n\times 1} \\ [\varPsi]_{n\times 1} \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{7}$$

where $[\Phi]_{n\times 1}$, $[\Psi]_{n\times 1}$ are $n\times 1$ vectors representing the fields $\Psi$ and $\Phi$, where $v_1, v_2$ also are $n\times 1$ vectors and where $X$ stands for various matrices of order $n\times n$. Instead of solving (7) directly, one can address (6) in the following way. For instance, one can take the exterior derivative of (6), in which case the LHS vanishes, and one is thus left with a homogeneous system of the form

$$\begin{pmatrix} X \; X \end{pmatrix} \begin{pmatrix} [\Phi]_{n\times 1} \\ [\Psi]_{n\times 1} \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix} \tag{8}$$

This gives, through one $n\times n$ inversion $\Phi$ as a function of $\Psi$ or vise versa. This can then be plugged into the coderivative of the original equation, which gives a final $n\times n$ equation of the form

$$\begin{pmatrix} X \end{pmatrix} \begin{pmatrix} [\Phi] \end{pmatrix} = \begin{pmatrix} v_1 \end{pmatrix} \tag{9}$$

In principle, one $2n\times 2n$ inversion is replaced with two $n\times n$ inversions. One should note that the price to pay is more matrix multiplications, which, however, are easy to parallelize. A more practical way of splitting (7) is to use identities of the type (4) and test (6) against suitable test function directly. Namely, if the test functions are co-gradients, i.e., vector fields of the form $\hat{\mathbf{n}}\times\nabla u$, testing against $n$ suitable functions will produce a homogeneous equation of the format (8). The final equation of the form (9) is then achieved when testing against test vector fields which are gradients.

## 4 Numerical Results

Although this approach has been reported earlier, the full implementation using the scalar potentials $\Phi$ and $\Psi$ has not been demonstrated so far. Considering the space available, we will only exemplify the above approach in the most familiar case, i.e., the sphere, although there is no particular restriction on the geometry except the assumption that the surface is homeomorphic to a sphere. We consider a sphere with radius $a$=1 m, which is illuminated with a plane wave along the z axis and with the electric field parallel to the x axis. In this example, the wavelength is 2/3 m, which gives a wave number of $3\pi$/m and a frequency of 450 MHz.

The resulting potentials and currents are illustrated in Fig. 1-4. In Fig. 5, a corresponding calculation with the commercial program FEKO, [6], is displayed. In Fig. 1, the calculated scalar potential $\Phi$ and the corresponding 'current' $\nabla\Phi$ is shown. $\Phi$ is displayed as an intensity map over the surface, where $|\Phi|$ is given by the brightness, and where the phase information in $\Phi$ is encoded in the colour. The gradient vector field $\nabla\Phi$ is shown with real (red) and imaginary (blue) parts. In Fig. 2, $\Psi$ and the co-gradient $\hat{\mathbf{n}}\times\nabla\Psi$ are displayed in an analogous manner. In Fig. 3, the real part of the total current $\mathbf{K} = \nabla\Phi + \hat{\mathbf{n}}\times\nabla\Psi$ is shown. From the fictive current $\mathbf{K}$, on gets the physical current $\mathbf{J}$ via $\mathbf{J}(\mathbf{r}) = e^{-ikz}\mathbf{K}(\mathbf{r})$, the real part of which is shown in Fig. 4.

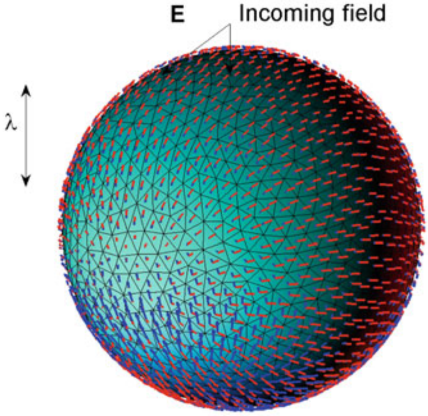


**Fig. 1:** Potential $\Phi$ and its current $\nabla_S \Phi \sim d\Phi$. $|\Phi|$ is given by the brightness of the surface, while the colour encodes the phase information in $\Phi$. The current vectors are displayed in *red* (real part) and *blue* (imaginary part). This current is irrotational

Because of the oscillatory nature of the factor $e^{-ikz}$, the current $\mathbf{J}$ is much more rapidly varying than $\mathbf{K}$, a fact that is easily noticeable by comparing Fig. 3 and

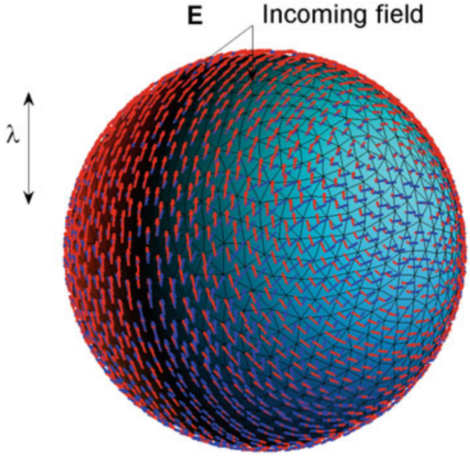


**Fig. 2:** Potential $\Psi$ and its current $\hat{n} \times \nabla_S \Phi$. $|\Psi|$ is given by the brightness of the surface, while the colour encodes the phase information in $\Psi$. The current vectors are displayed in *red* (real part) and *blue* (imaginary part). This current is divergence free

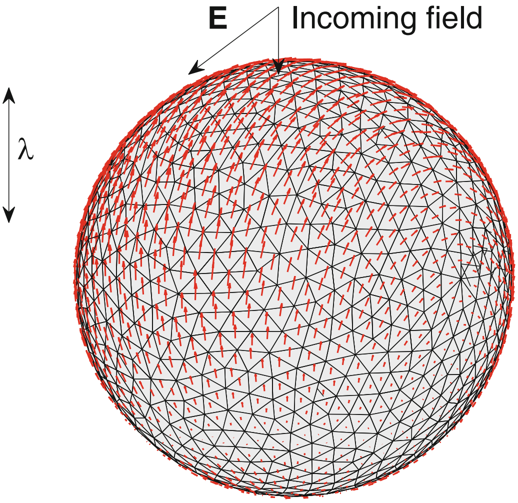


**Fig. 3:** Full current $\mathbf{K} = \nabla_S \Phi + \widehat{\mathbf{n}} \times \nabla_S \Psi$. Only the real part is shown

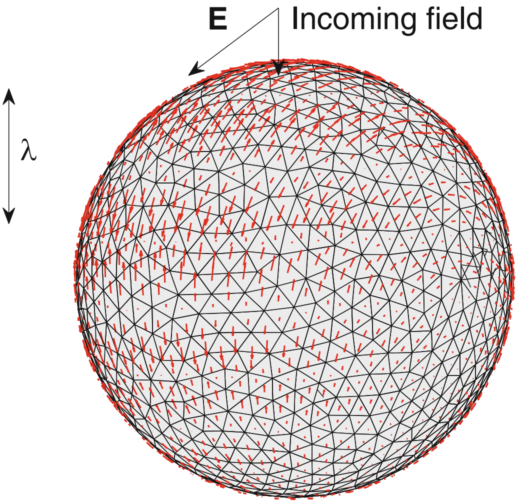


**Fig. 4:** Physical current $\mathbf{J} = e^{-\mathrm{i}kz}\mathbf{K}(\mathbf{r})$. Only the real part is shown

Fig. 4. It is this fact that allows for a much sparser sampling of **K** than a direct sampling of **J** requires, and therefore a smaller equation system to solve.

As a reference, we have also included the corresponding calculation done in the commercial program FEKO. The calculated current is shown in Fig. 5, and this current should be compared to the corresponding current in Fig. 4. As our calculations are still somewhat preliminary, it suffices at this stage to say that the RCS calculated from the current in Fig. 4 is 3.22 m$^2$ as compared to the RCS given by FEKO:

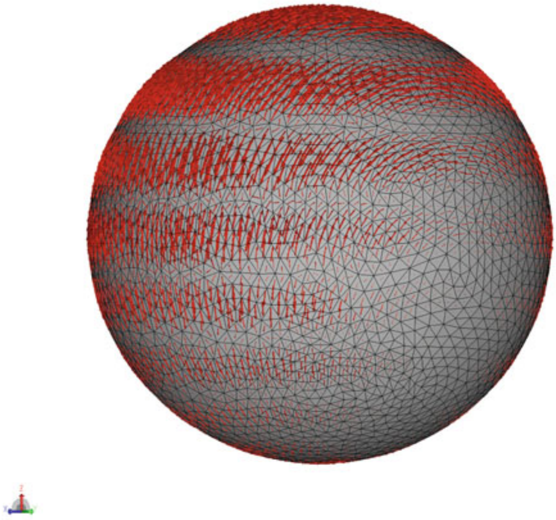


**Fig. 5:** Potential $\Phi$ and its current

3.47m$^2$. The most noticeable difference between Fig. 5 and Fig. 4. is obviously the fineness of the meshes. Although these mesh sizes are rather ad hoc, we have in the FEKO calculation used the suggested setting, which gave 9549 unknowns. For the potential method, we used the somewhat arbitrary choice of 1992 unknowns, the requirement being that the comparatively slowly varying fictive current $\mathbf{K}$ in Fig. 3 should still be well sampled.

It is possible to introduce spherical coordinates $r, \theta, \phi$ with respect to the orthonormal basis introduced in section 1. One can then decompose $\mathbf{J} = J_\theta \hat{\theta} + J_\phi \hat{\phi}$ so that both $J_\theta = J_\theta(\theta, \phi)$ and $J_\phi = J_\phi(\theta, \phi)$ are functions on the sphere. In Fig. 6 we have plotted $J_\theta(\theta, 0), \;\; 0 \leq \theta \leq \pi$ as calculated by both FEKO and the potential method. Most notable is the good agreement for small $\theta$ (the illuminated part of the sphere) as compared to large values of $\theta$ (the 'shadowed' part of the sphere). As stated earlier, conclusions at this stage should be drawn with care.

We believe that the arguments presented here should clearly demonstrate the potential benefits of the approach presented in the paper. Several further remarks and comments could be made. For instance, the splitting of the system of equations as described in section 3 has not been commented further. Also, in the present preliminary code, no emphasis has been put on the task of calculating the moment matrix. It is to be expected that the moment matrix in the potential method may be more costly to compute than in the traditional MoM. However, to what extent remains to be seen, and obviously, the computation of the moment matrix is easy parallelizable.

Another issue is the choice of basis functions in the discretization of $\Phi$ and $\Psi$. This is non-trivial, and in the present implementation, we have chosen global basis

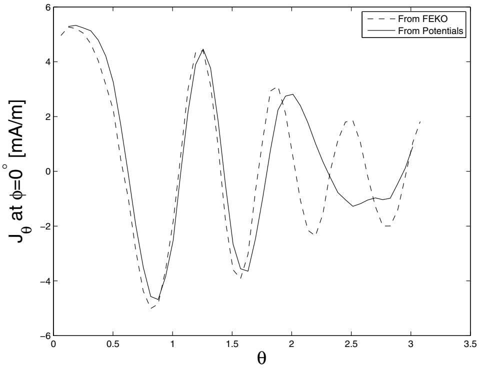


**Fig. 6:** Current components $J_\theta$ compared at the longitude $\phi = 0,\ 0 \leq \theta \leq \pi$. The reason for the rather poor agreement for large $\theta$ is, at the time of writing, under investigation

functions which can briefly be described as follows. For each node point $n_i$, a piecewise linear 'tent' function $\tau_i$ is considered, which is 1 at $n_i$ and zero at all other node points. Next a function $t_i = \tau_i + \gamma_i$ is formed, where $\gamma_i$ is constant and chosen so that $\int_S t_i \mathrm{d}S = 0$. Due to this condition, one can calculate $\Delta^{-1} t_i$ in the weak sense, and the resulting function serves as one of the basis functions for the potentials.

## 5 Discussion

We have described a new way of addressing EM scattering from closed PEC surfaces. By combining a 'down-sampling' of the problem and the Hodge decomposition theorem, this approach has the advantage of producing smaller systems of equations. The drawback may be that the moment matrix can be more costly to compute, but this should not be a serious defect since this computation is easy to parallelize. Preliminary calculations presented here illustrates these points, and the next natural step is a more detailed analysis.

## References

1. Harrington, R. F., *Field Computation by Moment Methods*, IEEE Press, 1993.
2. Chew, W. C. et al, editors, *Fast and efficient algorithms in computational electromagnetics*, Artech House, 2001.
3. Rao, S. M., Wilton, D.R., Glisson, A.W., Electromagnetic Scattering by Surfaces of Arbitrary Shape, *IEEE Trans. Antennas Propag.*, vol. **AP-30**, pp. 409-418, May 1982
4. Kristensson, G., *Spridningsteori med antenntillämpningar*, Studentlitteratur, Lund, 1999
5. Göckeler, M., Schücker, T., *Differential geometry, gauge theories, and gravity*, Cambridge University Press, 1987
6. www.feko.info

# New Trends in the Preconditioning of Integral Equations of Electromagnetism

David P. Levadoux*, Florence Millot, and Sébastien Pernet

**Invited speaker at the SCEE 2008 conference*

**Abstract** A new family of source integral equations is presented, dedicated to the solution of time-harmonic Maxwell scattering problems. Regardless of the composition of the obstacle – metallic, full dielectric or coated with an impedance layer – we show that a general methodology is able to guide the construction of some special equations whose the foremost feature is to be well-conditioned. Indeed, all of them are free of spurious modes and appear as some compact perturbations of positive operators (when it is not the identity), leading therefore to fast iterative solutions without the help of any preconditioner. These intrinsically well-conditioned equations open the way for interesting new developments in the field of boundary equation methods for Maxwell applications.

## 1 Motivation

The pertinence of integral equation methods for solving scattering Maxwell problems in harmonic regime requires no further proof. Using them in combination with a rapid multipole algorithm and an iterative solver makes them efficient and precise methods that can solve problems involving hundreds of thousands of unknowns. But the efficiency of iterative methods depends on the conditioning of the linear systems, so it is absolutely crucial to have either high-performance preconditioners or intrinsically well-conditioned integral formulations. It is in this field, a strategic one

David P. Levadoux
ONERA, Chemin de la Hunière, 91761 Palaiseau, France,
e-mail: david.levadoux@onera.fr
Mathematics Laboratory, University Paris XI, 91405 Orsay, France

Florence Millot, Sébastien Pernet
CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse, France,
e-mail: sebastien.pernet@cerfacs.fr

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_48,

because necessary for solving very large calculation configurations, that ONERA (the french aerospace lab) and CERFACS (European Centre for Advanced Scientific Computing) have undertaken a collaborative work. In this context, the authors' goal is to give a continuation to a powerful integral equation whose first tentative step was presented in [1] for the soft body problem of acoustic before to be extended in [2] to the perfect electrical conductor (PEC) problem of electromagnetism. The convincing results obtained in this last case were naturally calling for an adaptation of the method to more complex materials. In this direction, [3] brought a promising formulation dedicated to scattering problems with a Leontovich condition. Since then, the aim appears clearly to have in the future a well-conditioned integral equation, as efficient as the one built for the PEC case, but able to treat a realistic object made of dielectrics, metal and thin layer coatings, with a special regard to radar cross section applications.

The method we present here tries to fill this program and belongs to the class of source integral equation (SIE) methods, also known as indirect methods in the mathematical literature. Contrary to the more popular field (or direct) integral equations commonly used in industrial codes, whose unknowns have a clear physical meaning because being the Cauchy data of the electromagnetic field, the SIE methods lead to solutions (the sources) playing the role of pure mathematical currents on the boundary, which have to be radiated by a potential before producing the wanted induced electric currents. However, the desire to deal with unknowns with a physical interpretation imposes a constraint which severely restricts the possibilities to construct stable formulations. The equations derived from general potentials are much richer and provide tools for the composition of formulations with better properties than their "physical" counterparts.

At the heart of our method is the desire to find a way to incorporate in the integral formalism information on the solution that can be extracted beforehand, using for this the means offered by the pseudo-differential calculus of operators. Actually, in the case of the PEC problem the equation we are able to build depends on the choice of an operator $\widetilde{Y}_+$ whose purpose is to approximate the admittance of the diffracting body as best as possible. In the limit case where this approximation is exact, the integral operator to be inverted becomes the identity. We thus try to construct approximations of the admittance sufficiently accurate to produce after discretization a linear system that is by essence well conditioned.

Such an equation appears as a generalization of the combined source integral equation of Mautz-Harrigton [4] in which the coupling coefficient between two potentials is replaced by the $\widetilde{Y}_+$ operator. But the framework of this generalized combined source integral equation (GCSIE) is not specific to the PEC problem. Indeed, adopting a single and general setting to handle both boundary and transmission problems together, we present in section 2 a methodology to build a class of integral equations generalizing the former GCSIE to new scattering problems. Subsequent sections deal with the application of this general framework to PEC, impedance and transmission problems.

Related to our approach is some other connected works we refer to in [5]. Completely embedded in the GCSIE formalism is for instance the so-called generalized Brakhage-Werner equation studied in [6, 7] or the regularized equation of [8].

We warn the reader that due to space limitations he will not find neither any mathematical details – results being asserted without proof – nor a detailed presentation of the numerical schemes. We have preferred to give an overview of the subject, enlightening the generality of the methodology and focusing on some numerical results we hope convincing.

## 2 The General Framework of the GCSIE Formalism

All boundary or transmission problems we plan to solve, read formally as

$$\text{Find } w \in W \text{ such that } \gamma w = u_0 \tag{1}$$

the source term $u_0$ being a given distribution of currents on the boundary $\Gamma$ of a compact set $D$ (the obstacle), $W$ a functional space of admissible wave solutions usually propagating in $\mathbb{R}^3 \setminus D$ or $\mathbb{R}^3 \setminus \Gamma$, and $\gamma$ a boundary trace operator.

Moreover, $W$ can be parameterized with the help of a potential $\mathscr{C}$

$$w = \mathscr{C}(\gamma_c w) \ , \tag{2}$$

where $\gamma_c w$ stands for the Cauchy data of $w \in W$. Said differently, (2) means than any field $w \in W$ can be rebuilt from the knowledge of its Cauchy data $\gamma_c w$. The potential $\mathscr{C}$ linking any current on $\Gamma$ to a wave in $W$ is called Calderón potential.

Since the initial problem (1) we want to solve is supposed to be well-posed, there exists an operator $R$ (used hereafter as a regularizing one) defined by

$$R : \gamma w \mapsto \gamma_c w \ \ . \tag{3}$$

We have to keep in mind that $R$ is a boundary operator which by definition verifies the crucial relation

$$\gamma \mathscr{C} R = \mathrm{Id} \tag{4}$$

where Id is the identity operator on $\Gamma$.

Now, we are able to build, at least formally, a new class of boundary integral equations. Let $\widetilde{R}$ be an approximation of $R$. We decide to search the solution $w$ of the initial problem (1) under the form

$$w = \mathscr{C}\widetilde{R}u \tag{5}$$

where $u$ is a current distribution on $\Gamma$ acting like a source excitation of the potential $\mathscr{C}\widetilde{R}$. Therefore, in order to find a source $u$ radiating the field solution of our initial problem (1), we have to solve the resulting source (or indirect) integral equation

$$\gamma\mathscr{C}\widetilde{R}u = u_0 \ . \tag{6}$$

Because of the crucial relation (4), if $\widetilde{R} = R$ the new equation (6) becomes trivial. Therefore, we suspect that when $\widetilde{R}$ is a good approximation of $R$, the resulting equation is a "small" perturbation of identity which produces after discretization a well-conditioned linear system.

Although we won't treat the problem of the discretization of the GCSIE, it is worth to keep in mind that the vocation of these equations is to be solved iteratively. Hence, the problem to deal with a product of operators ($\gamma\mathscr{C}$ and $\widetilde{R}$) possibly non local, is not so sharp it could appear at first sight. Actually, in the context of an iterative method the $\widetilde{R}$ operator can be viewed as playing the role of a preconditioner.

## 3 Assumptions and Notations

Generic currents on $\Gamma$ (*i.e.* tangential vector-valued functions) are noted $\mathbf{u}$ or $\mathbf{v}$ and are supposed to belong to classical Sobolev spaces $H^s_{\mathrm{T}}(\Gamma)$. The space of finite energy currents ($s = 0$) is noted $L^2_{\mathrm{T}}(\Gamma)$.

A vector-valued function $\mathbf{E}$ is said to be an electric field if $\nabla\times\nabla\times\mathbf{E} - k^2\mathbf{E} = 0$. When such a field satisfies the well-known Sommerfeld radiation condition we say that it is radiating. The wave number $k$ is supposed to be constant on the exterior (resp. interior) domain $\Omega^+$ (resp. $\Omega^-$) of the obstacle.

Related to a given electric field $\mathbf{E}$ is the magnetic field $\mathbf{H} = \frac{1}{ik}\nabla\times\mathbf{E}$.

Given the unit outward normal $\mathbf{n}$ to $\Gamma$, the notation $\mathbf{n}\times$ means, following the context, either the $\pi/2$ rotation boundary operator on $\Gamma$, or the composition of the tangential trace operator followed by a rotation on the boundary. The tangential trace on $\Gamma$ of a given field $\mathbf{E}$ is noted $\mathbf{E}_{\mathrm{tan}}$.

The space $W^+$ (resp. $W^-$) is made of all radiating electric fields defined on $\Omega^+$ (resp. $\Omega^-$) and having tangential traces on $\Gamma$. The famous reconstruction formula (Stratton-Chu) valid for all $\mathbf{E} \in W^+$ is

$$\mathbf{E} = \mathscr{T}\mathbf{n}\times\mathbf{H} - \mathscr{K}\mathbf{n}\times\mathbf{E} \ , \tag{7}$$

where $\mathscr{T} = \frac{1}{ik}\nabla\times\nabla\times\mathscr{G}$, $\mathscr{K} = \nabla\times\mathscr{G}$ with $\mathscr{G}$ being the vector potential defined by $\mathscr{G}\mathbf{u}(x) = \frac{-1}{4\pi}\int_\Gamma \frac{e^{ik\|x-y\|}}{\|x-y\|}\mathbf{u}(y)\,\mathrm{d}y$ ($\|\ \|$ is the euclidean norm of $\mathbb{R}^3$).

Tangential traces took from the exterior domain and applied to the potentials $\mathscr{G}$, $\mathscr{K}$ and $\mathscr{T}$ define three pseudo-differential boundary operators $\mathbf{n}\times G = \mathbf{n}\times\mathscr{G}$, $\mathbf{n}\times K = \mathbf{n}\times\mathscr{K} + \mathrm{Id}/2$ and $\mathbf{n}\times T = \mathbf{n}\times\mathscr{T}$ of order $-1$, $-1$ and $1$ respectively.

## 4 Boundary Value Problems

Given an incident electric field $\mathbf{E}^{\text{inc}}$, a possible modelling of the scattered field $\mathbf{E}$ spread by an object coated with an impedance layer is

$$\text{Find } \mathbf{E} \in W^{+} \text{ such that } \mathbf{E}_{\text{tan}} + \alpha \mathbf{n} \times \mathbf{H} = -\mathbf{E}^{\text{inc}}_{\text{tan}} + \alpha \mathbf{n} \times \mathbf{H}^{\text{inc}} \ , \tag{8}$$

where $\alpha$ is a complex-valued function defined on $\Gamma$. Returning to the formal description of the problem (1) and wanting to derive the scattering problem (8), we set $W = W^{+}$ and the trace operator as

$$\gamma \mathbf{E} = \mathbf{n} \times \mathbf{E} - \alpha \mathbf{H}_{\text{tan}} \ . \tag{9}$$

The source excitation $u_0$ becomes $-\gamma \mathbf{E}^{\text{inc}}$, and because of the Stratton-Chu formula (7) we choose as Cauchy data trace operator $\gamma_c = (\mathbf{n}\times, \frac{1}{ik}\mathbf{n} \times \nabla\times)$ and as Calderón potential $\mathscr{C}(\mathbf{u},\mathbf{v}) = \mathscr{T}\mathbf{v} - \mathscr{K}\mathbf{u}$ which verify the abstract Green formula (2).

We want now to give an expression of $R$ more tractable than the definition (3) which requires to solve the initial problem (1). Given a current $\mathbf{u}$ on $\Gamma$, we consider the solution $\mathbf{E}$ of the problem (1) with $u_0 = \mathbf{u}$

$$\mathbf{u} = \mathbf{n} \times \mathbf{E} - \alpha \mathbf{H}_{\text{tan}} \ . \tag{10}$$

Applying the definition of $R$ given in (3) with our Cauchy data trace operator leads to $R\mathbf{u} = \gamma_c \mathbf{u} = (\mathbf{n} \times \mathbf{E}, \mathbf{n} \times \mathbf{H})$. So, writing $R$ in coordinates $(R_E, R_H)$ one has

$$R_E \mathbf{u} = \mathbf{n} \times \mathbf{E} \qquad\qquad R_H \mathbf{u} = \mathbf{n} \times \mathbf{H} \ . \tag{11}$$

Expanding $\mathbf{n} \times \mathbf{E}$ and $\mathbf{n} \times \mathbf{H}$ in (10) with (11) gives $R_E = \text{Id} - \alpha \mathbf{n} \times R_H$. Therefore $R = (\text{Id} - \alpha \mathbf{n} \times R_H, R_H)$ and we suggest to approach $R$ as $\widetilde{R} = (\text{Id} - \alpha \mathbf{n} \times \widetilde{R}_H, \widetilde{R}_H)$ where $\widetilde{R}_H$ is an approximation of $R_H$ to build. The GCSIE equation is

$$\gamma \mathscr{T} \widetilde{R}_H \mathbf{u} - \gamma \mathscr{K}(\mathbf{u} - \alpha \mathbf{n} \times \widetilde{R}_H \mathbf{u}) = -\mathbf{n} \times \mathbf{E}^{\text{inc}} + \alpha \mathbf{H}^{\text{inc}}_{\text{tan}} \ . \tag{12}$$

Noticing $Y_+$ the exterior admittance of $\Gamma$ linking $\mathbf{n} \times \mathbf{E}$ to $\mathbf{n} \times \mathbf{H}$ for all $\mathbf{E} \in W^{+}$ and using conjointly $\mathbf{n} \times \mathbf{E} = -Y_+(\mathbf{n} \times \mathbf{H})$ and right relation of (11), one obtains from (10) that $R_H = (\alpha \mathbf{n} \times \text{Id} - Y_+)^{-1}$. Therefore, it seems natural to search an approximation of $R_H$ under the form $\widetilde{R}_H = (\alpha \mathbf{n} \times \text{Id} - \widetilde{Y}_+)^{-1}$ where $\widetilde{Y}_+$ is an approximation of $Y_+$. In the special case of $\alpha = 0$ (PEC problem), using the fact that $Y_+^2 = -\text{Id}$, $R_H = Y_+$, we will prefer to take $\widetilde{R}_H = \widetilde{Y}_+$. Hence, we have

$$\widetilde{R}_H = \begin{cases} (\alpha \mathbf{n} \times \text{Id} - \widetilde{Y}_+)^{-1} \text{ if } \alpha \neq 0 \\ \widetilde{Y}_+ \text{ if } \alpha = 0 \end{cases} \tag{13}$$

and the next step is now to find good approximations of $Y_+$ leading to a well-posed equation (12)–(13).

### 4.1 A Direct Approximation of $Y_+$

A quite natural way to build approximations of the admittance is to pull back onto the boundary $\Gamma$ of the scatterer, the well known admittance of the tangent plane. Viewing the admittance of the plane as the trace of a potential, say $-2\mathbf{n}\times\mathscr{T}$, it is attractive to import crudely this formula onto the boundary and to consider the first approximation

$$\widetilde{Y}_+ = -2\mathbf{n}\times\mathscr{T} \quad . \tag{14}$$

This first attempt leads to a GCSIE (12)–(13) being a compact perturbation of identity on $H^s_{\mathrm{T}}$ but unfortunately, with spurious modes too. In order to remove it, one has to localize (14) with a quadratic partition of unity $(U_p,\chi_p)$[1] on the boundary, leading to the following pseudo-local approximation

$$\widetilde{Y}_+ = -2\sum_p \chi_p \mathbf{n}\times\mathscr{T}\chi_p \quad . \tag{15}$$

In the case of the PEC problem, the resulting equation is a one-to-one mapping but not a compact perturbation of identity! Actually, one can show that under assumption on the width of patches (which have to be not too small compared to the wavelength), the equation (read on $L^2_{\mathrm{T}}(\Gamma)$) is a compact perturbation of a positive and coercive operator [2, 9].

The question to know if, when $\alpha$ is not equal to 0 (impedant problem), the GCSIE is always well-posed with (15) is not answered for the moment. Anyway, in this case this technique raises a crucial problem of computation. Even if the support of the Schwartz kernel of $\widetilde{Y}_+$ is well narrowed around the diagonal, it is not rigorously a local operator. Hence, one could mind if the numerical cost of the iterative (or direct) inversion of such an operator would not be prohibitive in practice. The next construction of $\widetilde{Y}_+$ overcomes this problem.

### 4.2 An Indirect Approximation of $Y_+$ via the Helmholtz Potentials

Another technique to approximate $Y_+$ consists to use the Helmholtz decomposition. There exist two boundary operators $P_{\mathrm{loop}}$, $P_{\mathrm{star}}$ going from $H_{\mathrm{T}}(\Gamma)$ to $H(\Gamma)$ such that for all $\mathbf{u}\in H_{\mathrm{T}}(\Gamma)$

$$\mathbf{u} = -\mathbf{n}\times\nabla P_{\mathrm{loop}}\mathbf{u} + \nabla P_{\mathrm{star}}\mathbf{u} \quad .$$

If $A$ is an operator acting on vector fields of $\Gamma$, we can identify $A$ with a $2\times 2$ matrix of operators acting on scalar fields following

[1] $(U_p)_p$ is a set of patches recovering $\Gamma$, and $(\chi_p)_p$ a family of truncation functions with support in $U_p$ and such that $\sum_p \chi_p^2 = 1$.

$$A = \begin{pmatrix} -\mathbf{n}\times\nabla & \nabla \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} P_{\text{loop}} \\ P_{\text{star}} \end{pmatrix} \ .$$

We recall that $T$ can be read as $T = \frac{1}{ik}(\nabla G \div +k^2 G)$. On a plane, $\mathbf{n}$ and $\nabla$ are commuting with $G$, and hence the representation of $T$ with the Helmholtz potentials is

$$\mathbf{n}\times T = \frac{1}{ik} \begin{pmatrix} 0 & -G(\Delta + k^2\text{Id}) \\ k^2 G & 0 \end{pmatrix} \ .$$

Still in the plane, the Fourier transform of the kernel of $G$ is $\hat{G}(\xi) = \frac{1}{2i}(k^2 - \|\xi\|^2)^{-1/2}$. Therefore $G = \frac{1}{2i}(\Delta + k^2\text{Id})^{-1/2}$ and because $Y_+ = -2\mathbf{n}\times T$ on the plane, $Y_+$ is equal to

$$\widetilde{Y}_+ = \frac{1}{k} \begin{pmatrix} 0 & -(\Delta + k^2\text{Id})^{1/2} \\ k^2(\Delta + k^2\text{Id})^{-1/2} & 0 \end{pmatrix} \ . \tag{16}$$

So as before with the former formula (14), we have a representation of $Y_+$ on the plane throw a formula involving operators which make sense also on a general surface. Indeed, if $\Delta$ in (16) is viewed as the Laplace-Beltrami operator, this formula is able to define a $\widetilde{Y}_+$ operator on $\Gamma$ candidate to the GCSIE (12)–(13).

But as in the former construction, the resulting GCSIE suffers from spurious modes. Equivalent in spirit to the localization process used with the cut-off function, we have to localize $\widetilde{Y}_+$ in replacing $k$ with $k + i\varepsilon$ where $\varepsilon$ is a small damping parameter. Hence, the resulting GCSIE (12)–(13)–(16) appears as a well-posed equation being furthermore a compact perturbation of identity on $L^2_{\text{T}}(\Gamma)$ [3].

### 4.3 Discretization and Numerical Results

As ever said, the question of the discretization is out of the scope of this paper and we will be voluntary a little bit sketchy. The important is to show with numerical experiments that the GCSIE is truly a well-conditioned equation more powerful than the other ones which can be bought off the shelf.

For the PEC problem we have compared our equation to the classic EFIE and CFIE equations, and for the impedance problem to the ICFIE (Impedance CFIE) [10] and the BGLIE (Bachelot-Gay-Lange Integral Equation) [11]. At last, the choice of Raviart-Thomas finite elements of lower order was quite an evidence considering their popularity in all industrial codes.

About the iterative solver, the well-known GMRES has been practiced with a stopping criterion on the residue fixed to $10^{-4}$, and an optional SPAI (Sparse Pattern Approximation Inverse) preconditioner. When necessary, we used a so-called flexible GMRES (FGMRES) based on a deflation-like method.

Coming back to the the GCSIE (12) formally read as $(AR - BR')\mathbf{u} = \mathbf{u}_0$ ($A = \gamma\mathscr{T}$, $B = \gamma\mathscr{K}$, etc), we remark that the fundamental problem we have to face is to deal with the product $AR$ involving two non local operators. We can not use a Galerkin method. Not impossible in principle, the numerical computation of the matrix coefficients or even the matrix-vector product would be out of reach in practice. In other words, if we note $Q_h$ the $L^2$ projection onto the finite element space $X_h$, it is unrealistic to expect handle the discrete equation $Q_h\mathbf{u}_hAR - Q_hBR'\mathbf{u}_h = Q_h\mathbf{u}_0$ in order to have a discrete solution $\mathbf{u}_h$ of (12).

In the context of an iterative method, the challenge of the discretization of a GCSIE equation consists to find suitable projectors $P_h$, $P_h'$ onto $X_h$ such that given $n \in \mathbb{N}$, the $n^{\text{th}}$ iterate solution of

$$Q_hAP_hR\mathbf{u}_h - Q_hBP_h'R'\mathbf{u}_h = Q_h\mathbf{u}_0 \tag{17}$$

tends to the $n^{\text{th}}$ iterate solution of the continuous equation when the mesh-size parameter $h$ tends to 0. It is a very crucial step we can not develop here, but is explained in details in [12].

Nevertheless, discrete GCSIE (17) could seem to be still expensive, since we have to perform two independent products involving the non local operators $A$ and $B$. However, a judicious discretization of these operators by the fast multipole method (FMM) enables to gather the transfer and reconstruction phases for $A$ and $B$. Therefore, the additional cost beside the initial one represented by the sparse-like products with $R$ and $R'$ is just a single FMM product.

For the PEC problem, the techniques used to build stable $R$ operators ($R' = \text{Id}$) has been thoroughly studied in [13] and an outline of the most efficient implementation version is presented in [2] or [12]. One of its foremost features is that it leads to linear systems showing a condition number independent of both the mesh refinement and the frequency. Results are a significant speed up of the solution time. We point out, for instance, that the Channel cavity (Fig. 1, right), which models an aircraft air intake was processed at 7 GHz (300 000 unknowns) in half the computational time usually needed with a classical equation (Fig. 2). Same results stand for cavities in general as for instance the hollow sphere of Fig. 1 (left).

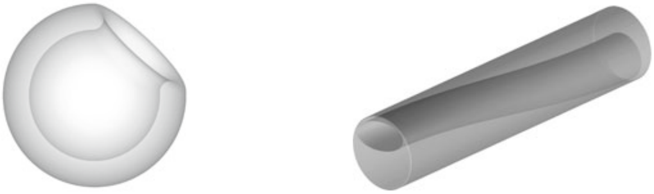


**Fig. 1:** Translucent view of the spherical cavity and the Channel air intake used for the tests

Concerning the indirect construction of $\widetilde{Y}_+$, the main problem to overcome is the synthesis of the square roots appearing in (16). The technique used is explained in [3] and is based on a Padé expansion of the square root took with a rotated branch cut $e^{i\pi/3}\mathbb{R}_-$ as prescribed in [14]. In practice, it leads to the construction of some

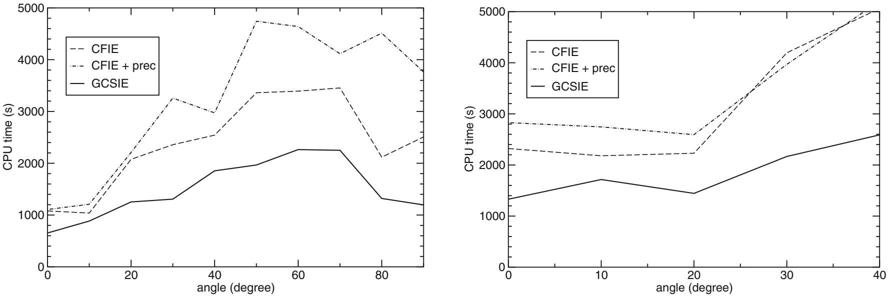


**Fig. 2:** Solution times as a function of the angle of presentation of PEC target. *Left*: spherical cavity at 2.8 GHz with 264,186 unknowns. *Right*: Channel at 7 GHz with 309,711 unknowns

additional sparse matrices efficiently factorized by the solver MUMPS (MUltifrontal Massively Parallel sparse direct Solver) [15]. The construction and the evaluation of $\widetilde{Y}_+$ is completely bearable since it represents at most 20% of the total CPU time. Comparisons with the others integral equations show that the GCSIE gives a similar accurate solution (Fig. 3) with a spectacular speed up of the CPU time (Tab. 1 and Fig. 3).

**Table 1:** Iteration counts and total CPU times for the Channel cavity PEC, full or partially coated at 5GHz with 153,033 unknowns

| Equation | Solver | Coating | Iterations | CPU time |
|---|---|---|---|---|
| GCSIE | GMRES | All surface | 22 | 43 min |
| BGLIE | GMRES + prec | All surface | No convergence | 14 h 12 min |
| BGLIE | FGMRES | All surface | 18 | 21 h 32 min |
| ICFIE | GMRES + prec | All surface | 35 | 3 h |
| GCSIE | GMRES | Inner surface | 22 | 43 min |
| BGLIE | FGMRES | Inner surface | 44 | 37 h |
| ICFIE | GMRES + prec | Inner surface | 37 | 3 h |
| GCSIE | GMRES | None (PEC) | 66 | 1 h 54 min |
| EFIE | FGMRES | None (PEC) | 50 | 28 h 32 min |
| CFIE | GMRES + prec | None (PEC) | 262 | 8 h 23 min |

## 5 Transmission Problems

Let us introduce some additional notations. To a given electric field $\mathbf{E}$ we associate the magnetic-like field $\mathbf{H}' = \nabla\times\mathbf{E}$. Let $\mathbf{n}^+ = \mathbf{n}$ and $\mathbf{n}^- = -\mathbf{n}$. Understanding $\mathbf{n}^+\times$ as a trace (took from $\Omega^+$) operator, we set $\gamma_E^+ = \mathbf{n}^+\times$ and $\gamma_{H'}^+ = \mathbf{n}^+\times\nabla\times$. Corresponding definitions for interior traces operators $\gamma_E^- = \mathbf{n}^-\times$ and $\gamma_{H'}^- = \mathbf{n}^-\times\nabla\times$ are

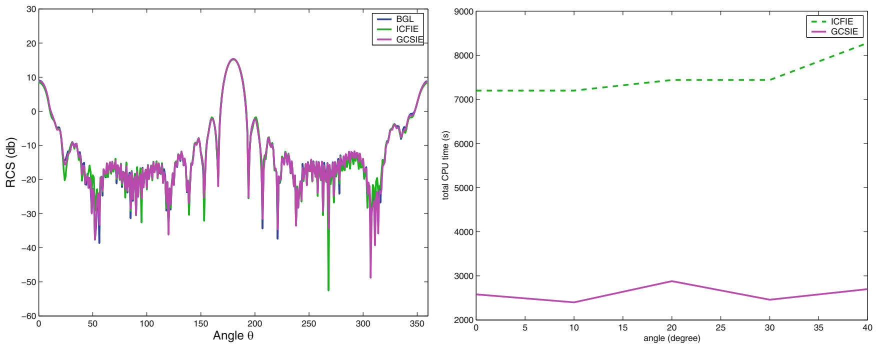


**Fig. 3:** *Left*: bistatic radar cross section of the partially coated channel cavity. *Right*: solution times as a function of the angle of presentation of the target

obvious. Because the wave number $k_+$ in $\Omega^+$ is different of $k_-$ in $\Omega^-$, we note $\mathscr{T}_+$, $\mathscr{K}_+$ (resp. $\mathscr{T}_-$, $\mathscr{K}_-$) the $\mathscr{T}$, $\mathscr{K}$ defined in section 3 with $k = k_+$ (resp. $k = k_-$).

In this section $W$ is the space of all electric fields $\mathbf{E}$ defined on $\mathbb{R}^3 \backslash \Gamma$ such that the restriction of $\mathbf{E}$ to $\Omega^+$ (resp. $\Omega^-$) is in $W^+$ (resp. $W^-$). For the sake of simplicity, the obstacle is supposed to be a dielectric whose the permeability is the same as the vacuum (*i.e.* non magnetic material). When an incident electric field $\mathbf{E}^{\text{inc}}$ collides with such an object it is usual to characterize the resulting transmitted/scattered field $\mathbf{E}$ as solution of the problem

$$\text{Find } \mathbf{E} \in W \text{ such that } \begin{cases} \mathbf{n}^+ \times \mathbf{E} + \mathbf{n}^- \times \mathbf{E} & = -\mathbf{n}^+ \times \mathbf{E}^{\text{inc}} \\ \mathbf{n}^+ \times \nabla \times \mathbf{E} + \mathbf{n}^- \times \nabla \times \mathbf{E} & = -\mathbf{n}^+ \times \nabla \times \mathbf{E}^{\text{inc}} \end{cases} \tag{18}$$

Abstract problem (1) becomes the concrete transmission problem (18) when $\gamma = (\gamma^+_{EH'} + \gamma^-_{EH'})$ with $\gamma^+_{EH'} = (\gamma^+_E, \gamma^+_{H'})$, $\gamma^-_{EH'} = (\gamma^-_E, \gamma^-_{H'})$ and $u_0 = -\gamma^+_{EH'}\mathbf{E}^{\text{inc}}$.

Cauchy data trace operator is chosen as $\gamma_c = (\gamma^+_{EH'}, \gamma^-_{EH'})$ related to which is the Calderón potential $\mathscr{C}(\mathbf{u}^+, \mathbf{v}^+, \mathbf{u}^-, \mathbf{v}^-) = \mathscr{C}^+(\mathbf{u}^+, \mathbf{v}^+) + \mathscr{C}^-(\mathbf{u}^-, \mathbf{v}^-)$, with $\mathscr{C}_\pm(\mathbf{u}, \mathbf{v}) = \frac{1}{ik^\pm}\mathscr{T}_\pm \mathbf{v} - \mathscr{K}_\pm \mathbf{u}$.

The translation of (3) is $(\gamma^+_{EH'} + \gamma^-_{EH'})\mathscr{R} = \text{Id}$, so if $R^+ = \gamma^+_{EH'}\mathscr{R}$ and $R^- = \gamma^-_{EH'}\mathscr{R}$

$$R^+ + R^- = \text{Id} \ . \tag{19}$$

If $\widetilde{R}_+$ is an approximation of $R^+$, it is natural to take $\text{Id} - \widetilde{R}_+$ as an approximation of $\widetilde{R}_-$. Noticing $C_+ = \gamma^+_{EH'}\mathscr{C}$ and $C_- = \gamma^-_{EH'}\mathscr{C}$, the GCSIE equation (6) becomes

$$\left(C_+\widetilde{R}^+ + C_-(\text{Id} - \widetilde{R}^+)\right)(\mathbf{u}, \mathbf{v}) = -(\gamma^+_{EH'} + \gamma^-_{EH'})\mathbf{E}^{\text{inc}} \ . \tag{20}$$

Now we explain how we build $\widetilde{R}_+$. Related to the Cauchy data $\gamma^+_{EH'}$ and $\gamma^-_{EH'}$ are the admittance operators $Y'_\pm : \mathbf{n}^\pm \times \mathbf{E} \mapsto \mathbf{n}^\pm \times \mathbf{H}'$. Reading $R^+$ as a $2 \times 2$ matrix of operators $R^+_{ij}$, the admittance operators $Y'_\pm$ allows to couple coefficients of $R^\pm$

between each others as

$$R^{\pm}_{21} = Y'_{\pm} R^{\pm}_{11} \qquad\qquad R^{\pm}_{22} = Y'_{\pm} R^{\pm}_{12} \ . \tag{21}$$

From (19) and (21), one has $R^{+}_{11} + R^{-}_{11} = \mathrm{Id}$ and $Y'_{+} R^{+}_{11} + Y'_{-} R^{-}_{11} = 0$, giving $R^{+}_{11}$ equal to

$$A = -(Y'_{+} - Y'_{-}) Y'_{-} \ . \tag{22}$$

Always from (19) and (21) one can express the coefficients of $R^{+}$ in function of $A$

$$R^{+} = \begin{pmatrix} A & -AZ'_{-} \\ Y'_{+}A & -Y'_{+}AZ'_{-} \end{pmatrix} \ . \tag{23}$$

Given $\widetilde{A}$, $\widetilde{Y}'_{+}$ and $\widetilde{Z}'_{-}$, approximations of resp. $A$, $Y'_{+}$ and $Z'_{-}$, one suggest to take as approximation of $R_{+}$

$$\widetilde{R}^{+} = \begin{pmatrix} \widetilde{R}^{+}_{E} \\ \widetilde{Y}'_{+} \widetilde{R}^{+}_{E} \end{pmatrix} \text{ where } \widetilde{R}^{+}_{E} = \widetilde{A}\left(\mathrm{Id} \ - \widetilde{Z}'_{-}\right) \ . \tag{24}$$

As approximation of $A$ we can take $\widetilde{A} = (\mathrm{Id} + \frac{\beta^2-1}{2}\Pi_{\mathrm{star}})/(\beta^2+1)$ (with $\beta = k_{+}/k_{-}$) because a pseudo-differential analysis shows that $A - \widetilde{A}$ is a $-1$ order operator. The problem to approximate $Y'_{+}$ and $Z'_{-}$ is the same as to approximate the admittance operators $Y_{+}$, $Y_{-}$ because $Y'_{+} = ik_{+}Y_{+}$ and $Z'_{-} = (ik_{-}Y_{-})^{-1} = -\frac{1}{ik_{-}}Y_{-}$. Therefore, $Y'_{+}$ and $Z'_{-}$ can be approached as the same manner as $Y_{+}$ in section 4.2. In this case, the resulting GCSIE (20) is a well-posed equation at any frequency. More precisely, the underlying operator is a one-to-one mapping and a compact perturbation of identity in $H^{-1/2}_{\mathrm{div}} \cap L^{2}_{\mathrm{T}}$.

To finish, let us give some promising results. On a spherical geometry, we have computed analytically the eigenvalues of the underlying operator of the GCSIE (20)–(24) equipped with the above $\widetilde{A}$ operator and the Padé approximations of admittance operators. Fig. 4 (left) reveals a spectrum well clustered around 1 although the Padé approximation is only of order 2. Fig. 4 (right) shows the GMRES convergence historical of the GCSIE compared to some others classical integral formulations (TENE-THNH, TENE-TH, TENE-NH) [16] used to treat the transmission problems. We point out a significant speed-up of the convergence rate. In the light of these first results, the new formulation seems very attractive.

## References

1. Levadoux, D.P., Michielsen, B.L.: Analysis of a boundary integral equation for high frequency Helmholtz problems. $4^{th}$ International Conference on Mathematical and Numerical Aspects of Wave Propagation pp. 765–767 (Golden, Colorado, 1–5 june 1998)

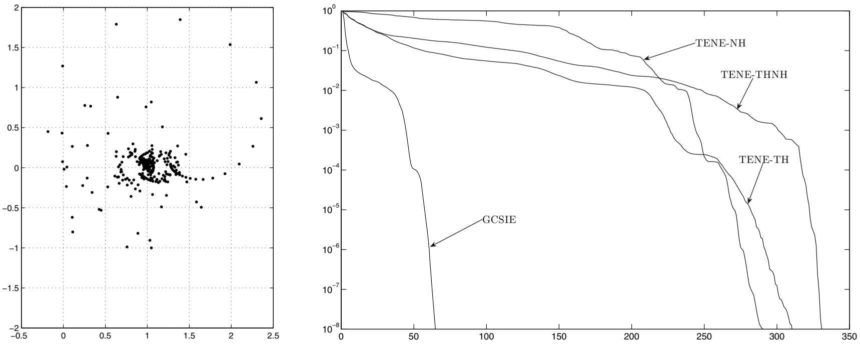


**Fig. 4:** Transmission problem: analytical results on a sphere for $k_+ = 50$ and $k_- = 70.71$. *Left*: spectrum of the GCSIE. *Right*: GMRES convergence historical of several formulations

2. Alouges, F., Borel, S., Levadoux, D.: A stable well-conditioned integral equation for electromagnetism scattering. J. Comp. Appl. Math **204**, 440–451 (2007)
3. Pernet, S.: A well-conditioned integral equation for iterative solution of scattering problems with a variable Leontovich boundary condition. Math. Model. Num. Anal. (submitted)
4. Mautz, J., Harrington, R.: A combined-source solution for radiation and scattering from a perfectly conducting body. IEEE Trans. Antennas Propag. **AP-27**(4), 445–454 (1979)
5. Levadoux, D.: Recent advances in the pre-conditioning of integral equations of electromagnetism. Oberwolfach Reports **5**, 56–59 (2007)
6. Antoine, X., Darbas, M.: Alternative integral equations for the iterative solution of acoustic scattering problems. Quart. J. Mech. Appl. Math. **58**(1), 107–128 (2005)
7. Darbas, M.: Generalized combined field integral equations for the iterative solution of the three-dimensional maxwell equations. Applied Mathematics Letters **19**(8), 834–839 (2006)
8. Bruno, O., Elling, T., Paffenroth, R., Turc, C.: Electromagnetic integral equations requiring small numbers of krylov-subspace iterations. Online preprint (2008)
9. Borel, S., Levadoux, D., Alouges, F.: A new well-conditioned integral formulation for Maxwell equations in three-dimensions. IEEE Trans. Antennas Propag. **53**(9), 2995–3004 (2005)
10. Collino, F., Millot, F., Pernet, S.: Boundary-integral methods for iterative solution of scattering problems with variable impedance surface condition. PIER **80**, 1–28 (2008)
11. Lange, V.: Equations intégrales espace-temps pour les équations de maxwell. calcul du champ diffracté par un obstacle dissipatif. Ph.D. thesis, Université de Bordeaux (1995)
12. Levadoux, D.: Stable integral equations for the iterative solution of electromagnetic scattering problems. C. R. Physique **7**(5), 518–532 (2006)
13. Borel, S.: Étude d'une équation intégrale stabilisée pour la résolution itérative de problèmes de diffractions d'ondes harmoniques en électromagnétisme. Ph.D. thesis, Université Paris XI (2006)
14. Milinazzo, F., Zala, C., Brooke, G.: Rational square-root approximations for parabolic equation algorithms. Journal of the Acoustical Society of America **101**(2), 760–766 (1997)
15. Multifrontal massively parallel solver. http://www.enseeiht.fr/lima/apo/MUMPS
16. Jung, B., Sarkar, T., Chung, Y.: A survey of various frequency domain integral equations for the analysis of scattering from three-dimensional dielectric objects. PIER **36**, 193–246 (2002)

# Simulation of Large Interconnect Structures Using ILU-Type Preconditioner

D. Harutyunyan, W. Schoenmaker, and W.H.A. Schilders

**Abstract** For a fast simulation of interconnect structures we consider preconditioned iterative solution methods for large complex valued linear systems. In many applications the discretized equations result in ill-conditioned matrices, and efficient preconditioners are indispensable to solve the linear systems accurately. We apply the dual threshold incomplete LU (ILUT) factorization as preconditioners for the BICGSTAB iterative solver. On complicated problems with a different range of frequencies we show that the BICGSTAB method with the ILUT preconditioner provides a very efficient solution for the linear systems.

## 1 Introduction

With the increasing complexity of on-chip interconnect structures more robust and fast simulation methods are necessary to understand the behavior of electromagnetic fields in such complex structures. For a better understanding of the performance of these structures field simulation approaches provide more insight about the behavior of the electromagnetic fields.

The governing equations of the electromagnetic fields are given by the Maxwell equations. For many applications the potential formulation of the Maxwell equations is used which has several advantages. In particular, for interconnect structures the potential formulation allows separate modeling of fields in dielectric, semiconductor and metallic regions, which reduces the computational time essentially [1,2].

D. Harutyunyan, W.H.A. Schilders
Technische Universiteit Eindhoven, Den Dolech 2, 5612 AZ Eindhoven, The Netherlands
NXP Semiconductors, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands, e-mail: d.harutyunyan@tue.nl, wil.schilders@nxp.com

W. Schoenmaker
Magwel NV, Martelarenplein 13, 3000 Leuven, Belgium, e-mail: wim.schoenmaker@magwel.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_49,

The differential operators are discretized using the usual finite-volume methods for the electric potential and the charge density. However, for the magnetic vector potential the finite-volume method is replaced by a 'finite-surface method' [3], whose origin is found in Stokes' theorem in contrary to the finite-volume methods that are rooted in Gauss' theorem [1]. This discretization method preserves important physical characteristics of the electromagnetic fields at the discrete level, and we obtain physically relevant solutions. After the discretization of the differential equations we obtain a linear system of equations of the form $Ax = b$, where the coefficient matrix $A$ is a large scale, sparse and complex valued. For large scale problems direct linear solvers are not always possible to implement, and Krylov subspace methods are common tools to solve linear systems approximately. The performance of Krylov subspace methods highly depends on the condition number of the matrix, and for complicated real life problems the resulting matrix $A$ is usually ill-conditioned. For such problems Krylov subspace methods either require too many iteration steps for the convergence or, in the worst case, they do not converge at all. To overcome these difficulties a good matrix preconditioner can significantly improve the convergence rate of the Krylov subspace methods.

In this paper we apply the BICGSTAB iterative solution algorithm [4] with the ILUT preconditioner [5, 6]. In two problems we show that the ILUT preconditioner improves the convergence rate of BICGSTAB algorithm significantly, and provides a very accurate solution for the linear system.

## 2 Potential Formulation of the Maxwell Equations

From the Maxwell equations it follows that there is a vector potential $\mathbf{A}$ and a scalar potential $V$ such that

$$\mathbf{B} = \nabla \times \mathbf{A}, \quad (1) \qquad \mathbf{E} = -\partial_t \mathbf{A} - \nabla V, \quad (2)$$

where $\mathbf{E}$ is the electric field and $\mathbf{B}$ is the magnetic flux density. The current density is $\mathbf{J} = \sigma\,\mathbf{E} + \mathbf{J}_{\text{diff}}$, where $\mathbf{J}_{\text{diff}}$ is the diffusive part of the carrier flows. This way of writing the current allows us to deal with metals in which the diffusive currents are negligible, as well as with semiconductors. In the latter case the first term represents the drift terms of the electron and hole currents. Then the potential formulation of the Maxwell equations in the frequency domain can be written as [1]

$$\nabla \times \mu^{-1} \nabla \times \mathbf{A} - (\sigma + \mathrm{j}\omega\varepsilon)(-\mathrm{j}\omega\,\mathbf{A} - \nabla V) = \mathbf{J}_{\text{diff}}, \tag{3a}$$

$$\nabla \cdot (\varepsilon(\nabla V + \mathrm{j}\omega\,\mathbf{A})) = -\rho, \qquad \text{in insulators and semiconductors,} \tag{3b}$$

$$\nabla \cdot ((\sigma + \mathrm{j}\omega\varepsilon)(\nabla V + \mathrm{j}\omega\,\mathbf{A})) = 0, \qquad \text{in metals,} \tag{3c}$$

where the dielectric permittivity $\varepsilon\,(=\varepsilon_0\varepsilon_r)$, the conductivity $\sigma$, and the magnetic permeability $\mu\,(=\mu_0\mu_r)$ are assumed to be space dependent positive definite tensors.

For the unique solution of (3) we use the following gauge condition which is linear in the scalar and vector potentials, namely

$$\frac{1}{\mu_0}\nabla(\nabla\cdot\mathbf{A})+\mathrm{j}\omega\varepsilon\xi\nabla V=0, \tag{4}$$

where $0 \le \xi \le 1$. The above equation resembles the Coulomb gauge for $\xi = 0$ and the Lorentz gauge for $\xi = 1$ and constant $\varepsilon$. In our applications the simulation domain consist of an interconnect structure extended by a region of air. Therefore two types of boundary conditions are defined, one on the air surface and the other on the interconnect (device) boundary. For the scalar potential $V$ it is straightforward to define voltage (Dirichlet) type boundary conditions on the metal terminals and Neumann type boundary conditions on the surface of the simulation domain. The boundary conditions for the vector potential $\mathbf{A}$ are more subtle and it is out of the scope of this paper to go into the details, instead we refer to [1,2].

For the discretization of (3) and (4) we use unstructured tetrahedral grid which is constructed by using commercial software Magwel. The vector potential $\mathbf{A}$ is computed by finite-surface method. From the differential geometrical point of view $\mathbf{A}$ is a 1-form and on the computational grid the corresponding degrees of freedoms are associated with the element edges. The finite volume method is applied for the computation of the scalar potential $V$, which is a 0-form and the degrees of freedom are associated with the element nodes. In this way we obtain physically relevant solutions. More details about the discretization scheme can be found in [1, 3]. In our applications the resulting linear systems are large and ill-conditioned, which requires a very robust and an efficient preconditioner for the iterative solvers.

## 3 Short Review on ILU Preconditioners

There are two class of general purpose preconditioners for large linear systems: ILU preconditioners and sparse approximate inverse preconditioners. Among ILU preconditioners a common approach is to use ILU(0) factorization which uses a fixed sparsity pattern. Although it is rather easy to construct the ILU(0) preconditioner, for large scale ill-conditioned problems it is well known that this factorization may not be robust and can lead to a very bad approximations and very poor convergence of the iterative methods. In particular, for our problems the ILU(0) preconditioner is not at all applicable because the resulting factorization is singular and can not be used as a preconditioner. An improvement is to use ILU($\tau$) factorization,which allows more fill-in depending on the drop tolerance $\tau$. The drawback of this method is that the memory requirement is unknown in advance. For ill-conditioned problems a small drop tolerance $\tau$ is required to construct a good ILU($\tau$) preconditioner, but for large problems the factorization is not possible because of memory limitations. The construction of a sparse approximate inverse for ill-conditioned problems is far more complicated and time consuming.

To overcome the memory limitation problems of an ILU($\tau$) preconditioner when applied to complicated problems we use a dual threshold ILUT($p,\tau$) preconditioner. Similar to the ILU($\tau$) preconditioner, the same dropping rule is applied based on the drop tolerance $\tau$, then only $p$ largest elements in the row of the L and U matrices are kept. In the course of factorization $\tau$ controls the computational cost while $p$ controls the computer memory, for details see [5].

The computational time required for the ILUT factorization can be reduced by proper reordering of the matrix elements. There are several reordering algorithms based on different methods. In our experiments we make a comparison between two common reordering methods. The first method is the symmetric reverse Cuthill-McKee reordering (SYMRCM) [7] and the second method is the approximate minimum degree (AMD) reordering [8].

## 4 Numerical Experiments

In all experiments the iterative procedure is stopped if the 2-norm of the relative residual (relative to the 2-norm of $b$) is reduced by a factor $10^{-12}$. We have chosen a small reduction factor in order to observe the validity range of the preconditioner.

We run all the numerical experiments under Linux machine with Intel Pentium IV with 3 GHz processor and 4 GB of RAM. Matrix reordering is done with Matlab built-in functions *amd* and *symrcm*. For the ILUT factorization ZITSOL [9] free software package written in C is applied with the BICGSTAB iterative solver, which has proved to be an efficient solver for the potential formulation of the Maxwell equations, see for example [10].

Several notations are used to show the properties of the preconditioner. The computational time in seconds required to construct the ILUT preconditioner is denoted by Pr-time and the corresponding required time for the BICGSTAB iterations is denoted by It-time. The density ratio of the preconditioned versus the original system is denoted by Ratio=$nnz(L+U)/nnz(A)$.

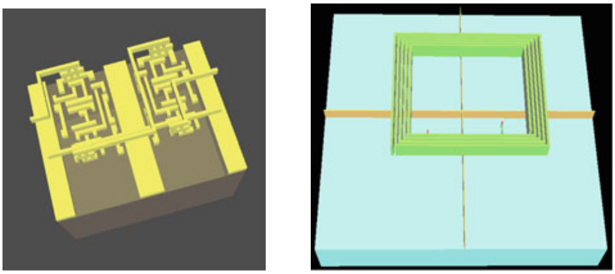


**Fig. 1:** *Left*: Interconnect structure of Test case 1. *Right*: On-chip inductor of Test case 2

## 4.1 Test Case 1

In this section we present numerical experiments on an interconnect structure, see Fig. 1 (left), with dimension in micrometers 4.4 × 5.5 × 4.24. This test case is provided by NXP semiconductors, where the operating frequency is 500 MHz.

First we consider a coarse mesh which results in a matrix of dimension 141513. The resulting matrix of the linear system is ill-conditioned and it is required to choose a small value for the drop tolerance $\tau$ to achieve convergence. Then a proper choice of $p$ is found depending on the complexity of the matrix and memory limitations. Detailed performance information about both methods is given in Table 1. Note that with the chosen values of the drop tolerances it was impossible to construct the $ILU(\tau)$ preconditioner because of memory limitations. With the AMD reordering the iterative method requires significantly less number of iterations as compared to the SYMRCM reordering. It is clear that the AMD reordering requires less time to construct the preconditioner, and the iteration time with the AMD reordering is much smaller as compared to the iteration time with the SYMRCM reordering. Furthermore, we note that with the AMD reordering the required fill-in of the preconditioner is less than that with the SYMRCM reordering. Because of the space limitation we do not show a similar table for the other experiments, but all of the above observations hold true for all our experiments. Distribution of the 25 smallest magnitude eigenvalues of the original matrix and the preconditioned matrix computed by the Jacobi-Davidson method is given in Fig. 2 (left). It is clear that the smallest magnitude eigenvalues of the preconditioned matrix are shifted away from the origin, which explains the good convergence behavior of the BICGSTAB method with the ILUT preconditioner.

We perform a similar experiment on a fine mesh which results in a matrix of dimension 428710. This case is more difficult and requires larger value of $p$ and smaller drop tolerance $\tau$ to obtain an accurate solution. The convergence diagram of the relative residual with the AMD reordering is given in Fig. 2 (right). Let us mention that we failed to obtain convergence with the SYMRCM reordering.

**Table 1:** Test case 1. Performance of the preconditioner with the two different reordering methods

| | | SYMRCM | | | | AMD | | | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $\tau$ | Pr-time | It-time | Its | Ratio | Pr-time | It-time | It | Ratio |
| 100 | $10^{-7}$ | 1029.53 | 325.65 | 430 | 9.39 | 618.13 | 111.80 | 189 | 6.24 |
| 100 | $10^{-8}$ | 1383.33 | 579.21 | 680 | 9.40 | 814.00 | 150.54 | 231 | 6.28 |
| 150 | $10^{-7}$ | 1867.49 | 300.55 | 303 | 13.94 | 1046.66 | 76.25 | 104 | 8.62 |
| 150 | $10^{-8}$ | 2231.82 | 273.94 | 273 | 13.99 | 1154.57 | 74.46 | 99 | 8.71 |
| 200 | $10^{-6}$ | 1807.44 | 600.37 | 472 | 18.21 | 986.45 | 122.82 | 147 | 10.57 |

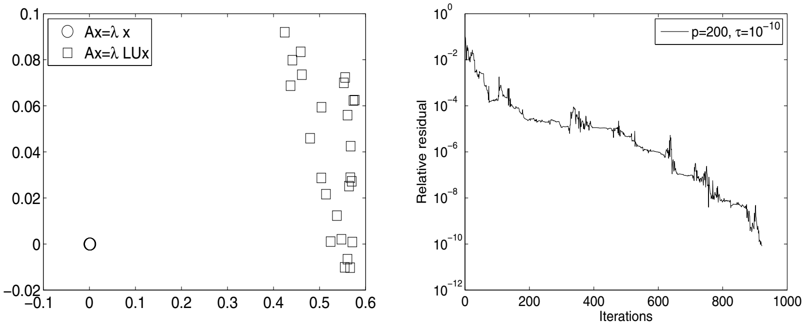


**Fig. 2:** Test case 1. *Left*: Distribution of the 25 smallest magnitude eigenvalues on the coarse mesh of the original matrix and the preconditioned matrix with the ILUT$(p,\tau)$ preconditioner with the AMD reordering. *Right*: Convergence diagram of the relative residual on the fine mesh with the ILUT$(p,\tau)$ preconditioner with the AMD reordering

## 4.2 Test Case 2

In the following numerical experiments we consider an on-chip inductor. The dimension of the structure in micrometers is $1000 \times 1000 \times 407$. The inductor with 4.5 windings is provided by austriamicrosystems and contains a pattern of n-well implants below the inductor in the active device layer. This pattern is mimicked here by the large cross in Fig. 1 (right). The goal of this pattern is to reduce eddy currents in the substrate. The inductor is processed in the M4 (the 4th metal layer) and the underpath is found in M3.

In our applications for the solution of the linear systems the AMD reordering has proved to be more efficient than the SYMRCM reordering, therefore in the following experiments we use only AMD reordering. Convergence diagrams of the relative residual of the BICGSTAB iterative method with the ILUT preconditioner for the frequencies of 1 GHz and 10 GHz are shown in Fig. 3. As it is expected, for higher frequencies more fill-in, smaller drop tolerance and more iterations are required to achieve the same order of accuracy.

In practice the choice of the parameters $p$ and $\tau$ is more based on the problem and experience, see also [5, 6]. In our applications we have made the following observations:

- For a fixed value of $p$ and $\tau_0$, for which a convergence is reached, further decreasing the drop tolerance $\tau < \tau_0$ the number of iterations does not decrease significantly but instead it requires much more time for the construction of the preconditioner.
- For a fixed value of $\tau$ (or $p$) by increasing the fill-in parameter far enough (or by decreasing the drop tolerance ) the required time for BICGSTAB iterations is almost constant and the most time is spent of the construction of the preconditioner.

Based on our experience we suggest in practical applications for difficult problems to start with $p \approx 50$ and $\tau \approx 10^{-5}$ and then follow how the error of the iterative method behaves. If convergence is not reached then based on the first observation we suggest at first to increase the fill-in parameter. If no convergence is reached then decrease also the drop tolerance.

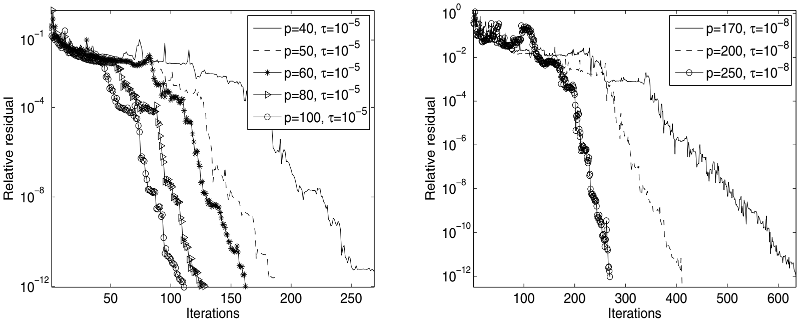


**Fig. 3:** Test case 2. Convergence diagram of the relative residual with the ILUT$(p,\tau)$ preconditioner. *Left*: Frequency is 1GHz, *right*: Frequency is 10 GHz

The goal of the electromagnetic field solving in this experiment is to compute the quality factor Q, the inductance L and the resistance R. These variables are extracted from the admittance matrix $Y$, which is computed by post-processing step after the vector and scalar potentials are computed. A very standard procedure for computing the $Y$ parameters is to apply a voltage at one port and ground all the other ports. Then compute output current at all the ports and calculate corresponding elements of $Y$ matrix. Repeat this procedure for all the ports and obtain the admittance matrix. Current through a surface $S$ is computed by $I_S = \int_S \mathbf{J} \cdot \mathrm{d}S$. The simulation results for the Y-parameters as well as the measurement results are shown in Fig. 4.

## 5 Conclusions

We discussed simulation of interconnect structures where the resulting linear systems after space discretization are large and ill-conditioned. We have shown that the ILUT preconditioner is well applicable for these large and difficult problems and provides a very efficient solution. The use of AMD reordering is necessary for such complicated problems. The performance of the preconditioner was demonstrated in two different structures.

**Acknowledgements** The first author is financed by the European Commission through the Marie Curie Actions of its Sixth Program under the project O-MOORE-NICE, contract number MTKI-

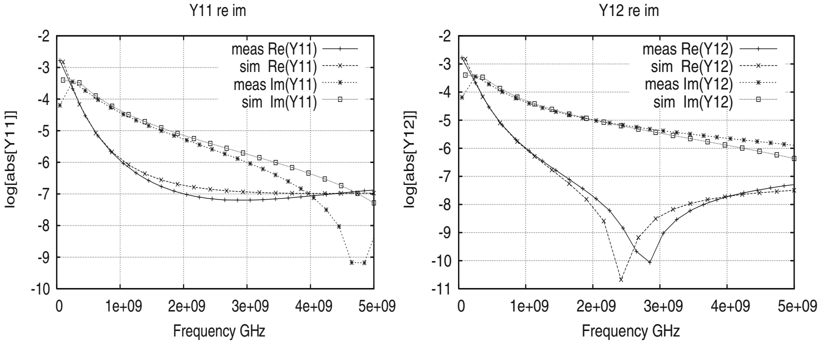


**Fig. 4:** Test case 2. Comparison of the real and imaginary parts of the Y11 and Y12 parameters

CT-2006-042477. The second and third authors acknowledge support from the EU FP5 and FP6 projects CODESTAR and CHAMELEON-RF.

## References

1. Meuris, P., Schoenmaker, W., Magnus, W.: Strategy for electromagnetic interconnect modeling. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **20**, 753–762 (2001)
2. Schoenmaker, W., Meuris, P.: Electromagnetic interconnects and passives modeling: software implementation issues. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **21**(5), 534–543 (2002)
3. Schoenmaker, W., Meuris, P.W.S., van der Kolk K.-J. van der Meijs N.: Maxwell equations on unstructured grids using finite-integration methods. proceedings of the 12th International Conference in Simulation of Semiconductor Processes and Devices (SISPAD 2007), Vienna Austria, September 2007 (2007)
4. van der Vorst, H.A.: Iterative Krylov methods for large linear systems, *Cambridge Monographs on Applied and Computational Mathematics*, vol. 13. Cambridge University Press, Cambridge (2003)
5. Saad, Y.: ILUT: a dual threshold incomplete *LU* factorization. Numer. Linear Algebra Appl. **1**(4), 387–402 (1994)
6. Saad, Y.: Iterative methods for sparse linear systems, second edn. Society for Industrial and Applied Mathematics, Philadelphia, PA (2003)
7. Cuthill, E., McKee, J.: Reducing the bandwidth of sparse symmetric matrices. In: Proc. 24th Nat. Conf. ACM, pp. 157–172 (1969)
8. Amestoy, P.R., Davis, T.A., Duff, I.S.: An approximate minimum degree ordering algorithm. SIAM J. Matrix Anal. Appl. **17**(4), 886–905 (1996)
9. Li, N., Suchomel, B., Osei-Kuffuor, D., Saad, Y.: Zitsol iterative solvers package. http://www-users.cs.umn.edu/~saad/software/ITSOL/
10. Haber, E., Ascher, U.M., Aruliah, D.A., Oldenburg, D.W.: Fast simulation of 3D electromagnetic problems using potentials. Journal of Computational Physics **163**(1), 150 – 171 (2000)

# High-Order Discontinuous Galerkin Methods for Computational Electromagnetics and Uncertainty Quantification

J.S. Hesthaven*, T. Warburton, C. Chauviere, and L. Wilcox

**Invited speaker at the SCEE 2008 conference*

**Abstract** We discuss the basics of discontinuous Galerkin methods (DG) for CEM as an alternative of emerging importance to the widely used FDTD. The benefits of DG methods include geometric flexibility, high-order accuracy, explicit time-advancement, and very high parallel performance for large scale applications. The performance of the scheme shall be illustrated by several examples. As an example of particular interest, we further explore efficient probabilistic ways of dealing with uncertainty and uncertainty quantification in electromagnetics applications. Whereas the discussion often draws on scattering applications, the techniques are applicable to general problems in CEM.

## 1 Introduction

The simplicity, robustness, and reasonable accuracy of the classical finite-difference time-domain (FDTD) method [14] for solving the time-domain Maxwell's equations has propelled this method to become the method of choice among engineers and scientist solving Maxwell's equations in the time-domain. The last decade has seen

J.S. Hesthaven
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA, e-mail: jan.hesthaven@brown.edu

T. Warburton
Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA, e-mail: timwar@rice.edu

C. Chauviere
Laboratoire de Mathématiques, Université Blaise Pascal, 63177 Aubière, France, e-mail: cedric.chauviere@math.univ-bpclermont.fr

L. Wilcox
Institute for Computational Engineering and Sciences (ICES), University of Texas at Austin, Austin, TX 78712, USA, e-mail: lucasw@ices.utexas.edu

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_50,

an explosion in applications and developments, many driven by the very influential texts by Taflove [11, 12].

By now it is also clear, however, that the FDTD methods have severe limitations, e.g., its inherent 2nd order accuracy severely limits its ability to correctly represent wave motion over long distances unless the grid is prohibitively fine. Furthermore, the simplicity of the method, on one hand its very strength, also becomes its most severe restriction by prohibiting the accurate representation of problems in complex geometries.

For the accurate and efficient modeling of large scale EM applications the shortcomings of low order methods render them impractical due to the need for fine grids to avoid prohibitive error accumulation. However, this understanding of the very source of the limitations also suggest that a high-order time-domain solution technique may offer the efficiency and accuracy required for future large scale CEM modeling capabilities. High-order methods are characterized by being able to accurately represent wave propagation over very long distances, using only a few points per wavelength and with an error accumulation rate that is significantly reduced as compared to 2nd order accurate schemes [9]. For three-dimensional applications, this translates into dramatic reductions in the required computational resources, i.e., memory and execution time, and promises to offer the ability to model problems of a realistic complexity and size.

In the following we discuss some of the basic elements of discontinuous Galerkin methods with an emphasis on time-domain electromagnetics. As we will see, these recent developments have paved the way for overcoming many of the restrictions associated with classical high-order methods. In contrast to high-order schemes based on classical finite element techniques, the approach taken here leads to fully explicit schemes.

## 2 The Discontinuous Galerkin Method

The time-dependent Maxwell's equations in the scattered field formulation are

$$\varepsilon \frac{\partial \mathbf{E}^s}{\partial t} = \nabla \times \mathbf{H}^s + \sigma \mathbf{E}^s + \mathbf{S}^E, \tag{1}$$

$$\mu \frac{\partial \mathbf{H}^s}{\partial t} = -\nabla \times \mathbf{E}^s + \mathbf{S}^H, \tag{2}$$

where, $\mathbf{E}^s$ and $\mathbf{H}^s$ denote the scattered electric and magnetic fields, $\varepsilon(\mathbf{x})$ and $\mu(\mathbf{x})$ are the local permittivity and permeability, $\sigma(\mathbf{x})$ is the conductivity of the media and $\mathbf{S}^E$ and $\mathbf{S}^H$ are source terms. Here we have not explicitly written the divergence constraints assuming that the initial conditions satisfy these constraints. Taking the divergence of equations (1)-(2) verifies that if the initial conditions satisfy the divergence constraints then the solution to Maxwell's equations (1)-(2) will also satisfy the divergence constraints.

Let the incident field $(\mathbf{E}^i, \mathbf{H}^i)$ be a solution to Maxwell's equations in a media with permittivity, permeability, and conductivity—$\varepsilon^i(\mathbf{x})$, $\mu^i(\mathbf{x})$, $\sigma^i(\mathbf{x})$, respectively. Along a perfect electric conductor (PEC), the boundary conditions on the total electric field $\mathbf{E}^t = \mathbf{E}^i + \mathbf{E}^s$ and the total magnetic field $\mathbf{H}^t = \mathbf{H}^i + \mathbf{H}^s$ are

$$\hat{\mathbf{n}} \times \mathbf{E}^t = \mathbf{0}, \ \ \mathbf{H}^t \cdot \hat{\mathbf{n}} = 0, \tag{3}$$

where $\hat{\mathbf{n}}$ is the outward pointing normal vector at the surface.

We now briefly describe the computational methods used for solving Maxwell's equations (1)-(2) in the physical space. A discontinuous Galerkin method is used as this offers a number of advantages over widely used alternatives (see [8] for a thorough discussion) and we shall simply sketch its main components. First, we rewrite Maxwell's equations (1)–(2) in conservation form

$$\mathbf{Q}\frac{\partial \mathbf{q}}{\partial t} + \nabla \cdot \mathbf{F}(\mathbf{q}) = \mathbf{S}, \tag{4}$$

where

$$\mathbf{q} = \begin{pmatrix} \mathbf{E} \\ \mathbf{H} \end{pmatrix}, \ \ \mathbf{F}_i(\mathbf{q}) = \begin{pmatrix} -\mathbf{e}_i \times \mathbf{H} \\ \mathbf{e}_i \times \mathbf{E} \end{pmatrix}, \tag{5}$$

signify the state vector $\mathbf{q}$ and the components of the tensor $\mathbf{F}$ and $\mathbf{e}_i$ denotes the Cartesian unit vectors. On the right-hand side of (4), $\mathbf{S} = [\mathbf{S}^E, \mathbf{S}^H]$ is the source term, which depends on the incident field, and the material matrix $\mathbf{Q}$ is a diagonal matrix with values $(\varepsilon, \varepsilon, \varepsilon, \mu, \mu, \mu)$ on its diagonal. We assume that the computational domain, $\Omega$, is tessellated by triangles in two spatial dimensions and tetrahedrons in three spatial dimensions, similar to what is done in a finite element/finite volume method.

Given an element $D$ of the tessellation, we represent the local solution $\mathbf{q}_N$ restricted to $D$ is given as

$$\mathbf{q}_N(\mathbf{x}, t) = \sum_{i=1}^{N} \widetilde{\mathbf{q}}_i(t) L_i(\mathbf{x}), \tag{6}$$

where $\widetilde{\mathbf{q}}_i$ reflects nodal values, defined on the element. The function $L_i(\mathbf{x})$ signifies an $n$th order Lagrange polynomial ($N = (n+1)(n+2)/2$ for triangles and $N = (n+1)(n+2)(n+3)/6$ for tetrahedrons), associated with grid points on the reference element as illustrated in Figure 1 (see [8] for details).

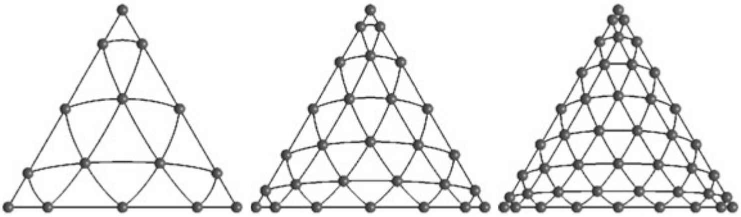


**Fig. 1:** Examples of nodal sets on the equilateral triangle for orders 4, 6, and 8

The discrete solution, $\mathbf{q}_N$, of Maxwell's equations is required to satisfy

$$\int_D \left( \mathbf{Q} \frac{\partial \mathbf{q}_N}{\partial t} + \nabla \cdot \mathbf{F}(\mathbf{q}_N) - \mathbf{S}_N \right) L_i(\mathbf{x}) \mathrm{d}\mathbf{x} = \oint_{\partial D} \hat{\mathbf{n}} \cdot [\mathbf{F}(\mathbf{q}_N) - \mathbf{F}^*] L_i(\mathbf{x}) \mathrm{d}\mathbf{x}. \qquad (7)$$

In (7), $\mathbf{F}^*$ denotes a numerical flux, the expression of which is given as

$$-[\hat{\mathbf{n}} \times \mathbf{H} - (\hat{\mathbf{n}} \times \mathbf{H})^*] = -\frac{1}{2\{\{Z\}\}} \hat{\mathbf{n}} \times \left[ Z^+ (\mathbf{H}^- - \mathbf{H}^+) - \alpha \hat{\mathbf{n}} \times (\mathbf{E}^- - \mathbf{E}^+) \right],$$

and

$$[\hat{\mathbf{n}} \times \mathbf{E} - (\hat{\mathbf{n}} \times \mathbf{E})^*] = \frac{1}{2\{\{Y\}\}} \hat{\mathbf{n}} \times \left[ Y^+ (\mathbf{E}^- - \mathbf{E}^+) + \alpha \hat{\mathbf{n}} \times (\mathbf{H}^- - \mathbf{H}^+) \right],$$

for the equations for the electric and magnetic fields, respectively. Here $\hat{\mathbf{n}}$ is an outward pointing unit vector defined at the boundary $\partial D$ of the element $D$. Using standard notation, $\{\{A\}\}$ signify the average across the interface.

In both cases, we have the possibility of the piecewise constant material coefficients, represented by

$$Z^\pm = \frac{1}{Y^\pm} = \sqrt{\frac{\mu^\pm}{\varepsilon^\pm}},$$

as the local impedance and conductance, respectively. The parameter $\alpha$ is a free parameter with $0 \leq \alpha \leq 1$. For $\alpha = 0$ the scheme is energy conserving but has a potential for nonphysical solutions in rare cases [8]. For $\alpha > 0$, the scheme is slightly dissipative.

Note that this is an entirely local formulation where the fluxes are responsible for coupling of the elements and interchange of information to ensure that the union of the local solutions provides the global solution. Relaxing the continuity of the elements decouples the elements, resulting in a block-diagonal global mass matrix which can be trivially inverted in preprocessing. After discretization of the operators and evaluation of the integrals appearing in (7), the problem can be rewritten in matrix-vector form (see [8])

$$\mathbf{QM} \frac{\mathrm{d}\mathbf{q}_N}{\mathrm{d}t} + \mathbf{S} \cdot \mathbf{F}_N - M\mathbf{S}_N = \mathbf{F}\hat{\mathbf{n}} \cdot [\mathbf{F}_N - \mathbf{F}^*]. \qquad (8)$$

The matrices $\mathbf{M}$, $\mathbf{S}$, and $\mathbf{F}$ represent the local mass-, stiffness-, and face-integration matrices, respectively, the exact entries of which only depend on the metric of the element. The local nature of the scheme allows for the use of an explicit solver for the time discretization of (8) and this is done using an explicit fourth-order Runge–Kutta method. Purely local time advancements are also possible.

The analysis of the scheme given above is complete and one can prove both stability and high-order accuracy is the solution is smooth enough [8]. In particular, for the dissipative upwind flux $\alpha = 1$, one can generally expect optimal accuracy of like $\|\mathbf{q} - \mathbf{q}_N\| \leq Ch^{n+1}$ for $h$ being a measure of the cell size.

To illustrate the performance of the scheme, we consider plane wave TM scattering of a $ka = 20\pi$ metallic cylinder. As simple as the case is, it allows for a thorough

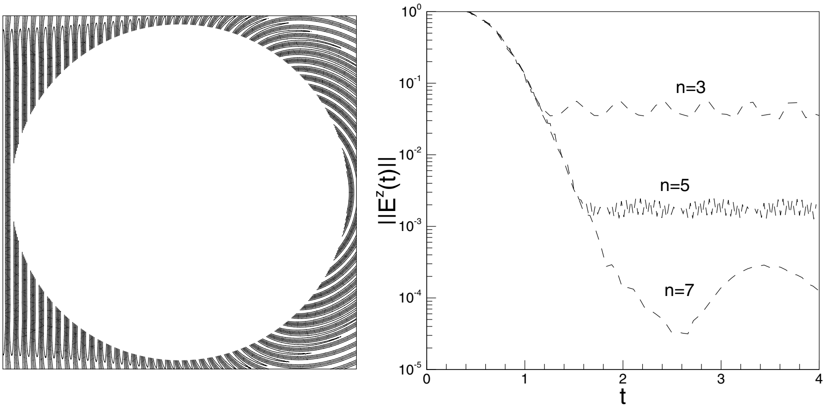


**Fig. 2:** On the *left* is shown plane wave TM scattering of a $ka = 20\pi$ metallic cylinder. The snapshot is for $E_z$. On the *right* we show the error in $E^z$ for plane wave TM scattering by a $ka = 20\pi$ metallic cylinder as a function of time for increasing resolution

validation through the exact solution. We use 950 elements and an high-order local boundary condition [4]. A snapshot of $E^z$ is shown in Fig. 2.

To measure the accuracy of the solution we compute the error in $E^z$ as a function of time for increasing resolution. The results are shown in Fig. 2. For 5th order polynomials ($n = 5$) there are 8-10 points per wavelength. The results confirm exponential convergence as expected. This is also a indication of the excellent performance of the high order local boundary conditions which introduces errors well below the approximation error.

As a considerably more challenging problem, let us consider scattering by a perfectly conducting business card sized metallic plate as illustrated in Fig. 3. The horizontally polarized plane wave impinges at the metallic plate at an almost grazing angle, causing the excitation of strong waves along the edges of the metallic plate as well as along the length of the plate. These waves contribute significantly to the scattering process and need to be resolved to accurately predict the far field scattering. In Fig. 3 we also show the comparison between the experimentally measured monostatic RCS [13] and a number of particular computed data points. We observe good agreement over the full azimuthal range with results well within the experimental error. We note in particular the good agreement in the backscatter region where the scattering is dominated by traveling waves.

Many further examples and validation tests can be found in [5–7].

## 3 Modeling Uncertainty in CEM

While computational methods have become increasingly refined and accurate, their reliance on exact data, e.g., complete descriptions of geometries, materials, sources

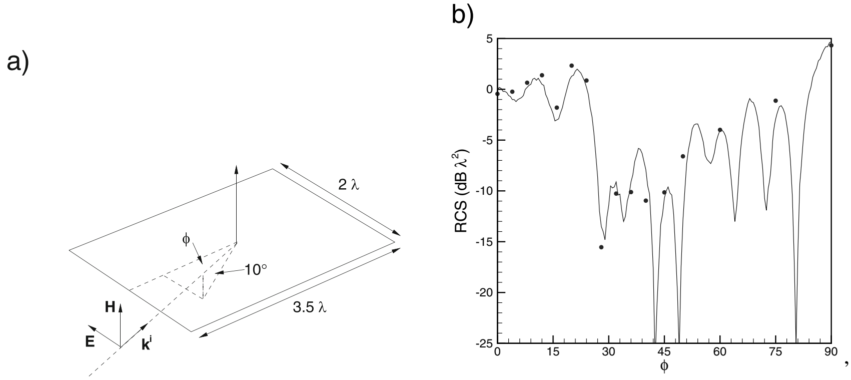


**Fig. 3:** In (**a**) we show the geometry for the plane wave scattering by a metallic business card while (**b**) shows the comparison between monostatic RCS experimental results [13] (*full line*) for horizontal polarization of the illuminating field and particular computed data points (*dots*)

etc, are emerging as a bottleneck in the modeling of problems of realistic complexity. For instance, if one attempts to model an experiment, a classic computational approach requires knowledge to a degree of detail which is unrealistic and often impossible to obtain, e.g., one can not hope to control all elements of an experiment, measure all details of an initial condition or geometry, know the microstructure of all materials etc.

The usual approach to deal with this lack of knowledge or uncertainty is to simply assume some mean parameters and compute the corresponding solution. If the solution is robust to parameter variation, this is indeed a reasonable approach. However, for general problems where the sensitivity of parts of the solution can be significant, a solution based on mean parameters is not likely to match very well with experiments and, thus, is not a good predictive tool. We would like to be able to model the impact of the uncertainty, assumed to have certain properties derived from experiments or otherwise, on the computed results, essentially resulting in an ensemble of possible solution values with an associated probabilities which would immediately enable the computation of statistical moments, e.g., means and variances.

As an advanced application of the computational framework presented above, let us here pursue this goal and present a systematic, accurate, and efficient way of addressing this type of problem, built on top of high-order accurate discontinuous Galerkin methods for solving the time-domain Maxwells equations.

The key result on which we shall rely is due to Wiener (1938) (see also Cameron and Martin [1] ) and shows that the Chaos expansion can be used to approximate any functional in $L^2(\Omega,\mathscr{P})$ where $\mathscr{P}$ is a Gaussian measure on $\Omega$. For such random processes $X(\theta)$, the Chaos expansion reads

$$X(\theta) = a_0 H_0 + \sum_{i_1=1}^{d} a_{i_1} H_1(\xi_{i_1}(\theta)) + \sum_{i_1=1}^{d}\sum_{i_2=1}^{i_1} a_{i_1 i_2} H_2(\xi_{i_1}(\theta),\xi_{i_2}(\theta)) + \dots, \qquad (9)$$

where $\xi = (\xi_1(\theta), ..., \xi_d(\theta))$ represents $d$ independent Gaussian variables with zero mean and unit variance, each depending on the random event $\theta$, and $H_n$ are the multivariate Hermite polynomials. Clearly the number of terms in the expansion (9) grows as

$$P = \frac{(n+d)!}{n!d!}, \tag{10}$$

where $n$ is the length of the Hermite expansion and $d$ is the dimension of the Gaussian random space. To model the impact of uncertainty on the propagation of electromagnetic waves, we include the randomness in the usual spatial-temporal dimensions, i.e., the electric field and the magnetic field become $\mathbf{E}(\mathbf{x}, t, \theta)$ and $\mathbf{H}(\mathbf{x}, t, \theta)$, reflecting that the fields are functions of $d$ independent random variables, $(\xi_{i_1}(\theta), ..., \xi_{i_d}(\theta))$.

In the following we shall discuss in some detail how this can be utilized to construct an efficient computational method. For simplicity of the discussion, we assume in the sequel that one Gaussian variable suffices to represent the process (i.e. $d = 1$). However, the formulation is general and applies to problems which require many random variables to describe the stochastic processes.

Using the Chaos expansion we can express $\mathbf{q}(\mathbf{x}, t, \theta) = (\mathbf{E}(\mathbf{x}, t, \theta), \mathbf{H}(\mathbf{x}, t, \theta))^T$ as

$$\mathbf{q}(\mathbf{x}, t, \theta) = \sum_{i=1}^{P} \mathbf{q}^i(\mathbf{x}, t) \Psi_i(\theta). \tag{11}$$

We can write the computational scheme, taking into account the randomness in a general setting, as

$$\begin{cases} Q(\theta) M \dfrac{\mathrm{d}\mathbf{q}_N}{\mathrm{d}t} + S \cdot \mathbf{F}_N - M\mathbf{S}(\theta)_N = F \hat{\mathbf{n}} \cdot [\mathbf{F}_N - \mathbf{F}^*] \\ \mathbf{q}_N(\mathbf{x}, t = 0, \theta) = \mathbf{f}(\mathbf{x}, \theta) \end{cases}, \tag{12}$$

where the initial conditions are given by the function $\mathbf{f} = \mathbf{f}(\mathbf{x}, \theta)$ and the unknown vector $\mathbf{q}_N$ is given by (11). As a first step, we discretize (12) in the random space using a Galerkin approach. Multiplying (12) by a test function $\Psi_k(\theta)$, replacing $\mathbf{q}_N$ by its Chaos expansion and using orthogonality under the Gaussian measure, we obtain

$$\forall k \in [1, P] : \sum_{i=1}^{P} \langle Q\Psi_i, \Psi_k \rangle M \frac{\mathrm{d}\mathbf{q}_N^i}{\mathrm{d}t} + k! S \cdot \mathbf{F}_N^k - M\mathbf{S}_N^k = F \sum_{i=1}^{P} \hat{\mathbf{n}} \cdot [\mathbf{F}_N^i - \mathbf{F}^{i*}]. \tag{13}$$

The initial conditions in (12) also need to be projected on to the Chaos basis to give an initial condition for each mode of $\mathbf{q}_N^i$ in the Chaos expansion, i.e.

$$\forall i \in [1, P] : \mathbf{q}_N^i(\mathbf{x}, t = 0) = \frac{1}{i!} \langle \mathbf{f}(\mathbf{x}, \theta), \Psi_i \rangle \ . \tag{14}$$

Considering Eq.(13) we observe that we have recast the general stochastic problem into a system of $P$ coupled deterministic problems which we can now discretize in space/time as discussed in Sec. 2.

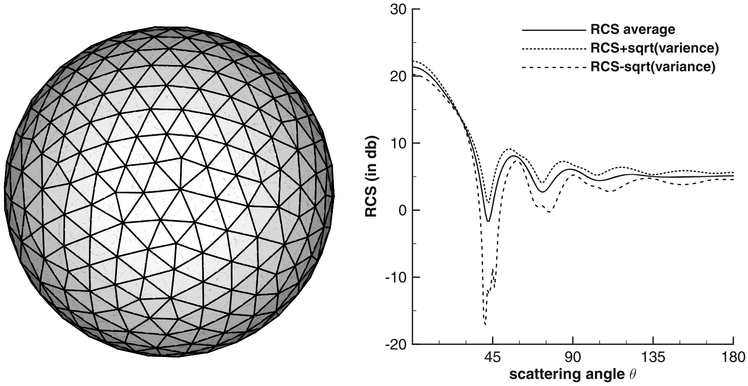


**Fig. 4:** On the *left* we show one sample of a surface mesh for the sphere with a random radius and the *right* illustrates the RCS with uncertainty in the radius of the sphere

Once the vectors $\{\mathbf{q}_N^i\}_{1\leq i\leq P}$ of the system (13) have been computed, we have available at every point in space an approximation to the probability density of the solution of the system. If we assume that we seek the moments of the solutions or a linear combination of them we can take advantage of the basis to obtain

$$\langle \mathbf{q}(\mathbf{x},t,\theta),1\rangle = \sum_{i=1}^{P} \mathbf{q}^i(\mathbf{x},t)\delta_{1i} = \mathbf{q}^1(\mathbf{x},t) \ , \tag{15}$$

i.e., the average is simply the first mode in the Chaos expansion. In a similar way, we can obtain the variance and higher moments. Often, however, we are interested in the statistics of some derived, possibly non-linear, functional, $F(\mathbf{q})$ of $\mathbf{q}(\mathbf{x},t,\theta)$, e.g., computation of the impact on the radar cross section (RCS) of the uncertainty in the scattering problem. To achieve this we consider

$$F(\mathbf{q}(\theta)) = \sum_{j=1}^{P} F(\mathbf{q}(\theta_j))L_j(\theta) \ .$$

i.e.,, we simply need to evaluate the general functionals at the values of $\theta_j$ and since we have already obtained full probabilistic information in the expansions we can use these results directly to obtain the required information and, thus, the probabilistic information on $F(\mathbf{q})$. All information of interest, e.g., moments, can now be extracted from this in the same way as for the simple variables. Naturally, one can evaluate the integrals using a classic Monte Carlo approach. This can be done at little cost since it only requires evaluation of the expansions and not solution of Maxwell's equations.

For the first experiment we consider the scattering of a plane wave, with normalized frequency $\omega = 1$, from a PEC sphere. We assume the sphere has a uniformly distributed random radius in the interval $[0.9\lambda, 1.1\lambda]$, where $\lambda$ is the wavelength of

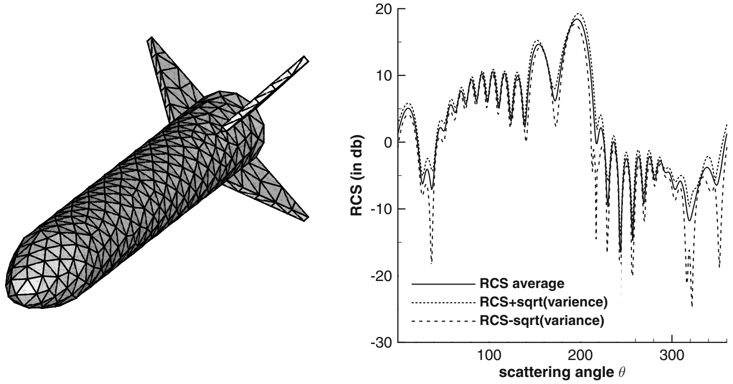


**Fig. 5:** On the *left* we show the surface mesh for the three-dimensional rocket and on the *right* the RCS for the three-dimension rocket problem. Results are shown with the mean RCS as well as ± one standard deviation

the incident field. For the spatial discretization we use fourth order elements and a sample mesh is presented in Figure 4(restricted to the surface of the sphere) and we show the average of the RCS and the possible variations around its average value.

For the second example we consider the scattering of a plane wave, with frequency $\omega = 1$, from a PEC rocket. The direction of the incident field is assumed to be unknown but uniformly distributed in the interval $[10, 20]$ degrees. For this calculation the physical space is discretized with degree five polynomials in each element. Figure 5 shows the mesh (restricted to the surface of the rocket) and the average of the RCS and the possible variations around its average value.

## 4 Final Remarks

The discontinuous Galerkin method is at this stage a robust, efficient, accurate and thoroughly validated alternative to the more classic FDTD method. It overcomes many of the problems with both FDTD methods and alternatives such as finite-volume and finite element methods. Furthermore, large scale software [10] is available for download and use and there are several examples of successful third party use. In this paper we have focused on PEC objects but there is nothing special about these. The method is entirely general and can accommodate general materials, including anisotropic and nonlinear materials as needed. Furthermore, the efficiency of the method has been demonstrated on large problems already.

We have also discussed the combination of these techniques with more recent developments to enable the modeling of PEC objects with random shapes and uncertainties in the incident field. The approach described can, however, equally well be used to account for others types of uncertainties as well as in connection with other computational techniques. For example, instead of being purely reflective, the object can be a material with a random shape. In this case, it is necessary to mesh the entire domain and define a permittivity $\varepsilon$ that takes some value inside the object and another value outside. For material objects, the shape of the objects can be moved randomly in the same way as a PEC object. In [2, 3], the uncertainty in the shape of a material object was studied. However, the approach used was limiting the uncertainty to be modeled to a single random variable. Other types of uncertainties were also studied (randomness of the source term to mimic a slight variation in the frequency of the source, randomness of the permittivity).

The combination of these two methods offers a unique ability to model large scale time-dependent EM problems at high accuracy and with the ability to accurately and efficiently account of sources of uncertainty, leading to sensitivity estimates of measures of interest, e.g., the radar cross section.

**Acknowledgements** The authors appreciate the partial support of AFOSR under grant FA9550-04-1-0072 and FA9550-07-1-0425.

## References

1. Cameron R.H., Martin, W.T.: The Orthogonal Development of Nonlinear Functionals in Series of Fourier-Hermite Functionals. Ann. Math. **48**, 385–392 (1947)
2. Chauvière, C., Hesthaven, J.S., Lurati, L.: Computational modeling of uncertainty in time-domain electromagnetics. SIAM J. Sci. Comput. **28**, 751–775 (2006)
3. Chauvière, C., Hesthaven, J.S., Wilcox, L.: Efficient Computation of RCS from Scatters of Uncertain Shapes. IEEE Trans. Antennas Propagat. **55**, 1437–1448 (2007)
4. Hagstrom, T., Warburton, T.: A New Auxiliary Variable Formulation of High-Order Local Radiation Boundary Conditions: Corner Compatibility Conditions and Extensions to First Order Systems. Wave Motion (2007) – to appear.
5. Hesthaven, J.S., Warburton, T.: High-order nodal methods on unstructured grids. I. Time-domain solution of Maxwell's equations. J. Comput. Phys. **181**, 1–34 (2002)
6. Hesthaven, J.S., Warburton, T.: Discontinuous Galerkin methods for the time-domain Maxwell's equations: An introduction. ACES Newsletter **19** 10-29 (2004)
7. Hesthaven, J.S., Warburton, T.: High Order Nodal Discontinuous Galerkin Methods for the Maxwell Eigenvalue Problem. Royal Soc. London Ser A **362** 493–524 (2004)
8. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications. Springer Texts in Applied Mathematics **54**, Springer Verlag, New York (2008)
9. Kreiss, H.O., Oliger, J.: Comparison of Accurate Methods for the Integration of Hyperbolic Problems. Tellus **24**, 199–215 (1972)
10. nudg++: www.nudg.org
11. Taflove, A.: Computational Electrodynamics. The Finite-Difference Time-Domain Method. Artech House, Boston (1995)
12. Taflove, A. (Ed.): Advances in Computational Electrodynamics: The Finite-Difference Time-Domain Method. Artech House, Boston (1998)
13. Volakis, J.L.: Benchmark Plate Radar Targets for the Validation of Computational Electromagnetics Programs. IEEE Antennas Propagat. Mag. **34**, 52–56 (1992)
14. Yee, K.S.: Numerical Solution of Initial Boundary Value Problems involving Maxwells Equations in Isotropic Media. IEEE Trans. Antennas Propag. **14**, 302–307 (1966)

# Efficient Simulation of Large-scale Dynamical Systems Using Tensor Decompositions

F. van Belzen and S. Weiland

**Abstract** Tensors are the natural mathematical objects to describe physical quantities that evolve over multiple independent variables. This paper considers the computation of empirical projection spaces by decomposing a tensor that can be associated with measured data. We show how these projection spaces can be used to derive reduced order models. The procedure is applied to a two-dimensional heat diffusion problem and a problem in fluid flow dynamics.

## 1 Introduction

Common model reduction techniques such as balanced truncation, Krylov methods, and Proper Orthogonal Decompositions (POD) [7,8] are projection based methods. In this paper, we examine the POD method to reduce the complexity of distributed systems in which signals evolve both in space and time. The POD method is particularly popular in computational fluid dynamics applications where it achieves substantial reductions of complexity while maintaining a high level of predictive power in reduced order models. The method leads to simplified models by applying a Galerkin projection on both the signals and the equation residuals of a distributed dynamical model. A key feature of the method is that projection spaces are determined from optimal low rank approximations of data. The corresponding algebraic tool is the singular value decomposition (SVD) of matrices that have the total mesh size of the spatial geometry as its dimension. For models with spatial geometries that are two, three or larger dimensional, such computations become particularly cumbersome when combined with fine gridded mesh-sizes. Indeed, applications in fluid dynamics with three dimensional spatial geometries easily lead to over $10^6$ grid cells and thus require an SVD of data objects of dimension $10^6$ at least.

F. van Belzen, S. Weiland
Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, e-mail: f.v.belzen@tue.nl, s.weiland@tue.nl

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_51,

To circumvent this problem, we propose a method to compute data-dependent projection spaces that leaves the Cartesian structure in multidimensional arrays of measured data intact. For this, we propose an extension of the concept of SVD to tensors and apply this to reduce the complexity of distributed systems with a Cartesian spatial geometry.

## 2 Model Reduction by Galerkin Projections

Consider an arbitrary linear distributed system described by the Partial Differential Equation (PDE)

$$R\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_N}\right) w = 0. \tag{1}$$

Here, $R \in \mathbb{R}^{\cdot \times 1}[\xi_1, \ldots, \xi_N]$ is a real matrix valued polynomial in $N$ indeterminates and (1) is viewed as a PDE in the signal $w : \mathbb{X} \subset \mathbb{R}^N \to \mathbb{R}$ that evolves over $N$ independent variables. A Galerkin projection of this model is generally defined as follows. First, the space $\mathbb{X}$ of independent variables is assumed to be a Cartesian product $\mathbb{X} = \mathbb{X}' \times \mathbb{X}''$ (typically the product of a spatial and a temporal domain). Second, a Hilbert space $\mathscr{H}$ of real-valued functions on $\mathbb{X}'$ is introduced with inner product $\langle \cdot, \cdot \rangle$. Any complete orthonormal basis $\{\varphi_n\}_{n=1,2\ldots}$ of $\mathscr{H}$ then allows solutions $w$ of (1) to be represented by a spectral expansion $w(x', x'') = \sum_n a_n(x'')\varphi_n(x')$ in which the modal coefficient $a_n$ is uniquely determined by $a_n = \langle w, \varphi_n \rangle$. For $r > 0$, the reduced order model is then defined by the collection of solutions $w_r(x', x'') = \sum_{n=1}^{r} a_n(x'')\varphi_n(x')$ that satisfy

$$\left\langle R\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_N}\right) w_r, \varphi \right\rangle = 0 \quad \forall \varphi \in \mathscr{H}_r \tag{2}$$

where $\mathscr{H}_r$ is the finite dimensional projection space $\mathscr{H}_r = \operatorname{span}\{\varphi_1, \ldots, \varphi_r\}$. If the spectral expansion of $w_r$ is substituted in (2) and $\mathbb{X}'' \subseteq \mathbb{R}$, then (2) becomes a system of $r$ ordinary differential equations in the modal coefficients $a_n$, $n = 1, \ldots, r$. Clearly, the quality of the reduced order model entirely depends on the choice of basis functions $\{\varphi_n\}$. In the POD method, the orthonormal basis functions $\varphi_n$ of $\mathscr{H}$ depend on data that have been either measured or inferred from the model (1). Specifically, for given data $w$ with $w(\cdot, x'') \in \mathscr{H}$, the basis functions $\varphi_n$ are the ordered normalized eigenfunctions of the data correlation operator $\Phi : \mathscr{H} \to \mathscr{H}$ that is defined as

$$\langle \psi_1, \Phi \psi_2 \rangle := \int_{\mathbb{X}''} \langle \psi_1, w(\cdot, x'') \rangle \cdot \langle w(\cdot, x''), \psi_2 \rangle \mathrm{d}\, x'' \qquad \psi_1, \psi_2 \in \mathscr{H}.$$

That is, the basis functions $\varphi_n$ satisfy $\Phi\varphi_n = \lambda_n \varphi_n$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. The data correlation operator $\Phi$ is a well defined linear, bounded, self-adjoint and non-negative operator on $\mathscr{H}$. In applications, the domains $\mathbb{X}'$ and $\mathbb{X}''$ are typically sampled by finite element methods so that $\mathscr{H}$ becomes finite dimensional and $\Phi$

becomes a symmetric non-negative definite matrix. The calculation of POD basis functions then becomes an algebraic eigenvalue or singular value decomposition problem.

## 3 Tensor Decompositions

In this paper we assume that the domain $\mathbb{X}$ of (1) has the Cartesian structure $\mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_N$. We propose a construction of the projection space $\mathscr{H}_r$ that is inferred from a measured or simulated solution $w$ of (1) but that reflects the Cartesian structure of the domain of independent variables in a more explicit way. More specifically, suppose that, for $n = 1, \ldots, N$, the domain $\mathbb{X}_n$ is gridded into a finite set of $L_n$ elements and let $X_n := \mathbb{R}^{L_n}$. Suppose that $w$ is a *Finite Element* (FE) solution of (1) that is defined on the $L = \Pi_{n=1}^{N} L_n$ grid elements. Then $w$ defines a multidimensional array $[[w]] \in \mathbb{R}^{L_1 \times \cdots \times L_N}$ in which the $(\ell_1, \ldots, \ell_N)$th entry is the sample $w_{\ell_1 \ldots \ell_N}$ on the Cartesian grid.

At a more abstract level $[[w]]$ defines a tensor. An *order-N tensor T* is a multilinear functional $T : X_1 \times \cdots \times X_N \to \mathbb{R}$ operating on $N$ vector spaces $X_1, \ldots, X_N$. The elements of $T$, $t_{\ell_1 \cdots \ell_N}$, are defined with respect to bases for $X_1, \ldots, X_N$ according to $t_{\ell_1 \cdots \ell_N} = T(e_1^{\ell_1}, \cdots, e_N^{\ell_N})$, where $\{e_n^{\ell_n},\ \ell_n = 1, \ldots, L_n\}$ is a basis for $X_n$, $n = 1, \ldots, N$. For example, $T(x_1, x_2) := \langle x_2, A x_1 \rangle$ defines an order-2 tensor whose element $t_{\ell_1, \ell_2}$ is the $(\ell_1, \ell_2)$th entry of the matrix $A$.

A FE solution $w$, or its associated multidimensional array $[[w]]$, therefore defines the tensor

$$W = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_N=1}^{L_N} w_{\ell_1 \cdots \ell_N} e_1^{\ell_1} \otimes \cdots \otimes e_N^{\ell_N} \tag{3}$$

where $e_1 \otimes \cdots \otimes e_N$ is shorthand for the *rank-1 tensor* $E : X_1 \times \ldots \times X_N \to \mathbb{R}$, defined by $E(x_1, \ldots, x_N) := \Pi_{n=1}^{N} e_n^{\top} x_n$ and where $w_{\ell_1 \cdots \ell_N}$ is the data element on the sample point with index $(\ell_1, \cdots \ell_N)$.

The tensor (3) associated with the FE solution defines suitable projection spaces by decomposing the tensor $W$ in rank-1 tensors as follows. For each of the vector spaces $X_n$, $n = 1, \ldots, N$ we propose the construction of an orthonormal basis $\{\varphi_n^{\ell_n},\ \ell_n = 1, \ldots, L_n\}$ such that a coordinate change of $W$ with respect to these bases achieves that the truncated tensor

$$W_r := \sum_{\ell_1=1}^{r_1} \cdots \sum_{\ell_N=1}^{r_N} \hat{w}_{\ell_1, \ldots, \ell_N} \varphi_1^{\ell_1} \otimes \ldots \otimes \varphi_N^{\ell_N} \tag{4}$$

with $r = (r_1, \ldots, r_N)$ and $r_n \leq L_n$, $n = 1, \ldots, N$, will minimize the error $\|W - W_r\|$, in a suitable tensor norm, [2,3,5].

For order-2 tensors (matrices) this problem is solved by the singular value decomposition. For higher-order tensors, it is not straightforward how to construct proper sets of orthonormal bases with this property. Different methods exist, including the Higher-Order Singular Value Decomposition [1] and the Tensor SVD [3].

As for the latter, the *singular values* of an order-$N$ tensor $T$, denoted $\sigma_k(T)$ are defined as follows. For $n = 1,\ldots,N$ let $\mathscr{S}_n^{(1)} := \{x \in X_n \mid \|x\|_n = 1\}$ be the unit sphere of elements in $X_n$. Define

$$\sigma_1(T) = \sup\,\{|T(x_1,\ldots,x_N)| \mid x_n \in \mathscr{S}_n^{(1)},\ 1 \le n \le N\} \tag{5}$$

Since $T$ is continuous and the Cartesian product $\mathscr{S}^{(1)} = \mathscr{S}_1^{(1)} \times \cdots \times \mathscr{S}_N^{(1)}$ of unit spheres is compact, an extremal solution of (5) exists and is attained by an $N$-tuple

$$(\varphi_1^{(1)},\ldots,\varphi_N^{(1)}) \in \mathscr{S}^{(1)}.$$

Subsequent singular values of $T$ are defined in an inductive manner by setting $\mathscr{S}_n^{(k)}$ the set of unit norm elements $x \in X_n$ for which $\langle x, \varphi_n^{(j)} \rangle = 0$ for $j = 1,\ldots,(k-1)$. The $k$th singular value is then defined as

$$\sigma_k(T) = \sup\,\{|T(x_1,\ldots,x_N)| \mid x_n \in \mathscr{S}_n^{(k)},\ 1 \le n \le N\}. \tag{6}$$

and its solution defines the singular vectors at level $k$ by the $N$-tuple

$$(\varphi_1^{(k)},\ldots,\varphi_N^{(k)}) \in \mathscr{S}^{(k)}.$$

Due to the iterative construction the singular values are positive and ordered, i.e. $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$. This construction leads to an orthonormal basis

$$\{\varphi_n^{\ell_n}, \ell_n = 1,\ldots,L_n\}, \quad n = 1,\ldots,N$$

for each of the $N$ vector spaces $X_n$. The representation of $T$ with respect to this basis is called the (tensor) singular value decomposition of $T$ and the numbers $\sigma_k(T)$ are referred to as the corresponding singular values. An important result on the approximation properties of this decomposition is the following theorem.

**Theorem 1.** *The tensor $T_1 := \sigma_1 \varphi_1^{(1)} \otimes \cdots \otimes \varphi_N^{(1)}$ is the optimal rank-1 approximation of $T$ in the sense that $\|T - T_1\|$ is minimal among all rank 1 approximations of $T$. Here $\|T\|^2 := \sum t^2_{\ell_1,\ldots,\ell_N}$ is the Frobenius norm.*

We refer to [4], [6] for more details on this decomposition.

## 4 Numerical Examples

The theory discussed in the previous sections will be applied to two examples. We will first show the reduced order modeling of a scalar field, namely heat diffusion on a rectangular plate. Secondly, we will show how tensor techniques can be employed to compute suitable projection spaces for a two-dimensional flow field.

Consider the following model of a heat transfer process on a rectangular plate of size $L_x \times L_y$:

$$0 = -\rho c_p \frac{\partial w}{\partial t} + \kappa_x \frac{\partial^2 w}{\partial x^2} + \kappa_y \frac{\partial^2 w}{\partial y^2}. \tag{7}$$

Here, $w(x,y,t)$ denotes temperature on position $(x,y)$ and time $t \in \mathbb{T} := [0,T_f]$ and the rectangular spatial geometry defines the Cartesian product $\mathbb{X} \times \mathbb{Y} := [0,L_x] \times [0,L_y]$. Let $\mathscr{H} = \mathscr{L}_2(\mathbb{X} \times \mathbb{Y})$ be the Hilbert space of square integrable functions on $\mathbb{X} \times \mathbb{Y}$ and let $\mathscr{H}_r = \mathscr{X}_{r_1} \times \mathscr{Y}_{r_2}$ with $\mathscr{X}_{r_1} \subseteq \mathscr{X} = \mathscr{L}_2(\mathbb{X})$ and $\mathscr{Y}_{r_2} \subseteq \mathscr{Y} = \mathscr{L}_2(\mathbb{Y})$ be finite dimensional subspaces spanned by $r_1$ and $r_2$ orthonormal bases functions $\{\varphi_{\ell_1}\}$ and $\{\psi_{\ell_2}\}$, respectively.

Solutions of the reduced model are then given by $w_r(x,y,t) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} a_{ij}(t)\varphi_i(x)\psi_j(y)$ with $a_{ij}(t) = [A(t)]_{ij}$ a solution of the matrix differential equation

$$0 = -\rho c_p \dot{A} + \kappa_x F A + \kappa_y A P. \tag{8}$$

Here, $F$ and $P$ are defined as:

$$F = \begin{bmatrix} \langle \varphi_1, \ddot{\varphi}_1 \rangle & \cdots & \langle \varphi_1, \ddot{\varphi}_{r_1} \rangle \\ \vdots & & \vdots \\ \langle \varphi_{r_1}, \ddot{\varphi}_1 \rangle & \cdots & \langle \varphi_{r_1}, \ddot{\varphi}_{r_1} \rangle \end{bmatrix}; \quad P = \begin{bmatrix} \langle \psi_1, \ddot{\psi}_1 \rangle & \cdots & \langle \psi_1, \ddot{\psi}_{r_2} \rangle \\ \vdots & & \vdots \\ \langle \psi_{r_2}, \ddot{\psi}_1 \rangle & \cdots & \langle \psi_{r_2}, \ddot{\psi}_{r_2} \rangle \end{bmatrix}$$

Alternatively, $a_{ij}(t)$ is the solution of the ordinary differential equation

$$\rho c_p \dot{a}_{ij}(t) = \kappa_x \sum_{\ell_1=1}^{r_1} a_{\ell_1 j}(t) \left\langle \ddot{\varphi}_{\ell_1}(x), \varphi_i(x) \right\rangle + \kappa_y \sum_{\ell_2=1}^{r_2} a_{i\ell_2}(t) \left\langle \ddot{\psi}_{\ell_2}(y), \psi_j(y) \right\rangle \tag{9}$$

for $1 \le i \le r_1$ and $1 \le j \le r_2$.

**Table 1:** PDE parameter values

| Parameter | $\rho C_p$ | $\kappa_x$ | $\kappa_y$ | $L_x$ | $L_y$ | $T_f$ | $\Delta_x$ | $\Delta_y$ | $\Delta_t$ |
|---|---|---|---|---|---|---|---|---|---|
| Value | 5 | 0.5 | 0.5 | 3 | 4 | 3.6 | 0.05 | 0.05 | 0.05 |
| Unit | $\frac{J}{m^3 \cdot K}$ | $\frac{W}{m \cdot K}$ | $\frac{W}{m \cdot K}$ | m | m | s | m | m | s |

A FE solution of (7) has been computed with physical and discretization parameters as given in Table 1. Time slices, including the initial condition, of the simulation data can be seen in Fig. 1. The boundary conditions are chosen such that the plate is insulated from its environment. The orthonormal bases $\{\varphi_{\ell_1}\}$ and $\{\psi_{\ell_2}\}$ have been computed using the Higher-Order Singular Value Decomposition (HOSVD) [1] and the Tensor SVD. Basis functions are displayed in Fig. 2. The reduced order model (8) has been simulated for different reduction orders, $r = (r_1, r_2)$. The errors with

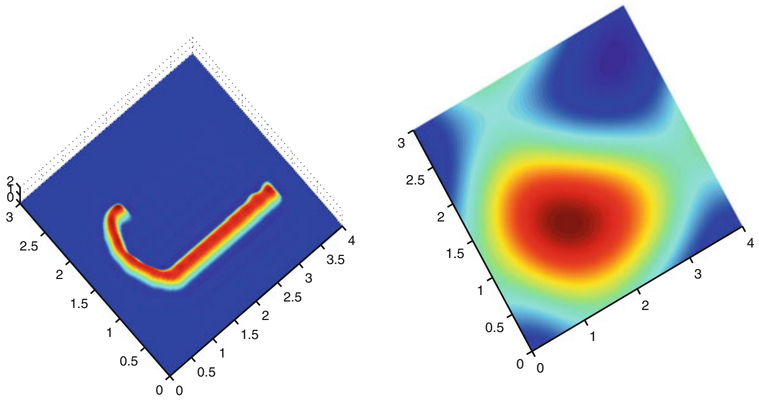


**Fig. 1:** First and final time slices of the FE solution of (7)

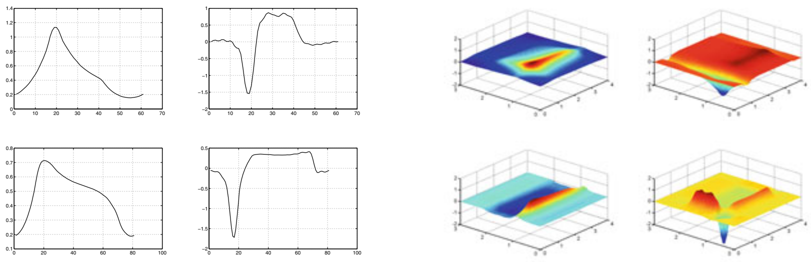


**Fig. 2:** First two basis functions for $\mathscr{X}$ (*top, left*) and $\mathscr{Y}$ (*bottom, left*), computed using Tensor SVD. On the *right* the multiplication of the basis functions is displayed, i.e. $\varphi_1(x) \otimes \psi_1(y)$ (*top left*), $\varphi_1(x) \otimes \psi_2(y)$ (*top right*), $\varphi_2(x) \otimes \psi_1(y)$ (*bottom left*) and $\varphi_2(x) \otimes \psi_2(y)$ (*bottom right*)

respect to the original simulation are given in Table 2. Simulation time of the original model is 17.22s, the reduced models have a simulation time of approximately 0.35s.

**Table 2:** Approximation Error Results, for HOSVD (left) and TSVD (right)

| $r$ | $\|W - W_r\|_F$ | $\frac{\|W-W_r\|_F}{\|W\|_F}$ | $\|W - W_r\|_F$ | $\frac{\|W-W_r\|_F}{\|W\|_F}$ |
|---|---|---|---|---|
| $(2,2)$ | 21.49 | 0.3434 | 22.88 | 0.366 |
| $(3,3)$ | 14.87 | 0.24 | 21.75 | 0.348 |
| $(5,5)$ | 7.634 | 0.122 | 14.95 | 0.239 |
| $(7,7)$ | 4.401 | 0.0704 | 10.85 | 0.174 |
| $(10,10)$ | 2.6319 | 0.0421 | 8.57 | 0.137 |

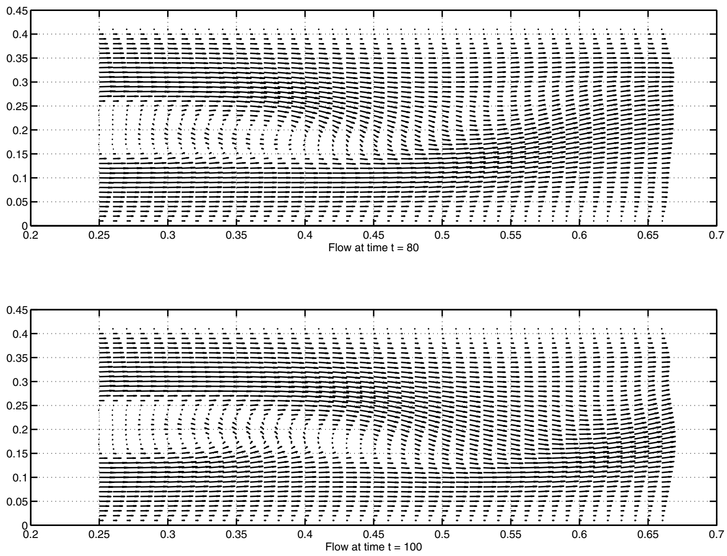


**Fig. 3:** Vector plot of the FE solution of the flow on two time instances

As a second example we consider the computation of projection spaces for a two-dimensional incompressible fluid flow. The flow is described by the Navier-Stokes equations in 2D

$$\rho\left(\frac{\partial \underline{v}}{\partial t}+(\underline{v}\cdot\nabla)\underline{v}\right) = \eta\Delta\underline{v}-\nabla P$$
$$\nabla\cdot\underline{v}=0$$

and defined on a rectangular spatial domain $\mathbb{X}\times\mathbb{Y}:=[0,L_x]\times[0,L_y]$. All boundary conditions are no-slip, except for the left boundary at $x=0$, where the system is excited through a time-varying boundary conditions. Furthermore, let $\mathscr{X},\mathscr{Y}$ be defined as $\mathscr{X}=\mathscr{L}_2(\mathbb{X})$ and $\mathscr{Y}=\mathscr{L}_2(\mathbb{Y})$.

A FE solution has been computed on a spatial grid of size $(L_1,L_2)=(42,42)$ and a temporal grid of size $L_3=193$. A vector plot of this solution on the two time instances $t=80$ and $t=100$ is displayed in Fig. 3. With this FE solution a tensor $W:\mathbb{R}^{L_1}\times\mathbb{R}^{L_2}\times\mathbb{R}^{L_3}\times\mathbb{R}^2\to\mathbb{R}$ can be associated. Computation of the SVD using the Tensor SVD algorithm gives the dominant spatial modes $\{\varphi_{\ell_1}\}$, $\{\psi_{\ell_2}\}$ , i.e. the projections spaces $\mathscr{X}_{r_1}=\operatorname{span}\{\varphi_1,\ldots,\varphi_{r_1}\}\subseteq\mathscr{X}$ and $\mathscr{Y}_{r_2}=\operatorname{span}\{\psi_1,\ldots,\psi_{r_2}\}\subseteq\mathscr{Y}$. In Fig. 4 the two most dominant patterns for $\mathscr{X}$ (left,top) and $\mathscr{Y}$ (left, bottom) are shown. The tensor products $\varphi_{\ell_1}(x)\otimes\psi_{\ell_2}(y)$ of these functions give the dominant patterns on the whole spatial domain, see Fig. 4 (right).

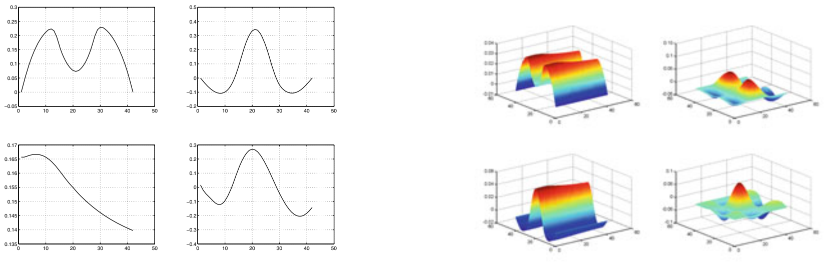


**Fig. 4:** First basis functions for $\mathscr{X}$ (*left, top*) and $\mathscr{Y}$ (*left, bottom*), computed using tensorial SVD [6]. On the *right* the multiplication of the basis functions is displayed, i.e. $\varphi_1(x) \otimes \psi_1(y)$ (*top left*), $\varphi_1(x) \otimes \psi_2(y)$ (*top right*), $\varphi_2(x) \otimes \psi_1(y)$ (*bottom left*) and $\varphi_2(x) \otimes \psi_2(y)$ (*bottom right*)

## 5 Conclusion

In this paper we considered model reduction for multidimensional systems using the POD method. For the computation of empirical projection spaces we proposed a method using tensor decompositions. The techniques proposed were applied to a two-dimensional heat diffusion problem and a problem in fluid flow dynamics. In the future, we plan to test the method on more complex examples and aim to compare different tensorial decompositions to assess accuracy, computational effort and reliability.

**Acknowledgements** This work is supported by the Dutch Technologiestichting STW under project number EMR.7851. The authors would like to thank Jaron Achterberg for providing the Finite Element implementation of the heat transfer process and the reviewers for their valuable input.

## References

1. de Lathauwer, L., et al.: A Multilinear Singular Value Decomposition. SIAM J. Matrix Anal. Appl. **21**(4) (2000)
2. Kolda, T.G.: Orthogonal tensor decompositions. SIAM J. Matrix Anal. Appl. **23**(1) (2001)
3. Belzen, F. van, Weiland, S., Graaf, J. de.: Singular value decompositions and low rank approximations of multi-linear functionals. In: Proc. 46th IEEE Conf. on Decision and Control (2007)
4. Belzen, F. van, Weiland, S.: Diagonalization and Low-Rank Appromixation of Tensors: a Singular Value Decomposition Approach. In: Proc. 18th Int. MTNS (2008)
5. Zhang, T., Golub, G.H.: Rank-one approximation to high order tensors. SIAM J. Matrix Analysis and Applications **23**(2), 534–550 (2001)
6. Weiland, S. and Belzen, F. van: Singular Value Decompositions and Low Rank Approximations of Tensors. Linear Algebra and its applications, submitted (2008)
7. Volkwein, S., Weiland, S.: An Algorithm for Galerkin Projections in both Time and Spatial coordinates. Proc. 17th MTNS (2006)
8. Kirby, M.: Geometric Data Analysis. John Wiley (2001)

# Robust FETI Solvers for Multiscale Elliptic PDEs

Clemens Pechstein and Robert Scheichl

**Abstract** Finite element tearing and interconnecting (FETI) methods are efficient parallel domain decomposition solvers for large-scale finite element equations. In this work we investigate the robustness of FETI methods in case of highly heterogeneous (multiscale) coefficients. Our main application are magnetic field computations where both large jumps and large variation in the reluctivity coefficient may arise. We give theoretical condition number bounds which are confirmed in numerical tests.

## 1 Introduction

Finite element tearing and interconnecting (FETI) methods due to Farhat and Roux [1, 2] are parallel solvers for large-scale finite element (FE) systems arising from partial differential equations (PDEs). Typically, the conditioning of such FE system matrices heavily suffers from the total number of degrees of freedom (DOFs). When the number of DOFs grows large, direct solvers are out of question and efficient preconditioners for iterative solvers are required. Additionally, the parallelization of numerical algorithms gets increasingly important to date. FETI methods are known to be parallel scalable and quasi-optimal with respect to the number of DOFs. For a comprehensive presentation of FETI and related methods we refer to the monograph by Toselli and Widlund [2]. As an additional advantage, one can easily couple finite and boundary element discretizations within the same framework, resulting in

Clemens Pechstein
Institute of Computational Mathematics, Johannes Kepler University, Altenberger Str. 69, 4040 Linz, Austria, e-mail: clemens.pechstein@numa.uni-linz.ac.at,

Robert Scheichl
Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK, e-mail: masrs@bath.ac.uk

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_52,

so-called coupled FETI/BETI methods, see [3–5]. Even exterior domains can be incorporated to model radiation conditions, see [6, 7].

Let us briefly describe the FETI method. As a model problem we consider the finite element discretization of the Poisson-type problem

$$-\nabla\cdot(\alpha\nabla u) \;=\; f \tag{1}$$

in the bounded domain $\Omega\subset\mathbf{R}^d$, $d=2$ or 3, subject to suitable interface and boundary conditions. In Section 4 we will consider a similar equation for 2D magnetostatics. The domain $\Omega$ is partitioned into $N$ non-overlapping subdomains $\Omega_i$, $i=1,\ldots,N$, cf. Fig. 1, right. Introducing separate unknowns $u_i$ on the subdomains including the DOFs on their boundaries, we obtain the saddle point problem

$$\begin{pmatrix} K_1 & & 0 & B_1^\top \\ & \ddots & & \vdots \\ 0 & & K_N & B_N^\top \\ B_1 & \cdots & B_N & 0 \end{pmatrix}\begin{pmatrix} u_1 \\ \vdots \\ u_N \\ \lambda \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_N \\ 0 \end{pmatrix}, \tag{2}$$

where $K_i$ are the subdomain stiffness matrices, and $f_i$ are the corresponding load vectors. The operators $B_i$ are signed Boolean matrices such that each row of the system

$$\sum_{i=1}^N B_i u_i \;=\; 0$$

has the form $u_i(x^h)-u_j(x^h)=0$ for a finite element node $x^h$ on the interface between the subdomains $\Omega_i$ and $\Omega_j$, thus enforcing the continuity of the solution $u$. The Lagrange multiplier $\lambda$ plays the role of a continuous flux on the subdomain interfaces. Introducing a special projection $P$, the dual problem to (2) can be written in the form

$$PF\lambda \;=\; d\,, \tag{3}$$

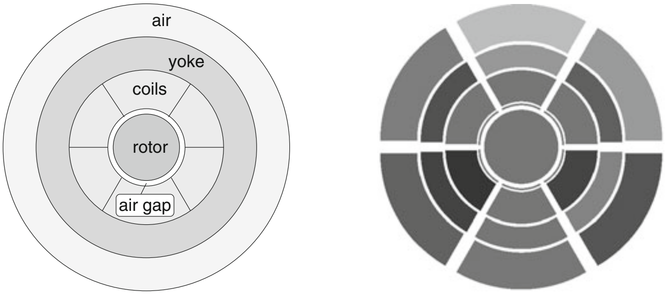


**Fig. 1:** *Left*: Model of an electric motor. *Right*: Possible subdomain partitioning (explosive view)

with $F = \sum_{i=1}^{N} B_i K_i^{\dagger} B_i^{\top}$, where the operators $K_i^{\dagger}$ correspond to the solution of (possibly) regularized Neumann problems on the subdomains. The FETI method is now a special projected preconditioned conjugate gradient (PCG) method to solve problem (3). The chosen preconditioner involves the solution of local Dirichlet problems, and the projection $P$ involves the solution of a coarse problem which corresponds to a sparse linear system of dimension $\mathscr{O}(N)$. Usually, one chooses the partition in a way that the local subdomain problems can efficiently be handled by *sparse direct* solvers, such as $LU$-factorization with suitable pivoting. The factorizations of the local system matrices can be computed in a preprocessing phase and kept in memory during the whole FETI method. Note that these local, decoupled problems can be parallelized in a straightforward manner, e. g., treating each subdomain on a different processor. Once problem (3) is solved, the actual solution $u$ can easily be determined from the Lagrange multiplier $\lambda$. The spectral condition number $\kappa$ of the preconditioned system can finally be estimated by

$$\kappa \;\leq\; C^*(\alpha)\,\max_{i=1}^{N}\big(1+\log(H_i/h_i)\big)^2\,, \tag{4}$$

where the constant $C^*(\alpha)$ is independent of the subdomain diameters $H_i$, the mesh parameters $h_i$, and the number $N$ of subdomains. If $\alpha$ is (globally) constant, then $C^*(\alpha) \sim 1$. As it is well known, the number of PCG iterations needed to achieve a given accuracy, is essentially determined by $\sqrt{\kappa}$. In a parallel scheme the total computational complexity of the FETI-PCG method is given by

$$\mathscr{O}\big((\mathscr{D}(N)+\mathscr{D}(N_{loc}))\,\log(\varepsilon^{-1})\,\sqrt{\kappa}\big)\,, \tag{5}$$

where $N_{loc} \sim \max_{i=1}^{N}(H_i/h_i)^d$ is the maximal number of DOFs per subdomain, $\mathscr{D}(\cdot)$ is the cost of the direct solver, and $\varepsilon > 0$ is the desired relative error reduction in the energy norm.

However, in many applications the original system matrix is ill-conditioned due to heterogeneous coefficient distributions. As we will discuss in Section 4, in magnetic field computations one may have

- large jumps in the reluctivity coefficient due to different materials, and
- smooth but large variation in the same coefficient due to nonlinear effects.

We are interested in the question whether/how the condition number $\kappa$ of the preconditioned FETI system is affected by this. If the heterogeneities are resolved by the subdomain partition (i. e., $\alpha$ constant on each $\Omega_i$), then, using a special diagonal scaling, Klawonn and Widlund [8] proved that $C^*(\alpha) \sim 1$. However, in general, using classical proof techniques, we only get

$$C^*(\alpha) \;\leq\; C\,\max_{i=1}^{N}\,\max_{x,y\in\Omega_i}\frac{\alpha(x)}{\alpha(y)}\,, \tag{6}$$

with $C$ independent of $\alpha$, i. e., the bound is proportional to the maximum local variation of $\alpha$ on the subdomains, which can be rather large. As noticed by several authors [5,9] this asymptotic bound is in general far too pessimistic, and robustness is observed for many special kinds of coefficient distributions.

The aim of the present contribution is to give more theoretical insight on the coefficient-dependency. We summarize our recent work [10] considering variation in subdomain interiors in Section 2, and we give an outlook to new theoretical results for the case of variation near the subdomain interfaces in Section 3. Finally, Section 4 deals with the application to magnetostatic problems.

## 2 Variation in Subdomain Interiors

In this section we give a sharper estimate than (6) for the case of variation in the subdomain interiors. On each subdomain $\Omega_i$ with diameter $H_i$ and discretization parameter $h_i$, we choose a width $\eta_i \in [h_i, H_i/2]$ and define the boundary layer $\Omega_{i,\eta_i}$ by the agglomeration of those finite elements which have distance at most $\eta_i$ from the boundary, cf. Fig. 2, left. Under suitable assumptions on the geometric setting and the subdomain partition, we can prove the bound

$$C^*(\alpha) \;\leq\; C \max_{j=1}^{N} \left(\frac{H_j}{\eta_j}\right)^2 \max_{i=1}^{N} \max_{x,y\in\Omega_{i,\eta_i}} \frac{\alpha(x)}{\alpha(y)}\,. \tag{7}$$

This bound involves only the variation of $\alpha$ in the boundary layer $\Omega_{i,\eta_i}$ and is independent of the variation of $\alpha$ in the subdomain interior $\Omega_i \setminus \Omega_{i,\eta_i}$. For $\eta_j \sim H_j$ we reproduce the known estimate (6), in particular our bound is still robust with respect to large jumps across the subdomain interfaces. However, if $\alpha$ exhibits large (even arbitrary) variation in the interior $\Omega_i \setminus \Omega_{i,\eta_i}$ of the subdomains, but varies little in the boundary layers, our new bound (7) is in general far better/sharper than (6). Moreover, if in addition the coefficient is larger in the interior $\Omega_i \setminus \Omega_{i,\eta_i}$ than in the boundary layer on each subdomain, then the quadratic factor $(H_j/\eta_j)^2$ reduces to a linear factor $H_j/\eta_j$. The detailed proof can be found in our recent paper [10].

In the following we give a two-dimensional numerical example. We partition the unit square into 25 congruent, square-shaped subdomains. The coefficient is chosen

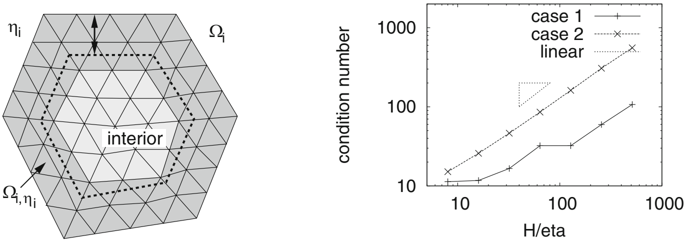


**Fig. 2:** *Left*: Subdomain boundary layer. *Right:* Estimated condition numbers $\kappa$ for varying width parameter $\eta$, fixed discretization parameter $h$ (logarithmic scales)

to be $\alpha = 10^5$ (Case 1) and $\alpha = 10^{-5}$ (Case 2) in the subdomain interiors, and $\alpha = 1$ on the rest. The distance between the "material" jump and the subdomain interfaces is denoted by $\eta$. We have used a globally uniform discretization with $H/h = 512$. Fig. 2, right, shows the condition numbers $\kappa$ of the preconditioned FETI systems (estimated by Lanczos' method) for different values of $\eta$. We see that our asymptotic bound is sharp for Case 1, but still slightly pessimistic for Case 2.

## 3 Interface Variation

In this section we would like to give an outlook on our work for interface variation which will be exposed in more detail in an upcoming paper. A key tool to the analysis of FETI methods is Poincaré's inequality,

$$\int_{\Omega_i} |w(x)|^2 \,\mathrm{d}x \;\leq\; C_P H_i^2 \int_{\Omega_i} |\nabla w(x)|^2 \,\mathrm{d}x,$$

which holds for all $w \in H^1(\Omega_i)$ with vanishing mean value, i. e., $\int_{\Omega_i} w(x)\,\mathrm{d}x = 0$. The constant $C_P > 0$ depends only on the shape of $\Omega_i$. A similar inequality holds if the average of $w$ over a part of the boundary $\partial\Omega_i$ vanishes. Concerning heterogeneous coefficients, we would be interested in an inequality of the same form but where the integrals are weighted with the coefficient $\alpha(x)$ and where the constant $C_P$ does not depend on $\alpha$, or at least only very mildly on the heterogeneity in $\alpha$. Such inequalities are not known in general, but we can show one for a special case.

Assume that each subdomain $\Omega_i$ consists of two connected subregions $\Omega_i^{(1)}$, $\Omega_i^{(2)}$ where $\alpha$ is mildly varying, i. e.,

$$\underline{\alpha}_i^{(k)} \;\leq\; \alpha(x) \;\leq\; \overline{\alpha}_i^{(k)} \qquad \text{for all } x \in \Omega_i^{(k)},\ k = 1,\ 2,$$

with moderate ratios $\overline{\alpha}_i^{(k)}/\underline{\alpha}_i^{(k)}$; we can think of two quasi-homogeneous materials within each subdomain. Using two separate Poincaré inequalities one can show that

$$\int_{\Omega_i} \alpha(x)\,|w(x)|^2 \,\mathrm{d}x \;\leq\; \Big\{ \max_{k=1,2} C_P^{(k)} \frac{\overline{\alpha}_i^{(k)}}{\underline{\alpha}_i^{(k)}} \Big\} H_i^2 \int_{\Omega_i} \alpha(x)\,|\nabla w(x)|^2 \,\mathrm{d}x, \qquad (8)$$

for all functions $w \in H^1(\Omega_i)$ which have vanishing mean value over the interface $\Lambda_i := \partial\Omega_i^{(1)} \cap \partial\Omega_i^{(2)}$, i. e., $\int_{\Lambda_i} w(x)\,ds_x = 0$. The constants $C_P^{(1)}$ and $C_P^{(2)}$ depend only on the shapes of the subregions $\Omega_i^{(1)}$ and $\Omega_i^{(2)}$ respectively, and on the relative shape of $\Lambda_i$. For a variant of FETI called *all-floating* FETI method [11–13], our Poincaré type inequality (8) finally allows a proof of the bound

$$C^*(\alpha) \;\leq\; C \max_{i=1}^{N} \max_{k=1,2} \frac{\overline{\alpha}_i^{(k)}}{\underline{\alpha}_i^{(k)}}, \qquad (9)$$

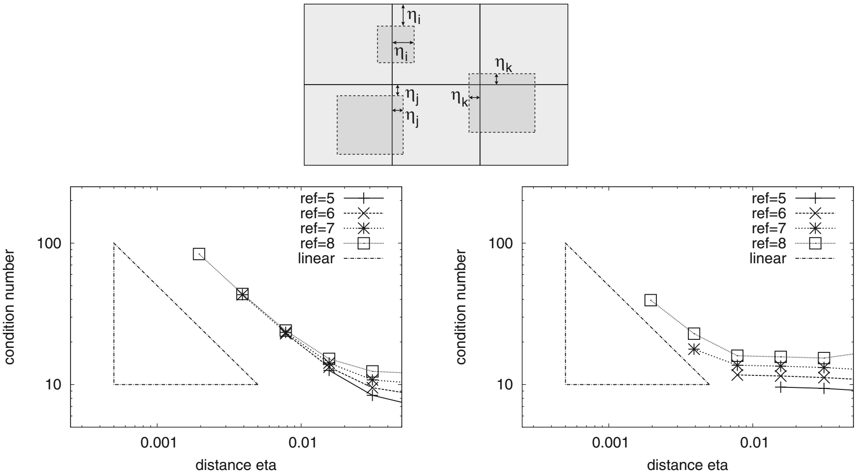


**Fig. 3:** *Upper*: Sketch of coefficient "islands" cutting through edges and crosspoints of the subdomain partitioning. *Lower left*: Condition numbers for "edge islands" for different levels of refinement, $H/h = 2^{\text{ref}}$. *Lower right*: Condition numbers for "crosspoint islands"

where the constant $C$ is independent of $H_i$, $h_i$, $N$, and $\alpha$, but it depends on the geometry of the subregions $\Omega_i^{(k)}$. Combining this idea with the theory from Section 2, one can even allow three qualitatively different subregions per subdomain:

- two connected subregions of mild variation in $\alpha$ that cover the boundary layer $\Omega_{i,\eta_i}$ of the subdomain, and
- a remaining part contained in the subdomain interior, where arbitrary variation of $\alpha$ can be allowed.

Under suitable assumptions on the shapes of these subregions it is again possible to give explicit bounds for $C^*(\alpha)$ involving (9) and the ratios $H_i/\eta_i$ similar to (7). For numerical examples we have tested so-called coefficient "islands" which cut through an *edge*, i. e., the interface of two subdomains, or which contain a *crosspoint* of four subdomains, cf. Fig. 3, upper. A suitable choice for $\eta$, the width of the boundary layer, is also indicated in that figure. Note, however, that we have only tested one island at a time. In each example we have set the coefficient $\alpha = 10^5$ in the island, and $\alpha = 1$ elsewhere. The estimated condition numbers for different values of $\eta$ and different levels of mesh refinement are depicted in Fig. 3, lower.

## 4 Application to Magnetostatic Problems

In the case of nonlinear magnetostatics in two dimensions (transverse magnetic mode), we have to solve

$$-\nabla \cdot [\nu_i(|\nabla u|)\nabla u] \,=\, f \qquad \text{in } \Omega_i\,, \tag{10}$$

subject to suitable interface and boundary conditions, where $u$ is the $z$-component of the magnetostatic vector potential, and $\nu_i$ is the reluctivity. For linear materials, $\nu_i$ is constant. For other materials, such as ferromagnetic ones, the reluctivity $\nu_i$ depends nonlinearly on the magnetic flux density $|\mathbf{B}| = |\nabla u|$, and it is defined by the material law $\mathbf{H} = \nu_i(|\mathbf{B}|)\mathbf{B}$ in $\Omega_i$, where $\mathbf{H}$ denotes the magnetizing force (note that we restrict ourselves to isotropic materials and neglect hysteresis). In our numerical computations we use realistic approximations of such material curves obtained from the interproximation method proposed in [14]. If we apply Newton's method to (10), the linearized system in each Newton step is of similar form as problem (1), only that we obtain a matrix-valued coefficient which depends on the current Newton iterate $u^{(k)}$, see, e. g., [5]. For many material curves, the variation of the coefficient depends mainly on the variation of $|\mathbf{B}|$. However, the flux density $|\mathbf{B}|$ may vary strongly along subdomain boundaries and large values of $|\mathbf{B}|$ appear mostly at singularities of the potential $u$, e. g., near material corners.

Contrary to the usual suggestion to choose subdomain partitions that resolve material interfaces in order to obtain robustness (for numerical examples see [5, 6]), our new bounds (7), (9) suggest that it might be more advantageous to put each peak of $|\mathbf{B}|$ and thus each material corner into the center of a subdomain. Fig. 4 shows two such examples. In both cases, the coefficient variation is approximately $7 \times 10^3$ but our FETI solver performs extremely well (Case 1: condition number 8.5, Case 2: condition number 13.7, compared to 8.3 for a globally constant coefficient). Our theory for interior variation (Section 2) can perfectly explain the low condition number in Case 1 since the boundary variation is small. Inspecting Case 2, we find that there are indeed two regions contained in the boundary layer with qualitatively different coefficients, see the jump in Fig. 4, lower right. Thus, Section 3 partially explains why the condition number is still quite robust with respect to the highly heterogeneous coefficient.

**Acknowledgements** We would like to thank Ulrich Langer for his encouragement. The first author acknowledges the financial support by the Austrian Science Funds (FWF), grant SFB F013.

## References

1. Farhat, C., Roux, F.X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. Int. J. Numer. Meth. Engrg. **32**(6), 1205–1227 (1991)
2. Toselli, A., Widlund, O.B.: Domain Decomposition Methods – Algorithms and Theory, *Springer Series in Computational Mathematics*, vol. 34. Springer, Berlin Heidelberg (2005)
3. Langer, U., Steinbach, O.: Boundary element tearing and interconnecting method. Computing **71**(3), 205–228 (2003)
4. Langer, U., Steinbach, O.: Coupled boundary and finite element tearing and interconnecting methods. In: Lecture Notes in Computational Sciences and Engineering, vol. 40, pp. 83–97. Springer, Heidelberg (2004)

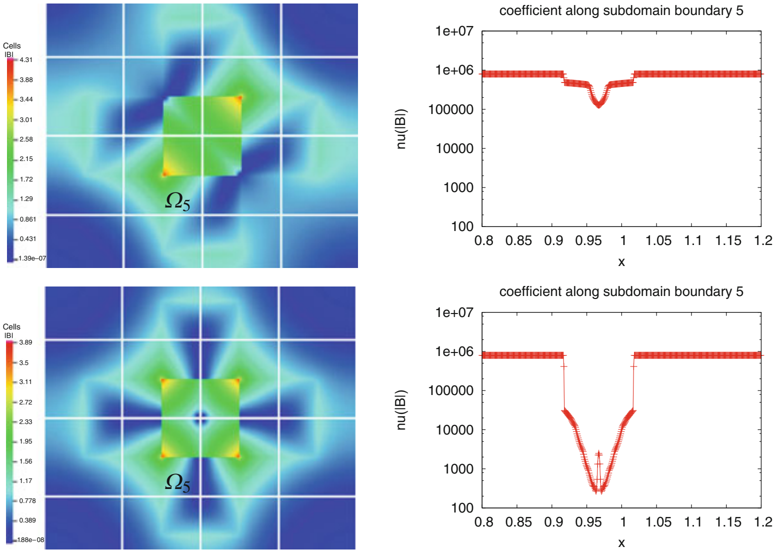


**Fig. 4:** *Upper*: Case 1. *Lower*: Case 2. *Left*: $|\mathbf{B}|$-field, subdomain partition. *Right*: Coefficient $\nu(|\mathbf{B}|)$ plotted along the boundary of subdomain $\Omega_5$

5. Langer, U., Pechstein, C.: Coupled finite and boundary element tearing and interconnecting solvers for nonlinear potential problems. Z. Angew. Math. Mech. **86**(12) (2006)
6. Langer, U., Pechstein, C.: Coupled FETI/BETI solvers for nonlinear potential problems in (un)bounded domains. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering 2006, *Mathematics in Industry: The European Consortium for Mathematics in Industry*, vol. 11, pp. 371–377. Springer, Berlin Heidelberg (2007)
7. Pechstein, C.: Boundary element tearing and interconnecting methods in unbounded domains (2008). To appear in *Applied Numerical Mathematics*
8. Klawonn, A., Widlund, O.B.: FETI and Neumann-Neumann iterative substructuring methods: Connections and new results. Comm. Pure Appl. Math. **54**(1), 57–90 (2001)
9. Rixen, D., Farhat, C.: A simple and efficient extension of a class of substructure based preconditioners to heterogeneous structural mechanics problems. Internat. J. Numer. Methods Engrg. **44**(4), 489–516 (1999)
10. Pechstein, C., Scheichl, R.: Analysis of FETI methods for multiscale PDEs. Numer. Math. **111**(2), 293–233 (2008)
11. Dostál, Z., Horák, D., Kučera, R.: Total FETI – An easier implementable variant of the FETI method for numerical solution of elliptic PDE. Commun. Numer. Methods Eng. **22**(12), 1155–1162 (2006)
12. Of, G.: BETI-Gebietszerlegungsmethoden mit schnellen Randelementverfahren und Anwendungen. Ph.D. thesis, Universität Stuttgart, Germany (2006)
13. Of, G.: The all-floating BETI method: Numerical results. In: Domain Decomposition Methods in Science and Engineering XVII, *Lecture Notes in Computational Science and Engineering*, vol. 60, pp. 295–302. Springer, Berlin Heidelberg (2008)
14. Pechstein, C., Jüttler, B.: Monotonicity-preserving interproximation of *B*-*H*-curves. Journal of Computational and Applied Mathematics **196**(1), 45–57 (2006)

# Nonlinear Models for Silicon Semiconductors

Salvatore La Rosa, Giovanni Mascali, and Vittorio Romano

**Abstract** In this paper we present exact closures of the 8-moment and the 9-moment models for the charge transport in silicon semiconductors based on the maximum entropy principle. The validity of these models is assessed by numerical simulations of an n-$^+$n-n$^+$ device. The results are compared with those obtained from the numerical solution of the Boltzmann Transport Equation both by Monte Carlo method and directly by a finite difference scheme.

## 1 Introduction

Simulation of modern electronic devices requires increasingly accurate models of charge transport in semiconductors in order to describe high-field phenomena such as hot electron propagation, impact ionization and heat generation. Moreover, in many applications in optoelectronics, it is necessary to describe the transient interaction of electromagnetic radiation with carriers in complex semiconductor materials: in these cases the characteristic times are of the order of the electron momentum or the energy flux relaxation times. These are some of the main reasons of the necessity of developing models which incorporate a number of moments of the distribution function higher than those in the drift-diffusion and the energy transport models.

These extended models, generally called hydrodynamical models, are usually derived from the infinite hierarchy of the moment equations of the Boltzmann Transport Equation (BTE) by suitable truncation procedures. One of the most successful

Salvatore La Rosa, Vittorio Romano
Dipartimento di Matematica e Informatica, Università di Catania, viale A. Doria 6, 95125 Catania, Italy, e-mail: larosa@dmi.unict.it, romano@dmi.unict.it

Giovanni Mascali
Dipartimento di Matematica, Università della Calabria and INFN-Gruppo c. Cosenza, 87036 Cosenza, Italy, e-mail: g.mascali@unical.it

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_53,

among these procedures is that based on the Maximum Entropy Principle (MEP) [1], see [2] for a complete review both for Si and GaAs semiconductors.

The models differ for the number of moments which are used and they usually comprise the balance equations of the electron density, the energy density, the average velocity, the energy flux and possibly also higher scalar and vector moments which do not have an immediate physical interpretation. In this paper we present the usual 8-moment model [3] together with a 9-moment model in which a further scalar moment is added: that corresponding to the squared microscopic electron energy. The two models are assessed by applying them to the benchmark problem of an n-$^+$n-n$^+$ silicon device.

## 2 Hydrodynamical Models with 8 and 9-Moments

In [3] we presented an 8-moment model for charge transport in semiconductors and we assessed its validity. In principle, one can try to improve this model by adding further scalar and vector moments as well as higher order tensor moments. Adding the scalar moment $nW_2$, one obtains a new model which is given by the following system of balance equations

$$\frac{\partial n}{\partial t}+\frac{\partial (nV^i)}{\partial x^i}=0, \tag{1}$$

$$\frac{\partial (nV^i)}{\partial t}+\frac{\partial (nU^{ij})}{\partial x^j}+neE_jH^{ij}=nC_{V^i}, \tag{2}$$

$$\frac{\partial (nW)}{\partial t}+\frac{\partial (nS^i)}{\partial x^i}+neV_iE^i=nC_W, \tag{3}$$

$$\frac{\partial (nS^i)}{\partial t}+\frac{\partial (nF^{ij})}{\partial x^j}+neE_jG^{ij}=nC_{S^i}. \tag{4}$$

$$\frac{\partial (nW_2)}{\partial t}+\frac{\partial (nS_2^i)}{\partial x^i}+2neE_iS^i=nC_{W_2}, \tag{5}$$

where $e$ is the absolute value of the electron charge and $\mathbf{E}$ the electric field. The macroscopic quantities, which are involved in the balance equations, are related to the electron distribution function $f(\mathbf{x},\mathbf{k},t)$ by the definitions

$$n=\int_{\mathbb{R}^3} f d\mathbf{k}, \qquad \text{electron density,}$$

$$W=\frac{1}{n}\int_{\mathbb{R}^3}\mathscr{E}(k)f d\mathbf{k}, \qquad \text{average electron energy,}$$

$$W_2=\frac{1}{n}\int_{\mathbb{R}^3}\mathscr{E}^2(k)f d\mathbf{k}, \qquad \text{average electron energy square,}$$

$$
\begin{aligned}
V^i &= \frac{1}{n}\int_{\mathbb{R}^3} f v^i \mathrm{d}\mathbf{k}, \quad \text{average velocity},\\
S^i &= \frac{1}{n}\int_{\mathbb{R}^3} f v^i \mathscr{E}(k) \mathrm{d}\mathbf{k}, \quad \text{energy flux},\\
S_2^i &= \frac{1}{n}\int_{\mathbb{R}^3} f v^i \mathscr{E}^2(k) \mathrm{d}\mathbf{k}, \quad \text{flux of the electron}\\
&\qquad\qquad\qquad\qquad\qquad \text{energy square},\\
U^{ij} &= \frac{1}{n}\int_{\mathbb{R}^3} f v^i v^j \mathrm{d}\mathbf{k}, \quad \text{velocity flux},\\
H^{ij} &= \frac{1}{n}\int_{\mathbb{R}^3} \frac{1}{\hbar} f \frac{\partial}{\partial k_j}(v^i) \mathrm{d}\mathbf{k}, \quad \text{(no physical interpretation)},\\
F^{ij} &= \frac{1}{n}\int_{\mathbb{R}^3} f v^i v^j \mathscr{E}(k) \mathrm{d}\mathbf{k}, \quad \text{flux of the energy}\\
&\qquad\qquad\qquad\qquad\qquad \text{flux},\\
G^{ij} &= \frac{1}{n}\int_{\mathbb{R}^3} \frac{1}{\hbar} f \frac{\partial}{\partial k_j}(\mathscr{E} v^i) \mathrm{d}\mathbf{k}, \quad \text{(no physical interpretation)},
\end{aligned}
$$

$$
\begin{aligned}
C_{V^i} &= \frac{1}{n}\int_{\mathbb{R}^3} \mathscr{C}[f] v^i \mathrm{d}\mathbf{k}, \quad \text{velocity production},\\
C_W &= \frac{1}{n}\int_{\mathbb{R}^3} \mathscr{C}[f] \mathscr{E}(k) \mathrm{d}\mathbf{k}, \quad \text{energy production},\\
C_{S^i} &= \frac{1}{n}\int_{\mathbb{R}^3} \mathscr{C}[f] v^i \mathscr{E}(k) \mathrm{d}\mathbf{k}, \quad \text{energy flux}\\
&\qquad\qquad\qquad\qquad\qquad \text{production},\\
C_{W_2} &= \frac{1}{n}\int_{\mathbb{R}^3} \mathscr{C}[f] \mathscr{E}^2(k) \mathrm{d}\mathbf{k}, \quad \text{electron energy}\\
&\qquad\qquad\qquad\qquad\qquad \text{square production},
\end{aligned}
$$

here $\mathscr{E}$ and $\mathbf{k}$ respectively are the electron energy in the conduction band and the wave vector, and $\mathscr{C}[f]$ is the collision operator which appears at the left hand of the BTE. These equations are coupled to the Poisson equation for the electric potential $\phi$

$$
\begin{aligned}
E^i &= -\frac{\partial \phi}{\partial x_i}, && (6)\\
\nabla \cdot (\varepsilon \nabla \phi) &= -e(N_+ - N_- - n), && (7)
\end{aligned}
$$

where $\varepsilon$ is the electric permittivity and $N_+$ and $N_-$ are the donor and acceptor density respectively (which depend only on the position).

The system (1)–(5) is not closed since the fluxes $\mathbf{S}_2, U, H, F, G$ and the production terms $C_{\mathbf{V}}, C_W, C_{\mathbf{S}}, C_{W_2}$ have to be expressed as functions of the fundamental variables $n, \mathbf{V}, W, \mathbf{S}$ and $W_2$. The closure can be achieved by means of MEP, using the distribution function which maximizes missing information (entropy) in order to evaluate the unknown moments. This distribution function is called the maximum entropy distribution function $f_{ME}$ [1, 2]. The MEP approach leads to a constrained optimization problem which is handled by resorting to the Lagrangian multipliers method, see [2] and references therein. In the present case, the constraints consist of the known moments $n, \mathbf{V}, W, \mathbf{S}$ and $W_2$ and they are used to express the Lagrange multipliers in terms of these moments. Actually, this is a highly non-linear problem, which, in the past, has been solved by assuming the distribution function to be slightly anisotropic and expanding it with respect to a suitable anisotropy parameter [2]. Recently [3] this problem has been solved numerically without resorting to asymptotic procedures. In this way the model is expressed in terms of the Lagrangian multipliers and the constitutive relations are given by integral expressions that do not allow an efficient numerical tabulation, but require the use of suitable quadrature formulas with respect to the microscopic energy. The interested reader is referred to [3] for the closure relations relative to the 8-moment model, the relations referring to the further quantities appearing in (5) being completely analogous. At the end apart from the Poisson equation, the resulting system is hyperbolic in the physically relevant region of the field variables.

It is important to notice that in the numerical integration of the models problems can arise due to the fact that there may exist moments that are not moments of the maximum entropy distribution [4]. In fact the set of the moments generated by $f_{ME}$ is a convex cone $\mathscr{M}$ [5]. In the 8-moment case $\mathscr{M}$ is generated by the Lagrangian multipliers such that

$$g(\lambda^W, \lambda^{\mathbf{S}}) = \lambda^W - \sqrt{\frac{1}{2\alpha m^*}} ||\lambda^{\mathbf{S}}|| > 0, \tag{8}$$

while in the 9-moment case the cone is generated by the Lagrangian multipliers which satisfy

$$\lambda^{W_2} > 0. \tag{9}$$

Here $\alpha$ is the non-parabolicity factor, $m^*$ the electron effective mass, $\lambda^W, \lambda^{\mathbf{S}}$ and $\lambda^{W_2}$ are the Lagrange multipliers which correspond to $W$, $\mathbf{S}$ and $W_2$, respectively. The conditions are obtained by requiring the integrability of $f_{ME}$.

## 3 Simulation of an n-$^+$n-n$^+$ Device

We have tested the 8 and 9-moment models by numerically solving them in the 1-D problem of an n-$^+$n-n$^+$ device, which is commonly used as a benchmark problem [6]. In this case the systems have the following form

$$\frac{\partial F^{(0)}(\Lambda)}{\partial t} + \frac{\partial F^{(1)}(\Lambda)}{\partial x} = G(\Lambda, E), \tag{10}$$

where $\Lambda$ is the vector of the unknown Lagrange multipliers, $F^{(0)}$ is the vector of the moments, $F^{(1)}$ is the vector of the fluxes and $G$ is the vector of the sources which takes into account both the effect of the scatterings that electrons suffer inside the device and the driving effect of the electric field. (10) is solved by using a splitting strategy, which consists of two successive steps [3]: the first step solves the system without sources (convection step), while the second step solves the system with the fluxes put equal to zero (relaxation step).

The convection step makes use of the Nessyahu–Tadmor scheme [7], which does not require the explicit knowledge of the characteristic structure of the system and is conservative and consistent. The latter two properties are necessary requirements for having correct shock capturing methods. The relaxation system is a system of ordinary differential equations, which can be solved by using an explicit Euler scheme.

The devices which have been considered are those reported in Table 1.

**Table 1:** $L_c$ length of the channel, doping concentration (respectively in the $n^+$ and $n$ regions) and $V_b$ applied voltage

| Channel length<br>$L_c$ ($\mu$m) | $n^+$<br>($10^{17}$ cm$^{-3}$) | $n$<br>($10^{17}$ cm$^{-3}$) | $V_b$<br>Volt |
|---|---|---|---|
| 0.2 | 10 | 0.1 | 1 |
| 0.1 | 10 | 0.1 | 1 |

The results of the two non-linear models presented here (indicated by 8 and 9-moment NLMEP models respectively) have been compared with those obtained by the direct solution of the BTE (DSBE), with Monte Carlo results (MC) and also with those derived by means of the model in which the closure is based on the asymptotic expansion (indicated as SLMEP model) [8]. The aim is threefold:

- to check the validity of the 8 and 9 moment models,
- to assess the relevance of the nonlinearity,
- to see if the integrability condition is always satisfied.

As regards the validity, we can say that the results of the 8-moment model are satisfactory. In fact, as can be seen from Figures 1 and 2, which refer to devices with channel length equal to $0.2\mu$m and $0.1\mu$m respectively, the 8-moment NLMEP model gives the solutions closest to those obtained both with the MC method and the direct integration of the BTE. This means that the anisotropy effects are not small when the channel is short and there are high electric fields inside the device. The solutions do not show any spurious oscillations which indicates that the assumed boundary conditions are compatible with the solutions of the problem: we have used Dirichlet conditions on the density and Neumann conditions on the Lagrange multipliers corresponding to the remaining moments, which are the fundamental variables of the model. Furthermore we also notice that the peak in the velocity near the second junction almost disappears in accordance with MC and DSBE results.

As regards the integrability, the problem is subtle. In fact in the transient there are wide oscillations which can bring the numerical solution out of the cone $\mathscr{M}$. As can be seen from Figure 3 left, for a device with a channel length of $0.2\mu$m this can be tackled by improving the precision of the numerical integration with respect to the microscopic energy in the closure relations. The numerical integration is effected by using the Gauss–Legendre formula and passing from 140 nodes to 310 nodes in the microscopic energy interval $[0\,\text{eV}, 1.6\,\text{eV}]$ the integrability is recovered. The situation is different when the channel length is $0.1\mu$m, in this case in fact there is a region near the first junction, see Figure 3 right, where the integrability does not improve even by increasing the number of nodes. The case of the 9-moment model is worse; in fact, as can be seen from Figure 4 right below we do not have integrability, independently on how precise the integration is. This is probably due to the fact that additional Lagrangian multipliers, associated to new moments corresponding to weight functions represented by powers of energy with an exponent greater than one, are zero at equilibrium states which are, therefore, located at the boundary of the realizability region. This implies that small perturbations can have both positive and negative sign causing a loss of integrability and limiting the validity of the non-linear models. As a consequence the solution of the 9-moment model, Figures $4_{1-3}$, is clearly unreliable.

## 4 Conclusion

In conclusion, the results, which we have obtained, make us affirm that a great attention has to be payed to whether the integrability condition is satisfied when using a completely non-linear model. The problem could be effectively solved by using a better approximation for the energy bands, in which the Brillouin zone, instead of being extended to all $\mathbb{R}^3$ as for the Kane dispersion relation, is a limited region as in the physical case.

**Acknowledgements** The authors acknowledge the financial support by M.I.U.R. (PRIN 2007 *Equazioni cinetiche e idrodinamiche di sistemi collisionali e complessi*), P.R.A. University of Catania and University of Calabria (ex 60 %), CNR (grant n. 00.00128.ST74), the EU Marie Curie RTN project COMSON grant n. MRTN-CT-2005-019417.

## References

1. Jaynes, E.T.: Information theory and Statistical Mechanics, Phys. Rev. **106**, 620–630 (1957)
2. Anile, A.M., Mascali, G., Romano, V.: Recent developments in Hydrodynamical Modeling of semiconductors. In : Anile A. M. , Allegretto, W., Ringhofer, C. (ed.): Mathematical Problems in Semiconductor Physics, Lecture Notes in Mathematics n. 1823, pp. 1–56, Springer, Berlin (2003)
3. La Rosa, S., Mascali, G., Romano, V.: Exact Maximum Entropy Closure of the Hydrodynamical Model for Si Semiconductors: the 8-Moment Case. Submitted (2008)
4. Junk, M.: Domain of definition of Levermore's five-moment system, J. Stat. Phys. **93**, 1143–1167 (1998)

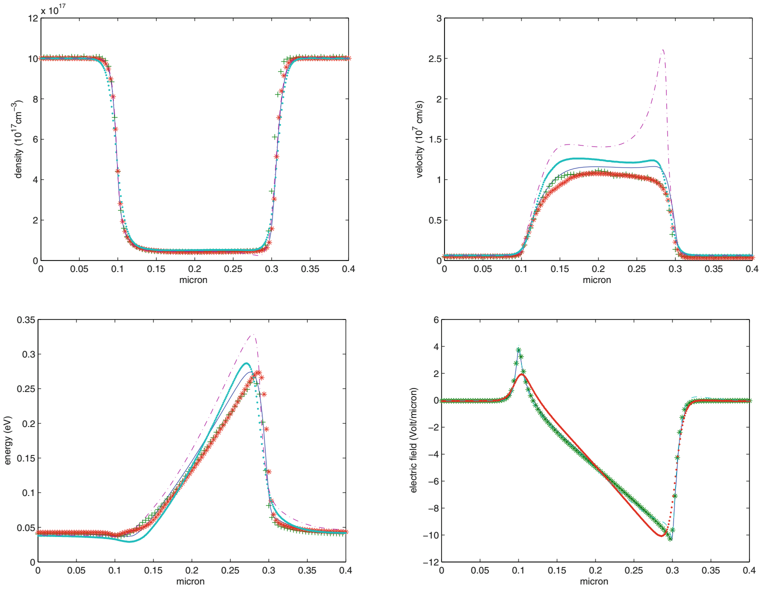


**Fig. 1:** $L_c = 0.2$ $\mu$m, stationary solution: 8 moment-NLMEP model (*continuous line*), SLMEP model (*dotted line*), MC simulation (*crossed line*), direct Boltzmann integration (*starry line*) and Baccarani Blotekjaer Wordeman (BBW) model (*dashed-dotted line*)

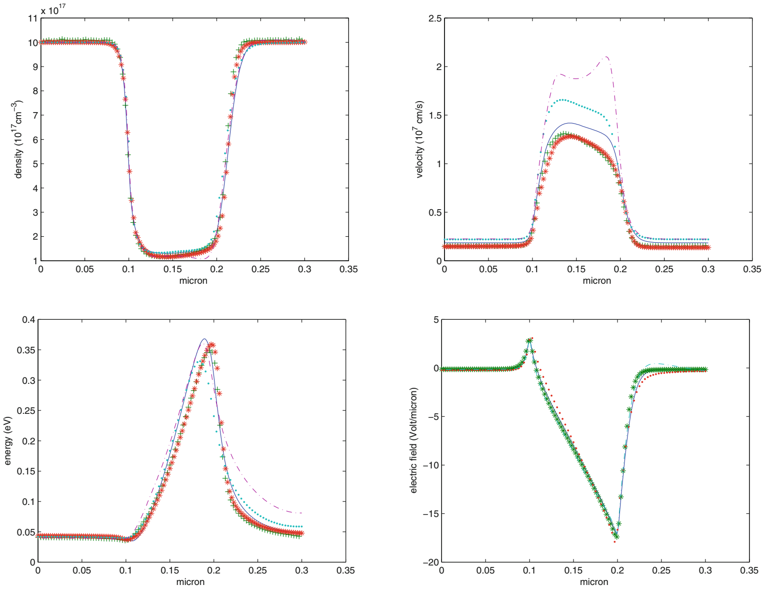


**Fig. 2:** $L_c = 0.1$ $\mu$m, stationary solution: 8 moment-NLMEP model (*continuous line*), SLMEP model (*dotted line*), MC simulation (*crossed line*), direct Boltzmann integration (*starry line*) and BBW model (*dotted line*)

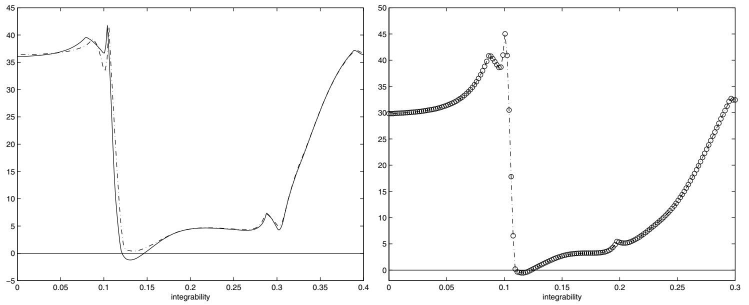


**Fig. 3:** $L_c = 0.2 \, \text{and} \, 0.1 \mu\text{m}$, integrability condition. *Left*: 140 nodes in the energy interval (in eV) [0,1.55] (*continuous line*), 310 nodes in [0,1.6] (*dashed-dotted line*). *Right*: 490 nodes in [0,3.39] (*circles*), 735 nodes in [0,3.39] (*dashed-dotted line*)

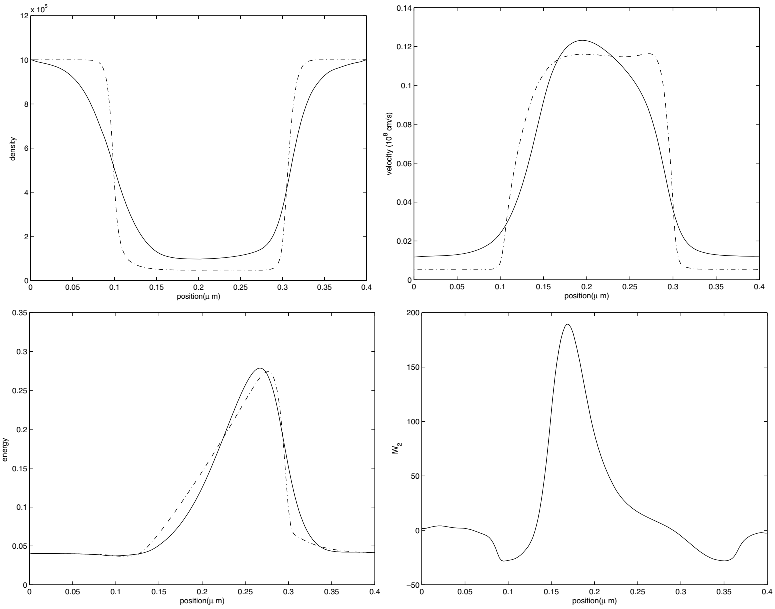


**Fig. 4:** $L_c = 0.2 \ \mu\text{m}$, stationary solution: 8 moment-NLMEP model (*dashed-dotted line*), 9 moment-NLME model (*continuous line*). Integrability: plot of $\lambda^{W_2}$ versus the position

5. Junk, M., Romano, V.: Maximum entropy moment system of the semiconductor Boltzmann equation using Kane's dispersion relation, Cont. Mech. Thermodyn., **17**, 247–267 (2005)
6. Anile, A.M., Carrillo, J.A., Gamba, I.M., Shu, C.W.: Approximation of the BTE by a relaxation-time operator. VLSI design Journal, **13**, 349–354 (2001)
7. Nessyahu, H., Tadmor, E.: Non-oscillatory central differencing for hyperbolic conservation law. J. Comp. Phys. **87**, 408–463 (1990)
8. Romano, V.: Nonparabolic band hydrodynamical model of silicon semiconductors and simulation of electron devices. Math. Meth. Appl. Sciences **24**, 439–471 (2001)

# Multiobjective Optimization Applied to Design of PIFA Antennas

Stefan Jakobsson, Björn Andersson, and Fredrik Edelvik

**Abstract** In this paper multiobjective optimization is applied to antenna design. The optimization algorithm is a novel response surface method based on approximation with radial basis functions. It is combined with CAD and mesh generation software, and electromagnetic solvers. To demonstrate the procedure we optimize the geometric design and feed position of a PIFA antenna located on a ground plane.

## 1 Introduction

In many engineering applications there are often, at least partly, conflicting requirements. In antenna design the requirements can be based on size, S-parameters, functions of the directivity of the antenna, bandwidth, input impedance and/or other characteristics of the antenna. The usual way of treating such problems is to employ a weight-based trial-and-error strategy, where the objectives are weighted to form a single objective function. This approach has several disadvantages - the weights are often highly sensitive and no trade-off discussion is possible.

A more attractive alternative is to avoid weights and instead optimize the objective functions simultaneously subject to certain constraints. In multi-objective optimization the ultimate goal is to find all Pareto optimal solutions: a solution is Pareto optimal if there is no other solution which is better in all objectives. The decision making process, when the antenna engineer decides which of the Pareto optimal designs that best meets the requirements, takes place when all possibilities and limitations are known.

In a current project we are developing new efficient optimization algorithms with the purpose of studying communication performance possibilities and limitations for multiple antennas within a limited area, such as a handheld terminal. The antenna

Stefan Jakobsson, Björn Andersson, Fredrik Edelvik
FCC–Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Göteborg, Sweden, e-mail: stefan.jakobsson@fcc.chalmers.se, bjorn.andersson@fcc.chalmers.se, fredrik.edelvik@fcc.chalmers.se

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_54,

elements are so-called printed inverted "F" antenna designs (PIFA) that have low profile, good radiation characteristics and wide bandwidths. This makes them an attractive choice for antenna designs for various wireless systems. The project is a collaboration between the Fraunhofer-Chalmers Centre and the Antenna Research Centre at Ericsson AB.

We have developed a multiobjective optimization algorithm based on radial basis functions to find an approximation of the *Pareto front* (the set of Pareto solutions). In this paper the algorithm is demonstrated for optimization of the design of a PIFA antenna on a ground plane. The objective functions are the maximum return loss ($|S_{11}|$ in dB) in a frequency band, the height and the enclosed area of the antenna element. The design parameters describe the geometry and feed position of the PIFA on the ground plane. The electromagnetic simulations are performed with the MoM solver from the software package efield® [1], which utilizes CADfix [2] for mesh generation.

## 2 A Multiobjective Optimization Algorithm

The optimization algorithm used, called *qualSolve*, is a response surface method based on interpolation/approximation with radial basis functions and is described in [3]. In each iteration of the optimization an interpolation/approximation (also called surrogate model) of every objective function is made based on all previous evaluations of the goal functions. By using for example evolutionary algorithms, an approximation of the Pareto front for the surrogate models is made. In the second step, a new evaluation point is chosen based on the approximate Pareto front. It is therefore crucial that the surrogate models are of as high quality as possible. Since the evaluations of the objective functions involve time-consuming simulations a reliable surrogate model can greatly improve efficiency compared to for example to genetic algorithms, as the number of evaluations of the objective functions are reduced.

### *2.1 Surrogate Models for Antenna Optimization*

It turns out that the output from antenna simulations (for example the $S$-parameters) have some characteristic properties that must be taken into account in order to achieve good approximations. Near resonance a small change in the design parameters often results in a large change in the output. We have tried different techniques for building surrogate models for the antenna data, including interpolation with Radial Basis Functions and approximation with rational functions. This led us to a new interpolation technique which we call rational radial basis function interpolation. Our experience is that direct application of standard interpolation with radial basis functions produce surrogate models which approximate the true objective functions poorly in the interesting regions in the parameter space, and rational approximation with polynomials has problems as the spatial dimension increases. The new interpolation method with Rational Radial Basis Functions is a combination of these two concepts which is both easy to handle and, as will be shown, produces accurate surrogate models.

### 2.1.1 Rational Interpolation/Approximation

Suppose we want to interpolate a function $f$ with a quotient of two functions $p$ and $q$ at the data points $\{\mathbf{x}_k\}_{k=1}^N, \mathbf{x}_k \in \mathbb{R}^d$ and $f(\mathbf{x}_k) = f_k \in \mathbb{C}$. Then we must have

$$f(\mathbf{x}_k) = \frac{p(\mathbf{x}_k)}{q(\mathbf{x}_k)}, \qquad k = 1, \ldots, N. \tag{1}$$

This does not determine the values of $p$ and $q$ at the data points. A good option is to choose them as simple as possible. For rational interpolation we choose $p$ and $q$ as polynomials of some order

$$p(x) = p_0 + p_1 x + \cdots + p_m x^m, \qquad q(x) = 1 + q_1 x + \cdots + q_n x^n.$$

The task is then to determine the coefficients of these polynomials so that the interpolation condition (2.1.1) holds for all $k = 1, \ldots, (m+n+1)$ (compare with Padé approximations for which the derivatives up to order $(m+n)$ agree with a given function's derivatives). If we have more data than coefficients, the coefficients can be chosen to minimize the least square error:

$$\min_{\substack{p \in \mathscr{P}_m, \\ q \in \mathscr{P}_n}} \sum_{k=1}^{N} \left| f_k - \frac{p(\mathbf{x}_k)}{q(\mathbf{x}_k)} \right|^2, \tag{2}$$

where $\mathscr{P}_m$ denotes the space of polynomials of order $m$. Rational approximation with polynomials works well for one dimensional data but cannot be generalized easily to higher dimensions. This is due to the fact that the number of coefficient to be determined increases very rapidly as the order of the polynomials and the spatial dimension increases. Our efforts so far have shown that rational approximation can be made to work well in two dimensions for the antenna data from our simulations but becomes too complicated in higher dimensions.

### 2.1.2 Radial Basis Functions

Interpolation with radial basis functions is a much used technique for approximation of scattered data in any dimension with a well developed theory, see for example by Buhmann or Wendland [4, 5]. A radial basis function (RBF) expansion is given by

$$s(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \phi(\|\mathbf{x} - \mathbf{x}_i\|) + p(\mathbf{x}),$$

where $\phi$ is the basis function and $p$ a polynomial of degree one. Common basis functions are $r^3$ (spline), $r^2 \log(r)$ (thin plate spline) and $\exp(-r^2)$ (Gaussian). The interpolation condition is $s(\mathbf{x}_k) = f_k$. An additional condition guarantees that $s = f$, if $f$ is a first degree polynomial. Connected to each basis function is the *Native space* $\mathcal{N}_\Phi$ and the corresponding *Native space norm* $\|\cdot\|_{\mathcal{N}_\Phi}$ which can be interpreted as a measure of the “bumpiness” of the function, see [5, Chapter 10]. The RBF interpolant is the unique function which interpolates all data and has the least native space norm. Compare with cubic splines which minimize the integral $\int |g''(t)|^2\, dt$ among all functions which interpolate the data.

In the next section we show one attempt of combining the flexibility of RBF expansions with the properties of rational interpolation.

## 2.2 Rational RBF Interpolation

Rational interpolation of scattered data in many dimensions have previously been investigated in [6]. We decided to use two RBF expansions, here called $p$ and $q$, as a basis for the rational interpolation. Hence we have

$$f(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}, \qquad \mathbf{x} \in \mathbb{R}^d.$$

The interpolation condition is

$$p(\mathbf{x}_k) = f_k q(\mathbf{x}_k), \qquad k = 1, \ldots, N$$

To define the values of $p$ and $q$ at the data points $\mathbf{x}_k$ we need an extra condition. It is natural to choose $p$ and $q$ as smooth as possible which means that the native space norm should be as small as possible relative to their values at the data points. The following minimization problem is a realization of this idea: Find the minimizer to the following problem:

$$\min_{\substack{p,q \in \mathcal{N}_\Phi, \\ \|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 = 1, \\ p(\mathbf{x}_k) = f(\mathbf{x}_k) q(\mathbf{x}_k).}} \left(\|p\|^2_{\mathcal{N}_\Phi} + \|q\|^2_{\mathcal{N}_\Phi}\right), \tag{3}$$

where $\|\cdot\|_{\mathcal{N}_\Phi}$ denotes the native space norm.

Figures 1 and 2 show results for the proposed algorithm in one and two dimensions. One reason for the good results is that $S_{11}$ behaves as a rational function of order one in these cases and such functions are interpolated exactly.

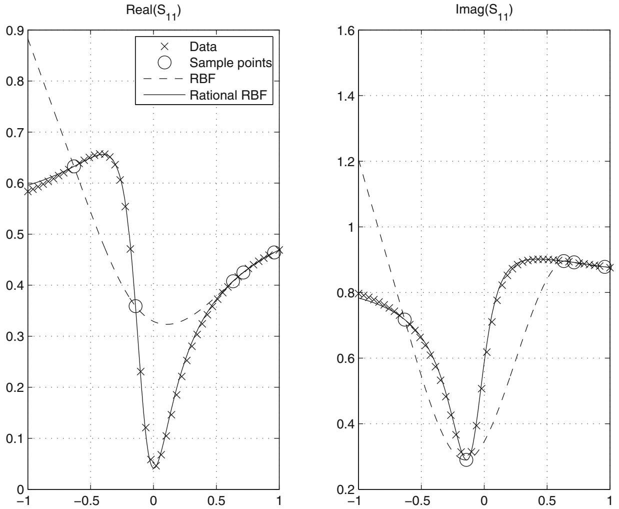


**Fig. 1:** A comparison of RBF interpolation and rational RBF interpolation for $S_{11}$ as a function of the design parameter $l_p$ (*cf.* Table 1) properly scaled

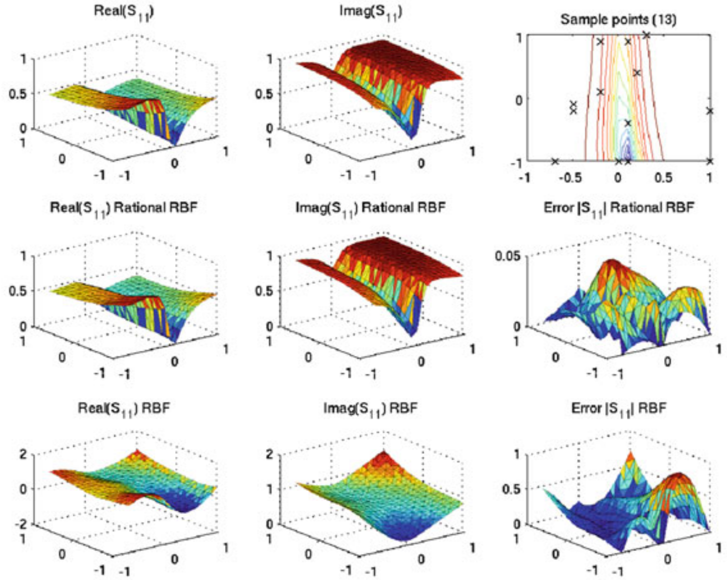


**Fig. 2:** A comparison of RBF interpolation and rational RBF interpolation for $S_{11}$ as a function of the two design parameters $l_p$ and $w_p$ (*cf.* Table 1) properly scaled. The data points used to build the interpolations are marked in the upper right subfigure

More details of this novel interpolation algorithm will be presented in a forthcoming paper.

## 3 Antenna Case

The optimization objective is to transmit as much electromagnetic energy as possible in a band of frequencies and at the same time minimize the footprint of the antenna. The antenna is defined by the parameters given in Table 1, where the first four parameters are the design variables in the optimization. A schematic view of the antenna can be seen in Figure 3.

These parameters describe the dimensions the antenna element as well as the feeding position. Note that upper limit of the feed position $x_p$ depends on the length of R, as we require the feed wire to be attached to the antenna. The feed itself is realized by means of a delta gap excitation with a characteristic impedance of 50 $\Omega$. As objective functions we choose the maximum return loss in the frequency band $f \in [2.5, 2.7]$ GHz, the area of R and the height of the antenna, all of which are to be minimized.

**Table 1:** Dimensions of the antenna. The first four are the design parameters in the optimization

| Parameter | Description | Value [mm] |
|---|---|---|
| $l_p$ | Length of R | $4.0 \leq l_p \leq 45.0$ |
| $w_p$ | Width of R | $3.0 \leq w_p \leq 15.0$ |
| $h_p$ | Height of the antenna | $2.0 \leq h_p \leq 10.0$ |
| $x_p$ | Feed position | $2.0 \leq x_p \leq l_p - 2.0$ |
| $l_g$ | Length of the ground plane | 100.0 |
| $w_g$ | Width of the ground plane | 45.0 |
| $t_g$ | Thickness of the ground plane | 2.0 |

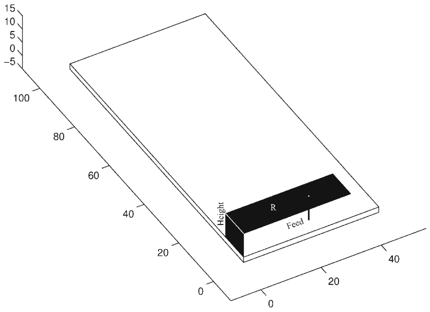
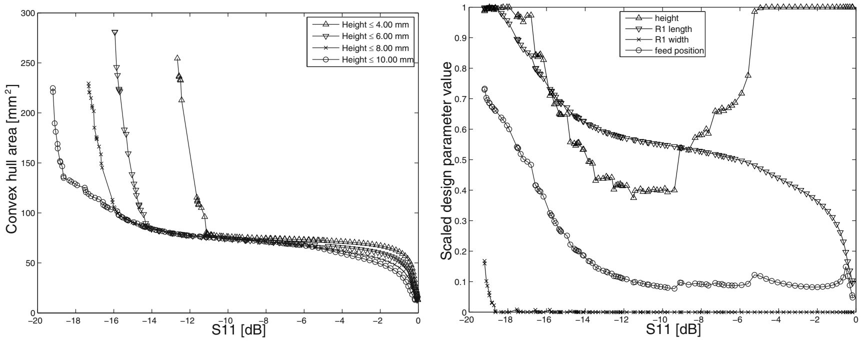


**Fig. 3:** Antenna to be optimized. The design parameters are the dimension of *R*, the height and the feed position

## 4 Results

The result of the optimization is plotted in Fig. 4a, where the area of the antenna element is plotted against $S_{11}$. The influence of the third objective, the height of the antenna, is visualized by four curves corresponding to different levels of constraints on the height. The design parameters corresponding to the black curve with $\circ$ markers are shown in Fig. 4b. A total of 1193 function evaluations were performed in order to extract the approximate Pareto front.



**(a)** Pareto front **(b)** Design parameters along the Pareto front

**Fig. 4:** To the *left figure* is the Pareto front and of the response surfaces. The four different lines corresponds to different constraints on the height. The *right figure* shows the design parameters along the front corresponding to the *black line with circle markers* in the *left figure*

We see that the three objectives are clearly opposing each other, except in a small area in the mid-range of $S_{11}$ where the four curves are close together and the area of the antenna element is relatively constant. We also see that the width of R is at its smallest allowed value at most of the front, which suggests that it might be beneficial to allow for smaller widths.

The convergence of the Pareto front can be seen in Fig. 5, where the front has been plotted for the full data set, the first 75%, and the first 50%. As there are only minor differences between the three curves it can be concluded that the algorithm has converged.

## 5 Conclusions

We have presented a multi-objective optimization algorithm based interpolation with rational radial basis functions. A key property of the algorithm is that the result is both a set of approximately Pareto-optimal solutions and also approximations of all objective function as expansions in radial basis function which can be used for

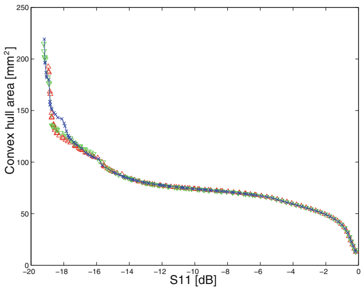


**Fig. 5:** Convergence of the Pareto front with no constraint on the height, corresponding to the *black line with circle markers* in Fig. 4a. Full data set with 1193 function evaluations (*blue crosses*), 75% (*green inverted triangles*), and 50% (*red triangles*)

further post-processing. This enables the algorithm to use less evaluations of the objective functions compared to e.g. genetic algorithms. Since the evaluations of the objective functions involve time-consuming simulations this fact can greatly improve efficiency. Another key aspect is that all numerical simulations contain errors (noise) and to replace interpolations with approximations is a feature of the algorithm to make it more robust. Finally, by avoiding a weight-based trial-and-error strategy, where the objectives are weighted to form a single objective function, the decision of the optimal solution is postponed until all possibilities and limitations are known.

The response surfaces shown for $S_{11}$ indicates how difficult this objective is to optimize, regardless of method. The minima are so sharp that a pure gradient based algorithm will have trouble finding its way down to the minima. We have also seen how difficult the true objective function is to interpolate correctly with standard methods. Future work includes studying a gradient based algorithm in combination with interpolation with the rational radial basis functions presented here.

**Acknowledgements** Financial support for this work has been provided by the Swedish Foundation for Strategic Research (SSF) through the Gothenburg Mathematical Modeling Centre (GMMC).

## References

1. efield®. http://www.efieldsolutions.com
2. TranscenData®. http://www.transcendata.com
3. Jakobsson, S., Patriksson, M., Rudholm, J., Wojciechowski, A.: A method for simulation based optimization using radial basis functions. To appear in Optimization and Engineering
4. Buhmann, M.D.: Radial basis functions, *Cambridge Monographs on Applied and Computational Mathematics*, vol. 12. Cambridge University Press, Cambridge (2003)
5. Wendland, H.: Scattered data approximation, *Cambridge Monographs on Applied and Computational Mathematics*, vol. 17. Cambridge University Press, Cambridge (2005)
6. Hu, X.G., Ho, T.S., Rabitz, H.: Rational approximation with multidimensional scattered data. Phys. Rev. E **65**(3), 035,701 (2002). DOI 10.1103/PhysRevE.65.035701

# Exploiting Model Hierarchy in Semiconductor Design Using Manifold Mapping

D.J.P. Lahaye and C.R. Drago

**Abstract** In this paper we solve an optimal doping profile control problem for semiconductors using the manifold mapping technique. As coarse and fine approximation we employ the drift diffusion and energy transport model, respectively. In this work the manifold mapping technique is applied for the first time to a problem in which the number of design variables varies with the finite element mesh points employed. The advantage of our approach is that it allows to optimize the energy transport model without having to implement an adjoint code while at the same preserving computational efficiency. Numerical results giving evidence of this claim for different values of the applied voltage will be shown.

## 1 Introduction

The interest in optimal control for semiconductor design has attracted considerable recent attention in both the engineering and applied mathematics community. A major objective in the optimal design is to improve the current flow over some contacts, for fixed applied voltages, by a slight change of the device doping profile. In most applications the design problem is addressed empirically and based on the knowledge and experience of electrical engineer. Although this problem can be clearly tackled by an optimization approach, only recently efforts have been made to solve the design problem via optimization techniques.

At first standard black box optimization methods requiring many solves of the forward model resulting in a high computational cost were applied [9]. Later the

D.J.P. Lahaye
Delft Institute of Applied Mathematics (DIAM), Technical University of Delft, Mekelweg 4, Delft, The Netherlands, e-mail: d.j.p.lahaye@tudelft.nl

C.R. Drago
Dipartimento di Matematica e Informatica, Università di Catania, Viale Andrea Doria 6, 95125 Catania, Italy, e-mail: drago@dmi.unict.it

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_55,

adjoint variable method was shown to drastically reduce the computational effort in the optimal control of the so-called drift diffusion model [8]. This approach was extended to the so-called energy transport model [5]. Comparisons between the drift-diffusion and the energy transport optimal designs were presented in [6].

Moreover in [6] the idea to exploit this classical model hierarchy to speed up the convergence of the optimization algorithms using an input space mapping algorithm was proposed [1, 2]. Herein the drift diffusion and energy transport models are the coarse and fine model, respectively. The advantage of this approach is that it allows to efficiently optimize the energy transport model without having to implementing its adjoint. The drawback is that the input space mapping solution does not necessarily coincide with the fine model optimum. In the manifold-mapping technique [4], the surrogate model is constructed in such a way that the solutions of the surrogate and fine model optimization problem do coincide.

In this work we capitalize in the above achievements by solving the design for the energy transport model using manifold-mapping while exploiting the drift-diffusion model as auxiliary model. In this work the manifold mapping technique is applied for the first time to a problem in which the number of design variables varies with the finite element mesh points employed. Another innovative aspect of this work is the fact that we employed the Comsol Multiphysics finite element simulation environment [3] to implement the drift diffusion and energy transfer models, as well as the adjoint of the former. The flexibility that this framework provides allows to extend this work to bi-polar, two-dimensional or physically more complex models in a straightforward manner.

## 2 Semiconductor Models

The drift diffusion model is the simplest and most popular semiconductor model and is widely used in commercial simulation packages. It is based on the assumption of isothermal motion and allows for an efficient numerical study of charge transport in many case of practical relevance. In today’s semiconductor technology however, the miniaturization of devices is ever progressing. The simulation of semiconductor devices on sub-micron scale therefore requires advanced transport models. Because of the presence of very high and rapidly varying electric fields, phenomena occur which cannot be described by means of drift-diffusion model. The energy-transport model on the other hand takes thermal effects related to the electron flow through the semiconductor crystal into account. It therefore allows for a more accurate physical description.

## 2.1 Drift-Diffusion and the Energy Transport Model

The stationary drift diffusion (DD) model in the unipolar case consists of a continuity equation for electron density $n$ coupled with a Poisson equation for the electrostatic potential $V$ [10]. Denoting the electron current density $J_n$ as

$$J_n = -(\nabla n - n\nabla V) \tag{1}$$

the equations in dimensionless form on the interval $\Omega = [0, L]$ read

$$\mathrm{div} J_n = 0 \tag{2a}$$

$$\lambda^2 \triangle V = n - C \tag{2b}$$

where $C$ and $\lambda^2 = \frac{\varepsilon_s U_T}{q C_m L^2}$ are the doping profile and the Debye length, respectively. Numerical values of the latter are given in Table 1.

Stratton's energy transport (ET) model consists of continuity equation for electron density $n$ and the temperature $T$ coupled to the same Poisson's equation for $V$ as used before [10]. Denoting the electron and the energy flux density as $J_n$ and $J_E$, one has that

$$J_n = -\left(\nabla n - \frac{n}{T}\nabla V\right) \tag{3a}$$

$$J_E = -\frac{3}{2}\left(\nabla(nT) - n\nabla V\right) \tag{3b}$$

and the equations in dimensionless form read

$$\mathrm{div} J_n = 0 \tag{4a}$$

$$\mathrm{div} J_E = J_n \cdot \nabla V + W(n,T) \tag{4b}$$

$$\lambda^2 \Delta V = n - C \tag{4c}$$

where $W(n,T) = -\frac{3}{2}\frac{n(T-1)}{\tau_w}$ is the energy production term and $\tau_w = \tau_0 \mu_0 U_T / L^2$ the scaled energy relaxation time (cf. Table 1).

Systems (2) and (4) have to be supplied with appropriate boundary conditions. We assume that on both endpoints of the one-dimensional domain $\Omega$ the following Dirichlet boundary conditions are imposed

$$n = n_D, \; T = T_D, \; V = V_D \qquad \text{for } x = 0 \text{ and } x = 1\,, \tag{5}$$

where the difference in $V_D$ between the left and rightmost endpoint is the applied voltage $\Delta V$. Having solved either the DD or ET model, the current can be obtained by integrating $J_n$ at the contact $x = 1$

$$I = J_n(x = 1)\,. \tag{6}$$

**Table 1:** Physical and dimensionless parameters

| Parameter | Physical meaning | Numerical value |
|---|---|---|
| q | Elementary charge | $1.6 \cdot 10^{-19}$As |
| $\varepsilon_s$ | Permittivity constant | $10^{-12}$AsV$^{-1}$cm$^{-1}$ |
| $\mu_0$ | (Low field) mobility constant | $1.4 \cdot 10^3$cm$^2$V$^{-1}$s$^{-1}$ |
| $U_T$ | Thermal voltage at $T_0 = 300$K | 0.026V |
| $\tau_0$ | Energy relaxation time | $0.4 \cdot 10^{-12}$s |
| L | Length of the device | $0.6\mu$m |
| $\lambda^2$ | Debye length | $9.0278 \times 10^{-5}$ |
| $\tau_w$ | Scaled energy relaxation time | $4 \times 10^{-3}$ |

## 3 Design Problem

The goal of optimal control problem we intend to solve is to increase the current computed at the contact at a particular voltage level by limited changes in the doping profile. More precisely, we assume a reference doping profile $\overline{C}$ for which a particular applied voltage $\Delta V$ gives rise to a reference current $\overline{I}$. We aim at changing $\overline{C}$ in such a way to increase the reference current $\overline{I}$ to obtain the target current $I^*$. In other words, we intend to minimize the cost functional [5, 6, 8]

$$F(y,C) = \frac{1}{2}(I - I^*)^2 + \frac{\gamma}{2}\int_\Omega |\nabla(C - \bar{C})|^2 \mathrm{d}x, \tag{7}$$

where $\gamma > 0$ allows to balance the relative weight of both terms in the right-hand side. The function $C$ enters as a source term in the DD and ET models and plays the role of design variable. The DD and ET models are interpreted as a constraint on the minimization problem that allow to determine the current density $J_n$ through the state variables $y = (n,V)$ and $y = (n,T,V)$, respectively.

## 4 Manifold-Mapping Technique

Let us consider an optimization problem with design variables $\mathbf{x}$ in the design space $\mathbf{x} \in X \subset \mathbb{R}^n$ and specifications $\mathbf{y} \in \mathbb{R}^m$. The accurate behavior of electromagnetic devices is often studied using models that have large computational costs, e.g., finite element models. In space-mapping terminology these models are called *fine* models. The fine model response is denoted by $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$. The problem we set out to solve in this work can be stated as

$$\text{find } \mathbf{x}_f^* \in X \text{ such that } \mathbf{x}_f^* = \underset{\mathbf{z} \in X}{\operatorname{argmin}} \, \|\mathbf{f}(\mathbf{z}) - \mathbf{y}\|, \tag{8}$$

where argmin denotes the argument of the minimum. Space-mapping needs a second, possibly less accurate but computationally cheaper model, called *coarse* model.

The coarse models are assumed to be defined over the same design space $X$. Their response is denoted by $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^m$. The auxiliary optimization problem can be formulated as

$$\text{find } \mathbf{x}_c^* \in X \text{ such that } \mathbf{x}_c^* = \underset{\mathbf{z}\in X}{\operatorname{argmin}} \, \|\mathbf{c}(\mathbf{z}) - \mathbf{y}\| \,. \tag{9}$$

The manifold-mapping technique [4] exploits coarse model information and defines a surrogate optimization problem whose solution does coincide with $x_f^*$. The key ingredient is the manifold-mapping function between the coarse and fine model image spaces $\mathbf{c}(X) \subset \mathbb{R}^m$ and $\mathbf{f}(X) \subset \mathbb{R}^m$. This function $\mathbf{S} : \mathbf{c}(X) \mapsto \mathbf{f}(X)$ maps the point $\mathbf{c}(\mathbf{x}_f^*)$ to $\mathbf{f}(\mathbf{x}_f^*)$ and the tangent space of $\mathbf{c}(X)$ at $\mathbf{x}_f^*$ to the tangent space of $\mathbf{f}(X)$ at $\mathbf{x}_f^*$. It allows the surrogate model $\mathbf{S}(\mathbf{c}(\mathbf{x}))$ and the manifold-mapping solution to be defined as follows

$$\text{find } \mathbf{x}_{mm}^* \in X \text{ such that } \mathbf{x}_{mm}^* = \underset{\mathbf{z}\in X}{\operatorname{argmin}} \, \|\mathbf{S}(\mathbf{c}(\mathbf{z})) - \mathbf{y}\| \,. \tag{10}$$

The manifold-mapping function $\mathbf{S}(\mathbf{x})$ is approximated by a sequence $\{\mathbf{S}_k(\mathbf{x})\}_{k\geq 1}$ yielding a sequence of iterands $\{\mathbf{x}_{k,mm}\}_{k\geq 1}$ converging to $\mathbf{x}_{mm}^*$. The individual iterands are defined by coarse model optimization

$$\text{find } \mathbf{x}_{k,mm}^* \in X \text{ such that } \mathbf{x}_{k,mm} = \underset{\mathbf{z}\in X}{\operatorname{argmin}} \, \|\mathbf{S}_k(\mathbf{c}(\mathbf{z})) - \mathbf{y}\| \,. \tag{11}$$

At each iteration $k$, the construction of $\mathbf{S}_k$ requires the singular value decomposition of the matrices $\triangle C_k$ and $\triangle F_k$ of size $m \times \min(k,n)$ whose columns span the coarse and fine model tangent space in the current iterand, respectively. Denoting these singular value decompositions by

$$\triangle C_k = U_{k,c} \, \Sigma_{k,c} \, V_{k,c}^T \text{ and } \triangle F_k = U_{k,f} \, \Sigma_{k,f} \, V_{k,f}^T \,, \tag{12}$$

we introduce the updated objective $\mathbf{y}_k$ as

$$\mathbf{y}_k = \mathbf{c}(\mathbf{x}_k) - \left[\triangle C_k \, \triangle F_k^\dagger + (I - U_{k,c} U_{k,c}^T)\right] (\mathbf{f}(\mathbf{x}_k) - \mathbf{y}) \,, \tag{13}$$

where superscript $^\dagger$ denotes the pseudo-inverse. With this notation, the problem (11) can shown be to be asymptotically equivalent to

$$\text{find } \mathbf{x}_{k,mm}^* \in X \text{ such that } \mathbf{x}_{k,mm} = \underset{\mathbf{z}\in X}{\operatorname{argmin}} \, \|\mathbf{c}(\mathbf{z}) - \mathbf{y}_k\| \tag{14}$$

Details on the construction of the matrices $\triangle C_k$ and $\triangle F_k$, on properties of the mapping function $\mathbf{S}(\mathbf{x})$, as well as on the conditions under which the iteration (11) does converge, can be found in [7]. The computation of $\triangle F_k$ does not require the fine model sensitivity. By construction $\mathbf{x}_{mm}^* = x_f^*$ holds. The innovative aspect of this work concerns the number of design variables. In previous publications the MM technique was applied to sizing optimization problems with a limited number of design variables. In this work in contrast, the design variable $\mathbf{x}$ is a grid function whose dimensions varies with the number of finite element mesh points employed.

## 5 Numerical Results

In this section we test the performance of the manifold mapping optimization for a one-dimensional $n^+ - n - n^+$ ballistic diode, which is a simple model for the channel of a MOS transistor. The semiconductor domain is given by the interval $\Omega = [0,L]$. In the $n^+$-regions a maximal doping concentration of $C_m = 5 \cdot 10^{17}$ cm$^{-3}$ is prescribed. In the $n$–channel the minimal doping density is $2 \cdot 10^{15}$ cm$^{-3}$. The length of the $n^+$-regions and of the channel is $0.1\mu$m and $0.4\mu$m, respectively.
For the coarse model optimization we have used the same gradient algorithm as in [6], with constant step size ($\alpha_{0.26V} = 10^{-3}; \alpha_{0.52V} = 10^{-3}; \alpha_{1V} = 10^{-4}; \alpha_{1.5V} = 10^{-4}$), where the gradient is obtained by means of the adjoint equations. The DD and ET models as well as the adjoint of the former were implemented in the Comsol Multiphysics finite element simulation environment [3]. We used a mesh consisting of 256 elements.

We aimed at achieving an amplification of the current of 50% for different value of the applied voltage of $0.26, 0.52, 1, 1.5V$. The convergence history of the manifold mapping algorithm for the four test cases can be found in Fig. 2. The overall performance of the algorithm is very promising because as little as 6 fine model evaluations are sufficient to reach the optimum. The optimal doping profiles are presented in Fig. 1 and the agreement between the current targets and the optimal ones is shown in Fig. 3. Note that for increasing biasing voltages the drift diffusion and energy transport model yield quite different responses.

The convergence statistics can be found in Table 2. Compared with the space mapping optimization approach presented in [6] we reduced furthermore the number of evaluations of the coarse drift diffusion model. In any case we emphasize that the computation of the coarse model gradient is much cheaper than evaluating the fine model gradient.

**Table 2:** Gradient steps in the coarse model optimization

| Volt | Step 1 | Step 2 | Step 3 | Step 4 | Step 5 | Step 6 |
|---|---|---|---|---|---|---|
| 0.26 | 1 | 4 | 3 | 1 | 1 | 1 |
| 0.52 | 3 | 3 | 4 | 1 | 1 | 1 |
| 1 | 10 | 9 | 8 | 1 | 1 | 1 |
| 1.5 | 5 | 5 | 4 | 1 | 1 | 1 |

## 6 Conclusions and Outlook

In this work we have demonstrates that exploiting model hierarchy in semiconductor design using manifold mapping results in an efficient computational procedure. It allows to solve the optimal control problem for the energy transport model using as

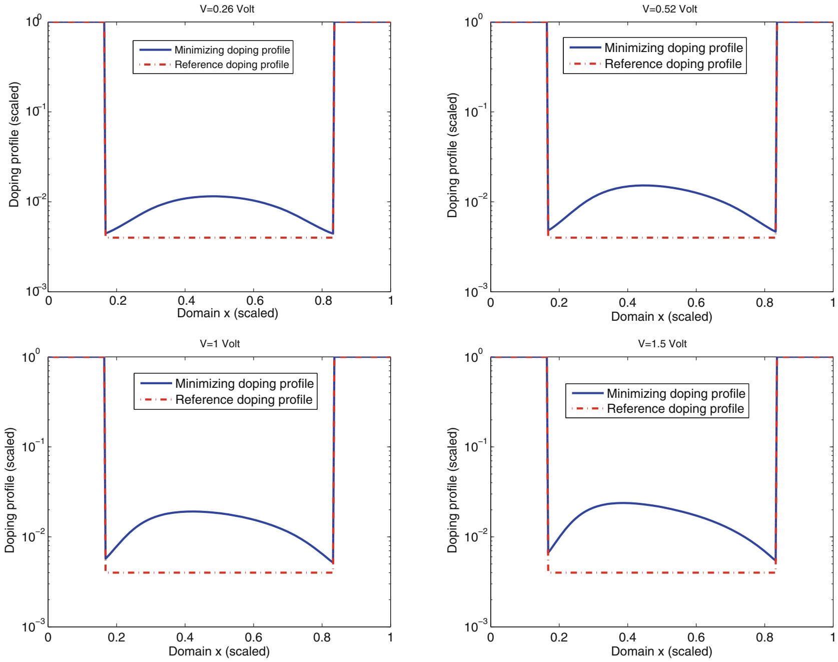


**Fig. 1:** Optimal doping profile for biasing voltages of 0.26, 0.52, 1, 1.5 V

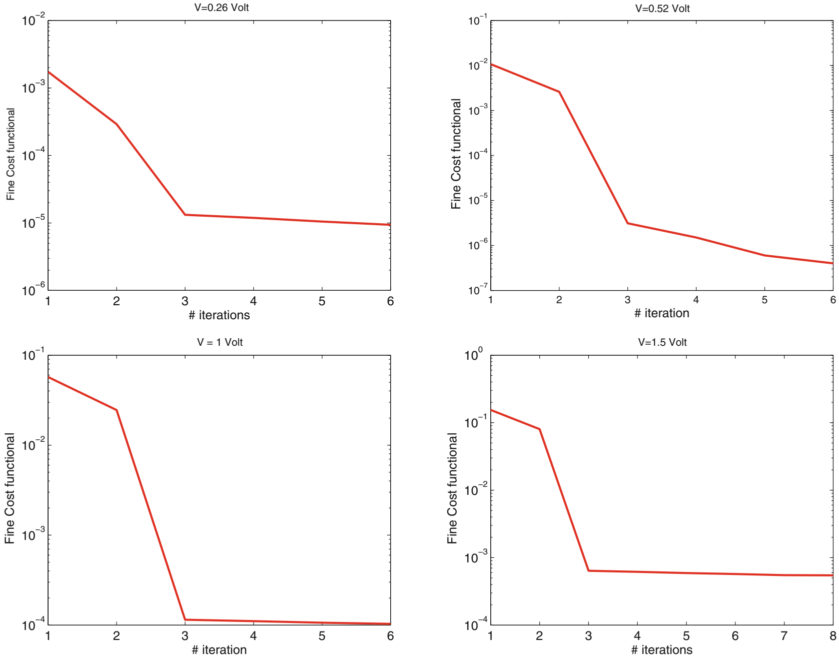


**Fig. 2:** Convergence history of the MM algorithm for biasing voltages of 0.26, 0.52, 1, 1.5 V

little as 6 finite element simulation. In doing so, we applied the manifold mapping technique to an optimization problem with many design variables for the first time. Future work will focus on extension to two-dimensional and bi-polar devices.

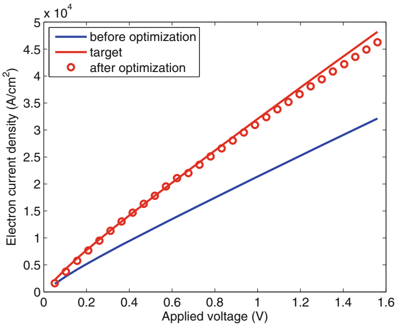


**Fig. 3:** Current-voltage curve

## Acknowledgement

This work is dedicated to Prof. A. M. Anile who made this collaboration possible.

## References

1. Bandler, J.W., Cheng, Q.S., Dakroury, S.A., Mohamed, A.S., Bakr, M.H., Madsen, K., Søndergaard, J.: Space Mapping: The State of the Art. IEEE Trans. on Microwave Theory and Techniques **52**(1), 337–361 (2004)
2. Bandler, J.W., Biernacki, R.W., Chen, S.H., Grobelny, P.A., Hemmer, R.H.: Space Mapping Technique for Electromagnetic Optimization. IEEE Trans. on Microwave Theory and Techniques **42**(12), 2536–2544 (1994)
3. www.comsol.com
4. Echeverría, D., Hemker, P. W.: Space Mapping and Defect Correction. Comp. Methods in Appl. Math. **5**(2), 107–136 (2005)
5. Drago, C.R., Anile, A.M.: An Optimal Control approach for an Energy Transport model in Semiconductor Design. In: A.M. Anile (ed.) Scientific Computing in Electrical Engineering SCEE 2004, *Mathematics in Industry*, vol. 9, pp. 323–331. Springer, Berlin Heidelberg New York (2006)
6. Drago, C.R., Pinnau, R.: Optimal Dopant Profiling based on Energy Transport Semiconductor Models. Math. Mod. Meth. Appl. Sc. **18**(2), 195–214 (2008)
7. Echevverría, D.: Multilevel Optimization: Space Mapping and Manifold Mapping. Universiteit van Amsterdam (2007)
8. Hinze, M., Pinnau, R.: An Optimal Control approach to Semiconductor Design. Math. Mod. Meth. App. Sc. **12**(1), 89–107 (2002)
9. Lee, W.R., Wang, S., Teo, K.L.: An optimization approach to a finite dimensional parameter estimation problem in semiconductor device design. Journal of Computational Physics **156**, 241–256 (1999)
10. Jüngel, A.: Quasi-hydrodynamic Semiconductor Equasions PNDEA, Birkhäuser (2001).

# Solving Inverse Problems by Space Mapping with Inverse Difference Method

Murat Şimsek and N. Serap Şengör

**Abstract** The surrogate methods have been used to ease the computational burden in various disciplines. In this work, a surrogate method based on space mapping is proposed to solve inverse problems. Even though the efficiency of space mapping and its variants has been demonstrated in numerous work, using it for inverse problems is addressed for the first time in this work. The efficiency of the proposed method is demonstrated solving the shape reconstruction of a conducting cylinder.

## 1 Introduction

The distinctive feature of surrogate methods are their capability of combining the computational efficiency of a coarse model with the accuracy of the fine model and in Space Mapping (SM) technique this is provided through a mapping from the fine model input space to the coarse model input space [1]. In Space Mapping with Difference (SM-D) method this mapping has been adjusted by enlarging the dimension of the domain of the mapping with the coarse model output. With this adjustment, the need to evaluate the fine model was reduced and the simulation results obtained for different applications revealed that the extrapolation capability of the models obtained with SM-D were improved [2,3]. In this work, the approach used in SM-D method is considered for inverse problems and a new method named Space Mapping with Inverse Difference (SM-ID) is proposed. As the proposed method deals with the problems that are instinctively inverse, the method has! ! two important features different than other SM techniques. Even though the mapping between coarse model and fine model parameter spaces are constructed similar to Linear Inverse Mapping (LISM) algorithm given in [4], the parameter extraction (PE) step needed in LISM and other SM techniques is no longer necessary in SM-ID to build an appropriate space mapping function $P(.)$. The other difference is using inverse coarse

Murat Şimsek, N. Serap Şengör
Electrical and Electronics Engineering Faculty, Istanbul Technical University, Maslak, Istanbul, Turkey, e-mail: simsekmu@itu.edu.tr, sengorn@itu.edu.tr

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_56,

model instead of coarse model. The inverse coarse model is generated as a multilayer perceptron in this work.

In the next section, SM-D method will be reviewed and in the third section the proposed method will be introduced. The simulation results obtained in solving the shape reconstruction problem of a conducting cylinder [5] by SM-ID will be given in the fourth section.

## 2 Space Mapping with Difference

In most of SM techniques as Aggressive Space Mapping (ASM) and SM a mapping, $P(.)$, from the fine model input space to the coarse model input space is constructed as following:

$$x_c = P(x_f) \tag{1}$$

such that

$$R_c(P(x_f)) \approx R_f(x_f) \tag{2}$$

where, $x_f$, $x_c$, $R_f(.)$ and $R_c(.)$ are fine and coarse model design parameters and fine and coarse model responses, respectively [1]. As it can be followed from the block diagram given in Fig. 1, in SM-D method [2] a mapping from $x_f$ to $x_c$ is formed as following:

$$x_c = P_d(Y_f, x_f) + x_f \tag{3}$$

Here, $P_d(.,.)$ maps the fine model response $Y_f = R_f(x_f)$ and the fine model design parameter $x_f$ to the difference between fine and coarse model design parameters $x_d$. Since the fine model response $Y_f = R_f(x_f)$ is already obtained, using it does not give rise to an extra computational burden. The steps of SM-D method to find

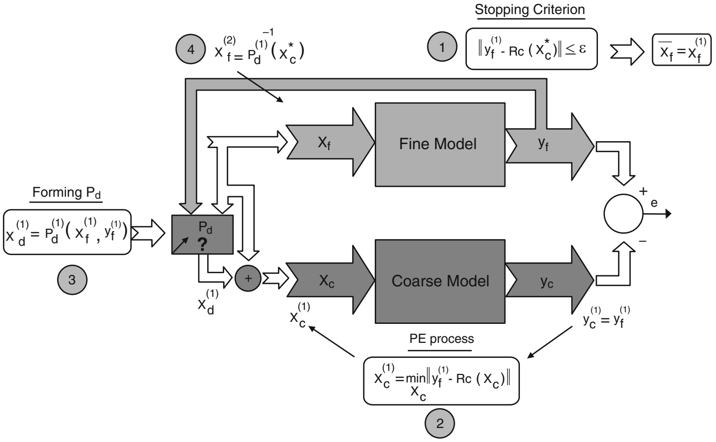

Stopping Criterion
1
$\|y_f^{(1)} - Rc\,(x_c^*)\| \leq \varepsilon$
$\overline{x_f} = x_f^{(1)}$
4
$x_f^{(2)} = P_d^{(1)^{-1}}(x_c^*)$
$x_f$
Fine Model
$y_f$
+
e
−
Forming $P_d$
$x_d^{(1)} = P_d^{(1)}(x_f^{(1)}, y_f^{(1)})$
3
$P_d$
?
+
$x_d^{(1)}$
$x_c$
Coarse Model
$y_c$
$x_c^{(1)}$
$y_c^{(1)} = y_f^{(1)}$
PE process
$x_c^{(1)} = \min_{x_c} \|y_f^{(1)} - Rc\,(x_c)\|$
2

**Fig. 1:** Block diagram of the SM-D method

the optimum design parameters of fine model $\overline{x_f}$ giving rise to optimum fine model response $\overline{Y_f}$ using coarse model responses $R_c(.)$ are the following:

- pre-step 1: choose $Y_c^* = \overline{Y_f}$
- pre-step 2: find $x_c^*$ from $x_c^* = \min_{x_c} \|\overline{Y_f} - R_c(x_c)\|$
- pre-step 3: set $x_f^{(1)} = x_c^*$
- pre-step 4: find $Y_f^{(1)} = R(x_f^{(1)})$ set $i = 1$
- step 1: if $\|Y_f^{(i)} - Y_c^*\| \leq \varepsilon$ then $\overline{x_f} = x_f^{(i)}$ else go to step 2
- step 2 (Parameter Extraction): find $x_c^{(i)}$ using $x_c^{(i)} = \min_{x_c} \|Y_f^{(i)} - R_c(x_c)\|$
- step 3: form $P_d^{(i)} = QD^\dagger$ where $Q \doteq [x_c - x_f]$, $D \doteq [1\ x_f\ Y_f]^T$ and set $i = i+1$
- step 4: set $x_f^{(i+1)} = {P_d^{(i)}}^\dagger (x_c^*)$ and go to step 1

The first four pre-steps summarizes initialization of $Y_f$ and the last four steps gives the algorithm of SM-D method.The $Q$ and $D$ matrices used here are given below for $i = m$ and the †'s denote the pseudo inverse:

$$Q = \begin{bmatrix} x_{c1}^{(1)} - x_{f1}^{(1)} & x_{c1}^{(2)} - x_{f1}^{(2)} & \dots & x_{c1}^{(m)} - x_{f1}^{(m)} \\ x_{c2}^{(1)} - x_{f2}^{(1)} & x_{c2}^{(2)} - x_{f2}^{(2)} & \dots & x_{c2}^{(m)} - x_{f2}^{(m)} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{cn}^{(1)} - x_{fn}^{(1)} & x_{cn}^{(2)} - x_{fn}^{(2)} & \dots & x_{cn}^{(m)} - x_{fn}^{(m)} \end{bmatrix}_{nXm} \qquad D = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{f1}^{(1)} & x_{f1}^{(2)} & \dots & x_{f1}^{(m)} \\ x_{f2}^{(1)} & x_{f2}^{(2)} & \dots & x_{f2}^{(m)} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{fn}^{(1)} & x_{fn}^{(2)} & \dots & x_{fn}^{(m)} \\ Y_f^{(1)} & Y_f^{(2)} & \dots & Y_f^{(m)} \end{bmatrix}_{(n+2)Xm} \tag{4}$$

These matrices expand till satisfactory result is obtained. Following the statement in step 3, given above, $P_d(.,.)$ mapping can be obtained as following where $P_d = QD^\dagger$ for each iteration:

$$P_d(Y_f, x_f) \doteq P_d \begin{bmatrix} 1 \\ x_f \\ Y_f \end{bmatrix} \tag{5}$$

## 3 Space Mapping with Inverse Difference

In design problems, the main concern is to determine the design parameters $x_{design}$ which minimize an objective function defined over responses $R(x_{design})$ of design parameters $x_{design}$. In inverse problems, main concern is to determine the some parameters $x$ of the problem given responses $R(x)$. As both problems are synthesis problems the solution is not unique and furthermore for the inverse problems it can be ill-possed. When feedforward neural network structure is used the ill-possed nature of the problem is dealt with regularization method in Tikhonov sense [6]. In order to deal with inverse problems, a new method Space Mapping with Inverse Difference (SM-ID), based on the idea of extending the inputs of mapping $P(.)$ as in SM-D method is proposed and its block diagram is given in Fig. 2.

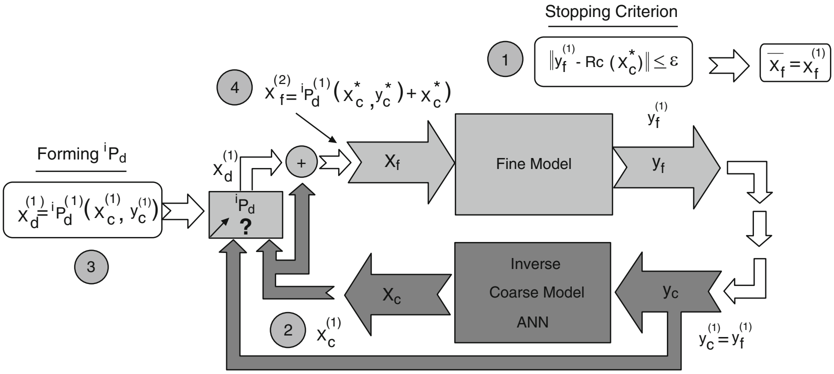

Stopping Criterion
1
$\|y_f^{(1)} - Rc\,(x_c^*)\| \leq \varepsilon$
$\overline{x_f} = x_f^{(1)}$
4
$x_f^{(2)} = {}^iP_d^{(1)}(x_c^*, y_c^*) + x_c^*)$
$y_f^{(1)}$
Forming ${}^iP_d$
$x_d^{(1)}$
+
$X_f$
Fine Model
$y_f$
$x_d^{(1)} = {}^iP_d^{(1)}(x_c^{(1)}, y_c^{(1)})$
${}^iP_d$
?
3
Inverse
Coarse Model
ANN
$X_c$
$y_c$
2
$x_c^{(1)}$
$y_c^{(1)} = y_f^{(1)}$

**Fig. 2:** Block diagram of the SM-ID method for $i = 1$

As it can be followed from Fig. 2, SM-D method is modified in two ways; first inverse coarse model is used instead of coarse model. In most applications, there will not be possibility of obtaining an inverse coarse model, in such cases using a well-known feedforward neural network structure as multilayer perceptron would be suitable. The second modification is in constructing the SM function $P(.)$. As the relation in SM-ID will be the inverse of the relation set up in SM-D, this function is denoted by ${}^iP_d(.)$ to imply that it builds a mapping between inverse coarse model output $Y_c$, inverse coarse model design parameters $x_c$ and fine model design parameters $x_f$. The SM function ${}^iP_d(.)$ maps $Y_c$ , $x_c$ to $x_d$ as following:

$$x_f =^i P_d(Y_c, x_c) + x_c \tag{6}$$

SM function ${}^iP_d(.)$ will resemble that of LISM algorithm [4] but since inverse coarse model is used there is no need for parameter extraction step [1,4]

The algorithm of SM-ID method is given in the following:

- step 1: if $\|Y_f^{(i)} - R_c(x_c^*)\| \leq \varepsilon$ then $\overline{x_f} = x_f^{(i)}$ else go to step 2
- step 2: set $Y_c = Y_f^{(i)}$ and find $x_c^{(i)}$ from inverse coarse model
- step 3: form ${}^iP_d^{(i)} = QD^\dagger$ where $Q \doteq [x_f - x_c]$, $D \doteq [1\ x_c\ Y_c]^T$ and set $i = i+1$
- step 4: set $x_f^{(i+1)} = {}^iP_d^{(i)}(x_c^*, Y_c^*) + x_c^*$

To obtain the new fine model parameter the following relation is used:

$$x_f^{(m+1)} = {}^iP_d \begin{bmatrix} 1 \\ x_c^* \\ Y_c^* \end{bmatrix} + x_c^* \tag{7}$$

In this work, the SM function ${}^{i}P_d(.,.)$ is a linear mapping as given in Equation 8 .

$$
{}^{i}P_d(.,.) = \begin{bmatrix} c_1 & b_{11} & \ldots & b_{1(n+1)} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ c_n & b_{n1} & \ldots & b_{n(n+1)} \end{bmatrix}_{nX(n+2)} \tag{8}
$$

The $Q$ and $D$ matrices used at each iteration to form ${}^{i}P_d(.,.)$ are given below for $i = m$.

$$
Q = \begin{bmatrix} x_{f1}^{(1)} - x_{c1}^{(1)} & x_{f1}^{(2)} - x_{c1}^{(2)} & \ldots & x_{f1}^{(m)} - x_{c1}^{(m)} \\ x_{f2}^{(1)} - x_{c2}^{(1)} & x_{f2}^{(2)} - x_{c2}^{(2)} & \ldots & x_{f2}^{(m)} - x_{c2}^{(m)} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ x_{fn}^{(1)} - x_{cn}^{(1)} & x_{fn}^{(2)} - x_{cn}^{(2)} & \ldots & x_{fn}^{(m)} - x_{cn}^{(m)} \end{bmatrix}_{nXm} \quad D = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_{c1}^{(1)} & x_{c1}^{(2)} & \ldots & x_{c1}^{(m)} \\ x_{c2}^{(1)} & x_{c2}^{(2)} & \ldots & x_{c2}^{(m)} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ x_{cn}^{(1)} & x_{cn}^{(2)} & \ldots & x_{cn}^{(m)} \\ Y_c^{(1)} & Y_c^{(2)} & \ldots & Y_c^{(m)} \end{bmatrix}_{(n+2)Xm} \tag{9}
$$

Since ${}^{i}P_d(.,.)$ is linear and there is no need for parameter extraction the computational burden is decreased compared to other SM methods.

# 4 Simulation Results for Reconstruction of a Conducting Cylinder

The inverse problem considered is the reconstruction of a conducting cylinder problem. To indicate the efficiency of SM-ID method the results obtained are compared first with the results obtained from conventional Artificial Neural Networks (ANN) structures. Also the results obtained from SM-ID are compared with other SM based methods as ASM and LISM.

The conventional ANN structure used is multilayer perceptron and it is trained with different number (50, 100, 200) of data. The ANN trained with 50 and 100 data is also used as inverse coarse model while implementing SM-ID method. In Figs. 3–8 the test set results are exposed. The ANN's trained have 20 inputs, nine outputs, where the inputs are the real amd imaginary components of scattered electric field obtained at 10 different positions and the outputs are the Fourier series coefficients of the geometrical shape of the conducting cylinder. It can be followed from the Table 1 and Figs. 3–8 that SM-ID results outperforms the ANN results. Also,as the number of data used increases, the iterations needed to construct ${}^{i}P_d$ decreases.

The results obtained for the reconstruction of conducting cylinder using SM-ID method are further compared with ASM and LISM methods and these are given in

**Table 1:** Comparison of SM-ID results with ANN

| | ANN-50 | SM-ID-50<br>iteration:5 | ANN-100 | SM-ID-100<br>iteration:4 | ANN-200 |
|---|---|---|---|---|---|
| Max error | 0.22067 | 0.03762 | 0.11889 | 0.00389 | 0.10386 |
| Mean error | 0.06715 | 0.00868 | 0.05266 | 0.00081 | 0.04285 |

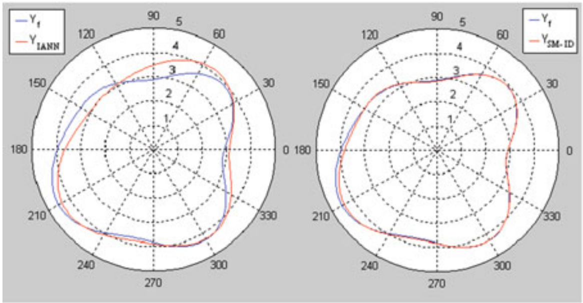


**Fig. 3:** On the *left* ANN result and on the *right* SM-ID result is given for 50 data

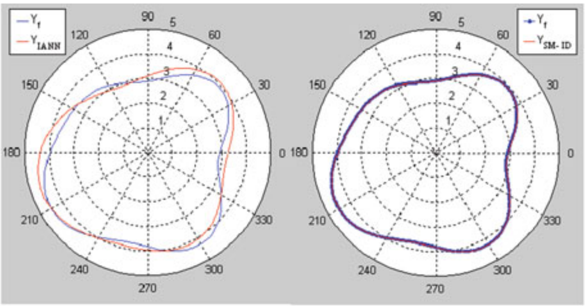


**Fig. 4:** On the *left* ANN result and on the *right* SM-ID result is given for 100 data. Since the *lines on the right* overlaps, the difference can be followed from Table 1

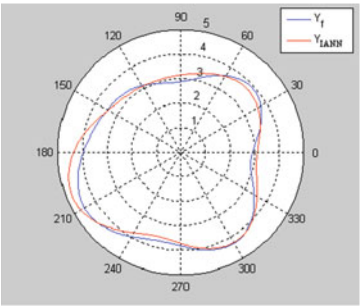


**Fig. 5:** Only the ANN result is given for 200 data as SM-ID fits almost perfectly for 100 data

Table 2. In ASM and LISM, ANN trained with 100 data for inverse problem is used as the coarse model thus no need for parameter extraction arisen. It can be followed from Fig. 6 that the LISM diverged after two iterations so only the result obtained with two iterations are given in Table 2.

**Table 2:** Comparison of SM-ID results with LISM, ASM and ANN

| | LISM<br>iteration:2 | ASM<br>iteration:13 | SM-ID-100<br>iteration:7 | ANN-100 |
|---|---|---|---|---|
| Max error | 0.11085 | 0.04053 | 0.00883 | 0.11194 |
| Mean error | 0.03459 | 0.01337 | 0.00237 | 0.03489 |

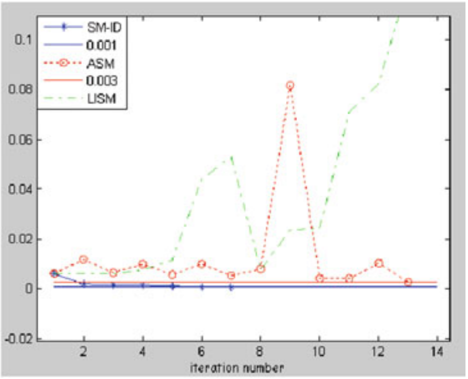


**Fig. 6:** Comparison of convergences of different methods. Two *straight lines* are shown to ease the comparison

## 5 Conclusion

In this work a novel SM based method is proposed to solve inverse problems and already existing methods based on SM technique as LISM and ASM are used for inverse problems. Due to the nature of the inverse problems in all three SM based methods instead of parameter extraction step inverse coarse model which is implemented by multilayer perceptron is used. It is shown that the proposed method gives better or results for reconstruction of conducting cylinder.

**Acknowledgements** The authors thank Necmi Serkan Tezel for introducing the inverse scattering problem.

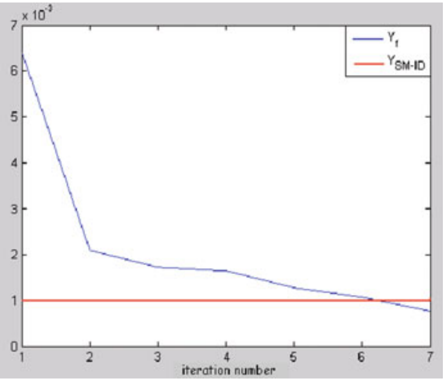


**Fig. 7:** Convergence of SM-ID method. The *straight line* denotes the stopping criterion

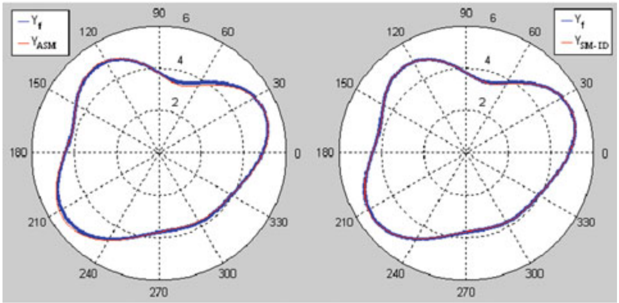


**Fig. 8:** Comparison of ASM with SM-ID. As the results for both methods fits the actual data, the difference between methods can be followed from Table 2

## References

1. Bandler, J.W., Cheng, Q.S., Dakroury, S.A., Mohamed, S.A., Bakir, M.H., Madsen K., Sondergaard, J.: Space Mapping : The State of the Art. IEEE Trans. on Microwave Theory and Tech. **52**, 337–361 (2004)
2. Simsek, M., Sengor, N.S.: A New Modelling Method Based on Difference between Fine and Coarse Models Using Space Mapping. Abstract Book of 2nd Int. Workshop SMSMEO-06, 64–65 (2006)
3. Simsek, M.: Novel Methods for Surrogate Optimization and Modeling Using Space Mapping Techniques. Ph.D. Thesis Report 2, Istanbul Technical University (2007)
4. Rayas-Sánchez, J.E., Lara-Rojo, F., Martinez-Guerrero, E.: A Linear Inverse Space-Mapping (LSIM) Algorithm to Design Linear and Nonlinear RF and Microwave Circuits. IEEE Trans. on Microwave Theory and Tech. **53**, 960–368 (2005)
5. Tezel, N.S., Şimşek, C.: Neural Network Approach to Shape Reconstruction of a Conducting Cylinder. Istanbul University, Journal of Electrical and Electronics Eng. **7**, 299–304 (2004)
6. Haykin, S.: Neural Networks-A Comprehensive Foundation. Prentice Hall (1999)

# Part V
# Model-Order Reduction

# Introduction to Part V

E. Jan W. ter Maten

Over the years, *model-order reduction* (MOR) always was greatly inspired by problems from the electronics industry. Especially problems from interconnect and from parasitics extraction offered a nice class of large, linear problems. MOR aims to compress large systems but requires that its input-output behavior is preserved (within tolerances). There are several techniques available. For good general references on MOR the reader is referred to [1–3].

As a result of several international and national co-operative research projects[1] in which MOR is a dedicated topic, MOR received more pronounced attention also at the SCEE conferences: SCEE 2006 had five presentations on MOR, including the invited talks by A.C. Antoulas [4] and L.M. Silveira [5]. The opening invited talk by P. Benner at SCEE 2008 was followed by as many as 13 presentations on MOR. This rising trend is also reflected in the programs of several other conferences like SIAM-CSE 2007, ICIAM 2007, and ECMI 2008, as well as at workshops at TU Eindhoven (2006, 2007, [6, 7]) and the University of Hamburg (2008, [8]), and at the CoMSON Autumn School on Future Developments in Model Order Reduction, in Terschelling, the Netherlands (2009, [9]).

MOR already had great potential to generate small, efficient models that approximate output results well while preserving several important properties like passivity, stability, and reciprocity. All techniques ranging from balanced truncation (BT) to Krylov-subspace methods, to methods based on singular-value decompositions

E. Jan W. ter Maten
NXP Semiconductors, Corp. I&T/DTF/A&M/Physical Design Methods, Mathematics, High Tech Campus 46, 5656 AE Eindhoven, The Netherlands, e-mail: jan.ter.maten@nxp.com

[1] Mentioned are:
- the EU-FP6-MCA-RTN project “Coupled Multiscale Simulation and Optimization in Nanoelectronics” (CoMSON), 2005–2009, http://www.comson.org/
- the EU-FP6-MCA-ToK project “Operational MOdel Order REduction for Nanoscale IC Electronics” (O-MOORE-NICE!), 2007–2010, http://www.tu-chemnitz.de/mathematik/industrie_technikprojekte/omoorenice/
- the German BMBF project “Systemreduktion für IC Design in der Nanoelektronik” (SyreNe), 2007–2010, http://www.syrene.org/

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_57,

(SVDs) or to modal expansions, work more or less satisfactorily for linear state-space methods in the single-input–single-output (SISO) case. The papers in this book address aspects of linear algebra to efficiently solve intermediate problems, but also steps for needed generalizations to make the methods of real interest for industrial problems: generalizations to systems of differential-algebraic equations (DAEs), treatment of the multiple-input–multiple-output (MIMO) case, inclusion of parameterizations, partitioning techniques with associated structure-preserving MOR methods, and, finally, treatment of nonlinearity.

The first 11 papers of Part V deal with MOR techniques for linear problems.

The known BT methods for linear state-space ordinary differential equations (ODEs) provide error bounds and guarantee stability. A special variant of BT called positive-real BT preserves passivity. Progress in linear algebra now allows the methods to deal with problem sizes of up to $10^6$. The invited paper by Benner extends the theory for BT, developed for linear ODEs, to linear descriptor systems (state-space formulations of DAEs). A block triangulation is obtained by the disk-function method followed by a block diagonalization by solving a generalized Sylvester equation. Clever linear algebra generates in a stable manner the necessary bases. This is one of Benner's focusing points: by concentrating on the generation of the necessary bases, the many operators required need not to be determined explicitly. Also, several properties are obtained in an implicit way. Attention is given to passivity preservation, sparsification, and synthesis of the reduced-order model.

Stykel and Reis consider a passivity-preserving MOR method for circuit equations after applying a Möbius transformation. They derive the so-called bounded real BT. It requires balancing two Gramians that satisfy the projected Lur'e equations. Under some assumptions such equations can be rewritten as the projected Riccati equations. This results in the passivity-preserving BT method for electrical circuits (PABTEC).

Ionutiu et al. guarantee passivity preservation by the spectral-zero method, introduced by A.C. Antoulas and D. Sorensen in 2005; this method needs a special eigenvalue algorithm to determine the most dominant spectral zeros as generalised eigenvalues of a special matrix pencil (a Hamiltonian eigenvalue problem). The eigenvalue method used in this paper is a generalisation of the dominant-pole algorithm, proposed by J. Rommes and N. Martins in 2006 for MOR using modal expansion. Synthesis of the reduced-order model is achieved.

The paper by Yetkin and Dag proposes MOR by approximating dominant poles in a modal expansion. The number of dominant eigenvalues of interest are determined by Gershgorin eigenvalue-inclusion methods. For the Gershgorin cluster closest to the imaginary axis, the eigentriples are computed by some eigenvalue-deflation algorithm. For each eigentriple, the correction to the transfer function determines the error-control to stop.

Roos et al. present a global-approximation-based order reduction (GABOR) that preserves passivity and reciprocity for RLC circuits. A two-sided moment matching using Laurent expansions around zero and at infinity is exploited. The specific algorithm may be improved further, if a more stable implementation, based on fully implicit moment matching, is found in the future.

The contribution by Feng and Benner exploits recent developments in parameterized MOR dealing with non-rational occurrences of frequency-dependent terms. The first author developed stable methods to deal with MOR of state-space systems where the matrices are expressed in series with respect to the parameters. In principle, frequency can be treated as a parameter as well. In the paper, two appropriate parameters that both depend on frequency, and hence are correlated, are identified. A recursive procedure is proposed that uses orthogonalization at each iteration.

The paper by Benner and Schneider considers the MIMO, or multi-terminal, problem. They consider truncated SVD methods to make some existing methods like SVDMOR and extended SVDMOR (ESVDMOR), which are based on effectively reducing the number of input and/or output terminals, more efficient for large-scale problems.

Ugryumova and Schilders start from a full circuit model described by Kirchhoff's laws in the frequency domain involving all voltages and branch currents. The model describes the electromagnetic properties of an interconnect system. The terminal voltage-to-current transfer can be described by an admittance matrix $Y(s)$. The reduction in size of this matrix is based on selecting several nodes as "super nodes" and effectively treating them as terminals as well. This set includes the original ports. The selection corresponds to a block partitioning. Given this set, the effective admittance matrix is written as sum of RL and of C contributions. The first part is approximated by a dominant-pole expansion, while the second term is reduced by an eigenvalue decomposition and setting the non-positive eigenvalues to zero. This ensures a stable and passive approximation that also is realizable.

The paper by Honkala et al. presents a hierarchical MOR flow, where the linear parts of the (flat or hierarchically defined) circuit are divided into independently reducable subcircuits by using the hMETIS graph-partitioning algorithms. A suitable MOR method can be then applied to the different types of subcircuits. In the paper, the PRIMA and Liao–Dai methods are used. The latter only approximates two moments. Hence, here the blocks to be reduced should not be too large.

Miettinen et al. propose a passive, stable, netlist-in–netlist-out-type MOR method suitable for the reduction of very large RL-circuit blocks. The method relies on partitioning the circuit into subcircuits that can be efficiently approximated with low-order macromodels. Here, expansion in $1/s$ is considered. The efficiency of the method is demonstrated with several simulations and comparison to the PRIMA method. By using a partitioning to match a small section of the original circuit with a low-order approximation, they avoid the possible ill-conditioning issues related to direct high-order macromodel matching approaches. Also here the hMETIS algorithms are used.

The paper by Rommes et al. considers MOR of large resistive MIMO networks. Full reduction of all internal unknowns would generate a system with a full matrix in which every terminal is connected to every other terminal. This smaller system may require more time to solve than the original large but sparse system. Hence one has to guarantee some level of sparsity in the reduced-order model. This is done by marking some key internal unknowns as important and to raise them to the status of terminal. This corresponds to a block partitioning of the original system. Each block is reduced in an exact way using Schur complements to guarantee that the same path resistances between input and output terminals as in the original problem are found. For the partitioning, the authors use concepts from matrix-reordering algorithms. They first bring the matrix to a balanced bordered block-diagonal (BBBD) form and next apply an approximate minimum-degree (AMD) ordering on each block to minimize fill-in. The partitioning can be graphically displayed by special tools.

The last three papers of Part V deal with nonlinear MOR.

In the paper by Mohaghegh et al., the trajectory piecewise-linear (TPWL) approach is adopted, in which, at specific time points (linearization points) along a time trajectory, a reduced-order model is generated in the solution space around a local linearization in the solution space. From the projections associated with each model, a global projection is defined in a larger subspace that encompasses all individual subspaces of the reduced-order model. This paper studies different linear MOR approaches with respect to their performance when used as a kernel for TPWL. The studied MOR approaches are PRIMA, SPRIM, and poor man's truncated balanced realization (PMTBR). The first two rely on Krylov-subspace methods. The last one exploits the direct relation between the multipoint rational projection framework and the TBR. During the training to define the linearization time points, errors should be measured in the full space and not in the reduced space. Also, the weighting procedure between different models is important. Finally, sensitivity with respect to slight changes of the input signal are considered.

In the contribution by Verhoeven et al., the proper orthogonal decomposition (POD) approach is studied for reducing nonlinear IC models. POD results in much more accurate models than obtained with TPWL, but also is more costly: without additional adaptions, the reduced-order model may even require more CPU time than the original unreduced model. This is due to the fact that replacing a linear operation $Ax$ by a nonlinear evaluation $f(x)$ prevents evaluating the projected version $W^T f(Vz)$ efficiently (in the linear case one can multiply the matrices in advance) and also the Jacobian. Missing-point estimation is a technique to select the most dominant state variables. The adapted POD presented in this paper applies different selections for $V$ and $W$, resulting in a significantly improved POD method.

The last paper by Condon and Grahovski studies MOR of perturbed nonlinear neural networks with feedback. It focusses on empirical BT-based MOR that builds empirical gramians. The nonlinearity is assumed to be written as a linear combination of linear and nonlinear parts (Hopfield model). Nonlinear neural networks fit in this framework. The feedback can be a nonlinear function of the output. Nonlinear controllability and observability gramians can be defined using the empirical BT approach. For perturbed Hopfield models, integral estimates for the perturbations, that guarantee the hyperstability property after MOR, are derived. The paper studies the qualitative behaviour of the solutions of the reduced perturbed model. Special attention is paid to the Popov hyper-stability properties.

## References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM Publications, Philadelphia (2005)
2. Benner, P., Mehrmann, V., Sorensen, D. (eds.): Dimension Reduction of Large-Scale Systems. *Lecture Notes in Computational Science and Engineering*, vol. 45. Springer, Berlin/Heidelberg (2005)
3. Schilders, W.H.A., van der Vorst, H.A., Rommes, J. (eds.): Model Order Reduction: Theory, Research Aspects and Applications. *Mathematics in Industry*, vol. 13. Springer, Berlin/Heidelberg (2008)
4. Ionutiu, R., Lefteriu, S., Antoulas, A.C.: Comparison of model reduction methods with applications to circuit simulation. In: Ciuprina, C., Ioan, D. (eds): Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 3–24. Springer, Berlin/Heidelberg (2008)
5. Silva, J.M.S., Fernández Villena, J., Flores, P., Silveira, L.M.: Outstanding issues in model order reduction. In: Ciuprina, C., Ioan, D. (eds): Scientific Computing in Electrical Engineering SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 139–152. Springer, Berlin/Heidelberg (2008)
6. Cluster Symposium on Model Order Reduction, TU Eindhoven, Dec. 6 and 13, 2006. http://www.win.tue.nl/casa/meetings/special/cluster/
7. Symposium on Recent Advances in Model Order Reduction, TU Eindhoven, Nov. 23, 2007. http://www.win.tue.nl/casa/meetings/special/mor07/
8. Workshop Model Reduction for Circuit Simulation, University of Hamburg, Oct. 30–31, 2008. http://www.math.uni-hamburg.de/spag/zms/syrene/
9. COMSON Autumn School on Future Developments in Model Order Reduction, organized by TU Eindhoven and NXP Semiconductors, Terschelling, Sept. 21–25, 2009. http://www.win.tue.nl/casa/meetings/special/mor09/

# Advances in Balancing-Related Model Reduction for Circuit Simulation

Peter Benner*

*Invited speaker at the SCEE 2008 conference

**Abstract** We discuss algorithms for balanced truncation (BT) based model reduction of linear systems. BT is known to have good global approximation properties and to preserve important system properties. A computable error bound allows to choose the order of the reduced-order model adaptively. We will emphasize those aspects that makes the application of BT to models arising in circuit simulation a non-straightforward task. In recent years, these issues have been addressed by several authors. We will survey some of these developments and demonstrate that BT is now suitable for linear descriptor systems encountered in circuit simulation.

## 1 Introduction

Model order reduction (MOR) is an indispensable tool in the design and analysis of integrated circuits (ICs) and circuit simulation in general. This is due to the fact that on the one hand, almost all IC design relies heavily on simulation and on the other hand, the complexity of the mathematical models used to replicate the behavior of an actual electronic circuit is growing more rapidly than computing resources. This is caused by the increased packing density and multi-layer technology which nowadays requires the modeling of thermic and other parasitic effects caused by the interconnect. In many situations, only the use of MOR techniques allows the numerical simulation of the usually very large systems of ordinary differential and differential-algebraic equations used to describe (parts of) complex circuit layouts. MOR has been particularly successful in reducing the complexity of large linear subcircuits modeling parasitic effects of interconnect and in small signal analysis, and it is becoming an increasingly useful tool also in other areas of circuit design [1].

Peter Benner
Mathematik in Industrie und Technik, Fakultät für Mathematik, TU Chemnitz, 09107 Chemnitz, Germany, e-mail: benner@mathematik.tu-chemnitz.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_58,

Linear circuit models can be described by linear descriptor systems of the form

$$E\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t), \tag{1}$$

where $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, and $x(t) \in \mathbb{R}^n, y(t) \in \mathbb{R}^p, u(t) \in \mathbb{R}^m$ denote generalized states, outputs, inputs, respectively. The corresponding transfer function

$$G(s) = C(sE - A)^{-1}B + D \tag{2}$$

results from describing the input-to-output map $u \to y$ in frequency domain.[1] One difficulty for balancing-related model reduction methods arises from $E$ being singular as it is usually the case in circuit simulation. In this paper we will mainly focus on advances made for resolving this issue.

The model reduction problem now consists of finding a reduced-order system,

$$\hat{E}\dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \hat{B}u(t), \quad \hat{y}(t) = \hat{C}\hat{x}(t) + \hat{D}u(t), \tag{3}$$

of order $r$, $r \ll n$, with the same numbers of inputs ($m$) and outputs ($p$), i.e., $\hat{A}, \hat{E} \in \mathbb{R}^{r\times r}$, $\hat{B} \in \mathbb{R}^{r\times m}$, $\ldots$, and associated transfer function $\hat{G}(s) = \hat{C}(s\hat{E} - \hat{A})^{-1}\hat{B} + \hat{D}$, so that for the same input function $u \in L_2(0,\infty;\mathbb{R}^m)$, we have $y(t) \approx \hat{y}(t)$.

The most popular MOR methods in circuit simulation are Padé(-type) approximations, also known as moment-matching methods. The $r$th Padé approximant $\hat{G}$ of $G$ is defined by the property $G(s) = \hat{G}(s) + \mathscr{O}((s-s_0)^{2r})$, i.e., $M_j = \widehat{M}_j$ for $j = 0, \ldots, 2r-1$, where the *moments* $M_j, \widehat{M}_j$ are the coefficients in a power (Laurent) expansion of $G, \hat{G}$, respectively, about some expansion point $s_0 \notin \Lambda(A,E)$[2]. Moment-matching and Padé approximation properties are obtained for methods based on the unsymmetric Lanczos process, called the *(matrix) Padé-via-Lanczos ((M)PVL) method* [2–4]. *Padé-type methods* are also based on the moment matching property, but the approximations need not match the maximum possible number of moments. One such method is PRIMA [5] which employs the Arnoldi process to compute the reduced-order model. PRIMA is a success story in MOR for circuit simulation as besides having moment-matching properties, it preserves stability and passivity of RLC circuit models.

Despite the success with Padé(-type) approximation techniques based on the moment-matching properties of Krylov subspace methods, some major difficulties of this approach persist:

1. So far there exists in general no computable error estimate or bound for $\|y - \hat{y}\|$ in some appropriate norm.
2. The reduced-order model provides good approximation quality only locally.
3. The preservation of physical properties like stability or passivity can only be shown in very special cases; usually some post processing which (partially) destroys the moment matching properties, is required.

[1] Note that frequently in the area of circuit simulation, different notation is used: there $E, A$, and $C$ become $C, G$, and $L^T$, respectively. The notation used here is standard in systems theory.

[2] $\Lambda(A,E)$ denotes the set of generalized eigenvalues of the matrix pencil $A - \lambda E$.

There are many recent advances with respect to items 1.–3. discussed in the recent literature, see, e.g., [6, 7], but due to space limitations we can not discuss all these new developments here.

All the above problems of moment-matching methods are avoided when using balanced truncation (BT) or its relatives. Computable error bounds or estimates exist and come essentially for free as by-product of the computational procedures for obtaining the reduced-order model. The methods have good global approximation properties and thus, the reduced-order models can serve as surrogate for a large frequency range. Stability of the linear system is preserved for all variants of BT, other properties like passivity (which is important for passive devices) can be preserved by a variant of BT called *positive-real BT (PRBT)* (see, e.g., [8, 9] and references therein). Note that the error estimate for PRBT given, e.g., in [9], needs a good estimate of the $\mathscr{H}_\infty$-norm (defined below) of $G(s)+D^T$ and thus is not as cheap to evaluate as, e.g., the BT error bound (7) below. On the other hand, for any reasonable approximation this quantity can be replaced without significant loss of information by the $\mathscr{H}_\infty$-norm of $\hat{G}(s)+D^T$ which can be computed at moderate cost.

It has been common belief until recently that BT-related methods are not applicable in circuit simulation due to the $\mathscr{O}(n^3)$ complexity required by matrix equation solvers used to solve the underlying Lyapunov or algebraic Riccati equations. But advances in numerical linear algebra nowadays allow to compute solutions to those Lyapunov and Riccati equations arising in BT-related methods for linear systems at a computational cost that scales with the cost for solving linear systems of equations with coefficient matrix $A+s_0E$. Thus, these methods can now be applied to systems of order $\mathscr{O}(10^6)$. Moreover, most of the difficulties resulting from a singular $E$ matrix have now also been overcome. Many of these developments are discussed in [6, 10] and references therein.

In the main part of this paper (Section 2), we will focus on one possibility to extend BT to descriptor systems. A parallel implementation of an earlier version of this algorithm is already described in [11]. This method does not make use of possible sparsity of the system matrices and can thus be applied in order to reduce fairly small linear subcircuits with up to a few thousands elements. We will comment briefly on extensions to the case of large-scale, sparse matrices in Section 3. Some further issues like sparsification of the reduced-order model and passivity preservation using balancing-related methods will also be discussed in Section 3.

## 2 A Balanced Truncation Algorithm for Descriptor Systems

The method described in this section is based on two stages. In the first stage, we decompose the transfer function of the descriptor systems into a part corresponding to all finite poles and a polynomial part. Standard BT can then be applied to the first part while the transfer function of the polynomial part is preserved, but may be realized by a system of smaller order.

First, we briefly explain how we apply BT to the part corresponding to the finite poles, then we present a method to compute the required decomposition of the transfer function. In subsection 2.3 we combine these algorithms to a BT algorithm for descriptor systems and some numerical results are reported in subsection 2.4.

## *2.1 Balanced Truncation for Generalized State-Space Systems*

In this section, we briefly describe BT for systems of the form (1) when $E$ is nonsingular. Such systems will be called *generalized state-space (GSS) systems* in the following. For more thorough descriptions and in particular the mathematical background of the method in case $E = I_n$ see [8, 12, 13].

Throughout this and the following sections, we always assume $\lambda E - A$ to be stable, i.e., to have all its (finite) eigenvalues in the open left half of the complex plane. We call a GSS system, realized by $(A,B,C,D,E)$ as in (1) with $E$ nonsingular *balanced*, if the solutions $P,Q$ of the dual Lyapunov equations

$$APE^T + EPA^T + BB^T \;=\; 0, \quad A^T QE + E^T QA + C^T C \;=\; 0, \tag{4}$$

satisfy

$$P = E^T QE = \operatorname{diag}(\sigma_1, \dots, \sigma_n) \quad \text{with} \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n > 0. \tag{5}$$

The $\sigma_j$ are the Hankel singular values (HSVs) of the GSS system.

*Remark 1. $P, E^T QE$* are the *controllability* and *observability Gramians* of the linear time-invariant system $\dot{x}(t) = E^{-1}Ax(t) + E^{-1}Bu(t)$, $y(t) = Cx(t) + Du(t)$, which is equivalent to (1). As our method is equivalent to applying BT to this standard state-space system, this definition appears to be quite natural here. Our algorithm to solve (4) computes $E^T QE$ directly rather than $Q$ — this has a certain advantage over using $Q$ as observability Gramian as in [14, 15]. On the other hand, the definition used in [14, 15] yields Gramians directly for the descriptor system (1) and turns out to be the appropriate approach in this case. Note also that in case $E$ is singular, in contrast to common belief in many references in the literature, BT can not be directly based on (4) as the Lyapunov equations may or may not have solutions [14, 15]. The BT method for descriptor systems developed in [14, 15] therefore makes use of so-called *projected Lyapunov equations*. It turns out that Algorithm 4 below is mathematically equivalent to this approach, but solves the projected Lyapunov equations only implicitly.

A balanced realization of a minimal GSS system can be computed via a *system equivalence transformation*

$$\begin{aligned} \mathcal{T} : (A,B,C,D,E) \mapsto\; & (LAT, LB, CT, D, LET) \\ & = \left( \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix}, D, \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \right), \end{aligned} \tag{6}$$

where $L,T$ are nonsingular matrices so that (5) is true for the transformed system. Such a transformation always exists which easily follows from the theory for standard systems [12].

Now if $\sigma_r > \sigma_{r+1}$ and the partitioning in (6) is chosen according to $r$, simple truncation leads to the reduced-order model $(\hat{A},\hat{B},\hat{C},\hat{D},\hat{E}) = (A_{11},B_1,C_1,D,E_{11})$ with some benign properties: first, it can be shown that the reduced-order model is again stable, and second, the error bound

$$\|G-\hat{G}\|_{\mathscr{H}_\infty} \leq 2\sum_{j=r+1}^{n} \sigma_j, \tag{7}$$

holds. Here, $\|\,.\,\|_{\mathscr{H}_\infty}$ denotes the $\mathscr{H}_\infty$-norm, i.e., the 2-induced Hardy operator norm of real rational matrix functions having no poles in the right half plane (see, e.g., [8] and references therein). Due to its nature as 2-induced operator norm, the bound (7) implies (using the Paley-Wiener theorem)

$$\begin{aligned}\|y-\hat{y}\|_{L_2(0,\infty;\mathbb{R}^p)} &= \|y-\hat{y}\|_{\mathscr{H}_2^p} = \|Gu-\hat{G}u\|_{\mathscr{H}_2^p} \leq \|G-\hat{G}\|_{\mathscr{H}_\infty}\|u\|_{\mathscr{H}_2^m} \\ &\leq 2\left(\sum_{j=r+1}^{n}\sigma_j\right)\|u\|_{\mathscr{H}_2^m} = 2\left(\sum_{j=r+1}^{n}\sigma_j\right)\|u\|_{L_2(0,\infty;\mathbb{R}^m)},\end{aligned}$$

where $\mathscr{H}_2^q$ is the frequency domain equivalent of $L_2(0,\infty;\mathbb{R}^q)$ obtained by the (normalized) Laplace transform. Thus, the output error in both, frequency and time domain, can be bounded. The existence of this bound is considered to be the main advantage of BT over other MOR methods, in particular as it can be computed as a by-product of the BT procedure without additional cost and allows to adaptively choose the order of the reduced-order model if it is requested that $\|y-\hat{y}\| \leq \tau\|u\|$ for a given tolerance $\tau$ and either one of the 2-norms in frequency or time domain.

It remains to show how to solve (4) and how to compute $L,T$ as in (6). First we note that it is actually not necessary to compute $P,Q$ and $L,T$ explicitly. Following the ideas for standard systems from [13], one can show that the reduced-order model can be computed (even for non-minimal systems) by the following procedure: as $P,Q$ are positive semidefinite, there exist matrices $S \in \mathbb{R}^{r_P\times n}, R \in \mathbb{R}^{r_Q\times n}$ (by $r_P, r_Q$ we denote the ranks of $P,Q$, respectively) so that $P = S^TS$ and $E^TQE = R^TR$. Now compute a singular value decomposition (SVD)

$$SR^T = [U_1, U_2]\begin{bmatrix}\Sigma_1 & \\ & \Sigma_2\end{bmatrix}\begin{bmatrix}V_1^T \\ V_2^T\end{bmatrix}, \quad \Sigma_1 = \operatorname{diag}(\sigma_1,\ldots,\sigma_r)$$

and set $\hat{L} = \Sigma_1^{-1/2}V_1RE^{-1} \in \mathbb{R}^{r\times n}$, $\hat{T} = S^TU_1\Sigma_1^{-1/2} \in \mathbb{R}^{n\times r}$. Then it is easy to verify that $\hat{L}$ and $\hat{T}$ are the first $r$ rows and columns of $L,T$ from (6) and thus the reduced-order model can equivalently be computed by

$$(\hat{A},\hat{B},\hat{C},\hat{D},\hat{E}) = (\hat{L}A\hat{T},\hat{L}B,C\hat{T},D,\hat{L}E\hat{T}).$$

**Algorithm 2** Coupled Newton iteration for dual Lyapunov equations

```
INPUT: $(A,B,C,E) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{n\times n}$ as in (4); a convergence tolerance $\tau$.
OUTPUT: Numerical full-rank factors such that $P = S^T S$, $E^T QE = R^T R$, where $P, Q$ are the solutions of (4).
$A_0 \leftarrow A$, $S_0 \leftarrow B$, $R_0 \leftarrow C$, $j = 0$.
while $\|A_j + E\|_1 > \tau$ do
    Determine scaling factor $c_j$.
    $S_{j+1} \leftarrow$ full-rank factor of $\frac{1}{\sqrt{2c_j}} \left[ S_j \;\; c_j E A_j^{-1} S_j \right]$.
    $R_{j+1} \leftarrow$ full-rank factor of $\frac{1}{\sqrt{2c_j}} \begin{bmatrix} R_j \\ c_j R_j A_j^{-1} E \end{bmatrix}$.
    $A_{j+1} \leftarrow \frac{1}{2c_j} \left( A_j + c_j^2 E A_j^{-1} E \right)$.
    $j \leftarrow j+1$.
end while
Solve $SE^T = S_j^T$ for $S$ and set $R := R_j$.
```

Note that $\hat{E} = I_r$ and thus the corresponding computations can be saved. Also observe that $\Sigma_1^{-1/2} V_1 R \in \mathbb{R}^{r\times n}$ and thus $\hat{L}$ can be obtained as the solution of the linear system of equations $\hat{L}E = \Sigma_1^{-1/2} V_1 R$ with only $r$ right-hand sides so that $E^{-1}$ needs not be formed explicitly.

In many cases, the numerical ranks of $P, Q$ are small ($r_P, r_Q \ll n$) and thus it is desirable to compute $S, R$ as above directly without first computing Cholesky factors of $P$ and $E^T QE$ as it is done in Hammarling's method for (4) [14–16]. A very efficient method to get $S, R$ directly can be based on the sign function method, for details see [17,18]. The resulting algorithm is given in Algorithm 2. There, the scalar $c_j$ is a scaling factor used to accelerate convergence of this iteration (which is ultimately quadratic). The full-rank factors are computed using rank-revealing LQ/QR factorizations (RRLQ/RRQR) with respect to a tolerance $\varepsilon$ for rank determination, without accumulation of orthogonal factors which makes their computation fairly cheap with a computational complexity bounded by $4n \max\{r_P, r_Q\}^2$ operations per iteration step. For details on the scaling parameter $c_j$ and the column/row compression step see [17]. As $\lim_{j\to\infty} A_j = -E$, the iteration can easily be stopped as soon as $\|A + E\| \leq \tau \cdot \|E\|$ for an appropriate convergence tolerance $\tau$ and an easy to compute matrix norm. After convergence, we obtain the desired full-rank factors of the Gramians as $S = \frac{1}{\sqrt{2}} (E^{-1} \lim_{j\to\infty} S_j)^T$, $R = \frac{1}{\sqrt{2}} \lim_{j\to\infty} R_j$. Note that again, there is no need to compute $E^{-1}$ as $S$ can be obtained by solving a system of linear equations with $r_P$ right-hand sides. In [17] a variant of this iteration is discussed that employs the $R$-factor of the QR factorization of $E$ in the iteration instead of $E$ itself. In this way, each iteration step becomes a lot cheaper and furthermore, the QR factorization can be used to solve the required linear systems of equations with coefficient matrix $E$ (when computing $\hat{L}$ and $S$) just by application of the transposed orthogonal factor of $E$ and backward substitution.

## 2.2 Additive Decomposition of the Transfer Function

In this section we show how to compute an explicit additive decomposition of the transfer function $G(s)$ as in (2) so that $G(s) = G_f(s) + G_\infty(s)$, where $G_f(s)$ and $G_\infty(s)$ have exclusively finite and infinite poles, respectively. Such an algorithm was already proposed in [19]. Here, we suggest a method which employs different computational kernels to achieve this decomposition. The required computations are particularly efficient on computer architectures where matrix multiplication can be performed (almost) at peak performance.

The additive decomposition is achieved by computing nonsingular matrices $U,V \in \mathbb{R}^{n\times n}$ that block-diagonalize $\lambda E - A$, i.e.,

$$\lambda \hat{E} - \hat{A} := U(\lambda E - A)V = \lambda \begin{bmatrix} E_0 & 0 \\ 0 & E_\infty \end{bmatrix} - \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix},$$

and setting $\hat{B} := UB =: \left[\begin{smallmatrix} B_f \\ B_\infty \end{smallmatrix}\right]$, $\hat{C} := CV =: \begin{bmatrix} C_f & C_\infty \end{bmatrix}$, $\hat{D} := D$. Then

$$\begin{aligned} G(s) &= C(sE-A)^{-1}B + D \;=\; \hat{C}(s\hat{E}-\hat{A})^{-1}\hat{B} + \hat{D} \\ &= \begin{bmatrix} C_f & C_\infty \end{bmatrix} \begin{bmatrix} sE_f - A_f & \\ & sE_\infty - A_\infty \end{bmatrix}^{-1} \begin{bmatrix} B_f \\ B_\infty \end{bmatrix} + D \\ &= \underbrace{C_f(sE_f - A_f)^{-1}B_f}_{=:G_f(s)} + \underbrace{C_\infty(sE_\infty - A_\infty)^{-1}B_\infty + D}_{:=G_\infty(s)}. \end{aligned} \tag{8}$$

Thus, we can apply balanced truncation as described in the previous subsection to $G_f$ in order to obtain a reduced-order system with transfer function $\hat{G}_f$.

The block-diagonalization is achieved using a two stage process. First, a block-triangularization of $\lambda E - A$ is computed using the disk function method as described next, then a block diagonalization is achieved by solving a certain generalized Sylvester equation.

**Block-triangularization using the disk function method.** The algorithm discussed here is adapted from [20], and is based on earlier work by Malyshev [21]. This algorithm is referred to as *disk function method* as it can be used to compute the disk function of a matrix pencil, for details see [22]. We also make use of improvements suggested in [23] to reduce its cost.

Given a regular matrix pencil $\lambda E - A$ having all finite eigenvalues inside the unit circle, Algorithm 3 provides an implementation of the disk function method which computes $\tilde{U}, \tilde{V}$ such that

$$\tilde{U}(\lambda E - A)\tilde{V} = \lambda \begin{bmatrix} E_f & W_E \\ 0 & E_\infty \end{bmatrix} - \begin{bmatrix} A_f & W_A \\ 0 & A_\infty \end{bmatrix}, \tag{9}$$

**Algorithm 3** Disk function method

```
INPUT: A matrix pencil $\lambda E - A$, $E, A \in \mathbb{R}^{n\times n}$ with no eigenvalues on the unit circle.
OUTPUT: Orthogonal $\tilde{U}, \tilde{V} \in \mathbb{R}^{n\times n}$ that block-triangularize $\lambda E - A$.
Set $E_0 = E$, $A_0 = A$.
for $j = 0, 1, \ldots$ until convergence do
    $\begin{bmatrix} E_j \\ -A_j \end{bmatrix} \rightarrow \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_j \\ 0 \end{bmatrix}$ (QR factorization),
    $A_{j+1} \leftarrow Q_{12}^T A_j$ and $E_{j+1} \leftarrow Q_{22}^T E_j$,
    $s = j+1$.
end for
Use the subspace extraction procedure from [23] in order to compute $\tilde{U}, \tilde{V}$.
```

where $E_f \in \mathbb{R}^{n_f\times n_f}, A_\infty \in \mathbb{R}^{n_\infty\times n_\infty}$ are nonsingular, $n_f$ is the number of eigenvalues inside the unit circle (here, this equals the number of finite eigenvalues), $n_\infty := n - n_f$ is the number of infinite eigenvalues, and $E_\infty \in \mathbb{R}^{n_\infty\times n_\infty}$ is of nilpotency index $\nu$ which is the index of $\lambda E - A$. (Note that in general, if there are also finite eigenvalues outside the unit circle, a block-triangularization is achieved where $\lambda E_\infty - A_\infty$ contains the finite eigenvalues of modulus larger than 1 and the infinite eigenvalues.)

Algorithm 3 is based on a generalized power iteration (see [23, 24] for more details) and the fact that (see [21, 24])

$$\lim_{j\to\infty} (A_j + E_j)^{-1} E_j = \mathscr{P}^0, \quad \lim_{j\to\infty} (A_j + E_j)^{-1} A_j = \mathscr{P}^\infty,$$

where $\mathscr{P}^0$ and $\mathscr{P}^\infty$ are projectors onto the right deflating subspaces of $A - \lambda E$ corresponding to the eigenvalues inside and outside the unit disk $\mathscr{D}_1(0)$. Convergence of the algorithm is usually checked based on the relative change in $R_j$. Note that the QR decomposition in Step 1 is unique if we choose positive diagonal elements as $\left[ E_j^T, -A_j^T \right]^T$ has full rank in all steps [25]. The convergence rate of the iteration in Algorithm 3 is globally quadratic [20] with deferred convergence in the presence of eigenvalues very close to the unit circle and stagnation in the limiting case of eigenvalues on the unit circle. Also, the method is proven to be numerically backward stable in [20]. Again, accuracy problems are related to eigenvalues close to the unit circle due to the fact that the spectral decomposition problem becomes ill-conditioned in this case.

It should be noted that for our purposes, neither the disk function nor the projectors $\mathscr{P}^0$ nor $\mathscr{P}^\infty$ need to be computed explicitly. All we need are the related matrices $\tilde{U}, \tilde{V}$ from (9). This only requires orthogonal bases for the range and nullspace of these projectors. These can be obtained using a clever subspace extraction technique proposed in [23]. Due to space limitations, we can not provide further details here.

In order to separate finite from infinite eigenvalues using the disk function method for a stable matrix pencil $\lambda E - A$, Algorithm 3 is applied to $(A, \alpha E)$, where $\alpha$ is the radius of a circle, centered at the origin, enclosing the finite eigenvalues of $\lambda E - A$. Sometimes, $\alpha$ can be estimated from the physical background, otherwise a generalization of the Geršgorin circles to matrix pencils [26, 27] may be employed.

**Block-diagonalization.** After block-triangularization as described above, the matrix pencil $\lambda E - A$ has the form (9). A block-diagonal form can now be obtained using the solution matrices $Y, Z$ of the generalized Sylvester equation

$$A_f Y + Z A_\infty + W_A = 0, \quad E_f Y + Z E_\infty + W_E = 0. \tag{10}$$

Then

$$\begin{aligned} \lambda \hat{E} - \hat{A} &:= U(\lambda E - A)V := \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix} \tilde{U}(\lambda E - A)\tilde{V} \begin{bmatrix} I & Y \\ 0 & I \end{bmatrix} \qquad (11) \\ &= \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix} \left( \lambda \begin{bmatrix} E_f & W_E \\ 0 & E_\infty \end{bmatrix} - \begin{bmatrix} A_f & W_A \\ 0 & A_\infty \end{bmatrix} \right) \begin{bmatrix} I & Y \\ 0 & I \end{bmatrix} = \lambda \begin{bmatrix} E_f & 0 \\ 0 & E_\infty \end{bmatrix} - \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix}. \end{aligned}$$

A significant simplification can be observed for matrix pencils of index $\nu = 1$: in this case, $E_\infty = 0$ so that (10) boils down to the subsequent solution of the two linear systems of equations

$$E_f Y = W_E, \quad Z A_\infty = -(W_A + A_f Y). \tag{12}$$

Otherwise, i.e., for $\nu > 1$, one can use an appropriate solver for generalized Sylvester equations, e.g., the Fortran 77 subroutine SB04OD from the Subroutine Library in Control Theory (SLICOT)[3] or its MATLAB gateway function `slgesg` from the SLICOT Basic Systems and Control Toolbox in order to solve (10).

## 2.3 Balanced Truncation for Descriptor Systems

In this section, we combine the algorithms from the previous two sections in order to derive a method for balanced truncation for descriptor systems. The resulting algorithm is mathematically equivalent to an algorithm proposed in [14, 15], but differs in the underlying computational routines employed. Our method may be more efficient in computing environments where matrix multiplication is very fast compared to the fine-grain computations required in the GUPTRI algorithm [28] employed in [14, 15], while our method may suffer from wrong rank decisions in situations when it is difficult to numerically distinguish finite and infinite eigenvalues of $\lambda E - A$.

Employing a minimal realization of $G_\infty$, the reduced-order descriptor system becomes $\hat{G}(s) = \hat{G}_f(s) + G_\infty(s)$. In [14, 15] it is shown that the order $\hat{n}_\infty$ of a minimal realization of $G_\infty$ satisfies $\hat{n}_\infty \leq \min\{\nu m, \nu p, n_\infty\}$. In case of $\nu = 1$, we get

$$G_\infty(s) \equiv \hat{D} := D - C_\infty A_\infty^{-1} B_\infty.$$

[3] See `www.slicot.org`.

**Algorithm 4** BT algorithm for descriptor systems

```
INPUT: A stable descriptor system realized by $(A,B,C,D,E)$ as in (1).
OUTPUT: A stable reduced-order model $(\hat{A},\hat{B},\hat{C},\hat{D},\hat{E})$ of order $r$ satisfying the error bound (7).
{Compute the additive decomposition of the transfer function.}
Compute $\alpha > 0$ so that $\Lambda(A, \alpha E) \subset \mathscr{D}_1(0)$.
Apply Algorithm 3 to $(A, \alpha E)$ in order to block-triangularize $\lambda E - A$ as in (9).
if $\nu = 1$ then
    Solve the linear systems of equations (12).
else
    Solve the generalized Sylvester equation (10).
end if
Compute the block-diagonalization as in (11).
Apply the resulting system equivalence transformation in order to obtain (8).
{Compute the reduced-order model.}
Apply BT as described in subsection 2.1 to $G_f$ and obtain $\hat{G}_f$.
if $\nu = 1$ then
    Set $\hat{D} := D - C_\infty A_\infty^{-1} B_\infty$ and $\hat{G}(s) = \hat{G}_f + \hat{D}$.
else
    Compute a minimal realization of $G_\infty$ and set $\hat{G}(s) = \hat{G}_f + G_\infty$.
end if
```

In case no feed-through term ("D term") is allowed in the simulation software for which the reduced-order model is generated, $\hat{G}(s)$ can then be realized as

$$\hat{G}(s) = \left[\hat{C},\, \hat{D}\right] \left( s \begin{bmatrix} \hat{E} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \hat{A} & 0 \\ 0 & -I_m \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{B} \\ I_m \end{bmatrix}.$$

Procedures for computing a minimal realization of $G_\infty$ in case of index $\nu > 1$ can be found in [14, 15] and amount to applying discrete-time balanced truncation with zero error to the polynomial part. The cost of this procedure is in general $\mathscr{O}(n_\infty^3)$. It can be reduced if the corresponding discrete Lyapunov equations can be solved for their low-rank factors directly similar to Algorithm 2, see, e.g., [29].

As $G_\infty$ is not reduced, just a different realization of possibly smaller order is employed so that $G(s) - \hat{G}(s) = G_f(s) - \hat{G}_f(s)$, the error bound (7) applies.

The resulting BT algorithm for descriptor systems is summarized in Algorithm 4.

## 2.4 Numerical Examples

Algorithm 4 was implemented as C subroutine in the circuit simulator TITAN[4] [30]. In the following, we will present simulation results for two examples provided by Qimonda AG, München, obtained by using reduced-order circuits computed by this subroutine within TITAN.

[4] Copyrighted software, developed by Qimonda AG, München

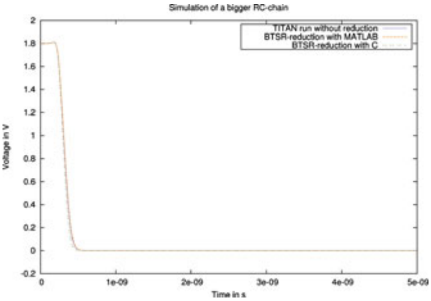


**Fig. 1:** TITAN simulation results for small nonlinear circuit, 1 linear subcircuit ($n = 297$) replaced by reduced-order model ($r = 31$)

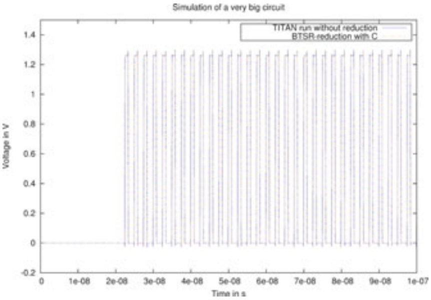


**Fig. 2:** TITAN simulation results for industrial circuit, 14 linear subcircuits are reduced

In the first example, a small nonlinear circuit model, designed for testing and verification of algorithms, is used. The circuit consists of 297 resistors, 268 capacitors, 4 voltage sources, and 8 MOSFETs. A linear subcircuit of order $n = 297$ was extracted and replaced by a model of order $r = 31$ computed by Algorithm 4. Simulation results are shown in Fig. 1. The figure shows results obtained by a MATLAB implementation of Algorithm 4 developed by the author and the C implementation from [30]. A slightly larger error results from using the C version which hints to an unresolved bug in the software. This is under current investigation.

In the second example, an industrial example with 14,677 resistors, 15,404 capacitors, 14 voltage sources, and 4,800 MOSFETs was investigated. The analysis showed that 14 linear subcircuit of varying order could be extracted and reduced. Simulation results for the original circuit and for a model where the 14 linear subcircuits were replaced by BT reduced-order models are shown in Fig. 2. Here we see again that the reduced-order model behaves well in time domain simulation.

## 3 Further Developments

Besides the aspect of singular $E$ often arising in circuit simulation, a number of further issues need to be addressed when applying BT for MOR in this area.

**Large-scale, sparse systems.** Large-scale Lyapunov equations can nowadays be solved by using, e.g., the low-rank ADI method, at a cost that scales with the cost of solving linear systems of equations with coefficient matrices $A-\mu E$; see the surveys [10, 31] and references therein. Thus, BT can be implemented at a cost proportional to Krylov-subspace based methods. Usually more sparse factorizations have to be computed using ADI methods, but the resulting MOR method has the advantageous properties of BT[5]. The ADI method for Lyapunov equations can also be extended to descriptor systems, see [32].

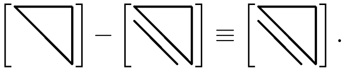
**Sparsification of reduced-order models.** BT is often criticized for producing dense reduced-order models. (Note: this is also true for most moment-matching methods like PRIMA, except for PVL-like methods.) Mostly, reduced-order models are used when solving linear systems of equations of the form

$$
\begin{aligned}
(\mathrm{i}\omega\hat{E}-\hat{A})x &= b \text{ in frequency-domain analysis,} && (13)\\
(\hat{E}-h_k\hat{A})x_{k+1} &= \hat{E}x_k+\ldots \text{ in implicit integrators (transient analysis,...).} && (14)
\end{aligned}
$$

The cost for solving the linear systems may not benefit from the smaller order, if efficient sparse direct solvers for the full-size sparse system matrices are available.

A significant reduction can be achieved by transforming $(\hat{A},\hat{E})$ to Hessenberg-triangular form [25, Algorithm 7.7.1], i.e., compute orthogonal $Q,Z$ such that

$$Q(\lambda\hat{E}-\hat{A})Z=\lambda$$



The new reduced-order system is then $(Q\hat{A}Z, Q\hat{B}, \hat{C}Z, \hat{D}, Q\hat{E}Z)$, the linear systems of equations (13) and (14) then have Hessenberg form, and can thus be solved using $r-1$ Givens rotations only! This only requires the introduction of a dedicated solver for Hessenberg systems in the simulation software.

**Passivity preservation.** An important physical property in circuit simulation is passivity as, e.g., RLC circuits only contain passive devices. Thus, the reduced-order model should preserve this property. For symmetric transfer functions as they are usually encountered in RLC circuit models, BT automatically preserves passivity. Other possibilities are balancing-related methods such as PRBT, see [8, 9] and references therein. A number of recent papers deal with the efficient implementation of PRBT, see [9] for a review. Current efforts are directed towards extending the method to large-scale descriptor systems with sparse coefficient matrices and

[5] Despite unavoidable errors, loss of the theoretical properties of BT is usually not observed in practice.

employing the structure of circuit matrices more efficiently so that explicit computation of projectors can be avoided[6].

**Synthesis.** BT variants based on split-congruence transformations as in [33] are under current investigation. As *split-congruence BT* preserves reciprocity of the transfer function, this allows synthesis of the reduced-order model as circuit. The basic idea here is to exploit the structure of RLC circuits, leading to a “symmetric” transfer function with (for networks without voltage sources)

$$sE-A = s\begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} + \begin{bmatrix} A_1 & A_2^T \\ -A_2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} = C^T, \quad D = 0,$$

where $A_1, E_1 \geq 0$, $E_2 > 0$. This structure can be preserved in the reduced-order model if the BT truncation matrices $\hat{L}, \hat{T}$ are embedded in a so-called split-congruence transformation [33]. The mathematical properties of this approach are not clear yet; we will report on this BT variant in the future.

**Acknowledgements** The author thanks René Günzel for providing some of the simulation results and graphics contained in Section 2.4 and Dr. Georg Denk from Qimonda AG, München, Germany, for providing the circuit models used in these simulations.

The work reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF), grant no. 03BEPAE1, and the EU project O-MOORE-NICE! funded with the Marie Curie Host Fellowships for the Transfer of Knowledge (ToK) Industry-Academia Partnership Scheme. Responsibility for the contents of this publication rests with the author.

## References

1. Tan, S., He, L.: Advanced Model Order Reduction Techniques in VLSI Design. Cambridge University Press, New York (2007)
2. Freund, R.: Model reduction methods based on Krylov subspaces. Acta Numerica **12**, 267–319 (2003)
3. Feldmann, P., Freund, R.: Efficient linear circuit analysis by Padé approximation via the Lanczos process. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **14**, 639–649 (1995)
4. Gallivan, K., Grimme, E., Van Dooren, P.: Asymptotic waveform evaluation via a Lanczos method. Appl. Math. Lett. **7**(5), 75–80 (1994)
5. Odabasioglu, A., Celik, M., Pileggi, L.: PRIMA: passive reduced-order interconnect macromodeling algorithm. In: Tech. Dig. 1997 IEEE/ACM Intl. Conf. CAD, pp. 58–65. IEEE Computer Society Press (1997)
6. Benner, P., Mehrmann, V., Sorensen, D. (eds.): Dimension Reduction of Large-Scale Systems, *Lecture Notes in Computational Science and Engineering*, vol. 45. Springer-Verlag, Berlin/Heidelberg, Germany (2005)
7. Schilders, W., van der Vorst, H., J.Rommes: Model Order Reduction: Theory, Research Aspects and Applications. Springer-Verlag, Berlin, Heidelberg (2008)
8. Antoulas, A.: Approximation of Large-Scale Dynamical Systems. SIAM Publications, Philadelphia, PA (2005)

[6] See `http://www.math.tu-berlin.de/~stykel` for a number of recent preprints dealing with this issue.

9. Benner, P., Faßbender, H.: Numerische Methoden zur passivitätserhaltenden Modellreduktion. at-Automatisierungstechnik **54**(4), 153–160 (2006)
10. Benner, P.: Numerical linear algebra for model reduction in control and simulation. GAMM Mitt. **29**(2), 275–296 (2006)
11. Benner, P., Quintana-Ortí, E., Quintana-Ortí, G.: Parallel model reduction of large-scale linear descriptor systems via Balanced Truncation. In: M. Daydé, J. Dongarra, V. Hernández, J. Palma (eds.) High Performance Computing for Computational Science – VECPAR 2004, no. 3402 in LNCS, pp. 340–353. Springer-Verlag, Berlin/Heidelberg (2005)
12. Moore, B.: Principal component analysis in linear systems: Controllability, observability, and model reduction. IEEE Trans. Automat. Control **AC-26**, 17–32 (1981)
13. Tombs, M., Postlethwaite, I.: Truncated balanced realization of a stable non-minimal state-space system. Internat. J. Control **46**(4), 1319–1330 (1987)
14. Stykel, T.: Analysis and numerical solution of generalized Lyapunov equations. Dissertation, TU Berlin (2002)
15. Stykel, T.: Gramian-based model reduction for descriptor systems. Math. Control Signals Systems **16**(4), 297–319 (2004)
16. Penzl, T.: Numerical solution of generalized Lyapunov equations. Adv. Comp. Math. **8**, 33–48 (1997)
17. Benner, P., Quintana-Ortí, E.: Solving stable generalized Lyapunov equations with the matrix sign function. Numer. Algorithms **20**(1), 75–100 (1999)
18. Benner, P., Claver, J., Quintana-Ortí, E.: Efficient solution of coupled Lyapunov equations via matrix sign function iteration. In: A. Dourado et al. (ed.) Proc. 3rd Portuguese Conf. on Automatic Control CONTROLO'98, Coimbra, pp. 205–210 (1998)
19. Kågström, B., Van Dooren, P.: A generalized state-space approach for the additive decomposition of a transfer matrix. J. Numer. Linear Algebra Appl. **1**(2), 165–181 (1992)
20. Bai, Z., Demmel, J., Gu, M.: An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems. Numer. Math. **76**(3), 279–308 (1997)
21. Malyshev, A.: Parallel algorithm for solving some spectral problems of linear algebra. Linear Algebra Appl. **188/189**, 489–520 (1993)
22. Benner, P., Byers, R.: Disk functions and their relationship to the matrix sign function. In: Proc. European Control Conf. ECC 97, Paper 936. BELWARE Information Technology, Waterloo, Belgium (1997). CD-ROM
23. Sun, X., Quintana-Ortí, E.: Spectral division methods for block generalized Schur decompositions. Mathematics of Computation **73**, 1827–1847 (2004)
24. Benner, P.: Contributions to the Numerical Solution of Algebraic Riccati Equations and Related Eigenvalue Problems. Logos–Verlag, Berlin, Germany (1997). *Also:* Dissertation, Fakultät für Mathematik, TU Chemnitz–Zwickau, 1997.
25. Golub, G., Van Loan, C.: Matrix Computations, third edn. Johns Hopkins University Press, Baltimore (1996)
26. Stewart, G.: Gershgorin theory for the generalized eigenvalue problem $Ax = \lambda Bx$. Math. Comp. **29**, 600–606 (1975)
27. Kostić, V.: Eigenvalue localization for matrix pencils. Presented at workshop *Applied Linear Algebra — in honor of Ivo Marek. Novi Sad, April 28–30, 2008*
28. Demmel, J., Kågström, B.: The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. Part I: Theory and algorithms. ACM Trans. Math. Software **19**, 160–174 (1993)
29. Benner, P., Quintana-Ortí, E., Quintana-Ortí, G.: Parallel algorithms for model reduction of discrete-time systems. Int. J. Syst. Sci. **34**(5), 319–333 (2003)
30. Günzel, R.: Balanced truncation for descriptor systems arising in interconnect modeling. Diplomarbeit, Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, FRG (2008)
31. Gugercin, S., Li, J.R.: Smith-type methods for balanced truncation of large systems. In: Benner et al. [6]. Chapter 2 (pages 49–82).
32. Stykel, T.: Low-rank iterative methods for projected generalized Lyapunov equations. Electr. Trans. Num. Anal. **30**, 187–202 (2008)
33. Kerns, K., Yang, A.: Preservation of passivity during RLC network reduction via split congruence transformations. IEEE Trans. CAD Integr. Circuits Syst. **17**(7), 582–591 (1998)

# Passivity-Preserving Balanced Truncation Model Reduction of Circuit Equations

Tatjana Stykel and Timo Reis

**Abstract** We consider passivity-preserving model reduction of circuit equations using the bounded real balanced truncation method applied to a Moebius-transformed system. This method is based on balancing the solutions of the projected Lur'e or Riccati matrix equations. We also discuss their numerical solution exploiting the underlying structure of circuit equations. A numerical example is given.

## 1 Introduction

A modified nodal analysis (MNA) for linear RLC circuits yields a linear system of differential-algebraic equations (DAEs)

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where

$$E = \begin{bmatrix} A_{\mathcal{C}}\mathcal{C}A_{\mathcal{C}}^T & 0 & 0 \\ 0 & \mathcal{L} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -A_{\mathcal{R}}\mathcal{R}^{-1}A_{\mathcal{R}}^T & -A_{\mathcal{L}} & -A_{\mathcal{V}} \\ A_{\mathcal{L}}^T & 0 & 0 \\ A_{\mathcal{V}}^T & 0 & 0 \end{bmatrix}, \quad B = -\begin{bmatrix} A_{\mathcal{I}} & 0 \\ 0 & 0 \\ 0 & I \end{bmatrix} = C^T. \tag{2}$$

Here $A_{\mathcal{C}} \in \mathbb{R}^{n_\eta, n_{\mathcal{C}}}$, $A_{\mathcal{L}} \in \mathbb{R}^{n_\eta, n_{\mathcal{L}}}$, $A_{\mathcal{R}} \in \mathbb{R}^{n_\eta, n_{\mathcal{R}}}$, $A_{\mathcal{V}} \in \mathbb{R}^{n_\eta, n_{\mathcal{V}}}$ and $A_{\mathcal{I}} \in \mathbb{R}^{n_\eta, n_{\mathcal{I}}}$ are incidence matrices describing the circuit topology, and $\mathcal{R}$, $\mathcal{L}$ and $\mathcal{C}$ are resistance, inductance and capacitance matrices, respectively. Linear RLC circuits are often used to model interconnects, transmission lines and pin packages in VLSI networks.

In the following we will assume that

- the matrix $A_{\mathcal{V}}$ has full column rank;
- the matrix $[A_{\mathcal{C}}, A_{\mathcal{L}}, A_{\mathcal{R}}, A_{\mathcal{V}}]$ has full row rank;
- the matrices $\mathcal{R}$, $\mathcal{L}$ and $\mathcal{C}$ are symmetric and positive definite.

Tatjana Stykel, Timo Reis
Institut für Mathematik, MA 4-5, TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, e-mail: stykel@math.tu-berlin.de, reis@math.tu-berlin.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_59,

These assumptions guarantee that the pencil $\lambda E - A$ is regular, i.e., $\det(\lambda E - A) \not\equiv 0$. Moreover, system (1), (2) is *passive*, i.e., it does not generate energy, and *reciprocal*, i.e., its transfer function $G(s) = C(sE-A)^{-1}B$ satisfies the symmetry relation $G(s) = S_{\mathrm{ext}} G(s)^T S_{\mathrm{ext}}$ with an external signature $S_{\mathrm{ext}} = \mathrm{diag}(I_{n_I}, -I_{n_{\mathcal{V}}})$, see [1]. Furthermore, passivity is equivalent to the *positive realness* of $G$ meaning that $G$ is analytic in the open right half-plane $\mathbb{C}_+$ and $G(s) + G^T(\bar{s})$ is positive semidefinite for all $s \in \mathbb{C}_+$, see [2].

The number $n = n_\eta + n_{\mathcal{L}} + n_{\mathcal{V}}$ of state variables in (1) is related to the number of circuit elements and usually very large. This makes the analysis and numerical simulation of circuit equations unacceptably time consuming. Therefore, model order reduction is of great importance.

A general idea of model reduction is to approximate the large-scale system (1) by a reduced-order model

$$\begin{aligned} \tilde{E}\dot{\tilde{x}}(t) &= \tilde{A}\tilde{x}(t) + \tilde{B}u(t), \\ \tilde{y}(t) &= \tilde{C}\tilde{x}(t), \end{aligned} \tag{3}$$

where $\tilde{E}$, $\tilde{A} \in \mathbb{R}^{\ell,\ell}$, $\tilde{B} \in \mathbb{R}^{\ell,m}$, $\tilde{C} \in \mathbb{R}^{m,\ell}$ and $\ell \ll n$. It is required that the approximate system (3) captures the input-output behaviour of (1) to a required accuracy and preserves passivity and reciprocity. The preservation of these properties allows a back interpretation of the reduced-order model (3) as an electrical circuit which has fewer electrical components than the original one [1, 2].

Krylov subspace based methods [3, 4] are mostly used model reduction methods in circuit simulation. Although these methods are efficient for very large sparse problems, stability and passivity are not necessarily preserved in the reduced-order model. Passivity-preserving model reduction methods based on Krylov subspaces have been developed for standard state space systems [5, 6] and also for structured generalized state space systems describing interconnect circuits [4, 7, 8]. Despite the successful application of these methods in circuit simulation, they provide only a good local approximation and, so far, there exist no global error bounds.

Balanced truncation is another model reduction approach commonly used in control design. In order to capture specific system properties, different balancing techniques have been developed for standard state space systems, e.g., [9, 10] and also for DAEs [11, 12]. An important property of balancing-related model reduction is the existence of computable error bounds. Balanced truncation is based on the transformation of the dynamical system into a balanced form whose controllability and observability Gramians are both equal to a diagonal matrix. Then a reduced-order model is determined by the truncation of the states corresponding to small diagonal elements of the balanced Gramians.

In this paper, we present a passivity-preserving model reduction method for circuit equations (1), (2) that is based on so-called bounded real balanced truncation applied to a Moebius-transformed system. It requires balancing two Gramians that satisfy the projected Lur'e equations. Under some assumptions such equations can be rewritten as the projected Riccati equations. We also discuss the numerical solution of these matrix equations via Newton's method and present some results of numerical experiments.

Throughout the paper $\mathbb{R}^{n,m}$ denotes the spaces of $n \times m$ real matrices and $A^T$ stands for the transpose of $A \in \mathbb{R}^{n,m}$. An identity matrix of order $n$ is denoted by $I_n$ or simply by $I$. Further, for symmetric matrices $X, Y \in \mathbb{R}^{n,n}$, we write $X > Y$ ($X \geq Y$) if $X - Y$ is positive (semi)definite. For a real diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$, we have $|D| = \operatorname{diag}(|d_1|, \ldots, |d_n|)$ and $\operatorname{sign}(D) = \operatorname{diag}(\operatorname{sign}(d_1), \ldots, \operatorname{sign}(d_n))$.

## 2 Passivity-Preserving Balanced Truncation

In this section, we present a passivity-preserving balanced truncation method for circuit equations. This method is based on the fact that the transfer function $G(s)$ is positive real if and only if the Moebius-transformed function

$$\mathscr{G}(s) = \mathscr{M}(G(s)) := \big(I - G(s)\big)\big(I + G(s)\big)^{-1}$$

is *bounded real*, i.e., $\mathscr{G}$ is analytic in $\mathbb{C}_+$ and $I - \mathscr{G}(s)\mathscr{G}^T(\overline{s})$ is positive semidefinite for all $s \in \mathbb{C}_+$, see [2]. Note that for $G(s) = C(sE - A)^{-1}B + D$ with a nonsingular matrix $I + D$, the transfer function $\mathscr{G}(s) = \mathscr{M}(G(s))$ can be represented as $\mathscr{G}(s) = \mathscr{C}(s\mathscr{E} - \mathscr{A})^{-1}\mathscr{B} + \mathscr{D}$, where

$$\begin{array}{lll} \mathscr{E} = E, & \mathscr{A} = A - B(I+D)^{-1}C, & \mathscr{B} = -\sqrt{2}B(I+D)^{-1}, \\ \mathscr{C} = \sqrt{2}(I+D)^{-1}C, & \mathscr{D} = (I-D)(I+D)^{-1}. & \end{array} \tag{4}$$

For system (1), (2), a passive reduced-order model (3) can be computed by the model reduction method presented in [11, 13]. First, we consider the Moebius-transformed system $\mathscr{G} = \mathscr{M}(G)$ and apply a bounded real balanced truncation method to $\mathscr{G}$, i.e., to (4). The obtained bounded real reduced-order system $\tilde{\mathscr{G}}$ is then transformed into $\tilde{G} = \mathscr{M}(\tilde{\mathscr{G}})$ which is positive real.

### 2.1 *Bounded Real Balanced Truncation*

The bounded realness of $\mathscr{G}$ implies that $\mathscr{G}$ is proper, i.e., there exists $M_0 = \lim\limits_{s\to\infty} \mathscr{G}(s)$. Furthermore, for $E$, $A$, $B$ and $C$ as in (2), the *projected Lur'e equations* [1]

$$\begin{array}{l} EX(A-BC)^T + (A-BC)XE^T + 2P_lBB^TP_l^T = -2K_cK_c^T, \\ EXC^T - P_lBM_0^T = -K_cJ_c^T, \quad J_cJ_c^T = I - M_0M_0^T, \quad X = P_rXP_r^T \geq 0, \end{array} \tag{5}$$

and

$$\begin{array}{l} E^TY(A-BC) + (A-BC)YE + 2P_r^TC^TCP_r = -2K_o^TK_o, \\ -E^TYB + P_r^TC^TM_0 = -K_o^TJ_o, \quad J_o^TJ_o = I - M_0^TM_0 \quad Y = P_l^TYP_l \geq 0, \end{array} \tag{6}$$

are solvable for $X \in \mathbb{R}^{n,n}$, $K_c \in \mathbb{R}^{n,m}$, $J_c \in \mathbb{R}^{m,m}$ and $Y \in \mathbb{R}^{n,n}$, $K_o \in \mathbb{R}^{m,n}$, $J_o \in \mathbb{R}^{m,m}$, respectively, see [13]. Here, $P_r$ and $P_l$ are the projectors onto the right and left deflating subspaces of the pencil $\lambda E - A + BC$ corresponding to the finite eigenvalues along the right and left deflating subspaces corresponding to the eigenvalue at

[1] These equations are named after the Russian mathematician and engineer A.I. Lur'e (1901-1980). In the literature, they are also known as Kalman-Yakubovich-Popov equations [14].

infinity. The minimal solutions $X_{\min}$ and $Y_{\min}$ of (5) and (6) that satisfy $0 \le X_{\min} \le X$ and $0 \le Y_{\min} \le Y$ for all symmetric solutions $X$ and $Y$ of (5) and (6), respectively, are called the *bounded real controllability Gramian* and the *bounded real observability Gramian* of $\mathscr{G}$.

In the bounded real balanced truncation method, we determine the Cholesky factors $R$ and $L$ of $X_{\min} = RR^T$ and $Y_{\min} = LL^T$, respectively, and compute the singular value decomposition

$$L^T ER = [U_1, U_2]\operatorname{diag}(\Pi_1, \Pi_2)[V_1, V_2]^T,$$

where $[U_1, U_2]$ and $[V_1, V_2]$ have orthonormal columns, $\Pi_1 = \operatorname{diag}(\pi_1 I_{l_1}, \ldots, \pi_r I_{l_r})$ and $\Pi_2 = \operatorname{diag}(\pi_{r+1} I_{l_{r+1}}, \ldots, \pi_q I_{l_q})$ with $\pi_1 > \ldots > \pi_r > \pi_{r+1} > \ldots > \pi_q$. The values $\pi_j$ are called the *characteristic values* of $\mathscr{G}$. They determine the importance of state variables. A reduced-order model for $\mathscr{G} = [\mathscr{E}, \mathscr{A}, \mathscr{B}, \mathscr{C}, I]$ as in (4) can be computed by projection onto the left and right subspaces corresponding to the dominant characteristic values. Such a model is given by $\tilde{\mathscr{G}} = [\tilde{\mathscr{E}}, \tilde{\mathscr{A}}, \tilde{\mathscr{B}}, \tilde{\mathscr{C}}, I]$ with

$$\tilde{\mathscr{E}} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \qquad \tilde{\mathscr{A}} = \begin{bmatrix} W^T(A - BC)T & 0 \\ 0 & I \end{bmatrix},$$

$$\tilde{\mathscr{B}} = \begin{bmatrix} -\sqrt{2}W^T B \\ B_\infty \end{bmatrix}, \qquad \tilde{\mathscr{C}} = \begin{bmatrix} \sqrt{2}CT, & C_\infty \end{bmatrix},$$

where $W = LU_1\Pi_1^{-1/2}$, $T = RV_1\Pi_1^{-1/2}$, and the matrices $B_\infty$ and $C_\infty$ are chosen such that $I - M_0 = C_\infty B_\infty$.

## 2.2 Application to Circuit Equations

By exploiting the structure of circuit equations, the model reduction procedure presented above can be made more efficient and accurate. Since the MNA matrices in (2) satisfy

$$E^T = S_{\text{int}} E\, S_{\text{int}}, \qquad A^T = S_{\text{int}} A\, S_{\text{int}}, \qquad B^T = S_{\text{ext}} C\, S_{\text{int}},$$

where $S_{\text{int}} = \operatorname{diag}(I_{n_\eta}, -I_{n_{\mathscr{L}}}, -I_{n_{\mathscr{V}}})$ and $S_{\text{ext}} = \operatorname{diag}(I_{n_I}, -I_{n_{\mathscr{V}}})$, we find that

$$P_l = S_{\text{int}} P_r^T S_{\text{int}}, \qquad X_{\min} = S_{\text{int}} Y_{\min} S_{\text{int}} = S_{\text{int}} LL^T S_{\text{int}}^T = RR^T.$$

Thus, for the linear circuit equations (1), (2), it is enough to compute only one projector and solve only one projected Lur'e equation. Another projector and also the solution of the dual Lur'e equation are given for free. Furthermore, we can show that $L^T ER = L^T ES_{\text{int}} L$ is symmetric. Then the characteristic values $\pi_j$ can be computed from an eigenvalue decomposition of $L^T ES_{\text{int}} L$ instead of a more expensive singular value decomposition. Finally, using the symmetry of $(I - M_0)S_{\text{ext}}$, we can determine $B_\infty$ and $C_\infty$ from the eigenvalue decomposition of $(I - M_0)S_{\text{ext}}$.

Summarizing, we obtain the following PAssivity-preserving Balanced Truncation method for Electrical Circuits (PABTEC).

**Algorithm 1** *Passivity-preserving balanced truncation for electrical circuits*
Given $G=[E,A,B,C]$ as in (2), compute a reduced-order model $\tilde{G}=[\tilde{E},\tilde{A},\tilde{B},\tilde{C}]$.

1. Compute the Cholesky factor $L$ of $Y_{\min}=LL^T$ that is the minimal solution of the projected Lur'e equation (6).
2. Compute the eigenvalue decomposition
$$L^T E S_{\text{int}} L=[U_1,\,U_2]\operatorname{diag}(\Lambda_1,\Lambda_2)[U_1,\,U_2]^T,$$
where $[U_1,U_2]$ is orthogonal, $\Lambda_1=\operatorname{diag}(\lambda_1 I,\ldots,\lambda_r I)$, $\Lambda_2=\operatorname{diag}(\lambda_{r+1}I,\ldots,\lambda_q I)$ and $|\lambda_1|>\ldots>|\lambda_r|>|\lambda_{r+1}|>\ldots>|\lambda_q|$.
3. Compute the eigenvalue decomposition $(I-M_0)S_{\text{ext}}=U_0\Lambda_0U_0^T$, where $U_0$ is orthogonal and $\Lambda_0=\operatorname{diag}(\hat{\lambda}_1,\ldots,\hat{\lambda}_m)$.
4. Compute the reduced-order system
$$\begin{gathered}\tilde{E}=\begin{bmatrix} I & 0\\ 0 & 0\end{bmatrix},\qquad \tilde{A}=\frac{1}{2}\begin{bmatrix} 2W^TAT & \sqrt{2}W^TBC_\infty\\ -\sqrt{2}B_\infty CT & 2I-B_\infty C_\infty\end{bmatrix},\\ \tilde{B}=\frac{\sqrt{2}}{2}\begin{bmatrix}\sqrt{2}W^TB\\ -B_\infty\end{bmatrix}\qquad \tilde{C}=\frac{\sqrt{2}}{2}\begin{bmatrix}\sqrt{2}CT,\ C_\infty\end{bmatrix},\end{gathered}\tag{7}$$
where $B_\infty=S_0|\Lambda_0|^{1/2}U_0^TS_{\text{ext}}$, $\quad C_\infty=U_0|\Lambda_0|^{1/2}$, $\quad S_0=\operatorname{sign}(\Lambda_0)$,
$W=LU_1|\Lambda_1|^{-1/2}$, $\quad T=S_{\text{int}}LU_1S_1|\Lambda_1|^{-1/2}$, $\quad S_1=\operatorname{sign}(\Lambda_1)$.

One can show that the reduced-order system (7) is passive and reciprocal [13]. Furthermore, we can estimate the $\mathbb{H}_\infty$-norm of the error defined as
$$\|\tilde{G}-G\|_{\mathbb{H}_\infty}=\sup_{s\in\mathbb{C}_+}\|\tilde{G}(s)-G(s)\|,$$
where $\|\cdot\|$ denotes the spectral matrix norm. If $\|I+G\|_{\mathbb{H}_\infty}(\pi_{r+1}+\ldots+\pi_q)<1$, then we have the following error bound
$$\|\tilde{G}-G\|_{\mathbb{H}_\infty}\le\frac{\|I+G\|^2_{\mathbb{H}_\infty}(\pi_{r+1}+\ldots+\pi_q)}{1-\|I+G\|_{\mathbb{H}_\infty}(\pi_{r+1}+\ldots+\pi_q)},\tag{8}$$
see [11] for details.

## 3 Computation of the Bounded Real Gramian

If $I-M_0^TM_0$ is nonsingular, then $I-M_0M_0^T$ is also nonsingular and the projected Lur'e equation (6) can be rewritten as the projected algebraic Riccati equation
$$E^TY\hat{A}+\hat{A}^TYE+E^TY\hat{B}\hat{B}^TYE+P_r^T\hat{C}^T\hat{C}P_r=0,\quad Y=P_l^TYP_l,\tag{9}$$
where $\hat{A}=A-BC-2P_lB(I-M_0^TM_0)^{-1}M_0^TCP_r$, $\hat{B}=\sqrt{2}P_lBJ_o^{-1}$, $\hat{C}=\sqrt{2}J_c^{-1}C$, $J_o^TJ_o=I-M_0^TM_0$ and $J_cJ_c^T=I-M_0M_0^T$. One can show that the minimal solution $Y_{\min}$ of (6) is at least a semi-stabilizing solution of (9) in the sense that all the finite eigenvalues of $\lambda E-\hat{A}-\hat{B}\hat{B}^TY_{\min}E$ are in the closed left half-plane. Thus, the bounded real Gramian $Y_{\min}$ can be computed by solving (9) via Newton's method.

**Algorithm 2** *Newton's method for the projected Riccati equation*

```
Given $E$, $\hat{A} \in \mathbb{R}^{n,n}$, $\hat{B} \in \mathbb{R}^{n,m}$, $\hat{C} \in \mathbb{R}^{m,n}$, projectors $P_r$, $P_l$ and a stabilizing initial guess $Y_0$, compute an approximate solution of the projected Riccati equation (9).
FOR $j = 1, 2, \dots, j_{\max}$
    1. Compute $K_j = \hat{B}^T Y_{j-1} E$ and $A_j = \hat{A} + \hat{B} K_j$.
    2. Solve the projected Lyapunov equation
       $E^T Y_j A_j + A_j^T Y_j E = -P_r^T (\hat{C}^T \hat{C} - K_j^T K_j) P_r, \qquad Y_j = P_l^T Y_j P_l.$
END FOR
```

Similarly to the standard state space case [15, 16], one can show that if all the finite eigenvalues of $\lambda E - \hat{A}$ have negative real part, then starting with $Y_0 = 0$, all $\lambda E - A_j$ have finite eigenvalues in the open left half-plane only and $\lim\limits_{j \to \infty} Y_j = Y_{\min}$. Some difficulties may occur if the pencil $\lambda E - \hat{A}$ has eigenvalues on the imaginary axis. This problem remains for future work.

If the eigenvalues of $Y_{\min}$ decay to zero very rapidly, then $Y_{\min}$ can be well approximated by a matrix of low rank. Such a low-rank approximation can be computed in factored form $Y_{\min} \approx \tilde{L}\tilde{L}^T$ with $\tilde{L} \in \mathbb{R}^{n,k}$, $k \ll n$. To determine the low-rank factor $\tilde{L}$ we can use the same approach as in [17]. Starting with $Y_{1,0} = Y_0$ and $Y_{2,0} = 0$, in each Newton iteration we compute $K_j = \hat{B}^T (Y_{1,j-1} - Y_{2,j-1}) E$, $A_j = \hat{A} + \hat{B} K_j$ and then solve two projected Lyapunov equations

$$E^T Y_{1,j} A_j + A_j^T Y_{1,j} E = -P_r^T \hat{C}^T \hat{C} P_r, \qquad Y_{1,j} = P_l^T Y_{1,j} P_l, \tag{10}$$

$$E^T Y_{2,j} A_j + A_j^T Y_{2,j} E = -P_r^T K_j^T K_j P_r, \qquad Y_{2,j} = P_l^T Y_{2,j} P_l, \tag{11}$$

for the low-rank factors $L_{1,j}$ and $L_{2,j}$ such that $Y_{1,j} \approx L_{1,j} L_{1,j}^T$ and $Y_{2,j} \approx L_{2,j} L_{2,j}^T$, respectively. Once the convergence is observed, an approximate solution $Y_{\min} \approx \tilde{L}\tilde{L}^T$ of the projected Riccati equation (9) can be computed in factored form by solving the projected Lyapunov equation

$$E^T Y \hat{A} + \hat{A}^T Y E = -P_r^T \hat{C}_0^T \hat{C}_0 P_r, \qquad Y = P_l^T Y P_l \tag{12}$$

with $\hat{C}_0 = [\,\hat{C}^T,\ E^T (Y_{1,j_{\max}} - Y_{2,j_{\max}}) \hat{B}\,]^T$. For computing low-rank factors of the solutions of the projected Lyapunov equations (10)–(12), we can use the generalized alternating direction implicit method [18]. Note that in this method we need to compute the products $(E^T + \tau A_j^T)^{-1} v$ with $\tau \in \mathbb{C}_-$ and $v \in \mathbb{R}^n$. Taking into account that $E + \tau A_j = E + \tau(A - BC) - \hat{B}\hat{K}_j$ with the low-rank matrices $\hat{B} \in \mathbb{R}^{n,m}$ and $\hat{K}_j = \tau(J_o^{-T} M_0^T C P_r - K_j) \in \mathbb{R}^{m,n}$ we can use the Sherman-Morrison-Woodbury formula [19, Section 2.1.3] to compute these products as

$$(E^T + \tau A_j^T)^{-1} v = v_1 + M_{\hat{K}} \left( (I_m - \hat{B}^T M_{\hat{K}})^{-1} \hat{B}^T \right) v_1,$$

where $v_1 = (E^T + \tau(A - BC)^T)^{-1} v$ and $M_{\hat{K}} = (E^T + \tau(A - BC)^T)^{-1} \hat{K}_j^T$. The latter can be determined by solving linear systems with the sparse matrix $E^T + \tau(A - BC)^T$ either by computing sparse LU factorization or by using iterative Krylov subspace methods [20].

A major difficulty in the numerical solution of the projected Lyapunov and Riccati equations with large matrix coefficients is that the matrix $M_0$ and the spectral projectors $P_l$ and $P_r$ are required. Fortunately, we can exploit the structure of the MNA matrices (2) to construct the required projectors in explicit form using a matrix chain approach from [21]. Furthermore, we can obtain an explicit formula for the matrix $M_0$ and derive necessary and sufficient conditions for invertibility of $I - M_0^T M_0$ in terms of the circuit topology, see [13] for details.

# 4 Numerical Example

In this section, we present some results of numerical experiments to demonstrate the feasibility of the PABTEC method.

**Example** This example describing a three-port RC circuit was provided by NEC Laboratories Europe. We have a passive system of order $n = 2007$. The minimal solution of the projected Riccati equation (9) was approximated by a low-rank matrix $Y_{\min} \approx \tilde{L}\tilde{L}^T$ with $\tilde{L} \in \mathbb{R}^{n,118}$ using Newton's method. Figure 1 shows that the characteristic values decay rapidly, so we can expect a good approximation by a reduced-order model. The original system was approximated by a model of order $\ell = 44$. The spectral norms of the frequency responses $\|G(\mathrm{i}\omega)\|$ and $\|\tilde{G}(\mathrm{i}\omega)\|$ for a frequency range $\omega \in [1, 10^{15}]$ are presented in Figure 2. We also display there the absolute error $\|\tilde{G}(\mathrm{i}\omega) - G(\mathrm{i}\omega)\|$ and the error bound (8).

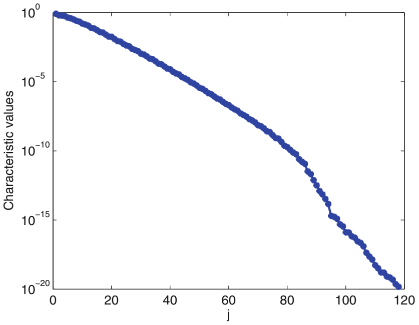


**Fig. 1** RC circuit: characteristic values of $\mathscr{G}$

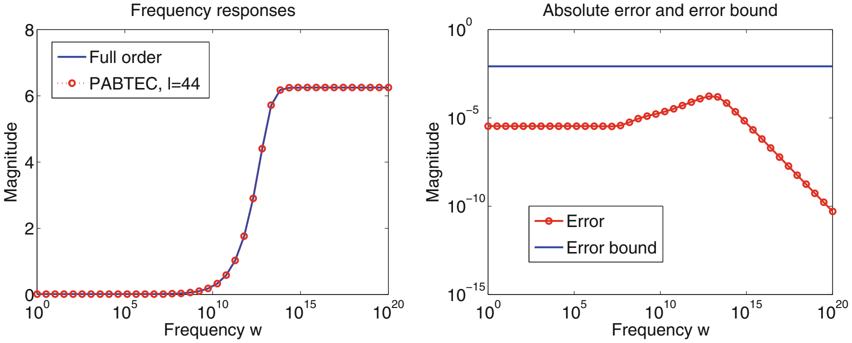


**Fig. 2:** RC circuit: (*left*) the frequency responses of the original and the reduced-order systems; (*right*) the absolute error and error bound

**Acknowledgements** This work is supported in part by the DFG Research Center MATHEON in Berlin and by the Research Network SYRENE - *System Reduction for Nanoscale IC Design* funded by the BMBF. The authors would like to thank Achim Basermann and Carsten Neff from NEC Laboratories Europe, IT Research Division for providing the circuit examples and for stimulating discussions.

## References

1. Reis, T.: Circuit synthesis of passive descriptor systems - a modified nodal approach. Internat. J. Circuit Theory Appl. (to appear)
2. Anderson, B., Vongpanitlerd, S.: Network Analysis and Synthesis. Prentice Hall, Englewood Cliffs, NJ (1973)
3. Feldmann, P., Freund, R.: Efficient linear circuit analysis by Padé approximation via the Lanczos process. IEEE Trans. Computer-Aided Design Integr. Circuit Syst. **14**, 639–649 (1995)
4. Odabasioglu, A., Celik, M., Pileggi, L.: PRIMA: Passive reduced-order interconnect macromodeling algorithm. IEEE Trans. Computer-Aided Design Integr. Circuits Syst. **17**(8), 645–654 (1998)
5. Antoulas, A.: A new result on passivity preserving model reduction. Systems Control Lett. **54**(4), 361–374 (2005)
6. Sorensen, D.: Passivity preserving model reduction via interpolation of spectral zeros. Systems Control Lett. **54**(4), 347–360 (2005)
7. Freund, R.: SPRIM: structure-preserving reduced-order interconnect macromodeling. In: Technical Digest of the 2004 IEEE/ACM International Conference on Computer-Aided Design, pp. 80–87. IEEE Computer Society Press, Los Alamos, CA (2004)
8. Freund, R., Feldmann, P.: The SyMPVL algorithm and its applications in interconnect simulation. In: Proceedings of the 1997 International Conference on Simulation of Semiconductor Processes and Devices, pp. 113–116. IEEE, New York (1997)
9. Gugercin, S., Antoulas, A.: A survey of model reduction by balanced truncation and some new results. Internat. J. Control **77**(8), 748–766 (2004)
10. Moore, B.: Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Automat. Control **26**(1), 17–32 (1981)
11. Reis, T., Stykel, T.: Passive and bounded real balancing for model reduction of descriptor systems. Preprint 25-2008, Institut für Mathematik, TU Berlin (2008)
12. Stykel, T.: Gramian-based model reduction for descriptor systems. Math. Control Signals Systems **16**, 297–319 (2004)
13. Reis, T., Stykel, T.: Passivity-preserving balanced truncation for electrical circuits. Preprint 32-2008, Institut für Mathematik, TU Berlin (2008).
14. Ionescu, V., Oară, C., Weiss M.: Generalized Riccati Theory and Robust Control: A Popov Function Approach. John Wiley and Sons, Chichester, UK (1999)
15. Benner, P.: Numerical solution of special algebraic Riccati equations via exact line search method. In: Proceedings of the European Control Conference (ECC97), Paper 786. BELWARE Information Technology, Waterloo, Belgium (1997)
16. Varga, A.: On computing high accuracy solutions of a class of Riccati equations. Control Theory Adv. Techn. **10**, 2005–2016 (1995)
17. Benner, P., Quintana-Ortí, E., Quintana-Ortí, G.: Efficient numerical algorithms for balanced stochastic truncation. Internat. J. Appl. Math. Comput. Sci. **11**(5), 1123–1150 (2001)
18. Stykel, T.: Low-rank iterative methods for projected generalized Lyapunov equations. Electron. Trans. Numer. Anal. **30**, 187–202 (2008)
19. Golub, G.H., Van Loan, C.F.: Matrix Computations. 3rd ed. Johns Hopkins University Press, Baltimore (1996)
20. Saad, Y.: Iterative Methods for Sparse Linear Systems. PWS Publishing Company, Boston, MA (1996)
21. März, R.: Canonical projectors for linear differential algebraic equations. Comput. Math. Appl. **31**(4/5), 121–135 (1996)

# A New Approach to Passivity Preserving Model Reduction: The Dominant Spectral Zero Method

Roxana Ionutiu, Joost Rommes, and Athanasios C. Antoulas

**Abstract** A new model reduction method for circuit simulation is presented, which preserves passivity by interpolating dominant spectral zeros. These are computed as poles of an associated Hamiltonian system, using an iterative solver: the subspace accelerated dominant pole algorithm (SADPA). Based on a dominance criterion, SADPA finds relevant spectral zeros and the associated invariant subspaces, which are used to construct the passivity preserving projection. RLC netlist equivalents for the reduced models are provided.

## 1 Introduction

The design of integrated circuits has become increasingly complex, thus electromagnetic couplings between components on a chip are no longer negligible. To verify coupling effects, on-chip interconnections are modeled as RLC circuits and simulated. As these circuits contain millions of electrical components, the underlying dynamical systems have millions of internal variables and cannot be simulated in full dimension. Model order reduction (MOR) aims at approximating the mathematical description of a large scale circuit with a model of smaller dimension, which replaces the original model during verification and speeds up simulation. The reduction method should preserve important properties of the original model (i.e., stability, passivity) and have an efficient, robust implementation, suitable for large-scale applications. RLC circuits describing the interconnect are *passive* systems, with *positive real* transfer functions [1], thus reduced models should also be passive. A passive reduced model can be synthesized back into an RLC circuit [1],

Roxana Ionutiu, Athanasios C. Antoulas
Department of Electrical and Computer Engineering, MS-380, William Marsh Rice University, P.O. Box 1892, Houston, TX 77251-1892, USA, e-mail: roxana.ionutiu@rice.edu, aca@rice.edu,
Jointly with the School of Engineering and Science, Jacobs University Bremen, 28725 Bremen, Germany, e-mail: r.ionutiu@jacobs-university.de

Joost Rommes
NXP Semiconductors, Corporate I&T/DTF, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands, e-mail: joost.rommes@nxp.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_60,

which is placed instead of the original in the simulation flow. Passive reduced circuits also guarantee stable simulations when integrated with the overall nonlinear macro-model [2–4] during later simulation stages.

The proposed *dominant spectral zero method (dominant SZM)* is a model reduction method which preserves passivity and stability, and is efficiently implemented using *the subspace accelerated dominant pole algorithm (SADPA)* [5,6]. Passivity preservation is ensured via a new approach, that of interpolation at *dominant spectral zeros*, a subset of spectral zeros of the original model. Dominant SZM reduces automatically all passive systems, including those with formulations unsuitable for PRIMA (first order susceptance-based models for inductive couplings (RCS circuits) [7] or models involving controlled sources, such as vector potential equivalent circuit (VPEC) [8] and partial element equivalent circuit (PEEC) models [9]). In comparison to *positive real balanced truncation (PRBT)* [10], dominant SZM efficiently handles systems with a possibly singular $\mathbf{E}$ matrix [see (1)]. Unlike *modal approximation (MA)* [5, 11] where interpolation is at dominant poles, our method matches the dominant spectral zeros of the original system, guaranteeing passivity.

The remainder of this article is structured as follows. The introduction continues with the mathematical setup of MOR in Sect. 1.1, and with a brief description of MOR via spectral zero interpolation in Sect. 1.2. Dominant SZM is presented concisely in Sect. 2.1 (following [12]). It is extended with the concept of dominance at $\infty$ (Sect. 2.2), and with an approach for converting the reduced models to circuit representations (Sect. 2.3). Numerical results follow in Sect. 3 and the paper concludes with Sect. 4.

### *1.1 Background on MOR*

The model reduction framework involves approximation of an original dynamical system described by a set of differential algebraic equations in the form:

$$\mathbf{E}\dot{\mathbf{x}}(t)=\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{u}(t),\ \mathbf{y}(t)=\mathbf{C}\mathbf{x}(t)+\mathbf{D}\mathbf{u}(t), \tag{1}$$

where the entries of $\mathbf{x}(t)$ are the system's internal variables, $\mathbf{u}(t)$ is the system input and $\mathbf{y}(t)$ is the system output, with dimensions $\mathbf{x}(t)\in\mathbb{R}^n$, $\mathbf{u}(t)\in\mathbb{R}^m$, $\mathbf{y}(t)\in\mathbb{R}^p$. Correspondingly, $\mathbf{E}\in\mathbb{R}^{n\times n}$, $\mathbf{A}\in\mathbb{R}^{n\times n}$, $(\mathbf{A},\mathbf{E})$ is a regular pencil, $\mathbf{B}\in\mathbb{R}^{n\times m}$, $\mathbf{C}\in\mathbb{R}^{p\times n}$, $\mathbf{D}\in\mathbb{R}^{p\times m}$. The original system $\Sigma(\mathbf{E},\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ is stable and passive and has *dimension n*, usually very large. We seek a reduced order model $\hat{\Sigma}(\hat{\mathbf{E}},\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}},\mathbf{D})$, which satisfies: $\hat{\mathbf{E}}\dot{\hat{\mathbf{x}}}(\mathbf{t})=\hat{\mathbf{A}}\hat{\mathbf{x}}(\mathbf{t})+\hat{\mathbf{B}}\mathbf{u}(\mathbf{t})$, $\hat{\mathbf{y}}(t)=\hat{\mathbf{C}}\hat{\mathbf{x}}(t)+\mathbf{D}\mathbf{u}(t)$, where $\hat{\mathbf{x}}\in\mathbb{R}^k$, $\hat{\mathbf{E}}\in\mathbb{R}^{k\times k}$, $\hat{\mathbf{A}}\in\mathbb{R}^{k\times k}$, $\hat{\mathbf{B}}\in\mathbb{R}^{k\times m}$, $\hat{\mathbf{C}}\in\mathbb{R}^{p\times k}$, $\mathbf{D}\in\mathbb{R}^{p\times m}$. $\widehat{\Sigma}$ is obtained by projecting the internal variables of the original system $\mathbf{x}$ onto a subspace *ColSpan* $\mathbf{V}\subset\mathbb{R}^{n\times k}$, along *Null* $\mathbf{W}^*\subset\mathbb{R}^{k\times n}$. The goal is to construct $\mathbf{V}$ and $\mathbf{W}$, such that $\widehat{\Sigma}$ is stable and passive. Additionally, $\mathbf{V}$ and $\mathbf{W}$ should be computed efficiently. The reduced matrices are obtained as follows:

$$\hat{\mathbf{E}}=\mathbf{W}^*\mathbf{E}\mathbf{V},\ \hat{\mathbf{A}}=\mathbf{W}^*\mathbf{A}\mathbf{V},\ \hat{\mathbf{B}}=\mathbf{W}^*\mathbf{B},\ \hat{\mathbf{C}}=\mathbf{C}\mathbf{V}. \tag{2}$$

### 1.2 MOR by Spectral Zero Interpolation

We revise the spectral zero interpolation approach for model reduction as proposed in [13, 14]. The ingredient for passivity preservation are the *spectral zeros* of $\Sigma(\mathbf{E},\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$, defined as follows:

**Definition 1.** For system $\Sigma$ with transfer function: $\mathbf{H}(s) := \mathbf{C}(s\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$, the spectral zeros are all $s\in\mathbb{C}$ such that $\mathbf{H}(s)+\mathbf{H}^*(-s)=0$, where $\mathbf{H}^*(-s)=\mathbf{B}^*(-s\mathbf{E}^*-\mathbf{A}^*)^{-1}\mathbf{C}^*+\mathbf{D}^*$.

According to [13, 14], model reduction via spectral zero interpolation involves forming rational Krylov subspaces:

$$\mathbf{V}=[(s_1\mathbf{E}-\mathbf{A})^{-1}\mathbf{B},\ \cdots,\ (s_k\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}],\ \ \mathbf{W}=[(-s_1^*\mathbf{E}^*-\mathbf{A}^*)^{-1}\mathbf{C}^*,\ \cdots,\ (-s_k^*\mathbf{E}^*-\mathbf{A}^*)^{-1}\mathbf{C}^*],\tag{3}$$

where $s_1\dots s_k, -s_1^*\dots -s_k^*$ are a subset of the spectral zeros of $\Sigma$. By projecting the original system with matrices (3) according to (2), the reduced $\hat{\Sigma}$ interpolates $\Sigma$ at the chosen $s_i$ and their mirror images $-s_i^*$, $i=1,\dots,k$ [1, 13]. Projection matrices $\mathbf{V}$ and $\mathbf{W}$ insure that the reduced system satisfies the positive real lemma [1, 13, 14], thus passivity is preserved. If in the original system $\mathbf{D}\neq\mathbf{0}$, the reduced system is strictly passive, and realizable with RLC circuit elements. In Sect. 2.2 we show one way of obtaining strictly passive reduced systems also when $\mathbf{D}=\mathbf{0}$.

## 2 The Dominant Spectral Zero Method

The new dominant spectral zero method (dominant SZM) is presented. The spectral zero method [13, 14] is extended with a dominance criterion for selecting finite spectral zeros. These are computed efficiently and automatically using the subspace accelerated dominant pole algorithm (SADPA) [5, 6]. We show in addition how, for certain RLC models, dominant spectral zeros at $\infty$ can also be easily interpolated.

### 2.1 Dominant Spectral Zeros and Implementation

In [14] it was shown that spectral zeros are solved efficiently from an associated Hamiltonian eigenvalue problem [15, 16]. In [13, 14] however, the selection of spectral zeros was still an open problem. We propose a solution as follows: we extend the concept of *dominance* from poles [6] to spectral zeros, and adapt the iterative solver SADPA for the computation of *dominant spectral zeros*. The corresponding invariant subspaces are obtained as a by-product of SADPA, and are used to construct the passivity preserving projection matrices $\mathbf{V}$ and $\mathbf{W}$. Essentially, dominant SZM is the SADPA-based implementation of modal approximation for the Hamiltonian system associated with $\mathbf{G}(s)=[\mathbf{H}(s)+\mathbf{H}^*(-s)]^{-1}$. Recalling Def. 1, the spectral zeros of $\Sigma$ are the poles of $\mathbf{G}(s)$, with partial fraction expansion: $\mathbf{G}(s)=\sum_{j=1}^{2n}\frac{\mathscr{R}_j}{s-s_j}$, where $s_i$ are the poles of $\mathbf{G}$ with associated residues $\mathscr{R}_j$ [5, 17]. The modal approximate of $\mathbf{G}(s)$ is obtained by truncating this sum: $\hat{\mathbf{G}}(s)=\sum_{j=1}^{2k}\frac{\mathscr{R}_j}{s-s_j}$. Dominant SZM together with the SADPA implementation is explained in detail in [12]. The procedure is outlined next.

1. Given $\Sigma(\mathbf{E},\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ with $\mathbf{D}=\mathbf{0}$, construct the associated Hamiltonian system[1] $\Sigma_s$, associated with transfer function $\mathbf{G}(s)$:
$$\mathbf{A}_s=\begin{pmatrix}\mathbf{A} & \mathbf{0} & \mathbf{B}\\ \mathbf{0} & -\mathbf{A}^* & -\mathbf{C}^*\\ \mathbf{C} & \mathbf{B}^* & \mathbf{0}\end{pmatrix},\ \mathbf{E}_s=\begin{pmatrix}\mathbf{E} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{E}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix},\ \mathbf{B}_s=\begin{pmatrix}\mathbf{B}\\ -\mathbf{C}^*\\ \mathbf{I}\end{pmatrix},\ \mathbf{C}_s=-\begin{pmatrix}\mathbf{C} & \mathbf{B}^* & \mathbf{I}\end{pmatrix} \tag{4}$$
2. Solve the Hamiltonian eigenvalue problem $(\Lambda,\mathbf{R},\mathbf{L})=\mathrm{eig}(\mathbf{A}_s,\mathbf{E}_s)$, i.e., $\mathbf{A}_s\mathbf{R}=\mathbf{E}_s\mathbf{R}\Lambda$, $\mathbf{L}^*\mathbf{A}_s=\Lambda\mathbf{L}^*\mathbf{E}_s$. $\mathbf{R}=[\mathbf{r}_1,\ldots,\mathbf{r}_{2n}]$, $\mathbf{L}=[\mathbf{l}_1,\ldots,\mathbf{l}_{2n}]$ and eigenvalues $\Lambda=\mathrm{diag}(s_1,\ldots,s_n,-s_1^*,\ldots,-s_n^*)$ are the spectral zeros of $\Sigma$.
3. Compute residues $\mathscr{R}_j$ associated with the stable[2] spectral zeros $s_j$, $j=1\ldots n$ as follows: $\mathscr{R}_j=\gamma_j\beta_j$, $\gamma_j=\mathbf{C}_s\mathbf{r}_j(\mathbf{l}_j^*\mathbf{E}_s\mathbf{r}_j)^{-1}$, $\beta_j=\mathbf{l}_j^*\mathbf{B}_s$.
4. Sort spectral zeros descendingly according to *dominance criterion* $\frac{\|\mathscr{R}_j\|}{|Re(s_j)|}$ [6, Chapter 3], and reorder right eigenvectors $\mathbf{R}$ accordingly.
5. Retain the right eigenspace $\widehat{\mathbf{R}}=[\mathbf{r}_1,\ \ldots,\ \mathbf{r}_k]\in\mathbb{C}^{2n\times k}$, corresponding to the stable $k$ most dominant spectral zeros.
6. Construct passivity projection matrices $\mathbf{V}$ and $\mathbf{W}$ from the rows of $\widehat{\mathbf{R}}$: $\mathbf{V}=\widehat{\mathbf{R}}_{[1:n,1:k]}$, $\mathbf{W}=\widehat{\mathbf{R}}_{[n+1:2n,1:k]}$, and reduce $\Sigma$ according to (2).

As explained in [12–14], by projecting with (2), $\widehat{\Sigma}$ interpolates the $k$ most dominant spectral zeros of $\Sigma$, guaranteeing passivity and stability. For large-scale applications, a full solution to the eigenvalue problem in step 2, followed by the dominant sort 3–4 is computationally unfeasible. Instead, the iterative solver SADPA [6, Chapter 3] is applied as explained in [12], with appropriate adaptations for spectral zero computation. SADPA implements steps 2–4 efficiently and automatically gives the $k$ most dominant spectral zeros and associated $2n\times k$ right eigenspace $\widehat{\mathbf{R}}$. The implementation requires performing an LU factorization of $(s_j\mathbf{E}-\mathbf{A})$ at each iteration. The relevant $s_j$ are nevertheless computed automatically in SADPA, which may have several advantages over other methods (see [12] for a more detailed cost analysis).

## 2.2 $\mathbf{D}=\mathbf{0}$ *and Dominance at* $s\to\infty$

Systems arising in circuit simulation often satisfy $\mathbf{D}=\mathbf{0}$ in (1). In this case, the projection (2), with $\mathbf{W}$ and $\mathbf{V}$ obtained in step 6 in Sect. 2.1, gives a lossless system [12]. This is because $\mathbf{W}$ and $\mathbf{V}$ only interpolate dominant finite spectral zeros, whereas the original system has spectral zeros at $\infty$, some of which may be dominant [18]. A strictly passive system (with all poles in the left half plane) can nevertheless be obtained by recovering this dominant behavior. For systems often occurring in circuit simulation this is achieved as follows. Consider the modified nodal analysis (MNA) description of an RLC circuit:

$$\underbrace{\begin{pmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathscr{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathscr{L}\end{pmatrix}}_{\mathbf{E}}\underbrace{\begin{pmatrix}\mathbf{v}_p\\ \mathbf{v}_i\\ \mathbf{i}_L\end{pmatrix}^{\!\cdot}}_{\dot{\mathbf{x}}}+\underbrace{\begin{pmatrix}\mathscr{G}_{\mathbf{11}} & \mathscr{G}_{\mathbf{12}} & \mathscr{E}_{\mathbf{1}}\\ \mathscr{G}_{\mathbf{12}}^* & \mathscr{G}_{\mathbf{22}} & \mathscr{E}_{\mathbf{2}}\\ -\mathscr{E}_{\mathbf{1}}^* & -\mathscr{E}_{\mathbf{2}}^* & \mathbf{0}\end{pmatrix}}_{-\mathbf{A}}\underbrace{\begin{pmatrix}\mathbf{v}_p\\ \mathbf{v}_i\\ \mathbf{i}_L\end{pmatrix}}_{\mathbf{x}}=\underbrace{\begin{pmatrix}\mathscr{B}_{\mathbf{1}}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}}_{\mathbf{B}}\mathbf{u}, \tag{5}$$

[1] For $\mathbf{D}\neq\mathbf{0}$ see [12] for the form of the Hamiltonian system; the algorithm follows as for $\mathbf{D}=\mathbf{0}$.

[2] $s\in\mathbb{C}$ is stable if $Re(s)<0$.

where $\mathbf{u}(t) \in \mathbb{R}^m$ are input currents and $\mathbf{y}(t) = \mathbf{C}\mathbf{x} \in \mathbb{R}^m$ are output voltages, $\mathbf{C} = \mathbf{B}^*$. The states are $\mathbf{x}(t) = [\mathbf{v}_p(t),\ \mathbf{v}_i(t),\ \mathbf{i}_L(t)]^T$, with $\mathbf{v}_p(t) \in \mathbb{R}^{n_p}$ the voltages at the input nodes (circuit terminals), $\mathbf{v}_i(t) \in \mathbb{R}^{n_i}$ the voltages at the internal nodes, and $\mathbf{i}_L(t) \in \mathbb{R}^{n_{i_L}}$ the currents through the inductors, $n_p + n_i + n_{i_L} = n$. $\mathscr{C}$ and $\mathscr{L}$ are the capacitor and inductor matrix stamps respectively. With (5) it is assumed that no capacitors or inductors are directly connected to the input nodes, thus $\mathbf{B} \in Null(\mathbf{E})$ and $\mathbf{C}^* \in Null(\mathbf{E}^*)$. As $\mathbf{B}$ and $\mathbf{C}$ are right and left eigenvectors corresponding to dominant poles (and spectral zeros) at $\infty$ [18], the modified projection matrices are:

$$\widetilde{\mathbf{W}} = [\mathbf{W}, \mathbf{C}^*],\ \widetilde{\mathbf{V}} = [\mathbf{W}, \mathbf{B}], \tag{6}$$

where $\mathbf{W}$ and $\mathbf{V}$ are obtained from step 6 in Sect. 2.1. With (6), the finite dominant spectral zeros are interpolated as well as the dominant spectral zeros at $\infty$, and the reduced system is strictly passive [18]. In [12] two alternatives were proposed for ensuring strict passivity for systems in the more general form (1) with $\mathbf{D} = \mathbf{0}$.

### 2.3 Circuit Representation of Reduced Impedance Transfer Function

Reduced models obtained with dominant SZM and other Krylov-type methods (PRIMA [2], SPRIM [3, 19], SPRIM/IOPOR [20, 21]) are mathematical abstractions of an underlying small RLC circuit. Circuit simulators however can only handle mathematical representations to a limited extent, and reduced models have to be synthesized with RLC circuit elements. We reduce all circuits with respect to the input impedance transfer function (i.e., the inputs are the currents injected into the circuit terminals and the outputs are the voltages measured at the terminals). After converting the reduced input impedance transfer function to netlist format, the reduced circuit can be driven easily by currents or voltages when simulated. Thus both the input impedance and admittance of an original model can be reproduced (see Sect. 3). Here, models obtained with dominant SZM are converted to netlist representations using the Foster impedance realization approach [22, 23]. Netlist formats for the SPRIM/IOPOR [3, 20, 21] reduced models are obtained via the RLCSYN unstamping procedure in [20]. With both approaches, the resulting netlists may still contain circuit elements with negative values, nevertheless this does not impede the circuit simulation. Obtaining realistic synthesized models with positive circuit elements only is still an open problem.

## 3 Numerical Results

Two transmission line models are reduced with the proposed dominant spectral zero method and compared with the input-output structure preserving method SPRIM/IOPOR [3, 20, 21]. For both circuits, the circuit simulators[3] yield systems in the form (5), thus the dominant SZM projection is (6). RLC netlist representations for the reduced models are obtained (see Sect. 2.3) and simulated with Pstar.

[3] Pstar and Hstar in-house simulators at NXP Semiconductors

The RLC transmission line with connected voltage controlled current sources (VCCSs) from [12] is reduced with dominant SZM, SPRIM/IOPOR [3, 20] and modal approximation (MA). The transfer function is an input impedance i.e., the circuit is current driven. Matlab simulations of the original and reduced models, as well as the Pstar netlist simulations are shown in Fig. 1: the model reduced with Dominant SZM gives the best approximation. Table 1 summarizes the reduction: the number of circuit elements and the number of states were reduced significantly and the simulation time was sped up.

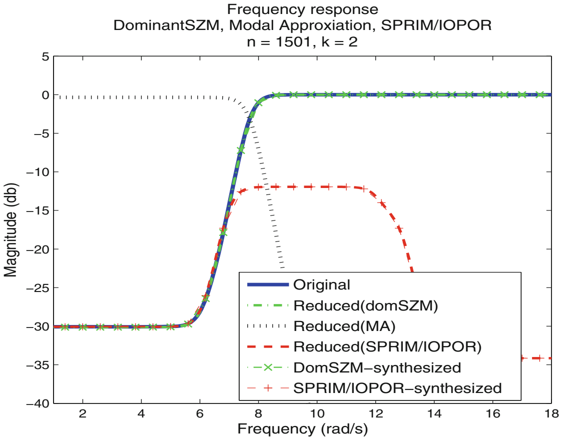


**Fig. 1:** Original, reduced and synthesized systems: Dominant SZM, SPRIM/IOPOR

**Table 1:** Transmission line with VCCSs: reduction and synthesis summary

| System | Dimension | R | C | L | VCCs | States | Simulation time |
|---|---|---|---|---|---|---|---|
| Original | 1501 | 1001 | 500 | 500 | 500 | 1500 | 0.5 $s$ |
| Dominant SZM | 2 | 3 | 2 | 0 | - | 4 | 0.01 $s$ |
| SPRIM/IOPOR | 2 | 6 | 3 | 1 | - | 4 | 0.01 $s$ |

In [12], the voltage driven *input admittance* of an RLC transmission line (consisting of cascaded RLC blocks) was reduced directly as shown in Fig. 3. Here we reduce and synthesize the underlying *input impedance* of the same transmission line (see Figures 2 and 4). When driving the reduced netlist by an input voltage during the actual circuit simulation, the same input admittance is obtained as if the input admittance had been reduced directly, as seen in Figures 3 and 5. Table 2 summarizes the reduction results. Although the reduced mathematical models have the same dimension ($k = 23$), the reduction effect can only be determined after obtaining the netlist representations. Although the SPRIM/IOPOR synthesized model has fewer states, it has more circuit elements than the dominant SZM model, i.e., the matrix stamp of the model is more dense. This suggests that simulation time is jointly determined by the number of states and the number of circuit elements. Thus for practical purposes it is critical to synthesize reduced models with RLC components.

**Table 2:** RLC transmission line: Input impedance reduction and synthesis summary

| System | Dimension | R | C | L | States | Simulation time |
|---|---|---|---|---|---|---|
| Original | 901 | 500 | 300 | 300 | 901 | 1.5 $s$ |
| Dominant SZM | 23 | 22 | 11 | 10 | 34 | 0.02 $s$ |
| SPRIM/IOPOR | 23 | 78 | 66 | 6 | 18 | 0.02 $s$ |

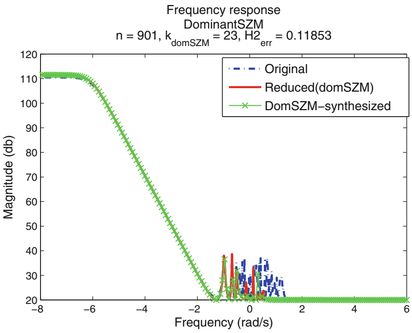


**Fig. 2:** Input impedance transfer function: original and reduced with Dominant SZM

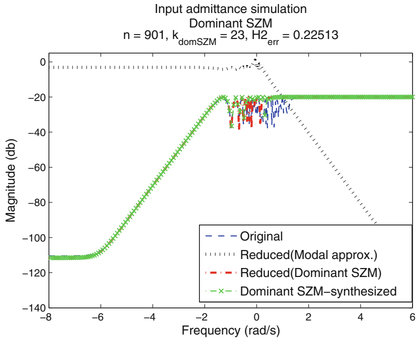


**Fig. 3:** Input admittance transfer function: original, synthesized Dominant SZM model

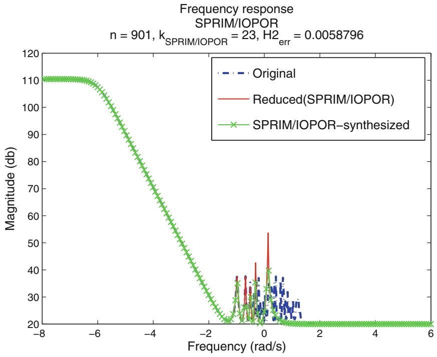


**Fig. 4:** Input impedance transfer function: original, reduced with SPRIM/IOPOR

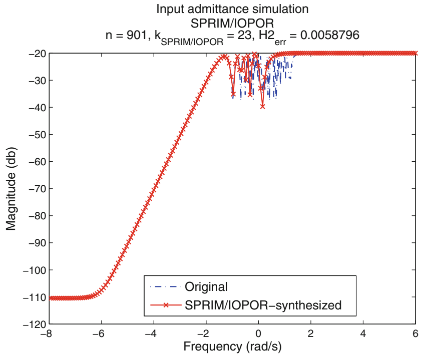


**Fig. 5:** Input admittance transfer function: original, synthesized SPRIM/IOPOR model

## 4 Concluding Remarks

A novel passivity preserving model reduction method is presented, which is based on interpolation of dominant spectral zeros. Implemented with the SADPA iterative solver, the method solves approximately an associated Hamiltonian eigenvalue problem, and constructs the passivity preserving projection. Netlist equivalents for the reduced models are simulated and directions for future work are revealed. Especially in model reduction of multi-terminal circuits, achieving structure preservation, sparsity and small dimensionality simultaneously is an open question. In this context, RLC synthesis with positive circuit elements will also be addressed.

**Acknowledgements** The first and third author thank the NSF Grants CCR-0306503, ACI-0325081, and CCF-0634902. The second author was supported by the O-MOORE-NICE! project, Marie Curie Actions of FP6, MTKI-CT-2006-042477. The first author also acknowledges the

Marie Curie Fellowship Programme, COMSON project MRTN-CT-2005-019417, for the work within TU/Eindhoven and NXP Semiconductors, Eindhoven, the Netherlands.

## References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Phil., PA (2005)
2. Odabasioglu, A., Celik, M., Pillegi, L.: Prima: Passive reduced-order interconnect macromodelling algorithm. IEEE Trans. CAD Circ. Syst. **17**, 645–654 (1998)
3. Freund, R.: Sprim: Structure-preserving reduced-order interconnect macromodeling. In: Proc. IEEE/ACM Int. Conf. on Comp. Aided Design, pp. 80–87. Los Alamitos, CA (2004)
4. Sheldon X. D. Tan, L.H.: Advanced Model Order Reduction Techniques in VLSI design. Cambridge University Press (2007)
5. Rommes, J., Martins, N.: Efficient computation of transfer function dominant poles using subspace acceleration. IEEE Trans. Power Syst. **21**(3), 1218–1226 (2006)
6. Rommes, J.: Methods for eigenvalue problems with applications in model order reduction. Ph.D. thesis, Utrecht University, Utrecht, The Netherlands (2007). URL http://rommes.googlepages.com/index.html
7. Zheng, H., Pileggi, L.: Robust and passive model order reduction for circuits containing susceptance elements. In: Proc. IEEE/ACM International Conference on Computer Aided Design ICCAD 2002, pp. 761–766 (2002)
8. You, H., He, L.: A sparsified vector potential equivalent circuit model for massively coupled interconnects. In: IEEE Intl. Symposium on Circuits and Systems, vol. 1, pp. 105–108 (2005)
9. Verbeek, M.E.: Partial element equivalent circuit (PEEC) models for on-chip passives and interconnects. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields **17**(1), 61–84 (2004)
10. Phillips, J., Daniel, L., Silveira, L.: Guaranteed passive balancing transformations for model order reduction. IEEE Trans. CAD Circ. Syst. **22**(8), 1027–1041 (2003)
11. Varga, A.: Enhanced modal approach for model reduction. Mathematical Modelling of Systems **1**, 91–105 (1995)
12. Ionutiu, R., Rommes, J., Antoulas, A.: Passivity preserving model reduction using dominant spectral zero interpolation. IEEE Trans. CAD Circ. Syst. **27**(12), 2250–2263 (2008)
13. Antoulas, A.C.: A new result on passivity preserving model reduction. Systems and Control Letters **54**, 361–374 (2005)
14. Sorensen, D.: Passivity preserving model reduction via interpolation of spectral zeros. Systems and Control Letters **54**, 347–360 (2005)
15. Kressner, D.: Numerical methods for general and structured eigenvalue problems. In: T. Barth, M.Griebel, D. Keyes, R. Nieminen, D. Roose, T.Schlick (eds.) Lecture Notes in Computational Science and Engineering. Springer (2005)
16. Watkins, D.S.: The Matrix Eigenvalue Problem: GR and Krylov subspace methods. SIAM (2007)
17. Kailath, T.: Linear Systems. Prentice-Hall (1980)
18. Ionutiu, R.: Passivity preserving model reduction in the context of spectral zero interpolation. Master's thesis, William Marsh Rice University, Houston, TX, USA (2008)
19. Freund, R.: Structure preserving model order reduction of RCL circuit equations. In: W. Schilders, H. van der Vorst, J. Rommes (eds.) Model Order Reduction, Theory, Research Aspects and Applications, *Mathematics in Industry*, vol. 13. Springer, Berlin, Germany (2008)
20. Yang, F., Zeng, X., Su, Y., Zhou, D.: RLC equivalent circuit synthesis method for structure-preserved reduced-order model of interconnect in VLSI. Commun. Comput. Phys **3**(2), 376–396 (2008)
21. Z. Bai R. Li, Y.S.: A unified Krylov projection framework for structure-preserving model reduction. In: W. Schilders, H. van der Vorst, J. Rommes (eds.) Model order reduction, *Mathematics in Industry*, vol. 11. Springer, Berlin, Germany (2008)
22. Guillemin, E.A.: Synthesis of passive networks, 2 edn. John Wiley (1959)
23. Ionutiu, R., Rommes, J.: Circuit synthesis of reduced order models. NXP-TN 2008/00316, NXP Semiconductors (2008)

# Applications of Eigenvalue Counting and Inclusion Theorems in Model Order Reduction

E. Fatih Yetkin and Hasan Dağ

**Abstract** We suggest a simple and an efficient iterative method based on both the Gerschgorin eigenvalue inclusion theorem and the deflation methods to compute a Reduced Order Model (ROM) to lower greatly the order of a given state space system. This method is especially efficient in symmetric state-space systems but it works for the other cases with some modifications.

## 1 Introduction

The computational cost of the simulation of today's technological equipments, especially those of the integrated circuits, can be very high. Hence the model order reduction methods have very wide application areas especially in sub-micron electronic device and microelectronic mechanical system (MEMS) modeling and simulation [1]. In general, a single input single output (SISO) system can be defined with the state equations in the standard form as below.

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + \mathbf{b}u \\ y &= \mathbf{c}^T\mathbf{x} + du \end{aligned} \tag{1}$$

where $A \in \mathscr{R}^{nxn}$, $\mathbf{b}, \mathbf{c} \in \mathscr{R}^n$ and $d$ is a scalar. Here, the dimension of the state space $n$ is very large and the model order reduction techniques are employed to build an $m^{\text{th}}$ order system where $m \ll n$. The reduced system can be given as below,

E. Fatih Yetkin
Computational Science and Engineering Program, Informatics Institute, Istanbul Technical University, Istanbul, Turkey, e-mail: fatih@be.itu.edu.tr

Hasan Dağ
Information Technologies Department, Kadir Has University, Istanbul, Turkey, e-mail: hasan.dag@khas.edu.tr

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_61,

$$\begin{aligned}\dot{\mathbf{z}} &= \hat{A}\mathbf{z} + \hat{\mathbf{b}}u \\ y &= \hat{\mathbf{c}}^T\mathbf{z} + du\end{aligned} \tag{2}$$

where $\hat{A} \in \mathscr{R}^{mxm}$, $\hat{\mathbf{b}}, \hat{\mathbf{c}} \in \mathscr{R}^m$ and $d$ is a scalar [2]. Use of the dominant eigenmodes is one of the ways to build a reduced state space model of the system under study. However, it is not computationally feasible due to high cost of computing all eigenmodes of the system at hand [3]. Therefore, computing only the dominant poles of a transfer matrix can also be an effective way for computing the reduced system matrices [4].

In this study, we suggest a new method based on the eigenvalue inclusion theorems such as Gerschgorin's. Gerschgorin discs are very useful and there are quite a few computationally efficient methods to determine the area of possible eigenvalue locations of a given matrix.

In our approach, a possible eigenvalue location and the number of dominant eigenvalues of interest are determined using Gerschgorin theorem automatically. Then, this number is selected as maximum iteration number for the algorithm. In each step of the algorithm the required number of eigenpairs are computed by using any eigenvalue deflation algorithms till the loop reaches the maximum iteration number or required error tolerance. The method works for symmetric systems. But one can use the Sturm sequences with the Wilf method to apply the algorithm to general systems [5].

The remaining of the paper is organized as follows. In the second section, the eigenvalue inclusion theorems are briefly introduced. In the third section, we present the suggested method. Some numerical examples are given in the fourth section. The conclusions and the future work are presented in the last section.

## 2 Eigenvalue Inclusion Theorems

### 2.1 Gerschgorin Theorem

Computing all eigenvalues of a matrix is not easy in most cases. In such cases one can use Gerschgorin's method to estimate the eigenvalues. Let $A$ be an $n \times n$ matrix and $a_{ij}$'s be its entries. One can define $C_i$ disks in the complex plane whose centers are values of diagonal entries of matrix $A$ as follows:

$$C_i = \{z \in C| \quad |z - a_{ii}| \leq R_{ii}\} \tag{3}$$

$$R_i = \sum_{j=1}^{n} |a_{ij}| \qquad i \neq j. \tag{4}$$

Here, the radius is the row sum in absolute values except the diagonal element.

**Gerschgorin Theorem:** All eigenvalues of the $A$ have to be in the union of these $C_i$ discs.

$$C = \bigcup_{i=1}^{n} C_i \tag{5}$$

Another important part of the Gerschgorin's theorem is that if $m$ of those disks do not touch to the remaining disks, then there exist exactly $m$ eigenvalues in these $m$ disks [6]. The proof of the theorem can be found in the literature such as [7].

### 2.2 *Modal Approximation*

Let us consider the standard state space system given in (1). The matrix $A$ can be expressed in a different way by using its eigenpairs

$$A = E\Lambda E^{-1}, \tag{6}$$

where $E$ contains eigenvectors of $A$ matrix in its columns and $\Lambda$ is a diagonal matrix containing the eigenvalues of $A$ matrix in its diagonal entries. If $Ew(t) = x(t)$ change of variables is applied then (1) can be written as follows:

$$\begin{aligned} \frac{\mathrm{d}w(t)}{\mathrm{d}t} &= E^{-1}AEw(t) + E^{-1}bu(t) \\ y(t) &= c^T Ew(t). \end{aligned} \tag{7}$$

The transfer function of the system in (7 and its pole-residue representation can be written as,

$$H(s) = \underbrace{c^T E}_{\hat{c}}(sI - \underbrace{\Lambda}_{\hat{A}})\underbrace{E^{-1}b}_{\hat{b}} \qquad H(s) = \sum_{i=1}^{N} \frac{\hat{b}_i \hat{c}_i}{s - \lambda_i} \tag{8}$$

where $\hat{A} = \Lambda$, $\hat{b} = E^{-1}b$ and $\hat{c} = c^T E$. In modal approximation methods, the terms having small residues and the terms having large negative part for $\mathrm{Re}(\lambda_i)$ are dropped from the pole-residue formulation [8]. Although the modal approximation method is conceptually familiar, it requires huge computational effort because of full eigen-decomposition.

## 3 Method and Algorithm

In our algorithm, modal approximation based method is used iteratively, but in order to avoid full eigen-decomposition the Gerschgorin theorem is used. Clusters of the Gerschgorin discs are determined first. Then the cluster, which is located nearest position to the $\mathrm{j}\omega$ axis, is selected. The number of the eigenvalues in it can be used as a limit for the iteration. In each step of the iteration, a deflation algorithm (Wielandt, etc.) is used to find the related eigenpairs of $A$ matrix [9]. For example, first $i$ eigenpairs are computed and the reduced system matrices are built by modal approximation approach in $i^{\text{th}}$ step. Then the difference between the standard Euclidian norm of the frequency response of $(i-1)^{\text{th}}$ order transfer function and the $i^{\text{th}}$ order one is computed. Until the difference of norms is smaller than the set tolerance value or the iteration number reaches its limit, iteration does not stop.

In the first step of the algorithm, the number of separated Gerschgorin discs is determined. To obtain this information, one can build a $n \times 2$ dimensional $\mathscr{T}$ matrix. Entries of the first column of $\mathscr{T}$ matrix are the center coordinates of Gerschgorin discs ($c_i$) and the entries of second one are the radius of the discs ($r_i$). Then the rows of the $\mathscr{T}$ matrix have to be reordered in the increasing order according to $c_i$. After reordering the $\mathscr{T}$ matrix, it has the information about the Gerschgorin discs from the most negative one to the most positive one. Then one can create a $\mathscr{P}$ matrix consisting of only 0 and 1 entries according to below relation

$$\mathscr{P}_{ij} = \begin{cases} 1 \text{ if} & c_i + r_i > c_j + r_j \\ 0 \text{ if} & c_i + r_i \leq c_j + r_j. \end{cases} \tag{9}$$

Because of the symmetry of the relation given in (9) we can say that, $\mathscr{P}_{ij} = \mathscr{P}_{ji}$. The rank of the $\mathscr{P}$ matrix is equal to the number of separated Gerschgorin discs. Moreover, the structure of the block matrices on the diagonal gives the number of the eigenvalues.

To illustrate this approach, one can find the matrix $\mathscr{P}$ from the given $A$

$$\mathbf{A} = \begin{bmatrix} -4.0 & 0.0 & 0.0 & -0.5 & -0.5 \\ -0.1 & -4.5 & -0.4 & 0.0 & -1.0 \\ 0.0 & -0.03 & -0.1 & -0.02 & -0.05 \\ -0.3 & 0.0 & -0.1 & -4.0 & -0.2 \\ -0.01 & -0.04 & 0.0 & 0.0 & -0.1 \end{bmatrix} \qquad \mathscr{P} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{10}$$

As we can see from the (10) matrices $\text{rank}(\mathscr{P}) = 2$ and then we can say that Gerschgorin discs of the $A$ matrix clustered in two different region of the complex plane. Another information arising from the $\mathscr{P}$ matrix, first disc cluster has three eigenvalues and second one has two eigenvalues. The Gerschgorin discs of the matrix $A$ are given in Fig.1.

The algorithm of the proposed method is given in Alg. 1. In the case of the $\text{rank}(\mathscr{P}) = 1$, the algorithm works up to convergence.

If $\mathscr{P}$ matrix is a rank-one matrix, the Gerschgorin based method does not give any information about the number of dominant eigenvalues of the $A$ matrix. In that case, more specific methods like modified Sturm sequences method or Gleyse-Jo method have to be used. The number of the eigenvalues in a specific geometric shape can be determined in these kinds of methods [10, 11].

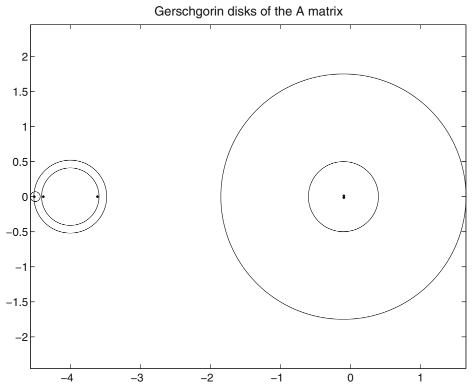


**Fig. 1:** Gerschgorin discs of $A$ matrix

**Algorithm 5**

**Require:** $A$, $b$, $c$ system matrices.
**Ensure:** Reduced system matrices $\hat{A}$, $\hat{b}$, $\hat{c}$.

```
Find the $\mathscr{P}$ matrix from $A$ to determine the number of Gerschgorin disc clusters.
Find the number of eigenvalues of the Gerschgorin disc cluster closest to the j$\omega$ axis and assign it as itnum.
for $r \leq itnum$ do
    Find first r eigenpairs with using any deflation method.
    Build r-dimensional system matrices using (8).
    if $||H_r - H_{r-1}||_2 < TOLERANCE$ then
        Give outputs as $\hat{A}$, $\hat{b}$ and $\hat{c}$ and exit.
    else
        r=r+1
    end if
end for
```

## 4 Numerical Applications

In this work we used two different type of benchmark examples. First one is a symmetric state-space model of a spiral inductor obtained by PEEC (Partial Element Equivalent Circuit) method . This inductor is intended as an integrated RF passive inductor first. For detailed information about technical properties and mathematical modelling of inductor we refer the reader to [12] and [13]. Original dimension of the system equals to 1434. But in our tests size of the system is reduced to 50 first. The spiral model has a symmetric structure. Hence its eigenvalues are all real.

If we apply the Gerschgorin method, we can find that $\text{rank}(\mathscr{P}) = 1$; thus the algorithm will work till it converges. We use the below relative error definition for the convergence criterion.

$$e_b = \frac{||H_N||_2 - ||H_n||_2}{||H_N||_2} \tag{11}$$

where $H_N$ is the original system frequency response and $H_n$ is the reduced system frequency response.

Frequency dependent resistance and inductance graphics of the spiral inductor for different reduction order is given in Fig.2.

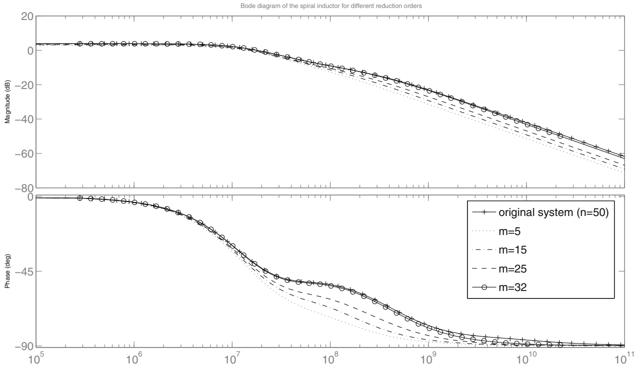


**Fig. 2:** Bode diagram of the spiral inductor example for different reduction order

An equivalent circuit of a simple interconnect geometry is used for the second numerical example. Two inductively coupled transmission line systems are selected as the example circuit. For the mathematical and electrical details of the example we refer the reader to [14]. The Bode diagram for different reduction order is given in Fig.3.

The convergence graphics of the method for both examples are given in Fig.4 and Fig. 5. In the spiral inductor example, relative error has an decreasing affinity as expected. It is obvious that if the number of the estimated eigenpairs is increased, the error will be reduced. On the other hand in the second example, the system has complex poles and the proposed method is not successful enough to estimate complex poles of the system. Although the method can achieve some accuracy, it is not robust. This is due to the complex eigenvalues of the $A$ matrix and the rank-one structure of the related $\mathscr{P}$ matrix. We have some suggestions to solve this problem, although, our suggestions are not conclusive yet. We plan to improve our method using more sophisticated algorithms like the Wilf method. That is, the method is expected behave the same as that of symmetric cases in the case of system matrices with complex eigenpairs.

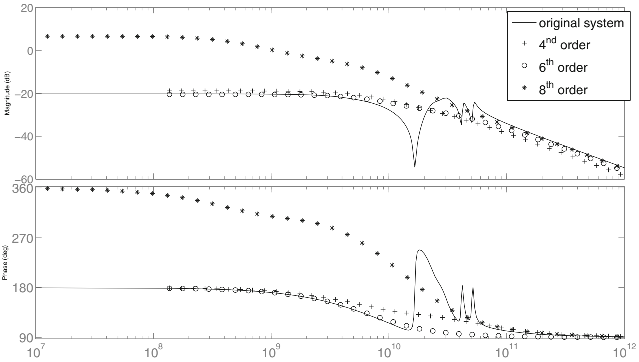


**Fig. 3:** Bode diagram of the transmission line example for different reduction order

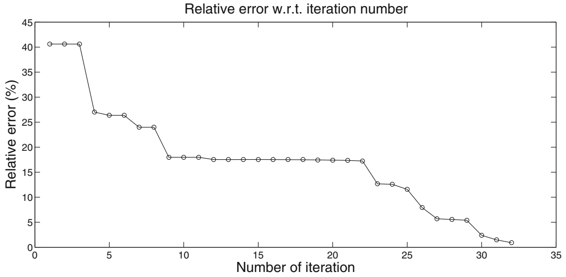


**Fig. 4:** Relative error of the method with respect to iteration number (also the dimension of the reduced system) for the spiral inductor example

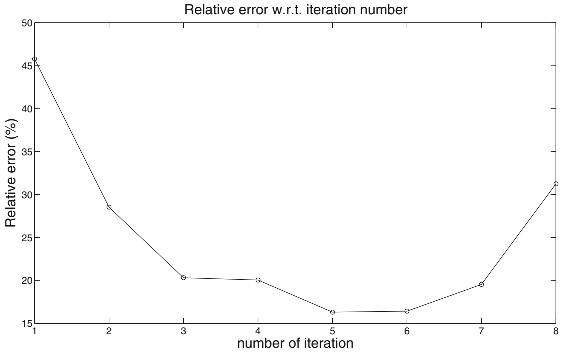


**Fig. 5:** Relative error of the method with respect to iteration number (also the dimension of the reduced system) for the transmission line example

## 5 Conclusions and Future Work

In this study, dominant poles of the linear state space system are estimated by using the Gerschgorin theorem and the Wielandt deflation algorithm in an iterative way. The method gives very accurate results especially for multi-time systems (means that the Gerschgorin discs of the **A** matrix clustered in different locations in complex plane or rank$(\mathscr{P}) \neq 1$). However, if all eigenvalues of system are within the same part of the complex plane and Gerschgorin discs are nested (means that rank$(\mathscr{P}) = 1$), the method becomes unstable. There is another pitfall in the suggested algorithm. If the system matrix $A$ has several complex eigenvalues, Wielandt (or another) deflation methods are not be sufficient to find these complex eigenvalues. In this case, the exact number of real and complex eigenvalues has to be known to get some satisfactory results from the algorithm. Therefore, a way has to be found to combine the Sturm sequences with the generalized bisection algorithm to get quick information about the approximate eigenvalues of the non-symmetric general system. Future work will be focused on these concepts.

## References

1. Meijs, N.P. , Smedes, T.: Accurate Interconnect Modeling: Towards Multi-million Transistor Chips As Microwave Circuits. *Int. Conf. On CAD, Proc. of ICCAD'96* p. 244-251, 1996.
2. Tan, S. X. D, L. He: Advanced Model Order Reduction Techniques in VLSI Design, Cambridge University Press, Cambridge, 2007.
3. Antoluas, A.C.: Approximation of Large-Scale Dynamical Systems, SIAM, Philadelphia, 2005.
4. Rommes, J.: Methods for Eigenvalue Problems with Applications in Model Order Reduction, Ph.D. Thesis, Univ. of Utrecht, 2007.
5. Wilf, H. S. : A Global Bisection Algorithm for computing the Zeros of Polynomials in the Complex Plane, J. Assoc. Comput. Mach. vol:25, p. 415-420, 1978.
6. Varga, R. S.: Gerschgorin-Type Eigenvalue Inclusion Theorems and Their Sharpness, Electronic Transactions on Numerical Analysis, Vol:12, p.113-133, 2001.
7. Varga, R. S.: Gerschgorin and His Circles, Springer, 2004.
8. Ogata, K.: Modern Control Engineering, Upper Saddle River, NJ, Prentice Hall, 2002.
9. Saad, Y.: Numerical Methods for Large Eigenvalue Problems, Manchester University Press, Manchester, UK, 1992.
10. Gleyse, B., Moflih, M.: Exact Computation of the Number of Zeros of a Real Polynomial in the Open Unit Disk by a Determinant Representation, Comput. Math. Appl., vol:38, p:257-263, 1999.
11. Jo, J. S., Jung, H.S., Ko, M. G., Lee, I. W.: Eigenvalue-counting Methods for Non-proportionally Damped Systems, Int. J. of Solids and Structures, Vol:40, p: 6457-6472, 2003.
12. Kamon, M., Tsuk, M. J., White, J.: Fasthenry: A Multipole-accelerated 3-D Inductance Extraction Program, IEEE Trans. on Microwave Theory and Techniques, vol:42(9), p:1750-1758, 1994.
13. Kamon, M., Wang, F., White, J.: Generating Nearly Optimal Compact Models from Krylov Subspace Based Reduced Order Models, IEEE Trans. on Circuits and Systems-II, vol:47(4), p:239-248, 2000.
14. Ionutiu, R., Lefteriu S., Antoluas, A. C.: Comparison of Model Reduction Methods with Application to Circuit Simulation, in: G. Ciuprina, D. Ioan (Eds.): Scientific Computing in Electrical Engineering, Series Mathematics in Industry, vol.11, Springer, p. 3-24, 2007.

# GABOR: Global-Approximation-Based Order Reduction

Janne Roos, Mikko Honkala, and Pekka Miettinen

**Abstract** This paper proposes a new approach for the Model-Order Reduction (MOR) of RLC circuits: Global-Approximation-Based Order Reduction (GABOR). GABOR preserves passivity and reciprocity, and matches the 'moments' of the underlying global approximation. However, GABOR has some problematic features, too. First, many matrices must be recursively precomputed into the memory space. Second, it is difficult to circumvent the singularity of the conductance matrix by any conventional frequency shifting. On the other hand, some tryouts for solving the second problem lead to finding interesting links between GABOR and other MOR methods. The correct operation of GABOR is verified with a simulation example.

## 1 Introduction

Typical Krylov subspace Model-Order Reduction (MOR) methods [1–3] are based on implicit moment matching. This approach results in a Taylor-series-like approximation, which is exact at the expansion point (e.g., at the origin of the complex plane), but which looses accuracy when moving far away (e.g., towards high frequencies). Therefore, one avenue to decrease (or, at least, to spread more equally) the approximation error could be to base the MOR on a global approximation. The RLC MOR method proposed in this paper, GABOR, is based on this idea.

PRIMA [1] operates with Y-parameters; SPRIM [2] and ENOR [3], in turn, are formulated using Z-parameters. Since circuit simulators use the Modified Nodal Analysis (MNA), Y-parameters are better suited for the (SPICE-netlist) synthesis of the reduced-order model [4]. However, for simplicity, we limit the discussion in this paper to the (frequency-domain) treatment of Z-parameters, only.

Janne Roos, Mikko Honkala, Pekka Miettinen
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland, e-mail: janne.roos@tkk.fi, mikko.a.honkala@tkk.fi, pekka.miettinen@tkk.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_62,

## 2 Derivation of GABOR

Let us consider an RLC circuit with $n = n_\mathrm{i} + N$ nodes, where $n_\mathrm{i}$ and $N$ are the number of internal nodes and external port nodes, respectively; typically, $N \ll n_\mathrm{i}$. Let the RLC circuit be excited by $N$ current sources for obtaining the Z-parameter matrix $\mathbf{Z}(s)$. Applying plain nodal analysis (which excludes voltage sources such that the $n$ node voltages are the only unknowns needed) to an RLC circuit, we obtain

$$\begin{aligned}\mathbf{Z}(s) &= \mathbf{L}^\mathrm{T}\left(\mathbf{G} + s\mathbf{C} + \frac{1}{s}\Gamma\right)^{-1}\mathbf{B}\\ &= \mathbf{L}^\mathrm{T}\left(\mathbf{I} + s\mathbf{G}^{-1}\mathbf{C} + \frac{1}{s}\mathbf{G}^{-1}\Gamma\right)^{-1}\mathbf{G}^{-1}\mathbf{B}\\ &\triangleq \mathbf{L}^\mathrm{T}\left(\mathbf{I} - s\mathbf{D} - \frac{1}{s}\mathbf{E}\right)^{-1}\mathbf{R}\end{aligned} \tag{1}$$

where $\mathbf{G}$, $\mathbf{C}$, and $\Gamma$ are $n$-by-$n$ symmetric semidefinite conductance, capacitance, and inverse-inductance matrices [3], respectively, and $\mathbf{L} = \mathbf{B}$ is an $n$-by-$N$ selector matrix. Also, we have denoted $\mathbf{D} = -\mathbf{G}^{-1}\mathbf{C}$, $\mathbf{E} = -\mathbf{G}^{-1}\Gamma$, and $\mathbf{R} = \mathbf{G}^{-1}\mathbf{B}$.

The idea behind GABOR is to approximate $\mathbf{Z}(s)$ in (1) by a (Laurent-series like) two-sided 'moment' series:

$$\begin{aligned}\mathbf{Z}(s) &= \mathbf{L}^\mathrm{T}\left(\cdots + \mathbf{N}_{-2}\frac{1}{s^2} + \mathbf{N}_{-1}\frac{1}{s} + \mathbf{N}_0 + \mathbf{N}_1 s + \mathbf{N}_2 s^2 + \dots\right)\\ &= \cdots + \mathbf{M}_{-2}\frac{1}{s^2} + \mathbf{M}_{-1}\frac{1}{s} + \mathbf{M}_0 + \mathbf{M}_1 s + \mathbf{M}_2 s^2 + \dots\end{aligned} \tag{2}$$

where the $N$-by-$N$ 'moments', $\mathbf{M}_i$, are related to the $n$-by-$N$ matrices, $\mathbf{N}_i$, by

$$\mathbf{M}_i = \mathbf{L}^\mathrm{T}\mathbf{N}_i, \;\; i = \dots, -2, -1, 0, 1, 2, \dots \tag{3}$$

Combining the last row of (1) and the first row of (2) and some algebra gives

$$\left(\mathbf{I} - s\mathbf{D} - \frac{1}{s}\mathbf{E}\right)\left(\cdots + \mathbf{N}_{-2}\frac{1}{s^2} + \mathbf{N}_{-1}\frac{1}{s} + \mathbf{N}_0 + \mathbf{N}_1 s + \mathbf{N}_2 s^2 \dots\right) = \mathbf{R} \tag{4}$$

Equating the negative/zero/positive powers of $s$ results in the following matrix equation (truncated to ±3 terms for notational convenience):

$$\begin{bmatrix} \mathbf{I} & -\mathbf{D} & & & & & \\ -\mathbf{E} & \mathbf{I} & -\mathbf{D} & & & & \\ & -\mathbf{E} & \mathbf{I} & -\mathbf{D} & & & \\ & & -\mathbf{E} & \mathbf{I} & -\mathbf{D} & & \\ & & & -\mathbf{E} & \mathbf{I} & -\mathbf{D} & \\ & & & & -\mathbf{E} & \mathbf{I} & -\mathbf{D} \\ & & & & & -\mathbf{E} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{N}_3 \\ \mathbf{N}_2 \\ \mathbf{N}_1 \\ \mathbf{N}_0 \\ \mathbf{N}_{-1} \\ \mathbf{N}_{-2} \\ \mathbf{N}_{-3} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{R} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{5}$$

Equation (5) can be solved semi-analytically by a two-sided Gaussian elimination process, resulting in two sets of recursions (shown for ±3 terms):

$$\begin{array}{llll}
\downarrow \mathbf{A}_3 & = \mathbf{D} & \uparrow \mathbf{N}_3 & = \mathbf{A}_3\mathbf{N}_2 \\
\downarrow \mathbf{A}_2 & = (\mathbf{I}-\mathbf{E}\mathbf{A}_3)^{-1}\mathbf{D} & \uparrow \mathbf{N}_2 & = \mathbf{A}_2\mathbf{N}_1 \\
\downarrow \mathbf{A}_1 & = (\mathbf{I}-\mathbf{E}\mathbf{A}_2)^{-1}\mathbf{D} & \uparrow \mathbf{N}_1 & = \mathbf{A}_1\mathbf{N}_0 \\
\rightarrow & & \rightarrow \mathbf{N}_0 & = (\mathbf{I}-\mathbf{D}\mathbf{A}_{-1}-\mathbf{E}\mathbf{A}_1)^{-1}\mathbf{R} \\
\uparrow \mathbf{A}_{-1} & = (\mathbf{I}-\mathbf{D}\mathbf{A}_{-2})^{-1}\mathbf{E} & \downarrow \mathbf{N}_{-1} & = \mathbf{A}_{-1}\mathbf{N}_0 \\
\uparrow \mathbf{A}_{-2} & = (\mathbf{I}-\mathbf{D}\mathbf{A}_{-3})^{-1}\mathbf{E} & \downarrow \mathbf{N}_{-2} & = \mathbf{A}_{-2}\mathbf{N}_{-1} \\
\uparrow \mathbf{A}_{-3} & = \mathbf{E} & \downarrow \mathbf{N}_{-3} & = \mathbf{A}_{-3}\mathbf{N}_{-2}
\end{array} \tag{6}$$

That is, starting from $\mathbf{A}_{-3}$ and $\mathbf{A}_3$, one recursively obtains $\mathbf{A}_{-1}$ and $\mathbf{A}_1$ that are used to solve $\mathbf{N}_0$. Then, starting from $\mathbf{N}_0$, one obtains all the $\mathbf{N}_{-i}$ and $\mathbf{N}_i$ terms, which, together with (3), give the negative and positive 'moments' up to the desired order. Note that with (2), (3), and (6), a global approximation is created, since the term $\mathbf{M}_0 = \mathbf{L}^{\mathrm{T}}\mathbf{N}_0$ (and all the other 'moments', $\mathbf{M}_i$) depend on the predefined number of 'moments', $(k_-, k_+)$. There is no expansion point: $\mathbf{M}_0 \neq \mathbf{Z}(0)$ and $\mathbf{M}_0 \neq \mathbf{Z}(\infty)$.

Here, let us point out that although the $\mathbf{N}_{-i}$ and $\mathbf{N}_i$ terms *could* be calculated explicitly to obtain a global approximation for $\mathbf{Z}(s)$, this is *not* done in GABOR; instead, the relations $\mathbf{N}_{-i} = \mathbf{A}_{-i}\mathbf{N}_{-i+1}$ and $\mathbf{N}_i = \mathbf{A}_i\mathbf{N}_{i-1}$ between the successive $\mathbf{N}_{-i}$ and $\mathbf{N}_i$ terms, respectively, are just used to span a Krylov subspace.

The next step is to find the projection matrices needed for MOR. To start, let us define the following Krylov subspace:

$$\mathrm{Kr}(\mathbf{N}_i, k_-, k_+) \equiv \mathrm{colspan}\{\mathbf{N}_{-k_-}, \ldots, \mathbf{N}_{-2}, \mathbf{N}_{-1}, \mathbf{N}_0, \mathbf{N}_1, \mathbf{N}_2, , \ldots, \mathbf{N}_{k_+}\} \tag{7}$$

In GABOR, a 'bidirectional' block-Arnoldi method is used to obtain

$$\mathbf{X} = [\mathbf{X}_{-k_-}, \ldots, \mathbf{X}_{-2}, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k_+}] \tag{8}$$

such that the orthonormalized $n$-by-$N$ blocks $\mathbf{X}_i$ span $\mathrm{Kr}(\mathbf{N}_i, k_-, k_+)$. Finally, the desired reduced-order matrices of GABOR are obtained by setting

$$\tilde{\mathbf{G}} = \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}, \;\; \tilde{\mathbf{C}} = \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}, \;\; \tilde{\Gamma} = \mathbf{X}^{\mathrm{T}}\Gamma\mathbf{X}, \;\; \tilde{\mathbf{B}} = \mathbf{X}^{\mathrm{T}}\mathbf{B}, \;\; \tilde{\mathbf{L}} = \mathbf{X}^{\mathrm{T}}\mathbf{L} \tag{9}$$

It can be shown [3] that any reduced-order model represented by (9) preserves passivity and reciprocity of the original RLC circuit. Moreover, GABOR matches the Z-parameter block 'moments' $\mathbf{M}_{-k_-}, \ldots, \mathbf{M}_{-2}, \mathbf{M}_{-1}, \mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_{k_+}$ of the underlying global approximation:

$$\tilde{\mathbf{M}}_i = \mathbf{M}_i, \;\; i = -k_-, \ldots, -2, -1, 0, 1, 2, \ldots, k_+ \tag{10}$$

A proof of the moment-matching property of GABOR is given in the Appendix.

Note that it is also possible to write (1) as a Neumann series:

$$\mathbf{Z}(s) = \mathbf{L}^{\mathrm{T}} \left[ \mathbf{I} - \left( s\mathbf{D} + \frac{1}{s}\mathbf{E} \right) \right]^{-1} \mathbf{R} = \mathbf{L}^{\mathrm{T}} \sum_{i=0}^{\infty} \left( s\mathbf{D} + \frac{1}{s}\mathbf{E} \right)^{i} \mathbf{R} \tag{11}$$

However, if we assume predefined $k_-$ and $k_+$, then GABOR evaluates the 'moments' $\mathbf{M}_{-k_-}, \ldots, \mathbf{M}_0, \ldots, \mathbf{M}_{k_+}$ 'exactly', while with the Neumann-series approach, an infinite number of terms (from which the 'moments' could be obtained after multiplications and grouping of $s^i$ and $1/s^i$ terms) would be required for the same accuracy. Also, to study the connections between these two approaches, one could expand each of the terms $\mathbf{A}_{-2}, \mathbf{A}_{-1}, \mathbf{N}_0, \mathbf{A}_1, \mathbf{A}_2$ in (6) in a Neumann series.

Here, it is fair to mention that GABOR has some problematic features, too. First, in (6), the recursive calculation of the $\mathbf{A}_i$ terms and $\mathbf{N}_0$ and the associated matrix inversions may result in numerical problems. Second, due to the global nature of the approximation, all the $\mathbf{A}_i$ terms (or the related LU factorizations) must be precomputed into the memory space. Third, the conductance matrix $\mathbf{G}$ in (1) is often singular for real-life RLC interconnect circuits; the last problem is treated in more detail in the next section.

## 3 Frequency Shifting

The possible singularity of the conductance matrix $\mathbf{G}$ in (1) is an issue in other nodal-formulation-based MOR methods, too. For GABOR, we *would like* to find such a frequency shifting (and/or scaling) that converts (1) into the following form:

$$\mathbf{Z}(z) = \mathbf{L}^{\mathrm{T}} \left( \hat{\mathbf{G}} + z\hat{\mathbf{C}} + \frac{1}{z}\hat{\Gamma} \right)^{-1} \mathbf{B} = \mathbf{L}^{\mathrm{T}} \left( \mathbf{I} - z\hat{\mathbf{D}} - \frac{1}{z}\hat{\mathbf{E}} \right)^{-1} \hat{\mathbf{R}} \tag{12}$$

where $z$ is the shifted frequency variable and $\hat{\mathbf{G}}$ is an invertible matrix, preferably

$$\hat{\mathbf{G}} = \mathbf{G} + s_0\mathbf{C} + \frac{1}{s_0}\Gamma \tag{13}$$

and $s_0$ is an appropriate real frequency (e.g., $s_0 = 10^9$ rad/s).

First, set $s = s + s_0 - s_0$ and $1/s = 1/s + 1/s_0 - 1/s_0$ and use (13) to obtain:

$$\mathbf{Z}(s) = \mathbf{L}^{\mathrm{T}} \left[ \hat{\mathbf{G}} + (s - s_0)\mathbf{C} + \left( \frac{1}{s} - \frac{1}{s_0} \right) \Gamma \right]^{-1} \mathbf{B} \triangleq \mathbf{L}^{\mathrm{T}} \left( \hat{\mathbf{G}} + z_1\mathbf{C} + \frac{1}{z_2}\Gamma \right)^{-1} \mathbf{B} \tag{14}$$

This approach can *not* be used with GABOR, since we have $z_1 \neq z_2$, thus violating (12). However, just to show an interesting link, (14) could be further processed as

$$\mathbf{Z}(s) = \mathbf{L}^{\mathrm{T}}\left[\mathbf{I} - \left(\frac{s}{s_0} - 1\right)\left(-s_0\hat{\mathbf{G}}^{-1}\mathbf{C}\right) - \left(\frac{s_0}{s} - 1\right)\left(-\frac{1}{s_0}\hat{\mathbf{G}}^{-1}\Gamma\right)\right]^{-1}\hat{\mathbf{G}}^{-1}\mathbf{B}$$
$$\triangleq \mathbf{L}^{\mathrm{T}}\left(\mathbf{I} - \sigma_1\hat{\mathbf{D}} - \sigma_2\hat{\mathbf{E}}\right)^{-1}\hat{\mathbf{R}} = \mathbf{L}^{\mathrm{T}}\sum_{i=0}^{\infty}\left(\sigma_1\hat{\mathbf{D}} + \sigma_2\hat{\mathbf{E}}\right)^{i}\hat{\mathbf{R}} \tag{15}$$

from which one could continue with the Krylov-subspace techniques of Ref. [5].

Second, we consider the specific frequency shifting and scaling that was an integral part of ENOR [3]:

$$z = -\frac{s - s_0}{s_0} \Rightarrow s = s_0(1 - z) \Rightarrow \frac{1}{s} = \frac{1}{s_0}\frac{1}{1-z} = \frac{1}{s_0}(1 + z + z^2 + \ldots) \tag{16}$$

Unfortunately, the expression for $1/s$ (that was not converted, explicitly, into a power series in Ref. [3]) is *not* in a convenient form for GABOR. In fact, if we insert these expressions for $s$ and $1/s$ in (1), require that $\mathbf{Z}(z) = \mathbf{M}_0 + \mathbf{M}_1 z + \mathbf{M}_2 z^2 + \ldots$, equate the powers of $z$, and do some algebra, it turns out that we have just obtained an alternative way to derive the recursion formulas for the ENOR method.

In the course of this work, the above and many other scalings were tried with GABOR. Unfortunately, it seems that any *consistent* scaling 'breaks the symmetry' of (1), and thus cannot convert it into the form of (12). Therefore, a 'dirty trick' was applied; $s = s_0(1 - z)$ was, still, used, but $1/s$ was very roughly approximated as

$$\frac{1}{s} = \frac{1}{s_0}\frac{1}{1-z} \approx -\frac{1}{s_0}\frac{1}{z} \tag{17}$$

The other 'dirty trick' was to ensure the invertibility of the enhanced $\mathbf{G}$ matrix by introducing an auxiliary diagonal perturbation, $g\mathbf{I}$. Inserting these in (1) results in

$$\mathbf{Z}(z) = \mathbf{L}^{\mathrm{T}}\left[(\mathbf{G} + s_0\mathbf{C} + g\mathbf{I}) - z\,(s_0\mathbf{C}) - \frac{1}{z}\left(\frac{1}{s_0}\Gamma\right)\right]^{-1}\mathbf{B} \tag{18}$$

which is (nearly) in the form of (12), and thus can be processed by applying the GABOR formulas (2)–(9).

## 4 Simulation Example

A dispersive transmission line was modeled with 50 LRCG sections, each having $L = 1\,\mathrm{nH}$, $R = 1\,\mathrm{m}\Omega$, $C = 1\,\mathrm{pF}$, and $G = 1\,\mathrm{mS}$. This two-port RLC circuit was reduced using a MATLAB/C implementation of (the frequency scaled) GABOR with $(k_-, k_+) = (5, 5)$, $s_0 = 5 \cdot 10^9$ rad/s $\Rightarrow f_0 = 5/(2\pi)$ GHz $\approx 0.796$ GHz, and $g = 10^{-9}\,$S. (Without the term $g\mathbf{I}$, the MOR fails.) Figure 1 shows $|Z_{21}(f)|$ in the frequency range $]0, 5]\,$GHz; the match is quite good up to 2 GHz. The original and reduced circuit resulted in 101-by-101 and 22-by-22 circuit matrices, respectively.

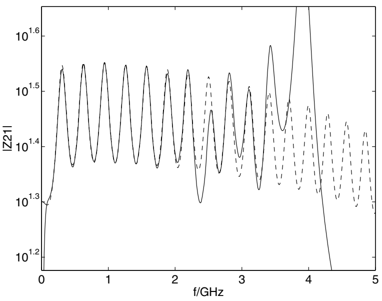


**Fig. 1:** Original (*dashed*) and reduced (*solid*) $|Z_{21}(f)|$

## 5 Conclusions

This paper proposed a Global-Approximation-Based Order Reduction (GABOR), which preserves the passivity and reciprocity of the original RLC circuit, and matches the ‘moments’ of the underlying global approximation. Also, the problems associated with GABOR were thoroughly discussed, and links between GABOR and other MOR methods were identified. While GABOR in itself is *not* a competitive MOR method for RLC circuits, the concept of global-approximation-based MOR might be worth further studies.

## Appendix

**Theorem 1.** *GABOR matches the Z-parameter block ‘moments’:*

$$\tilde{\mathbf{M}}_i = \mathbf{M}_i, \;\; i = -k_-, \ldots, -2, -1, 0, 1, 2, \ldots, k_+$$

*where the global-approximation ‘moments’ $\mathbf{M}_i$ and $\tilde{\mathbf{M}}_i$ are obtained from the original RLC circuit (1) and the reduced-order model (9), respectively.*

*Proof.* (The proof is partially based on some ideas of Ref. [1].) Let $\mathbf{I}_n$, ($\mathbf{0}_n$), $\mathbf{I}_N$ ($\mathbf{0}_N$), and $\mathbf{I}_q$ ($\mathbf{0}_q$) denote $n$-by-$n$, $N$-by-$N$, and $q$-by-$q$ unit (zero) matrices, respectively, where $n$ is the number of nodes, $N$ that of ports, and $q = (k_- + 1 + k_+) \cdot N$ is the order of reduction. Let us, due to lack of space but without loss of generality, derive formulas for $\mathbf{M}_1$ and $\tilde{\mathbf{M}}_1$ in the case where $(k_-, k_+) = (1, 1)$. By applying (1)–(5) for the original $\mathbf{M}_1$, we can write

$$\begin{aligned}\mathbf{M}_1 &= \mathbf{L}^{\mathrm{T}}\mathbf{N}_1 = \mathbf{L}^{\mathrm{T}}[\mathbf{I}_n\;\mathbf{0}_n\;\mathbf{0}_n]\begin{bmatrix}\mathbf{N}_1\\ \mathbf{N}_0\\ \mathbf{N}_{-1}\end{bmatrix} = \mathbf{L}^{\mathrm{T}}[\mathbf{I}_n\;\mathbf{0}_n\;\mathbf{0}_n]\begin{bmatrix}\mathbf{I} & \mathbf{G}^{-1}\mathbf{C} & \\ \mathbf{G}^{-1}\Gamma & \mathbf{I} & \mathbf{G}^{-1}\mathbf{C}\\ & \mathbf{G}^{-1}\Gamma & \mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}_{n\times N}\\ \mathbf{G}^{-1}\mathbf{B}\\ \mathbf{0}_{n\times N}\end{bmatrix}\\ &= \mathbf{L}^{\mathrm{T}}[\mathbf{I}_n\;\mathbf{0}_n\;\mathbf{0}_n]\begin{bmatrix}\mathbf{G} & \mathbf{C} & \\ \Gamma & \mathbf{G} & \mathbf{C}\\ & \Gamma & \mathbf{G}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}_n\\ \mathbf{I}_n\\ \mathbf{0}_n\end{bmatrix}\mathbf{B} \triangleq \mathbf{L}^{\mathrm{T}}[\mathbf{I}_n\;\mathbf{0}_n\;\mathbf{0}_n]\mathbf{G}_{\mathrm{T}}^{-1}\mathbf{B}_{\mathrm{T}} \triangleq \mathbf{L}^{\mathrm{T}}[\mathbf{I}_n\;\mathbf{0}_n\;\mathbf{0}_n]\mathbf{N}\end{aligned}$$

where we have denoted

$$\mathbf{G}_{\mathrm{T}} = \begin{bmatrix} \mathbf{G} & \mathbf{C} & \\ \Gamma & \mathbf{G} & \mathbf{C} \\ & \Gamma & \mathbf{G} \end{bmatrix}, \quad \mathbf{B}_{\mathrm{T}} = \begin{bmatrix} \mathbf{0}_n \\ \mathbf{I}_n \\ \mathbf{0}_n \end{bmatrix} \mathbf{B}, \quad \mathbf{G}_{\mathrm{T}}^{-1}\mathbf{B}_{\mathrm{T}} = \mathbf{N} = \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_0 \\ \mathbf{N}_{-1} \end{bmatrix}$$

Next, an appropriate expression is derived for $\tilde{\mathbf{M}}_1$:

$$\begin{aligned}
\tilde{\mathbf{M}}_1 &= \tilde{\mathbf{L}}^{\mathrm{T}}[\mathbf{I}_q \; \mathbf{0}_q \; \mathbf{0}_q] \begin{bmatrix} \tilde{\mathbf{G}} & \tilde{\mathbf{C}} & \\ \tilde{\Gamma} & \tilde{\mathbf{G}} & \tilde{\mathbf{C}} \\ & \tilde{\Gamma} & \tilde{\mathbf{G}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0}_q \\ \mathbf{I}_q \\ \mathbf{0}_q \end{bmatrix} \tilde{\mathbf{B}} \\
&= (\mathbf{X}^{\mathrm{T}}\mathbf{L})^{\mathrm{T}}[\mathbf{I}_q \; \mathbf{0}_q \; \mathbf{0}_q] \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X} & \\ \mathbf{X}^{\mathrm{T}}\Gamma\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X} \\ & \mathbf{X}^{\mathrm{T}}\Gamma\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0}_q \\ \mathbf{I}_q \\ \mathbf{0}_q \end{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{B} \\
&= \mathbf{L}^{\mathrm{T}}\mathbf{X}[\mathbf{I}_q \; \mathbf{0}_q \; \mathbf{0}_q] \left( \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & & \\ & \mathbf{X}^{\mathrm{T}} & \\ & & \mathbf{X}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{G} & \mathbf{C} & \\ \Gamma & \mathbf{G} & \mathbf{C} \\ & \Gamma & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{X} & & \\ & \mathbf{X} & \\ & & \mathbf{X} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{0}_q \\ \mathbf{I}_q \\ \mathbf{0}_q \end{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{B} \\
&\triangleq \mathbf{L}^{\mathrm{T}}\mathbf{X}[\mathbf{I}_q \; \mathbf{0}_q \; \mathbf{0}_q] \left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1} \begin{bmatrix} \mathbf{0}_q \\ \mathbf{I}_q \\ \mathbf{0}_q \end{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{B} \\
&= \mathbf{L}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}[\mathbf{I}_n \; \mathbf{0}_n \; \mathbf{0}_n] \begin{bmatrix} \mathbf{X} & & \\ & \mathbf{X} & \\ & & \mathbf{X} \end{bmatrix} \left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1} \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & & \\ & \mathbf{X}^{\mathrm{T}} & \\ & & \mathbf{X}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{0}_n \\ \mathbf{I}_n \\ \mathbf{0}_n \end{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{B} \\
&= \mathbf{L}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}[\mathbf{I}_n \; \mathbf{0}_n \; \mathbf{0}_n]\mathbf{X}_{\mathrm{D}} \left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1} \mathbf{X}_{\mathrm{D}}^{\mathrm{T}} \begin{bmatrix} \mathbf{0}_n \\ \mathbf{I}_n \\ \mathbf{0}_n \end{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{B}
\end{aligned}$$

where we have denoted

$$\mathbf{X}_{\mathrm{D}} = \begin{bmatrix} \mathbf{X} & & \\ & \mathbf{X} & \\ & & \mathbf{X} \end{bmatrix}$$

Now, starting from the right, the long expression of $\tilde{\mathbf{M}}_1$ is simplified step by step:

$$\begin{aligned}
\mathbf{X}_{\mathrm{D}}^{\mathrm{T}} \begin{bmatrix} \mathbf{0}_n \\ \mathbf{I}_n \\ \mathbf{0}_n \end{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{B} &= \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & & \\ & \mathbf{X}^{\mathrm{T}} & \\ & & \mathbf{X}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{0}_n \\ \mathbf{I}_n \\ \mathbf{0}_n \end{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{B} = \begin{bmatrix} \mathbf{0}_n \\ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{X}^{\mathrm{T}}\mathbf{B} \\ \mathbf{0}_n \end{bmatrix} = \begin{bmatrix} \mathbf{0}_n \\ \mathbf{X}^{\mathrm{T}}\mathbf{B} \\ \mathbf{0}_n \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & & \\ & \mathbf{X}^{\mathrm{T}} & \\ & & \mathbf{X}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{0}_n \\ \mathbf{I}_n \\ \mathbf{0}_n \end{bmatrix} \mathbf{B} = \mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}}
\end{aligned}$$

$$\Rightarrow \tilde{\mathbf{M}}_1 = \mathbf{L}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}[\mathbf{I}_n \; \mathbf{0}_n \; \mathbf{0}_n]\mathbf{X}_{\mathrm{D}} \left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1} \mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}}$$

As a preprocessing step for simplifying the term $\mathbf{X}_{\mathrm{D}}\left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}}$ in the above $\tilde{\mathbf{M}}_1$ formula, let us, first, treat an auxiliary term $\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N}$. Let $\mathbf{N}_i = \mathbf{X}_i\mathbf{T}_i$ be the factorization (QR decomposition) of $\mathbf{N}_i$ into an orthonormal $n$-by-$N$ matrix, $\mathbf{X}_i$ (with $\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i = \mathbf{I}$) and upper triangular $N$-by-$N$ matrix, $\mathbf{T}_i$; here, the matrices $\mathbf{X}_i$ are the block columns of the congruence-transform matrix $\mathbf{X}$, see (8). Now we can write

$$\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} = \begin{bmatrix} \mathbf{X} & & \\ & \mathbf{X} & \\ & & \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{X}^{\mathrm{T}} & & \\ & \mathbf{X}^{\mathrm{T}} & \\ & & \mathbf{X}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_0 \\ \mathbf{N}_{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_1 \\ \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_0 \\ \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_{-1} \end{bmatrix}$$

where, e.g.,

$$\begin{aligned} \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_1 &= [\mathbf{X}_{-1}\ \mathbf{X}_0\ \mathbf{X}_1] \begin{bmatrix} \mathbf{X}_{-1}^{\mathrm{T}} \\ \mathbf{X}_0^{\mathrm{T}} \\ \mathbf{X}_1^{\mathrm{T}} \end{bmatrix} \mathbf{X}_1\mathbf{T}_1 = [\mathbf{X}_{-1}\ \mathbf{X}_0\ \mathbf{X}_1] \begin{bmatrix} \mathbf{X}_{-1}^{\mathrm{T}}\mathbf{X}_1\mathbf{T}_1 \\ \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1\mathbf{T}_1 \\ \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\mathbf{T}_1 \end{bmatrix} \\ &= [\mathbf{X}_{-1}\ \mathbf{X}_0\ \mathbf{X}_1] \begin{bmatrix} \mathbf{0}_N\mathbf{T}_1 \\ \mathbf{0}_N\mathbf{T}_1 \\ \mathbf{I}_N\mathbf{T}_1 \end{bmatrix} = [\mathbf{X}_{-1}\ \mathbf{X}_0\ \mathbf{X}_1] \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{T}_1 \end{bmatrix} = \mathbf{X}_1\mathbf{T}_1 = \mathbf{N}_1 \end{aligned}$$

$$\text{Also, } \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_0 = \mathbf{N}_0 \text{ and } \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_{-1} = \mathbf{N}_{-1} \Rightarrow \mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} = \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_0 \\ \mathbf{N}_{-1} \end{bmatrix} = \mathbf{N}$$

Now, we can write

$$\begin{aligned} & \mathbf{N} = \mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} \\ \Leftrightarrow\ & \mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{N} = \mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} \\ \Leftrightarrow\ & \mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}} = (\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}})\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} \\ \Leftrightarrow\ & \left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}} = \mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} \\ \Leftrightarrow\ & \mathbf{X}_{\mathrm{D}}\left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}} = \mathbf{X}_{\mathrm{D}}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{N} \\ \Leftrightarrow\ & \mathbf{X}_{\mathrm{D}}\left(\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{G}_{\mathrm{T}}\mathbf{X}_{\mathrm{D}}\right)^{-1}\mathbf{X}_{\mathrm{D}}^{\mathrm{T}}\mathbf{B}_{\mathrm{T}} = \mathbf{N} \end{aligned}$$

$$\Rightarrow \tilde{\mathbf{M}}_1 = \mathbf{L}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}[\mathbf{I}_n\ \mathbf{0}_n\ \mathbf{0}_n]\mathbf{N} = \mathbf{L}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}[\mathbf{I}_n\ \mathbf{0}_n\ \mathbf{0}_n] \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_0 \\ \mathbf{N}_{-1} \end{bmatrix} = \mathbf{L}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{N}_1 = \mathbf{L}^{\mathrm{T}}\mathbf{N}_1 = \mathbf{M}_1$$

Similarly, $\tilde{\mathbf{M}}_i = \mathbf{M}_i,\ \ i = -k_-, \ldots, -2, -1, 0, 1, 2, \ldots, k_+$ Q.E.D.

## References

1. Odabasioglu, A., Celik, M., Pileggi, L.T.: PRIMA: passive reduced-order interconnect macromodeling algorithm. IEEE Trans. CAD of Int. Circ. and Syst., **17**, 645–654 (1998)
2. Freund, R.W.: SPRIM: structure-preserving reduced-order interconnect macromodeling. In: Proc. ICCAD 2004, pp. 80–87. San Jose, November 7–11 (2004)
3. Sheehan, B.N.: ENOR: model order reduction of RLC circuits using nodal equations for efficient factorization. In: Proc. DAC 1999, pp. 17–21. New Orleans, June 21–24 (1999)
4. Palenius, T., Roos, J.: Comparison of reduced-order interconnect macromodels for time-domain simulation. IEEE Trans. Microwave Theory and Tech., **52**, 2240–2250 (2004)
5. Feng. L., Benner, P.: Model order reduction for systems with coupled parameters. In: SCEE 2008 Book of Abstracts, pp. 23–24. Espoo, Sept. 28 – Oct. 3 (2008)

# Model Order Reduction for Systems with Non-Rational Transfer Function Arising in Computational Electromagnetics

Lihong Feng and Peter Benner

**Abstract** We consider model order reduction of a system described by a non-rational transfer function. The systems under consideration result from the discretization of electromagnetic systems with surface losses [1]. In this problem, the frequency parameter $s$ appears nonlinearly. We interpret the nonlinear functions containing $s$ as parameters of the systems and apply parametric model order reduction (PMOR) to the system. Since the parameters are functions of the frequency $s$, they are coupled to each other. Nevertheless, PMOR treats them as individual parameters. We review existing PMOR methods, and discuss their applicability to the problem considered here. Based on our findings, we propose an optimized method for the parametric system considered in this paper. We report on numerical experiments performed with the optimized method applied to real-life data.

## 1 Problem Description

The transfer functions of the systems considered here take the form

$$H(s) = sB^{\mathrm{T}}(s^2 I_n - 1/\sqrt{s}D + A)^{-1}B, \tag{1}$$

where $A, D$ are $n \times n$ matrices, $B$ is an $n \times p$ matrix ($p \ll n$), and $I_n$ is the identity matrix of size $n$. These transfer functions result from the spatial discretization of electromagnetic field equations, i.e., the Maxwell equations, describing the electrodynamical behavior of microwave devices, when surface losses are included in the physical model. For details see [1].

Lihong Feng, Peter Benner
Mathematics in Industry and Technology, Faculty of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany, e-mail: lihong.feng@mathematik.tu-chemnitz.de, benner@mathematik.tu-chemnitz.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_63,

Due to the symmetry inherent in the model under consideration, we will focus here on one-sided projection methods for model order reduction. That is, the reduced-order model is obtained by finding a projection matrix $V \in \mathbb{R}^{n\times q}$ and $V^TV = I_r$, such that

$$\hat{H}(s) = s\hat{B}^T(s^2 I_r - 1/\sqrt{s}\hat{D} + \hat{A})^{-1}\hat{B} \approx H(s)$$

is the reduced-order transfer function, with $\hat{D} = V^TDV$, $\hat{A} = V^TAV \in \mathbb{R}^{q\times q}$, $\hat{B} = V^TB \in \mathbb{R}^{q\times p}$, and $q \ll n$.

Since $H(s)$ contains not only $s$, but also two nonlinear functions of $s$, namely $s^2$ and $1/\sqrt{s}$, we intend to apply parametric model order reduction (PMOR) methods (e.g., [2,3]) in order to compute $\hat{H}(s)$. Notice that model reduction of a similar transfer function as in (1) is also discussed in [1], where a conventional non-parametric model order reduction method is considered. There, the generated projection matrix $V$ depends not only on some expansion point $s_0$, but also on some value $\tilde{s}$ used for fixing the term $1/\sqrt{s}$ to $1/\sqrt{\tilde{s}}$ in order to obtain a standard rational transfer function. This may cause poor approximation properties of $\hat{H}(s)$ at values $s$ where $1/\sqrt{s}$ is not close to $1/\sqrt{\tilde{s}}$. In our situation, the basic difference of PMOR compared to non-parametric model reduction is that the computed projection matrix $V$ only depends on the expansion point $s_0$, but not on any fixed value $\tilde{s}$ other than $s_0$ as the term $1/\sqrt{s}$ is treated as free parameter. Usually, this strategy produces a reduced-order model (transfer function) with evenly distributed small error for a large frequency range.

In order to apply PMOR methods as those suggested in [2, 3], $H(s)$ has to be expanded into a power series, then the projection matrix $V$ is computed based on the coefficients of the series expansion. Different ways of computing $V$ constitute different PMOR algorithms. Therefore, in the following we first consider various series expansions of $H(s)$ and discuss their suitability for PMOR.

The first possibility for an expansion of $H(s)$ into a Neumann series is obtained when $A$ is nonsingular as follows:

$$\begin{aligned} H(s) &= sB^T(s^2 I_n - 1/\sqrt{s}D + A)^{-1}B \\ &= sB^T(s^2A^{-1} - \tfrac{1}{\sqrt{s}}A^{-1}D + I_n)^{-1}A^{-1}B \\ &= sB^T[I_n - (\tfrac{1}{\sqrt{s}}A^{-1}D - s^2A^{-1})]^{-1}A^{-1}B \\ &= sB^T\sum_{i=0}^{\infty}(\tfrac{1}{\sqrt{s}}A^{-1}D - s^2A^{-1})^iA^{-1}B, \end{aligned}$$

where the last equality only holds if $\|\frac{1}{\sqrt{s}}A^{-1}D - s^2A^{-1})\| < 1$ for a suitably chosen matrix norm. However, in the considered application, $A$ is a singular matrix. Thus, the series expansion above cannot be used.

Extracting $s^2$ from the inverse in (1),

$$H(s) = \frac{1}{s}B^T(I_n - \frac{1}{s^2\sqrt{s}}D + \frac{1}{s^2}A)^{-1}B, \tag{2}$$

a second possibility for a Neumann series expansion is obtained:

$$H(s) = \frac{1}{s}B^T(I_n - \frac{1}{s^2\sqrt{s}}D + \frac{1}{s^2}A)^{-1}B = \frac{1}{s}B^T\sum_{i=0}^{\infty}\underbrace{(\frac{1}{s^2\sqrt{s}}D - \frac{1}{s^2}A)^i}_{:=Q(s)}B. \qquad (3)$$

Next we check whether the above Neumann series expansion is convergent or not. Although the interesting frequency in applications of (1) is relatively high ($s = \mathrm{j}\omega$, $\omega \approx 10^9$ Hz for the real-life data used in Section 4), ill-scaling of the matrices $D$ and $A$ also often encountered in practice prevents convergence of the above series expansion. (Note: in the example considered in Section 4, $D_{ij} \sim 10^{27}$, $A_{ij} \sim 10^{23}$.) Therefore, the series expansion in (3) is not applicable in practice either. (One would need $|s| > 10^{12}$ in order to achieve convergence of the Neumann series!)

Finally, we study a third alternative for power series expansion of (1). We also use (2) and define $s_1(s) := \frac{1}{s^2\sqrt{s}}$, $s_2(s) := \frac{1}{s^2}$. If we choose a nonzero expansion point $s_0$ and let $s_1(s) = s_1(s_0) + \sigma_1(s)$, $s_2(s) = s_2(s_0) + \sigma_2(s)$, then we get

$$\begin{aligned} H(s) &= \tfrac{1}{s}B^T(I_n - s_1(s)D + s_2(s)A)^{-1}B \\ &= \tfrac{1}{s}B^T(\underbrace{I_n - s_1(s_0)D + s_2(s_0)A}_{:=G} - \sigma_1 D + \sigma_2 A)^{-1}B \\ &= \tfrac{1}{s}B^T[I_n - (\sigma_1 G^{-1}D - \sigma_2 G^{-1}A)]^{-1}G^{-1}B \\ &= \tfrac{1}{s}B^T\sum_{i=0}^{\infty}(\underbrace{\sigma_1 G^{-1}D - \sigma_2 G^{-1}A}_{:=\mathscr{Q}})^i G^{-1}B \end{aligned} \qquad (4)$$

Simple calculations show that the entries of $\mathscr{Q}$ are small enough to guarantee $\|\mathscr{Q}\| < 1$ for small $\sigma_1, \sigma_2$, so that the series is convergent. Therefore, we will use the series expansion (4) in the following analysis. We will treat $\sigma_1$ and $\sigma_2$ as two individual parameters, although, they are both functions of $s$, and hence are coupled parameters. For ease of notation, in the following we will use $B_M := G^{-1}B$, $M_1 := G^{-1}D$, and $M_2 := -G^{-1}A$. The difference between PMOR methods lies in the computation of the projection matrix $V$. Therefore, in the next section, we review two different methods for computing $V$ and analyze drawbacks of these methods. Based on these considerations, we propose an improved method in Section 3. We will report on numerical experiments with these methods in Section 4.

## 2 Different Methods for Computing the Projection Matrix $V$

### *2.1 Directly Computing* $V$

A simple and direct way of obtaining $V$ is to compute the coefficient matrices in the series expansion:

$$\begin{aligned} H(s) = \tfrac{1}{s}B^{\mathrm{T}}[&B_M + M_1 B_M \sigma_1 + M_2 B_M \sigma_2 \\ &+ M_1^2 B_M \sigma_1^2 + (M_1 M_2 + M_2 M_1)B_M \sigma_1 \sigma_2 + \ldots + M_1^3 B_M \sigma_1^3 + \ldots]. \end{aligned} \qquad (5)$$

by direct matrix multiplication, then orthogonalize these coefficients to get the matrix $V$ [2] as below:

$$\text{range}\{V\} = \text{orth}\{B_M, M_1B_M, \ldots, (M_1M_2 + M_2M_1)B_M, \ldots\}. \tag{6}$$

Unfortunately, the coefficients quickly become linearly dependent due to numerical instability (see the analysis, e.g., in [3, 4]). In the end, the matrix $V$ is often so inaccurate that it does not possess the expected theoretical properties.

### 2.2 Recursively Computing $V$

A recursive method for computing $V$ is proposed in [3], which is based on certain recursions between the coefficients of the series expansion. The series expansion (4) can also be written in the following form:

$$H(s) = \frac{1}{s}[B_M + (\sigma_1M_1 + \sigma_2M_2)B_M + \ldots + (\sigma_1M_1 + \sigma_2M_2)^iB_M + \ldots]. \tag{7}$$

Using (7), we define

$$\begin{aligned} R_0 &= B_M, \\ R_1 &= [M_1R_0, M_2R_0], \\ &\vdots \\ R_j &= [M_1R_{j-1}, M_2R_{j-1}], \\ &\vdots \end{aligned} \tag{8}$$

We see that $R_0, R_1, \ldots, R_j, \ldots$ include all the coefficient matrices in the series expansion (7). Therefore, we can use $R_0, R_1, \ldots$ to generate the projection matrix $V$:

$$\text{range}\{V\} = \text{colspan}\{R_0, R_1, \ldots, R_m\}. \tag{9}$$

Here, $V$ can be computed by employing the recursive relations between $R_j,\ j = 0, 1, \ldots, m$ combined with the modified Gram-Schmidt process [3].

A disadvantage of this approach is that coefficients of the same powers of $\sigma_1, \sigma_2$ are treated separately and are orthogonalized sequentially using, e.g., the modified Gram-Schmidt process. This may lead to reduced-order models of larger order than the direct approach.

## 3 An Optimized Method for Computing the Projection Matrix $V$

In this section, we will overcome the drawback of the recursive method at least for the situation with two parameters as encountered in the electromagnetics model considered in the introduction. We will explain our approach using the first "mixed moment" in the series expansion (4).

For this purpose, note that the coefficients $M_1M_2B_M$ and $M_2M_1B_M$ are treated as two individual terms in (8). Observing that they are actually both coefficients of $\sigma_1\sigma_2$, they can be considered as one term during the computation as in (6). There are some similar coefficients which are computed separately in (8), but which are added up in (6), such as $M_2M_1^2B_M, M_1^2M_2B_M, M_1M_2M_1B_M$, since they are all coefficients of $\sigma_1^2\sigma_2$. Their treatment as individual matrices in (8) will, in some cases, result in more columns in the final projection matrix $V$ as compared to (6). This will produce reduced-order models which are not as small as possible. Next we develop a new set of recursions for the coefficient matrices in (6), such that the coefficients of $\sigma_1^i\sigma_2^j$, $i,j>0$, yield only one term. Furthermore, the modified Gram-Schmidt algorithm is applied to these recursions in order to compute the matrix $V$ in a numerically robust way which will result in an accurate reduced-order model.

Using the coefficient matrices in (6), we define new matrices $V^i$ as below, where $i$ is to be understood as an index, not as a power. If $J=i+j$, corresponding to the powers of the monomial $\sigma_1^i\sigma_2^j$, then

$$\begin{aligned}
J=0: &\quad & B_M &:= V^1 = [V_1^1]\\
J=1: &\quad & [M_1B_M, M_2B_M] &:= V^2 = [V_1^2, V_2^2]\\
J=2: &\quad & [M_1^2B_M, (M_1M_2+M_2M_1)B_M, M_2^2B_M] &:= V^3 = [V_1^3, V_2^3, V_3^3]\\
J=3: &\quad & [M_1^3B_M, (M_2M_1^2+M_1^2M_2+M_1M_2M_1)B_M, &\\
& & (M_2M_1M_2+M_2^2M_1+M_1M_2^2)B_M, M_2^3B_M] &:= V^4 = [V_1^4, V_2^4, V_3^4, V_4^4]\\
& & &\vdots\\
J=m: &\quad & [M_1^mB_M, \cdots, M_2^mB_M] &:= V^{m+1} = [V_1^{m+1}, \cdots, V_{m+1}^{m+1}]
\end{aligned}$$

The above definitions are based on the observation that $J=i$ actually corresponds to $i+1$ coefficient matrices. Let $V_0^i=0$ and $V_{i+1}^i=0$, $i=1,2,\ldots,m+1$, we derive the following recursions:

$$V_j^i = M_2V_{j-1}^{i-1} + M_1V_j^{i-1},\ j=1,2,\ldots,i;\ i=2,3,\ldots,m+1. \tag{10}$$

Based on the recursions in (10), we propose an algorithm which computes $V$ as in (6) using a modified Gram-Schmidt process. We describe this algorithm as in Algorithm 1. In this algorithm,

- $p$ is the number of columns of $B_M$, i.e., the number of inputs.
- $B_M(:,i)$ is the $i$th column of $B_M$, $i=1,2,\ldots,p$,
- *tol* is a tolerance determining linear dependency of the next computed column vector of $V$.
- $V_j^i(:,l)$ is the $l$th column in matrix $V_j^i$, $s_0$ is the expansion point in (4).

**Algorithm 1**

```
1. Apply the modified Gram-Schmidt process to the columns of $V^1 = B_M$:
   $V_1^1(:,1) = B_M(:,1)/\|B_M(:,1)\|$;
```

```
for $i = 2 : p$
  $w = B_M(:, i)$;
  for $j = 1 : i-1$
    $w = w - ((V_1^1(:, j))^T * w) * V_1^1(:, j)$;
  end
  if $\|w\| > tol$
    $V_1^1(:, i) = w/\|w\|$;
  else
    $V_1^1(:, i) = 0$;
  end
end
```

2. *Apply the modified Gram-Schmidt process to the columns in* $V_j^i$, $i = 2, 3, \ldots, m+1$, $j = 1, \ldots, i$

```
for $i = 2 : m+1$, $j = 1 : i$, $l = 1 : p$, do
  $w = M_2 V_{j-1}^{i-1}(:, l) + M_1 V_j^{i-1}(:, l)$;                    ($\sharp$)
  for $t_i = 1, \ldots, i-1$, $t_j = 1, \ldots, t_i$, $t = 1, \ldots, p$, do
    $w = w - ((V_{t_j}^{t_i}(:, t))^T * w) * V_{t_j}^{t_i}(:, t)$;
  end
  for $t_j = 1, \ldots, j-1$, $t = 1, \ldots, p$, do
    $w = w - ((V_{t_j}^{i}(:, t))^T * w) * V_{t_j}^{i}(:, t)$;
  end
  for $t = 1, \ldots, l-1$, do
    $w = w - ((V_j^i(:, t))^T * w) * V_j^i(:, t)$;
  end
  if $\|w\| > tol$
    $V_j^i(:, l) = w/\|w\|$;
  else
    $V_j^i(:, l) = 0$;
  end
end
```

3. *Delete the zero columns in* $V^i$, *yielding* $\tilde{V}^i$, $i = 1, 2, \ldots, m+1$.
4. $V = \{\tilde{V}^1, \tilde{V}^2, \cdots, \tilde{V}^{m+1}\}$.
5. *if* $s_0$ *is not a real number,* $\text{range}\{V\} = \text{span}\{\text{real}(V), \text{imag}(V)\}$.

*Remark 1.* a) In step 2. ($\sharp$), the multiplication with zero columns can be avoided, thus saving a matrix multiplication and the application of $G^{-1}$. Moreover, if both $V_{j-1}^{i-1}(:, l)$ and $V_j^{i-1}(:, l)$ are zero, the whole orthogonalization procedure can be skipped. This is implemented with conditional statements and careful bookkeeping for the zero columns (which must be kept until step 3. for dimension consistency). The necessary statements are not shown in Algorithm 1 to keep the algorithmic description brief.

b) A more efficient variant of Algorithm 1 would apply modified Gram-Schmidt orthogonalization only on level $i$ of the recursion and then run modified Gram-Schmidt again at step 4. in order to obtain $V$ with orthonormal columns. Whether this results in noticeable numerical inaccuracies requires further investigation.

It is shown in the next section that the order of the reduced-order model computed by Algorithm 1 is smaller than the order of the one derived by (9). The improved algorithm is well suited for the parametric system from computational electromagnetics considered in this paper. Its performance will be illustrated using an industrial test example in the next section.

## 4 Simulation Results

In this section, we compare the three introduced methods for computing $V$ when applied to an industrial test case.[1] For convenience, we call the method directly computing $V$ *DirectV*. The recursive method in Section 2.2 is named *RecV*, the optimized method proposed in Section 3 is called *ImRecV*, since it is an *im*proved method based on both *DirectV* and *RecV*. The order of the original system is $n = 29,295$. We show the numerical instability of *DirectV* in Table 1. In Table 2, we compare the orders of the reduced-order models derived by *RecV* and *ImRecV*.

In Table 1 and Table 2, $J$ is defined as above, i.e., the coefficients corresponding to $\sigma_1^i \sigma_2^j$, $i+j = 0,1,2,\dots,J$ are used to compute $V$. $q$ is the number of columns in the final projection matrix $V$ and thus also the order of the reduced-order model. From Table 1 we see that although *DirectV* and *ImRecV* use the same coefficient matrices (6) to compute $V$, the number of columns of $V$ computed by *DirectV* is smaller than for $V$ computed by *ImRecV*. The 4th column of Table 1 shows the difference in the number of columns of $V$. This difference increases with $J$ which is due to the numerical instability of *DirectV*, resulting in linear dependency of computed columns of $V$ that would be linearly independent if they were computed exactly.

**Table 1:** Numerical instability of *DirectV*

| $J$ | $q$ of *DirectV* | $q$ of *ImRecV* | Number of columns deleted by *DirectV* |
|---|---|---|---|
| 2 | 12 | 12 | 12-12=0 |
| 4 | 32 | 38 | 38-32=6 |
| 5 | 44 | 58 | 58-44=14 |
| 6 | 52 | 82 | 82-52=30 |

In Table 2, we compare *ImRecV* with *RecV*. Recall that *RecV* computes $V$ according to (9), where some coefficient matrices are computed separately, which results in a larger reduced-order model for the system considered here. The second row of Table 2 shows that *RecV* computes more columns than *ImRecV*, therefore, the reduced-order model computed by *RecV* is less compact than that computed by *ImRecV* and thus less efficient for numerical simulations.

[1] Provided by Dr. S. Reitzinger, CST Computer Simulation Technology, Darmstadt, Germany.

The accuracy of *ImRecV* is illustrated in Fig. 1, where the order of the reduced transfer function $\hat{H}(s)$ computed by *ImRecV* is $q = 38$. The solid line in Fig. 1 is the magnitude of the original transfer function $H(\mathrm{j}\omega)$ in the frequency range of interest in the application. The star markers correspond to $\hat{H}(s)$, and match the solid line very well. Furthermore, the CPU time for evaluating $H(\mathrm{j}\omega)$ at 1000 frequency points for $\omega \in [4 \times 10^9,\ 8 \times 10^9]$ is around 8 hours, while the CPU time for evaluating $\hat{H}(\mathrm{j}\omega)$ of order $q = 38$ at the same frequency points is only 0.84 seconds. All simulations are run on an IBM notebook with Intel CPU T2400, 1.83GHz, 1GB RAM.

**Table 2:** Compactness of reduced-order models computed by *RecV* and *ImRecV*

| $J$ | 2 | 4 | 5 | 6 |
|---|---|---|---|---|
| $q$ (*RecV*) | 24 | 116 | 242 | 494 |
| $q$ (*ImRecV*) | 12 | 38 | 58 | 82 |

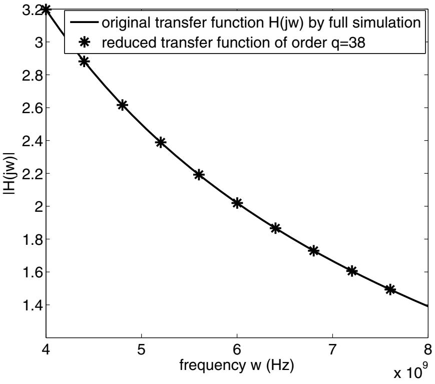


**Fig. 1:** Comparison of transfer functions

**Acknowledgements** This research was supported by the Alexander von Humboldt-Foundation and by the DFG grant BE 2174/7-1, *Automatic, Parameter-Preserving Model Reduction for Applications in Microsystems Technology.*

## References

1. Wittig, T., Schuhmann, R., Weiland, T.: Model order reduction for large systems in computational electromagnetics. Linear Algebra Appl. **415**(2-3), 499–530 (2006)
2. Daniel, L., Siong, O., Chay, L., Lee, K., White, J.: A multiparameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **22**(5), 678–693 (2004)
3. Feng, L., Benner, P.: A robust algorithm for parametric model order reduction. Proc. Appl. Math. Mech. **7**(1), 1021,501–1021,502 (2008)
4. Feldmann, P., Freund, R.: Efficient linear circuit analysis by Padé approximation via the Lanczos process. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **14**, 639–649 (1995)

# Model Order and Terminal Reduction Approaches via Matrix Decomposition and Low Rank Approximation

Peter Benner and André Schneider

**Abstract** We discuss methods for model order reduction (MOR) of linear systems with many input and output variables, arising in the modeling of linear (sub) circuits with a huge number of nodes and a large number of terminals, like power grids. Our work is based on the approaches SVDMOR and ESVDMOR proposed in recent publications [1–5]. In particular, we discuss efficient numerical algorithms for their implementation. Only by using efficient tools from numerical linear algebra, these methods become applicable for truly large-scale problems.

## 1 Introduction

Nowadays, MOR is an important and conventional step in the preprocessing of circuit simulation. The original model resulting from methods like modified nodal analysis has to be simplified due to its complexity. One issue of this simplification for VLSI design is the MOR of parasitic linear interconnect circuits. These circuits form substructures in the design of ICs and contain linear elements with comparatively little or no influence on the result of the simulation.

In some applications, the structure of these parasitic linear subcircuits has recently changed in the following sense. So far, the number of elements in these interconnect circuits was significantly larger than the number of connections to the whole circuit, the so-called pins or terminals. This assumption is no longer valid in all cases. Circuits with a lot of elements need extra power supply networks, so-called power grids [6,7]. In clock distribution networks, the clock signal is distributed from a common point to all the elements that need it for synchronization [8]. For simulating these circuits new methods are needed. Often, a lot of their terminals behave similar so that it is possible to compress the input-/output matrices in such a way that

Peter Benner, André Schneider
Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany, e-mail: benner@mathematik.tu-chemnitz.de, andre.schneider@mathematik.tu-chemnitz.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_64,

the I/O behavior can be realized through a few so-called virtual inputs/outputs [1–5]. As a consequence we deal with these virtual terminals, the number of which is much less than the original number of terminals. This allows the use of well known MOR methods like balanced truncation or Krylov subspace methods to reduce the number of inner nodes.

The intention of this paper is to explain the existing (E)SVDMOR approaches [1,4] and show improvements within the implementation in particular for large scale systems. In the following section, we review the fundamentals of the underlying approaches. We introduce the moments of a transfer function of the circuit describing system and show how to use the information in these moments in order to reduce the number of terminals. Later, we point out the weak point of the algorithm for really large scale systems and present a solution for this problem. After the introduction of this efficient algorithm to achieve a very compact model we show and discuss first numerical results in Section 3.

## 2 SVDMOR and ESVDMOR

Recent studies [1–5] have shown that we can make use of correlations between the plurality of input and output terminals. We use the singular value decomposition (SVD) based method SVDMOR [1,5] as well as the extended version of SVDMOR, the so-called ESVDMOR [2–5], which is the foundation for our work and will be explained in the following.

### *2.1 Extended-SVDMOR*

We assume that the linear system to be reduced has the following transfer function in frequency domain:

$$H(s) = L(sC+G)^{-1}B, \tag{1}$$

with $C, G \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m_{in}}$, and $L \in \mathbb{R}^{m_{out}\times n}$.

The number of inputs $m_{in}$ is not necessarily equal to the number of outputs, here $m_{out}$. Consider the $i$-th block moment of (1) defined as

$$\mathbf{m_i} = L(-G^{-1}C)^i G^{-1}B, \tag{2}$$

in terms of $\mathbf{m_i}$ as an $m_{out} \times m_{in}$ matrix

$$\mathbf{m_i} = \begin{bmatrix} m^i_{1,1} & m^i_{1,2} & \dots & m^i_{1,m_{in}} \\ m^i_{2,1} & m^i_{2,2} & \dots & m^i_{2,m_{in}} \\ \vdots & \vdots & \ddots & \vdots \\ m^i_{m_{out},1} & m^i_{m_{out},2} & \dots & m^i_{m_{out},m_{in}} \end{bmatrix}. \tag{3}$$

Note that the moments in (2) are equal to the coefficients of the Taylor series expansion of (1) in $s = 0$. The expansion in $s = s_0$ leads to frequency shifted moments defined as

$$\mathbf{m_i}(s_0) = L(-(s_0C+G)^{-1}C)^i(s_0C+G)^{-1}B. \tag{4}$$

The ESVDMOR approach uses the information of a combination of these moments to create a decomposition of (1) in the following way. To allow terminal reduction for inputs and outputs separately, $r$ different block moments forming two moment matrices are used: the input response matrix $M_I$ and the output response matrix $M_O$ defined as

$$M_I = \begin{bmatrix} \mathbf{m_0} \\ \mathbf{m_1} \\ \vdots \\ \mathbf{m_{r-1}} \end{bmatrix}, \qquad M_O = \begin{bmatrix} \mathbf{m_0}^T \\ \mathbf{m_1}^T \\ \vdots \\ \mathbf{m_{r-1}}^T \end{bmatrix}, \tag{5}$$

where column $k$ of $M_I$ represents the coefficients (moments) of the series expansion of (1) at all outputs due to input $k$. Similarly, each column $k$ of $M_O$ represents the coefficients of output $k$ due to all inputs. Note, that we expect the number of rows in each matrix to be larger than the number of columns so that the rank is determined by the column vectors. If not, $r$ has to be increased.

Applying the SVD to these matrices, we can obtain a low rank approximation

$$M_I = U_I \Sigma_I V_I^T \approx U_{I_{r_i}} \Sigma_{I_{r_i}} V_{I_{r_i}}^T, \qquad M_O = U_O \Sigma_O V_O^T \approx U_{O_{r_o}} \Sigma_{O_{r_o}} V_{O_{r_o}}^T, \tag{6}$$

where

- $\Sigma_{I_{r_i}}$ is an $r_i \times r_i$ diagonal matrix,
- $\Sigma_{O_{r_o}}$ is an $r_o \times r_o$ diagonal matrix,
- $V_{I_{r_i}}^T$ and $V_{O_{r_o}}^T$ are orthogonal $r_i \times m_{in}$ and $r_o \times m_{out}$ matrices that contain the dominant column subspaces of $M_I$ and $M_O$
- $U_{I_{r_i}}$ and $U_{O_{r_o}}$ are $rm_{out} \times r_i$ and $rm_{in} \times r_o$ matrices that are not used any further,
- $r_i$ and $r_o$ are the numbers of significant singular values as well as the numbers of the reduced virtual input and output terminals.

Equations (6) are the crucial points for our improvements described in Section 2.2.

Due to the fact that the important information about the dependencies of the I/O-ports is hidden in the matrices $V_{I_{r_i}}^T$ and $V_{O_{r_o}}^T$, approximations of $B$ and $L$ using the results of (6) lead to

$$B \approx B_r V_{I_{r_i}}^T \ \mathit{and} \ L \approx V_{O_{r_o}} L_r, \tag{7}$$

where $B_r \in \mathbb{R}^{n \times r_i}$ and $L_r \in \mathbb{R}^{r_o \times n}$ are consequences of applying the Moore-Penrose pseudoinverse (denoted by $(\cdot)^+$) of $V_{I_{r_i}}^T$ and $V_{O_{r_o}}$ (which are isometric) to $B$ and $L$, respectively. In detail, we have

$$B_r = BV_{I_{r_i}}(V_{I_{r_i}}^T V_{I_{r_i}})^{-1} = BV_{I_{r_i}}^{T+} = BV_{I_{r_i}} \tag{8}$$

and

$$L_r = (V_{O_{ro}}^T V_{O_{ro}})^{-1} V_{O_{ro}}^T L = V_{O_{ro}}^{+} L = V_{O_{ro}}^T L, \tag{9}$$

where $B_r \in \mathbb{R}^{n \times r_i}$ and $L_r \in \mathbb{R}^{r_o \times n}$. Consequently, we get a new internal transfer function $H_r(s)$,

$$H(s) \approx \widehat{H}(s) = V_{O_{ro}} \underbrace{L_r(G+sC)^{-1}B_r}_{:=H_r(s)} V_{I_{r_i}}^T. \tag{10}$$

This terminal reduced transfer function is now reduced to

$$\tilde{H}_r(s) = \tilde{L}_r(\tilde{G}+s\tilde{C})^{-1}\tilde{B}_r \approx H_r(s) = L_r(G+sC)^{-1}B_r \tag{11}$$

by some well known established MOR method, e. g., balanced truncation or a Krylov subspace method. At the end we get a very compact terminal and reduced-order model

$$H(s) \approx V_{O_{ro}} \tilde{H}_r(s) V_{I_{r_i}}^T. \tag{12}$$

Note that SVDMOR can be considered as a special case of ESVDMOR, using only one moment and one SVD, e. g. $r = 1$, and using $\mathbf{m_0}$ as moment.

## 2.2 Drawbacks and Solutions

For very large subcircuits the (E)SVDMOR methods are not suitable due to the use of the SVD. Suppose we have a matrix with dimension $n = 10^6$ and a modern CPU with 3 GHz. The computation of an SVD needs about $22n^3$ flops. This would mean $22 \cdot 10^{18}$ flops and therefore a total CPU time of approximately 230 years. Obviously, this is computationally too expensive. Hence, we combine the (E)SVDMOR approaches with cheaper matrix decomposition methods, like the truncated SVD (TSVD), which computes the needed singular values and the corresponding singular vectors only, see (6). Also other ideas to compute a truncated SVD-like decomposition cheaply can be used [9–11].

Furthermore, an explicit computation of the moments in (2) would be numerically unstable and too expensive. Without loss of generality we explain the decomposition of $M_I$, so that

$$M_I \approx U_{I_{r_i}} \Sigma_{I_{r_i}} V_{I_{r_i}}^T = \sum_{j=1}^{r_i} \sigma_j u_j v_j^T. \tag{13}$$

Recall that $r_i \ll m_{in}$ denotes the number of significant singular values and vectors. We do not know that number so we specify it depending on the error tolerance of the approximation. Unfortunately, there is no global error bound for the whole reduction yet (this is the topic of current research). We therefore simply use $\sigma_{r+1} < tol\sigma_1$ for a user-defined tolerance. Naturally, it is helpful to have a rapid decrease of the singular values $\sigma_j$, that means a lot of dependencies within the ports and enables a gainful terminal reduction, see the examples in Section 3.

The TSVD can be computed in several ways [9, 10, 12]. Consider the augmented matrix $A \in \mathbb{R}^{r \cdot m_{out}+m_{in} \times r \cdot m_{out}+m_{in}}$ of the form

$$A = \begin{pmatrix} 0 & M_I \\ M_I^T & 0 \end{pmatrix}. \tag{14}$$

One possibility is to compute the eigenvalues of matrix $A$ by the implicitly restarted Arnoldi method [13, 14]. It can be shown that the positive eigenvalues of $A$ are equal to the square roots of the eigenvalues of $M_I^T M_I$, and those square roots are equal to the singular values of $M_I$. Using an established algorithm we only need to provide a function applying the matrix $A$ to a vector $\mathbf{x}$ to build the needed Krylov subspace in order to determine the eigenvalues. This functions input arguments are a vector $\mathbf{x} \in \mathbb{R}^{r \cdot m_{out}+m_{in}}$ and a scalar $r$, which is equal to the number of used moments $r$, see (5). Output argument is a vector $\mathbf{y} \in \mathbb{R}^{r \cdot m_{out}+m_{in}}$,

$$A\mathbf{x} =: \mathbf{y} = ((y^1)^T, (y^2)^T, \ldots, (y^{r+1})^T)^T, \tag{15}$$

where for $i = 1, \ldots, r$

$$y^i = \begin{pmatrix} y_{(i-1)\cdot m_{out}+1} \\ \vdots \\ y_{i \cdot m_{out}} \end{pmatrix} \qquad \text{and} \qquad y^{r+1} = \begin{pmatrix} y_{r \cdot m_{out}+1} \\ \vdots \\ y_{r \cdot m_{out}+m_{in}} \end{pmatrix}. \tag{16}$$

Please note that we use the analog notation for the components $x^j$, $j = 1, \ldots, r+1$ of vector $\mathbf{x}$. If we insert (14) and (5) into (15) we get

$$\mathbf{y} = \begin{pmatrix} 0 & \begin{bmatrix} \mathbf{m_0} \\ \mathbf{m_1} \\ \vdots \\ \mathbf{m_{r-1}} \end{bmatrix} \\ \begin{bmatrix} \mathbf{m_0}^T & \mathbf{m_1}^T & \cdots & \mathbf{m_{r-1}^T} \end{bmatrix} & 0 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{r+1} \end{pmatrix}. \tag{17}$$

After a simple step of matrix multiplication we get the components $y^i$ for $i = 1, \ldots, r$ and $y^{r+1}$ of vector $\mathbf{y}$ as

$$y^i = \mathbf{m_{i-1}} x^{r+1} \qquad \text{and} \qquad y^{r+1} = \mathbf{m_0}^T x^1 + \cdots + \mathbf{m_{r-1}}^T x^r. \tag{18}$$

To compute these components efficiently we replace the block moments by their factors. In fact, we compute the $r+1$ parts of $\mathbf{y}$ by repeatedly applying the same factors to parts of $\mathbf{x}$, depending on whether it is a part of (18a) or (18b). We want to emphasize that we use the same factors each time. According to (2) the computation for (18a) follows Algorithm 6. The computation of (18b) is more involved, but follows the same recursive principle laid out in Algorithm 6. The computation of the decomposition of $M_O$ works analogously. These methods become numerically

**Algorithm 6** Computation of the components $y^i$

```
$a = Bx^{r+1}$
$a = G^{-1}a$
for $i = 1$ to $r$ do
    $y^i = La$
    $a = Ca$
    $a = -G^{-1}a$
end for
```

unstable for large $r$ but in practice $r$ often is small. For linear circuits with the same number of inputs and outputs, mostly one moment of the transfer function in (5), i. e., $r = 1$ so that we use the SVDMOR approach, is sufficient. Summarizing, this is a quite easy way which allows us to apply the SVD to large scale systems in a truncated way.

## 3 Numerical Results and Conclusions

The decay of the singular values of the moment used for computing the SVD is essential for (E)SVDMOR, so we firstly concentrate on this issue. Figure 1 shows the decrease of the 500 largest singular values of a circuit provided by the NEC Laboratories Europe, IT Research Division, NEC Europe Ltd. in St. Augustin, Germany. The circuit is called *circuit3* and consist of 3916 nodes, 1905 of them are terminals. We choose about 130 singular values to be significant based on the tolerance $\sigma_{r+1} < 10^{-2}\sigma_1$. That means, after the reduction we have 130 virtual input and output pins instead of 1905 terminals originally. Figure 2 shows the range of the 30 largest singular values of another circuit. It was provided by the Qimonda AG,

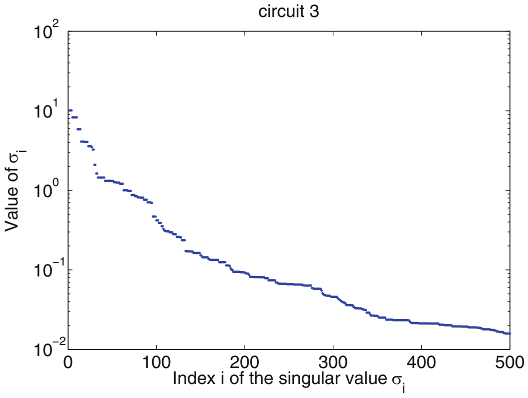


**Fig. 1:** Range of the largest 500 singular values of $\mathbf{m_0}$ of circuit *circuit3*

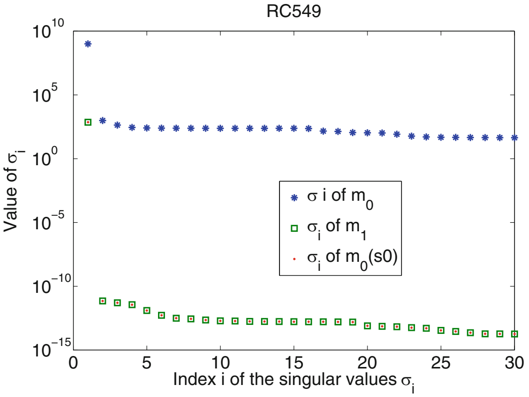


**Fig. 2:** Range of the largest 30 singular values of $\mathbf{m_0}$, $\mathbf{m_1}$ and $\mathbf{m_0}(s_0 = 10^8)$ of circuit *RC549*

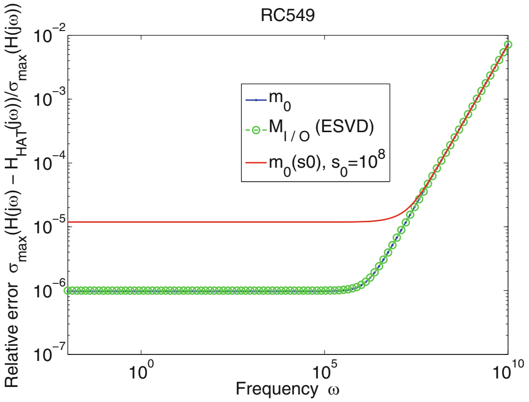


**Fig. 3:** Relative error $\varepsilon_{rel}$ of *RC549* by using SVDMOR with $\mathbf{m_0}$ and $\mathbf{m_0}(s_0 = 10^8)$ and ESVDMOR with $M_{I/O}$ consisting the information of the first 3 moments $\mathbf{m_0}$, $\mathbf{m_1}$, and $\mathbf{m_2}$

Munich, Germany. It is a test circuit called *RC549* and consists of 141 nodes and therefrom 70 terminals. Figure 2 points out clearly one significant singular value. Consequently, we reduce the system to one virtual terminal. The relative approximation error for circuit *RC549* is shown in Figure 3. We can observe that the error is sufficiently small up to the Gigahertz range which is enough for the application behind this problem (subcircuit of a memory chip).

Finalizing we would like to draw a few conclusions. If the pencil $sC + G$ of (1) is stable and a stability preserving MOR methods is used in (11), then the whole MOR algorithm described is stability preserving. Also, for typical classes of RLC circuits, the procedure is passivity preserving if the inner MOR method in (11) is. Due to space limitation, we will elaborate on this aspect elsewhere. In the future we want to present a global error bound as well as other approaches to perform the decomposition in (6) and (13) efficiently.

**Acknowledgements** The work reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF), grant no. 03BEPAE1. We also want to thank our industrial project partners Qimonda AG and NEC Laboratories Europe, IT Research Division, NEC Europe Ltd. Responsibility for the contents of this publication rests with the authors.

## References

1. Feldmann, P., Liu, F.: Sparse and efficient reduced order modeling of linear subcircuits with large number of terminals. In: ICCAD ’04: Proceedings of the 2004 IEEE/ACM Intl. Conf. Computer-aided design, pp. 88–92. IEEE Computer Society, Washington, DC, USA (2004)
2. Liu, P., Tan, S.X.D., Li, H., Qi, Z., Kong, J., McGaughy, B., He, L.: An efficient method for terminal reduction of interconnect circuits considering delay variations. In: ICCAD ’05: Proceedings of the 2005 IEEE/ACM International Conference on Computer-Aided Design, pp. 821–826. IEEE Computer Society, Washington, DC, USA (2005)
3. Liu, P., Tan, S.X.D., Yan, B., Mcgaughy, B.: An extended SVD-based terminal and model order reduction algorithm. In: Proceedings of the 2006 IEEE International, Behavioral Modeling and Simulation Workshop, pp. 44–49 (2006)
4. Liu, P., Tan, S.X.D., Yan, B., McGaughy, B.: An efficient terminal and model order reduction algorithm. Integr. VLSI J. **41**(2), 210–218 (2008)
5. Tan, S.X.D., He, L.: Advanced Model Order Reduction Techniques in VLSI Design. Cambridge University Press, New York, NY, USA (2007)
6. Tan, S.X.D., Shi, C.J.R., Lungeanu, D., Lee, J.C., Yuan, L.P.: Reliability-Constrained Area Optimization of VLSI Power/Ground Networks Via Sequence of Linear Programmings. In: in Proceedings of the ACM/IEEE Design Automation Conference, pp. 78–83 (1999)
7. Singh, J., Sapatnekar, S.: Congestion-aware topology optimization of structured power/ground networks. Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on **24**(5), 683–695 (2005)
8. Lin, Z., Carpenter, A., Ciftcioglu, B., Garg, A., Huang, M., Hui, W.: Injection-locked clocking: A low-power clock distribution scheme for high-performance microprocessors. IEEE Transactions on Very Large Scale Integration (VLSI) Systems **16**(9), 1251–1256 (2008)
9. Stoll, M.: A Krylov-Schur approach to the truncated SVD. Preprint, available at http://www.comlab.ox.ac.uk/files/721/NA-08-03.pdf (2008)
10. Hochstenbach, M.E.: A Jacobi–Davidson type SVD method. SIAM J. Sci. Comput. **23**(2), 606–628 (2001)
11. Berry, M.W., Pulatova, S.A., Stewart, G.W.: Algorithm 844: Computing sparse reduced-rank approximations to sparse matrices. ACM Trans. Math. Softw. **31**(2), 252–269 (2005)
12. Chan, T.F., Hansen, P.C.: Computing truncated singular value decomposition least squares solutions by rank revealing QR-factorizations. SIAM J. Sci. Stat. Comput. **11**(3), 519–530 (1990)
13. Sorensen, D.C.: Implicit application of polynomial filters in a k-step Arnoldi method. SIAM J. Matrix Anal. Appl. **13**(1), 357–385 (1992)
14. Lehoucq, R., Sorensen, D., Yang, C.: ARPACK user’s guide. Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods. Software - Environments - Tools, 6. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics. 142 p. (1998)

# Stability and Passivity of the Super Node Algorithm for EM Modeling of IC's

M.V. Ugryumova and W.H.A. Schilders

**Abstract** The super node algorithm performs model order reduction based on physical principles. Although the algorithm provides us with compact models, its stability and passivity have not thoroughly been studied yet. The loss of passivity is a serious problem because simulations of the reduced network may encounter artificial behavior which render the simulations useless. In this paper we explain why the algorithm delivers not passive reduced order models and present a way in order to overcome this problem.

## 1 Introduction

To increase their performance, the characteristic dimensions of interconnection systems are decreased and will decrease even further in the future. Higher speed makes the effect of higher frequency modes on the interconnection more important. Therefore, the analysis of the signal propagation on the interconnect system is important. However, this requires the solution of Maxwell's equations which is rather demanding from the point of view of which can hardly be used in conventional circuit simulators.

To be able to work with models for interconnect structures, a technique known as reduced order modeling is employed (for the various techniques, see [1]). One application where it is used is Fasterix. Fasterix is a layout simulation tool for electromagnetic behavior of interconnect systems such as PCBs, IC packages, filters and passive ICs [2]. As a first step in Fasterix a geometry preprocessor subdivides conductor into quadrilateral elements. In the lumped model derived directly from these elements, referred to here as the *original* (*full*) circuit model, the number of components in the circuit is of the order of the square of the number of elements.

M.V. Ugryumova, W.H.A. Schilders
Eindhoven University of Technology, Den Dolech 2 Postbus 513, 5600 MB Eindhoven, The Netherlands, e-mail: m.v.ugryumova@tue.nl

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_65,

However, this full circuit model is inefficient, because of computer memory and CPU limitations imply that the interconnect system cannot realistically be simulated. The principle model in Fasterix is a *reduced* circuit model, which is derived from the full model by the *super node algorithm*. Such model runs much faster and has been shown to be equally accurate in frequency domain. The algorithm employs a small subset of the original nodes, so called *super nodes* [2]. The number of supernodes depends on the user-defined maximum frequency, i.e. the highest frequency at which the model has to be valid.

The advantage of the super node algorithm is that it is inspired by physical insight into the models, and produces reduced RLC circuits depending on the maximum predefined frequency. Although the algorithm provides us with compact models, some of them suffer from instabilities which can be observed during time domain simulations. Therefore investigation of stability and passivity properties of the algorithm is primary important.

The paper is build up as follows. In section 2, 3 and 4, we briefly show the concept of the super node algorithm. In section 5 stability and passivity properties applied to the algorithm are discussed whereas in section 6 a technique to preserve passivity of the reduced models is presented. In the last section, a numerical example is considered.

## 2 Full and Reduced Order Models Used in Fasterix

Fasterix translates electromagnetic properties of the interconnect system into a full circuit model which is described by the system of Kirchhoff's equations [3]:

$$(\mathbf{R}+s\mathbf{L})I-\mathbf{P}V=0 \tag{1}$$

$$\mathbf{P}^T I+s\mathbf{C}V=J \tag{2}$$

where $\mathbf{R}\in\mathbb{R}^{\varepsilon\times\varepsilon}$ is the resistance matrix, $\mathbf{L}\in\mathbb{R}^{\varepsilon\times\varepsilon}$ is the inductance matrix, $\mathbf{P}\in\mathbb{R}^{\varepsilon\times\eta}$ is an incidence matrix, $\mathbf{C}\in\mathbb{R}^{\eta\times\eta}$ is the capacitance matrix, $I\in C^{\varepsilon}$ is a vector of currents flowing in the branches, $V\in C^{\eta}$ is a vector of voltages at the nodes. Vector $J\in C^{\eta}$ collects the terminal currents flowing into the interconnection system. Value $s$ is a complex number with negative imaginary part: $s=-j\omega$. Matrices **R**, **L**, **C** are symmetric and positive definite. Matrices **R**,**L**,**C**,**P** are calculated by Fasterix. Example of the circuit with $\eta=3$ and $\varepsilon=2$ is shown in Figure 1. Components $R_i$, $L_i$ and $C_{ij}$ are corresponding elements of the matrices **R**, **L** and **C**.

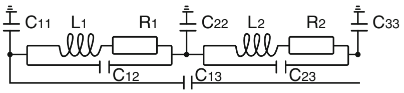


**Fig. 1** Example of the original RLC circuit described by (1)-(2)

From (1)-(2) one can obtain the voltage to current transfer with admittance matrix $\mathbf{Y} : \mathbb{C}^\eta \to \mathbb{C}^\eta$

$$J = \underbrace{\left(\mathbf{P}^T(\mathbf{R}+s\mathbf{L})^{-1}\mathbf{P}+s\mathbf{C}\right)}_{\mathbf{Y}(s)} V. \tag{3}$$

It simply says that if $V$ is given then $J$ can be calculated using $\mathbf{Y}(s)$ for some $s = s_0$. Admittance matrix $\mathbf{Y}(s)$ describes the behavior of the full circuit.

The goal is to obtain a circuit of order $\eta_1$ (preferably $\eta_1 \ll \eta$). The ports of the original model are kept in the reduced one. The original and reduced circuits should have approximately the same behavior at these ports.

In order to obtain admittance matrix of the reduced circuit, Fasterix subdivides the set of all nodes in the circuit into two subsets $N \in Z^{\eta_1}$ and $N' \in Z^{\eta_2}$. Evidently $\eta = \eta_1 + \eta_2$. Set $N$ contains super nodes, i.e. nodes which will be retained in the reduced circuit, and $N'$ contains other nodes. Due to this, vectors $V$, $J$ and matrices $\mathbf{P}$, $\mathbf{C}$ can be partitioned into blocks, see [2], [3] (chapter 8). Block matrix $\mathbf{P}_{N'}$ has full column rank. It is supposed that $J_{N'}$ consists of zeros.

If we consider the voltage in the super nodes as an input $V_N$, and currents flowing into the system through them as an output $J_N$, we come to the following system:

$$\left( \underbrace{\begin{pmatrix} \mathbf{R} & -\mathbf{P}_{N'} \\ \mathbf{P}_{N'}^T & \mathbf{0} \end{pmatrix}}_{\mathbf{G}} + s \underbrace{\begin{pmatrix} \mathbf{L} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{N'N'} \end{pmatrix}}_{\mathbf{C}} \right) x = \underbrace{\begin{pmatrix} \mathbf{P}_N \\ -s\mathbf{C}_{N'N} \end{pmatrix}}_{\mathbf{B}_i(s)} V_N, \tag{4}$$

$$J_N = \underbrace{\left( \mathbf{P}_N^T \; s\mathbf{C}_{N'N}^T \right)}_{\mathbf{B}_o^T(s)} x + s\mathbf{C}_{NN} V_N, \tag{5}$$

where $x = \left(I \;, V_{N'}\right)^T$. It should be noted that in (4) matrix $\mathbf{G}$ is positive real, and matrix $\mathbf{C}$ is positive semi-definite. From (4)-(5) it follows that $J_N$ is linearly related to $V_N$, i.e.

$$J_N = \underbrace{\left(\mathbf{B}_o^T(s)(\mathbf{G}+s\mathbf{C})^{-1}\mathbf{B}_i(s) + s\mathbf{C}_{NN}\right)}_{\mathbf{Y}_1(s)} V_N, \tag{6}$$

where $\mathbf{Y}_1(s)$ is admittance matrix of the reduced circuit. Expression (6) can be rewritten in the matrix form: $\mathbf{J}_N = \mathbf{Y}_1(s)\mathbf{V}_N$, where $\mathbf{V}_N = (\, V_N^1 \, \dots \, V_N^{\eta_1} \,)$ is a matrix of predescribed vectors of voltages and $\mathbf{J}_N = (\, J_N^1 \, \dots \, J_N^{\eta_1} \,)$ is a matrix of correspondent vectors of current. Further we assume that $\mathbf{V}_N$ is given and equals identity matrix. Therefore $\mathbf{J}_N = \mathbf{Y}_1(s)$.

In order to obtain the concrete RLC circuit described by $\mathbf{Y}_1(s)$, two approximations of $\mathbf{Y}_1(s)$ have to be performed. Derivation of them can be found in [3]. In this paper we will refer to them as $\mathbf{Y}_2(s)$ and $\mathbf{Y}_3(s)$. The last one will be considered in detail.

## 3 Admittance Matrix for the Full Frequency Range

In [3] the second approximation of $\mathbf{Y}_1(s)$ is constructed as

$$\mathbf{Y}_3(s) = \underbrace{\mathbf{P}_N^T \Psi \left(\Psi^T (\mathbf{R} + s\mathbf{L})\Psi\right)^{-1} \Psi^T \mathbf{P}_N}_{\mathbf{Y}_{RL}(s)} + s\mathbf{Y}_C, \tag{7}$$

where $\Psi$ is a null space of $\mathbf{P}_N^T$. Term $\mathbf{Y}_{RL}(s)$ stays for the contribution of resistances and inductances in the circuit. Term $s\mathbf{Y}_C$ comes from the high frequency range approximation and stays for the capacitance contribution [3]. $\mathbf{Y}_{RL}(s)$ can be presented in the pole-residue form as

$$\mathbf{Y}_{RL}(s) = \sum_{i=1}^{n} \frac{\mathbf{H}_i}{(s-\lambda_i)} = \sum_{i=1}^{n} \frac{\left(\Psi^T \mathbf{P}_N x_i\right)\left(y_i^* \mathbf{P}_N^T \Psi\right)}{(s-\lambda_i)}, \; n = \varepsilon - \eta_2. \tag{8}$$

where $\lambda_i$ are the eigenvalues of the matrix pencil $(\Psi^T \mathbf{L}\Psi, -\Psi^T \mathbf{R}\Psi)$. Since $\Psi^T \mathbf{L}\Psi$ and $\Psi^T \mathbf{R}\Psi$ are positive definite then $\lambda_i \in \mathbb{R}$ and $\lambda_i < 0$. $y_i$, $x_i \in \mathbb{R}^{\eta_1}$ are left and right eigenvectors respectively [4].

## 4 Realization

In this section we will show how $\mathbf{Y}_3(s)$ in (7) can be translated into RLC circuit. The network described by $\mathbf{Y}_3(s)$ has branches between all nodes and ground and between all nodes. Each branch is calculated as follows [5]. Branch between node $i$ and ground:

$$\mathbf{y}_{3,ii} = \sum_{j=1}^{n} \mathbf{Y}_{3,ij}. \tag{9}$$

Branch between node $i$ and node $j$:

$$\mathbf{y}_{3,ij} = -\mathbf{Y}_{3,ij}, i \neq j. \tag{10}$$

All elements of $\mathbf{Y}_3(s)$ have the same poles $\lambda_i$, and these become the poles for the network branches when calculated by (9) and (10). Each branch in (9) and (10) is given as a rational function $\sum_{i=1}^{n} \frac{c_i}{s-\lambda_i} + se$. Using Foster's canonical form [5], the branch can be represented by an electrical network as shown in Figure 2. $C$, $R_i$, $L_i$ are calculated as $C = e$, $R_i = -\lambda_i / c_i$, $L_i = 1/c_i$. Similar to the above, symmetric admittance matrix can be realized exactly by using a $\Pi$-structure template [6]. An example of the $\Pi$-structure template is shown in Figure 3, where each branch admittance is realized by the Foster's canonical form shown in Figure 2.

However Fasterix does not use straightforwardly this way of realization. Since calculation of all eigenvalues $\lambda_i$ in (8) may be time consuming process, Fasterix first approximates $\mathbf{y}_3(s)$ with $m$ $(m < n)$ terms. It is done as following. The set

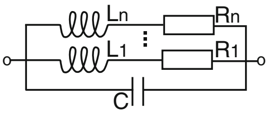


**Fig. 2:** Synthesis by electrical network

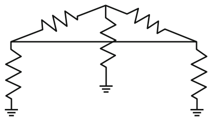


**Fig. 3:** A tree-port realization of the admittance matrix 3 by 3 based on $\Pi$-structure

of $m+1$ match frequencies, $s_k$, is chosen. This set consists of some large negative values between maximum predefined frequency $-\Omega$ and $-\max(\lambda_i)$, and some small negative values between $-\min(\lambda_i)$ and 0. For each $s_k$, corresponding admittance matrix has to be calculated. Elements of $\mathbf{Y}_3(s)$ approximate elements of $\mathbf{Y}_2(s)$ in frequency domain well therefore $\mathbf{Y}_2(s_k)$ instead of $\mathbf{Y}_3(s_k)$ can be used.

Solving the following set of $m+1$ equations

$$s_k \mathbf{y}_{C,ij} + \sum_{l=1}^{m} \frac{\widetilde{\mathbf{H}}_{l,ij}}{(s_k - \lambda_l)} = \mathbf{y}_{2,ij}(s_k),\ k = 1,...m+1. \tag{11}$$

for the coefficients $\mathbf{y}_{C,ij}$ and $\widetilde{\mathbf{H}}_{l,ij}$ is equivalent to determine the approximation of $\mathbf{y}_3(s)$ with $m < n$ terms. Like it was shown above, the reduced circuit consists of branches between every pair of circuit nodes. Each branch consists of $m$ parallel connections of a series resistor $R$ and inductor $L$, in parallel with a capacitor $C$. Thus for the branch between the circuit nodes $i$ and $j$

$$R_l = -\lambda_l \widetilde{\mathbf{H}}_{l,ij}^{-1},\ L_l = \widetilde{\mathbf{H}}_{l,ij}^{-1},\ C = \mathbf{y}_{C,ij}. \tag{12}$$

Evidently $m$ influences at the computational time of simulations. Fasterix chooses $m$ depending on the size of the model. Usually $m \leq 8$. For carrying out simulations of the circuit we used PSTAR which is the Philips circuit simulator program.

## 5 Stability and Passivity

Circuits constructed using rational functions need to satisfy the stability and passivity conditions for a linear time-invariant passive system. The stability condition requires that for a stable system, the output response be bounded for a bounded input excitation [7]. Hence, the rational function representing a stable system has to satisfy the following stability conditions: (1) the poles lie on the left half of the $s$ plane; (2) the rational function does not contain multiple poles along the imaginary axis of the $s$ plain.

The passivity condition requires that a passive circuit does not create energy. Since non-passive models combined with a stable circuit can generate an unstable time-domain response, this condition becomes important when model need to be combined with other circuit for time-domain simulations.

Passivity is closely related to positive realness of the admittance matrix. The admittance matrix $\mathbf{Y}(s)$ is positive real if (1) $\mathbf{Y}(s)$ is analytic for all $s$ with $Re(s) > 0$, (2) $\mathbf{Y}^*(s) = \mathbf{Y}(\bar{s})$ for all $s \in \mathbb{C}$, and (3) $\mathbf{Y}(s) + \mathbf{Y}^*(s) \geq 0$ for all $s$ with $Re(s) > 0$.

Condition (1) means that the system is stable. Condition (2) refers to the system that has real response. And condition (3) is equivalent to that the real part of $\mathbf{Y}(s)$ is a positive semidefinite matrix at all frequencies.

In the super node algorithm, admittance matrix plays a role of a system function. Notice that $\mathbf{Y}_3(s)$ in (7) is stable (all poles $\lambda_i < 0$) but not positive real since $\mathbf{Y}_C$ is an indefinite matrix. However the following theorem holds.

**Theorem 1.** *Admittance matrix $\mathbf{Y}_{RL}(s)$ in (7) is positive real.*

**Proof.** In section 3 it was shown that all poles $\lambda_i < 0$ therefore the system is stable. It is trivial to check out the second condition of positive realness. Let $\mathbf{B}^T = \mathbf{P}_N^T \Psi$. We will show that the third one is satisfied:

$$\mathbf{Y}_{RL}^*(s) + \mathbf{Y}_{RL}(s) = \mathbf{B}^T \left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right)^{-*} \mathbf{B} + \mathbf{B}^T \left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right)^{-1} \mathbf{B} = \tag{13}$$

$$= \mathbf{B}^T \left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right)^{-*} \left(\left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right) + \left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right)^*\right) \left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right)^{-1} \mathbf{B} =$$

$$= \mathbf{y}^* \left(\left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right) + \left(\tilde{\mathbf{R}} + s\tilde{\mathbf{L}}\right)^*\right) \mathbf{y},$$

with $\mathbf{y} = (\tilde{\mathbf{R}} + s\tilde{\mathbf{L}})^{-1}\mathbf{B}$. Thus it is sufficient to prove the positive realness for $\mathbf{W}(s) = \tilde{\mathbf{R}} + s\tilde{\mathbf{L}}$. For $s = \sigma + i\omega$ with $\sigma > 0$ we have:

$$\mathbf{W}^*(s) + \mathbf{W}(s) = (\tilde{\mathbf{R}} + s\tilde{\mathbf{L}})^* + \tilde{\mathbf{R}} + s\tilde{\mathbf{L}} = 2\tilde{\mathbf{R}} + 2\sigma\tilde{\mathbf{L}},$$

which is nonnegative definite. Thus, $\mathbf{Y}_{RL}(s)$ is positive real. ■

It is known [6] that a *Π*-structure template for realization of positive real admittance matrix guarantees construction of the passive circuit. However the important observation is that in the super node algorithm realization by the *Π*-structure template is applied to the approximation of $\mathbf{Y}_3(s)$ at a few frequency points $s_k$ and not directly to $\mathbf{Y}_3(s)$. So if $\mathbf{Y}_3(s)$ was positive real, the constructed RLC circuit might not be passive. In the next section, a way to obtain positive real $\mathbf{Y}_3(s)$ will be suggested.

## 6 Passivity Enforcement

In this section we present a technique in order to obtain positive real $\mathbf{Y}_3(s)$ which is efficient for the further realization. If both terms in (7) are positive real then $\mathbf{Y}_3(s)$ is positive real as well.

First we consider the term $s\mathbf{Y}_C$. Matrix $\mathbf{Y}_C$ is indefinite. Following the eigendecomposition $\mathbf{Y}_C = \mathbf{V}\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\eta_1})\mathbf{V}^{-1}$, all negative eigenvalues are set to zero. Subsequently, the matrix is reconstructed through the operation
$\tilde{\mathbf{Y}}_C = \mathbf{V}\mathrm{diag}(\tilde{\sigma_1}, \tilde{\sigma_2}, \ldots, \tilde{\sigma_{\eta_1}})\mathbf{V}^{-1}$ where the modified quantities are denoted with "~". This procedure allows us to get $\mathbf{Y}_C$ positive definite and positive real $s\mathbf{Y}_C$.

Above it was shown that $\mathbf{Y}_{RL}$ is positive real. However the number of terms in $\mathbf{Y}_{RL}(s)$ is related to the number of RL elements in the circuit as $O(n\eta_1^2)$. Taking it into account, we are interested to obtain an efficient approximation of $\mathbf{Y}_{RL}(s)$ which consists of $k < n$ terms and determines the effective admittance function behavior. Positive realness of the new approximation must be preserved. One effective way to achieve it is to use modal approximation [4]. Modal approximation requires selection of dominant eigenvalues and these can be computed via full null space methods (QR, QZ) or iterative subspace methods [4].

A pole $\lambda_j$ that corresponds to a residue $\mathbf{H}_j$ with relatively large $||\mathbf{H}_j||_2/|Re(\lambda_j)|$ is called a dominant pole, i.e. a pole that is well observable and controllable in the admittance function. In our case all $\lambda_i$ are real and negative. An approximation of $\mathbf{Y}_{RL}(s)$ that consists of $k < n$ terms with $||\mathbf{H}_j||_2/|Re(\lambda_j)|$ above some value, determines the effective admittance function behavior [4]:

$$\tilde{\mathbf{Y}}_{RL}(s) = \sum_{i=1}^{k} \frac{\mathbf{H}_i}{s-\lambda_i}. \tag{14}$$

Since $\lambda_i < 0$ and $\mathbf{H}_i = (\mathbf{P}_N^T \Psi x_i)(y_i^* \Psi^T \mathbf{P}_N) > 0$, with $x_i = y_i$, then it follows that (14) is positive real. Thus applying a $\Pi$-structure template for realization of $\tilde{\mathbf{Y}}_3(s) = \tilde{\mathbf{Y}}_{RL}(s) + s\tilde{\mathbf{Y}}_C$ ensures construction of passive RLC circuit.

## 7 Numerical Example

Fasterix model consists of two printed striplines, which are parallel to each other. The striplines are 1 mm wide and the length is 15 mm. For the maximum frequency 5 GHz , Fasterix generates mesh with 28 elements. Then this model is interpreted as a full RLC circuit with $\eta = 28$ nodes and $\varepsilon = 26$ RL-branches. In order to build reduced circuit, Fasterix chooses 15 super nodes and applies the super node algorithm.

For transient analysis, a trapezoidal pulse having rise/fall times of 1 ps and pulse width of 1 ns is applied to the pins of the lower strip. A 50 $\Omega$ resistor $R_{out}$ is connected between two ports of the upper strip. The voltage is measured over $R_{out}$ and regarded as output.

The transient response at the resistor $R_{out}$ is given in Figure 4. It can be seen that the time response is unstable since initially the super node algorithm does not preserve passivity. However, the super node algorithm with proposed passivity enforcement preserves passivity. Shown in Figure 5 the two waveforms of the original and reduced circuits match very well. Table 1 shows a comparison between original and reduced models. The reduced model has large amount of RLC elements. Nevertheless, when the original circuit is of high order, the simulation time is reduced. This happens because the number of mutual inductances is zero. For this particular example $\mathbf{Y}_{RL}(s)$ contains $n = 25$ terms and it was truncated till $k = 4$ terms with the most dominant poles.

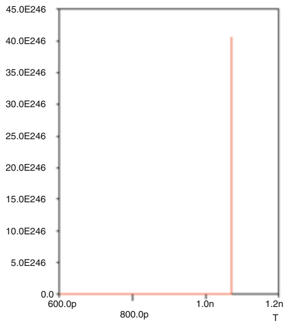


**Fig. 4:** Simulation in time domain

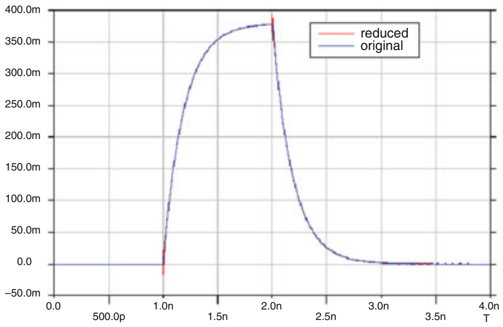


**Fig. 5:** Comparison of the original and reduced models

**Table 1:** Comparison of the original and the reduced models

| System | Dimension | $R$ | $L$ | $C$ | $L_{mutual}$ |
|---|---|---|---|---|---|
| Original | 28 | 26 | 26 | 91 | 245 |
| Reduced | 15 | 420 | 420 | 120 | 0 |

# 8 Conclusions

In this paper an overview of a reduction technique, the super node algorithm, used in the EM tool Fasterix has been presented. This algorithm delivers stable models, however we have shown that passivity is not preserved. As a remedy, a technique for passivity enforcement based partly on the modal approximation was introduced. Realization was performed by using a $\Pi$-structure template. This strategy solves the problem of preserving passivity. However the time complexity of the modified version of the super node algorithm still needs to be investigated.

# References

1. Schilders, W.H.A., van der Vorst, H., Rommes, J.: Model Order Reduction. Theory, Research Aspects and Applications. Springer, Berlin (2008)
2. Cloux, R.D., Maas, G.P.J.F.M., Wachters, A.J.H.: Quasi-static boundary element method for electromagnetic simulations of pcbs. Philips J. Res. **48**, 117–144 (1994)
3. Schilders, W.H.A., ter Maten, E.J.W.: Special volume : numerical methods in electromagnetics. Elsevier, Amsterdam (2005)
4. Rommes, J.: Methods for eigenvalue problems with application in model order reduction. Ph.D. thesis, Universiteit Utrecht, Utrecht (2007)
5. Gustavsen, B.: Computer code for rational approximation of frequency dependent admittance matrices. IEEE Transactions on Power Delivery **17**, 1093–1098 (2002)
6. Liu, P., Qi, Z., Tan, S.X.D.: Passive hierarchical model order reduction and realization of rlcm. In: Proc. 6th Int. Symp. on Quality Electronic Design (ISQED’05) (2005)
7. Kuo, F.F.: Network analysis and synthesis. Wiley, New York (1962)

# Hierarchical Model-Order Reduction Flow

Mikko Honkala, Pekka Miettinen, Janne Roos, and Carsten Neff

**Abstract** This paper presents a hierarchical model-order reduction (HMOR) flow, where the linear parts of a hierarchically defined circuits are divided into independently reducable subcircuits. The impact of the hierarchical structure and circuit partitioning on two MOR methods is discussed and some simulation results are presented.

## 1 Introduction

In this paper, a Hierarchical MOR (HMOR) method for very large linear blocks of nonlinear circuits is considered. In practice, such large linear blocks arise, e.g., from interconnect models. The HMOR approach proposed fully utilizes and preserves the hierarchy of the SPICE netlist defined by the designer and, in addition, further divides user-defined subcircuits into smaller subcircuits that can be independently reduced and then put together. The benefits of HMOR are:

1. Very large circuits can be reduced with limited computer resources.
2. Circuit partitioning makes possible to apply parallel processing in a natural way.
3. The computational cost of reduction can be minimized by reducing repeated structures (same linear subcircuit used several times in the overall circuit) only once.
4. The most suitable MOR method can be chosen for each subcircuit independently.

Mikko Honkala, Pekka Miettinen, Janne Roos
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland, e-mail: mikko.a.honkala@tkk.fi, pekka.miettinen@tkk.fi, janne.roos@tkk.fi

Carsten Neff
NEC Laboratories Europe, Rathausallee 10, 53757 St. Augustin, Germany, e-mail: neff@it.neclab.eu

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_66,

The idea of using hierarchy in MOR is not new and has been studied previously, e.g., in Refs. [1, 2]. In this paper, it is especially shown how to use different RC and RLC MOR methods for different subcircuits and what the impact of the circuit partitioning is on two MOR methods.

## 2 Hierarchical MOR Flow

Here, a hierarchical netlist-in–netlist-out MOR flow is presented. The flow utilizes hMETIS [3] graph-partitioning algorithms that are an extension of METIS [4] algorithms for circuit partitioning and, next, uses the MOR methods in a hierarchical manner. A suitable MOR method can be then applied to the different types of subcircuits. The MOR methods considered here are PRIMA [5], a Krylov subspace method for RLC circuits, and (modified) Liao–Dai [6], a low-order RC macromodel method. Of course, there are many other MOR methods that can be applied in this flow, e.g., [7] for RL subcircuits, and [8] for R subcircuits.

The HMOR flow proposed is briefly outlined in the following, but each step is discussed in more detail in Sections 3 and 4.

1. *Netlist parsing*: extract the graph representations of all the RLC circuits from the 'messy' hierarchical SPICE netlist containing linear and nonlinear elements.
2. *Circuit partitioning*: Separate disconnected parts, divide each RLC graph (using hMETIS) into appropriate subcircuit graphs, and map the graphs back onto circuit netlists.
3. *Matrix construction*: build, for each subcircuit, **G**, **C**, and **B**, the conductance, capacitance, and selector matrix of the MNA formulation, respectively.
4. *Model-order reduction*: use an appropriate MOR method for each subcircuit; here, PRIMA for RLC blocks and Liao–Dai for RC blocks.
5. *Macromodel realization*: synthesize each reduced subcircuit using resistors, capacitors, and, if needed, voltage-controlled current sources (VCCSes).
6. *Netlist reconstruction*: include each macromodel in the proper position in the final netlist to achieve the original hierarchical structure.

Note that now the original hierarchy is preserved. All the sources and nonlinear elements are untouched, and only the linear parts are reduced.

## 3 Hierarchy and Circuit Partitioning

### 3.1 Circuit Hierarchy

Consider a netlist consisting of linear and nonlinear components defined in hierarchical subcircuits (see Fig. 1). The netlist has, in addition to the main-level circuit (treated as a subcircuit equal to other subcircuits in the following), five different subcircuits, from which the subcircuit 4 is used twice.

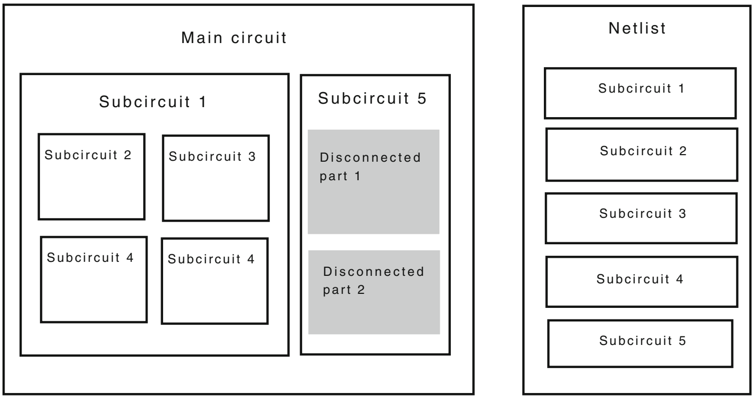


**Fig. 1:** Hierarchical structure of a circuit and the netlist of the corresponding circuit

Each subcircuit may have sources, nonlinear components (MOSFETs, diodes, etc.), and linear RLC components, and can be placed at any level of hierarchy. The netlist usually contains some plot and analysis commands, too. The HMOR flow automatically extracts the linear parts from each subcircuit separately, such that those nodes connected to nonlinear elements, sources, or user defined subcircuit ports, or are listed in plot commands are considered as external port nodes.

The subcircuit may have several disconnected RLC parts, e.g. two different RLC interconnections before and after nonlinear transistor circuit that are not connected to each other (e.g., see subcircuit 5 in Fig. 1). If disconnected parts are not separated into different subcircuits, numerical problems may arise, e.g. nonzero components are produced between disconnected parts. Even if no problems would occur, the disconnection is, in any case, a natural location for further partitioning.

If the netlist is processed hierarchically, each subcircuit needs to be reduced only once, compared to a typical approach, where the whole netlist is first flattened with all the hierarchical structures written out for each reference. For example, subcircuit 4 in Fig. 1 needs to be reduced only once. The same subcircuit can be placed several times in different levels of hierarchy.

## 3.2 *Circuit Partitioning*

In the HMOR-approach proposed, circuit partitioning is applied to each RLC block extracted from user-defined subcircuits. This partitioning has two significance: 1) large subcircuit is divided into smaller ones, such that it can be handled more easily, 2) some MOR methods are based on partitioning, like, e.g., Liao–Dai.

For methods for which partitioning is not an essential part of the algorithm (e.g., PRIMA), the partitioning is useful for very large subcircuits in order to be able to reduce them at all with limited computer resources.

The goal of the circuit partitioning regarding MOR is to obtain such subcircuits that have a large number of internal nodes, $n_\mathrm{i}$, compared to external port nodes, $n_\mathrm{e}$. Partitionings fulfilling this criterion are best suited for MOR, since the voltages of internal nodes are not of interest, and they may be reduced. Full elimination of internal nodes may destroy sparsity of the reduced model.

The MOR flow uses hMETIS [3] graph-partitioning algorithms for circuit partitioning. METIS [4] is an algorithm package for partitioning large irregular graphs and large meshes and for computing fill-in reducing orderings of sparse matrices. hMETIS [3] is an extension of METIS, which uses hypergraphs instead of graphs.

# 4 MOR Methods

This section briefly presents the two methods (PRIMA and Liao–Dai) that are used within the HMOR flow, and their applicability in a hierarchical manner is discussed.

In practice, of these methods PRIMA is suitable for RLC circuits while Liao–Dai is a plain RC MOR method. In theory, if there are only a few inductances (e.g., 10 inductances compared to 10000 capacitances and resistances) in the subcircuit, they can be left out from the linear part to be reduced, and, then, RC MOR methods can be used for the remaining RC-only part. However, omitting the inductances increases the number of terminals, making the overall process more complicated.

## *4.1 PRIMA*

The passive reduced-order interconnect macromodeling algorithm (PRIMA) [5] is based on the block Arnoldi algorithm and employs congruence transformations to project a large system of equations onto a smaller subspace, so that passivity is preserved during reduction. To this end, PRIMA uses the Arnoldi iteration as a numerically stable method of generating the Krylov subspace to match $\lfloor q/N \rfloor$ block moments of the $N$-port, where $q$ is the order of reduction.

After the circuit division, the circuit equations of each subcircuit are needed. Both PRIMA and Liao–Dai operate on the modified nodal analysis (MNA) equations.

The MNA equations of an $N$-port can be expressed as follows:

$$\begin{cases} \mathbf{C}\dfrac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = -\mathbf{G}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \\ \mathbf{i}_\mathrm{p}(t) = \mathbf{L}^\mathrm{T}\mathbf{x}(t), \end{cases} \tag{1}$$

where $\mathbf{x}(t)$ contains nodal voltages and branch currents of ports and inductances ($\mathbf{x}(0) = 0$), $\mathbf{u}$ and $\mathbf{i}_\mathrm{p}$ denote the port voltages and currents. $\mathbf{B} = \mathbf{L}$ where $\mathbf{B} \in \Re^{n\times N}$

is a selector matrix consisting of ones, minus ones and zeroes. $n$ is the total number of unknowns.

$$\mathbf{C} \equiv \begin{bmatrix} \mathbf{Q} & 0 \\ 0 & \mathbf{H} \end{bmatrix}, \ \mathbf{G} \equiv \begin{bmatrix} \mathbf{N} & \mathbf{E} \\ -\mathbf{E}^{\mathrm{T}} & 0 \end{bmatrix}, \ \mathbf{x} \equiv \begin{bmatrix} \mathbf{v} \\ \mathbf{i} \end{bmatrix}. \tag{2}$$

$\mathbf{N}$, $\mathbf{Q}$, and $\mathbf{H}$ are symmetric non-negative definite matrices containing the stamps from resistors, capacitors, and inductances, respectively. Vector $\mathbf{v}$ is the nodal voltage vector and $\mathbf{i}$ contains the branch currents of ports and inductances. The matrices $\mathbf{G} \in \Re^{n\times n}$ and $\mathbf{C} \in \Re^{n\times n}$.

Define $\mathbf{A} \equiv -\mathbf{G}^{-1}\mathbf{C}$ and $\mathbf{R} \equiv \mathbf{G}^{-1}\mathbf{B}$. Taking the Laplace transformation of (1) and solving for the port current variables, the $y$-parameter matrix $\mathbf{Y}(s)$ is

$$\mathbf{Y}(s) = \mathbf{L}^{\mathrm{T}}(\mathbf{I} - s\mathbf{A})^{-1}\mathbf{R}, \tag{3}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. The block moments of $\mathbf{Y}(s)$ are defined as the coefficients of the Taylor expansion of $\mathbf{Y}$ around $s = 0$:

$$\mathbf{Y}(s) = \mathbf{M}_0 + \mathbf{M}_1 s + \mathbf{M}_2 s^2 + \cdots. \tag{4}$$

The block moments can be computed using the relation

$$\mathbf{M}_i = \mathbf{L}^{\mathrm{T}}\mathbf{A}^i\mathbf{R}. \tag{5}$$

PRIMA transforms (1) into

$$\begin{cases} \tilde{\mathbf{C}}\dfrac{\mathrm{d}\tilde{\mathbf{x}}(t)}{\mathrm{d}t} = -\tilde{\mathbf{G}}\tilde{\mathbf{x}}(t) + \tilde{\mathbf{B}}\mathbf{u}(t), \\ \mathbf{i}(t) = \tilde{\mathbf{L}}^{\mathrm{T}}\tilde{\mathbf{x}}(t), \end{cases} \tag{6}$$

where

$$\tilde{\mathbf{C}} = \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}, \ \tilde{\mathbf{G}} = \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}, \ \tilde{\mathbf{B}} = \mathbf{X}^{\mathrm{T}}\mathbf{B}, \ \tilde{\mathbf{L}} = \mathbf{X}^{\mathrm{T}}\mathbf{L}. \tag{7}$$

These types of transformations are known as congruence transformations. The matrix $\mathbf{X}$ is an $n \times q$ matrix, which is obtained after $q/N + 1$ iterations of the block Arnoldi algorithm (the extra step is not necessary if $q/N$ is an integer).

The reduced MNA equations can be synthesized with various macromodels [9]. This work uses the macromodel proposed by Matsumoto [10].

## 4.2 Liao–Dai Method

The detailed description of the Liao–Dai method is presented in [6] and our modified version in [11]. Here, only a brief description of the method is given.

In principle, the Liao–Dai method can be divided into three steps.

1. Divide the circuit into smaller subcircuits. (HMOR divides each extracted RC subcircuit, because the partitioning is an integral part of the method.)
2. Compute the first two moments, $\mathbf{M}_0$ and $\mathbf{M}_1$, of $y$-parameters of each subcircuit.

3. Realize the $y$-parameters of each subcircuit by matching the moments of the low-order RC macromodel that preserves the two first moments.

The the low-order macromodel from port to port is such an RC-circuit that it match only the two first moment, i.e. its size is constant. Therefore, if the subcircuits to be reduced are small enough (the same order as the RC macromodels) once they are recombined, the overall behavior of the circuit is preserved. On the other hand, if the subcircuits to be reduced are too large, much precision is lost, i.e. the error of reduction is controlled by the size of partition. Thus, the quality of partitioning is crucial to the MOR method. Since only the first two block moments are used in the reduction, the method supports RC circuits only.

The reduction produces RC circuits with positive element values and thus the macromodels are passive and stable.

In this paper, a modified version [11] of Liao–Dai method is used: the S-parameter-based circuit partitioning algorithm used in the original Liao–Dai method (see [6] for details) was replaced with the hMETIS algorithm, and the moment computation of $y$-parameters is calculated as presented in (5).

## 5 Simulation Examples

The HMOR flow has been implemented using C and MATLAB. In the following, the effect of circuit partitioning on PRIMA and the Liao–Dai method is studied. Even though the program is capable of using different methods for each subcircuit, only one method is used for all. The simulations are run using APLAC circuit simulator.

To test the partitioning with the PRIMA algorithm, the three port RLC circuit [9] with 1081 nodes is reduced. The linear part has three external port nodes. The circuit was divided with three different partitionings. Each subcircuit is reduced with three different orders of reduction.

The partitioning is chosen to produce approximately equal sized partitions, such that the ratio $n_\mathrm{e}/n_\mathrm{i}$ is small. In some cases, the equal sized partitioning produces a poor ration, but here this is not the case. The size of a partition is defined by the number of components per partition. Little deviation from this number is possible.

The original/reduced circuit was run using AC analysis with a frequency sweep from 1 Hz to 1 GHz. The results obtained are listed in Table 1, where $q$, $N_\mathrm{p}$, R, C, L, VCCS are the order of reduction, number of partitions, number of resistors, capacitors, inductors, and VCCSes, respectively. Furthermore, $E_\mathrm{ac}$/%, $T_\mathrm{ac}$/s, and $t_\mathrm{ac}$ denote the normalized AC-analysis error, AC-analysis CPU time, and normalized AC-analysis time, respectively. The error is calculated as follows:

$$E_\mathrm{ac} = 100\% \cdot \sqrt{\frac{1}{n_\mathrm{samp}N_\mathrm{out}} \sum_{k=1}^{n_\mathrm{samp}} \sum_{i=1}^{N_\mathrm{out}} \left( \frac{u_i^k - d_i^k}{d_{\mathrm{max},i} - d_{\mathrm{min},i}} \right)^2}, \tag{8}$$

where $n_{\mathrm{samp}}$ and $N_{\mathrm{out}}$ stand for the number of samples of output voltages and number of output ports. $d$, $d_{\max}$, and $d_{\min}$ are the desired output voltages and their maximum and minimum values, respectively. In these simulations $n_{\mathrm{samp}} = 301$.

**Table 1:** PRIMA results for RLC circuit with partitioning

| $q$ | $N_{\mathrm{p}}$ | R | C | L | VCCS | $E_{\mathrm{ac}}/\%$ | $T_{\mathrm{ac}}/\mathrm{s}$ | $t_{\mathrm{ac}}$ |
|---|---|---|---|---|---|---|---|---|
| Orig | - | 363 | 369 | 360 | - | - | 0.86 | 1.00 |
| 6 | 1 | 6 | 7 | - | 36 | 13.417 | 0.15 | 0.17 |
|  | 3 | 14 | 18 | - | 68 | 15.269 | 0.16 | 0.19 |
|  | 5 | 30 | 39 | - | 188 | 7.347 | 0.21 | 0.24 |
| 10 | 1 | 10 | 13 | - | 61 | 2.835 | 0.16 | 0.19 |
|  | 3 | 30 | 43 | - | 141 | 0.399 | 0.19 | 0.22 |
|  | 5 | 50 | 69 | - | 301 | 1.951 | 0.27 | 0.31 |
| 30 | 1 | 30 | 40 | - | 180 | 0.009 | 0.21 | 0.24 |
|  | 3 | 90 | 131 | - | 420 | 0.001 | 0.34 | 0.40 |
|  | 5 | 150 | 219 | - | 900 | 0.001 | 0.61 | 0.75 |

As can be seen from Table 1, the simulation time increases with the number of partitions, and it seems that the error grows with number of partitions although the simulation time remains about the same.

The Liao–Dai algorithm is tested with a industrial RC circuit having 12 external port nodes and 1525 internal nodes. The results with several partitions are presented in Table 2. Also, some reference simulations with PRIMA reduced circuits are presented in Table 3. The comparison of results in Table 2 and Table 3 shows that with the same simulation time the RC reduction method gives a smaller simulation error than PRIMA.

**Table 2:** Liao–Dai results for RC circuit with several partitions

| $N_{\mathrm{p}}$ | $n$ | R | C | $E_{\mathrm{ac}}/\%$ | $T_{\mathrm{ac}}/\mathrm{s}$ |
|---|---|---|---|---|---|
| Orig. | 1537 | 10432 | 197 | - | 10.29 |
| 106 | 556 | 8515 | 573 | 1.083 | 12.49 |
| 53 | 237 | 3261 | 255 | 1.078 | 3.16 |
| 17 | 66 | 516 | 88 | 0.837 | 0.39 |
| 13 | 34 | 137 | 54 | 0.839 | 0.24 |
| 1 | 15 | 44 | 29 | 3.085 | 0.22 |

## 6 Conclusions

The hierarchical MOR flow was presented. It was shown how to reduce very large circuits using circuit partitioning, where the linear parts of the hierarchically defined

**Table 3:** PRIMA results for RC circuit (no partitioning)

| $q$ | $n$ | R | C | VCCS | $E_{ac}/\%$ | $T_{ac}/s$ |
|---|---|---|---|---|---|---|
| Orig. | 1537 | 10432 | 197 | - | - | 10.29 |
| 15 | 30 | 27 | 17 | 360 | $9.0 \cdot 10^{15}$ | 0.30 |
| 20 | 35 | 32 | 24 | 480 | 40.8 | 0.35 |
| 25 | 40 | 37 | 34 | 600 | 9.9 | 0.40 |

circuit are extracted, partitioned, and then reduced with appropriate methods. The reduced subcircuits are then put together so that the original hierarchy is preserved. The flow preserves the stability and passivity of linear circuits. Without partitioning the circuit may be so large that it is impossible to manage the computational cost. The results showed how hierarchy and partitioning affects the performance of two MOR methods. The Liao–Dai method is a circuit division oriented RC reduction method and, thus, very suitable for HMOR. The circuit division makes it possible to apply PRIMA to very large circuits.

## References

1. Lee, Y.M., Chen, C.C.P.: Hierarchical model order reduction for signal-integrity interconnect synthesis. In: Proceedings of the 11th Great Lakes symposium on VLSI, pp. 109–114 (2001)
2. Lee, Y.M., Cao, Y., Chen, T.H., Wang, J.M., Chen, C.C.P.: HiPRIME: hierarchical and passivity preserved interconnect macromodeling engine for rlck power delivery. IEEE Transactions on Computer-Aided design of Integrated Circuits and Systems **24**, 797–806 (2005)
3. Karypis, G., Kumar, V.: hMETIS, a hypergraph partitioning package version 1.5.3
4. Karypis, G., Kumar, V.: METIS, a software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices version 4.0
5. Odabasioglu, A., Celik, M., Pileggi, L.T.: PRIMA: passive reduced-order interconnect macromodeling algorithm. IEEE Transactions on Computer-Aided design of Integrated Circuits and Systems **17**, 645–654 (1998)
6. Liao, H., Dai, W.W.M.: Partitioning and reduction of RC interconnect networks based on scattering parameter macromodels. In: Digest of Technical Papers of IEEE/ACM International Conference on Computer Aided Design, pp. 704–709 (1995)
7. Miettinen, P., Honkala, M., Roos, J.: Partioning-based RL-in-RL-out MOR method. In: SCEE 2008 Book of Abstracts, pp. 119–120 (2008)
8. Rommes, J., Lenaers, P., Schilders, W.H.A.: Model order reduction for large resistance networks. In: SCEE 2008 Book of Abstracts, pp. 27–28 (2008)
9. Palenius, T., Roos, J.: Comparison of reduced-order interconnect macromodels for time-domain simulation. IEEE Transactions on Microwave Theory and Techniques **52**(9), 2240–2250 (2004)
10. Matsumoto, Y., Tanji, Y., Tanaka, M.: Efficient SPICE-netlist representation of reduced-order interconnect model. In: Proceedings of ECCTD'01, vol. 2, pp. 145–148. Espoo, Finland (2001)
11. Miettinen, P., Honkala, M., Roos, J., Neff, C., Basermann, A.: Study and development of an efficient RC-in–RC-out mor method. In: Proceedings of ICECS 2008, pp. 1277–1280 (2008)

# Partitioning-Based RL-In–RL-Out MOR Method

Pekka Miettinen, Mikko Honkala, and Janne Roos

**Abstract** This paper proposes a passive, stable, netlist-in–netlist-out-type Model-Order Reduction (MOR) method suitable for the reduction of very large RL circuit blocks. The method relies on partitioning the circuit into subcircuits that can be efficiently approximated with low-order macromodels. The efficiency of the method is demonstrated with several simulations and comparison to the PRIMA method.

## 1 Introduction

Although the study of linear MOR with interconnect circuits has been centered mainly on RC and RLC circuits, some pure RL circuit problem definitions have also been presented, e.g., in [1–3]. The demand for RL MOR arises in certain situations, such as when modeling a conductor's skin effect, magnetic diffusion in a magnetic rod, or eddy currents in a magnetic lamination with lumped elements.

Furthermore, one important motivation for the RL MOR presented in this paper is the possibility to use it (linked with circuit partitioning) on a single RL block appearing inside a much larger RLC circuit. In this case, the RL MOR method is used as one of the many specialized methods in a complete MOR tool.

The basic idea behind the proposed RL MOR method is to partition the circuit into smaller subcircuits, which may then be approximated with relatively simple fixed-size low-order macromodels, and finally combined back together. The concept of low-order macromodels via partitioning was first presented in [4] for RC circuits, which was further studied and refined in [5]. In [6] this was expanded with the support for RLC circuits by also using PRIMA [7] for the partitioned subcircuits.

Pekka Miettinen, Mikko Honkala, Janne Roos
Department of Radio Science and Engineering, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland, e-mail: pekka.miettinen@tkk.fi, mikko.a.honkala@tkk.fi, janne.roos@tkk.fi

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_67,

In this paper, a stable MOR method for the special case of RL circuits is proposed, inspired by the original RC MOR method in [4]. Compared to, e.g., many projection-based methods, the proposed RL-in–RL-out MOR method generates an RL netlist with positive elements as an output. The method may be conceptually divided into the three steps shown in Fig. 1: circuit partitioning (description in Sect. 2), calculation of *y*-parameter moments (Sect. 3), and macromodel synthesis (Sect. 4). After the reduced macromodels for each partition are generated, the macromodels are combined back together according to the partitioning. As a final step, parallel elements of adjacent partitions may be also combined together by applying basic circuit theory, to further reduce the number of generated elements.

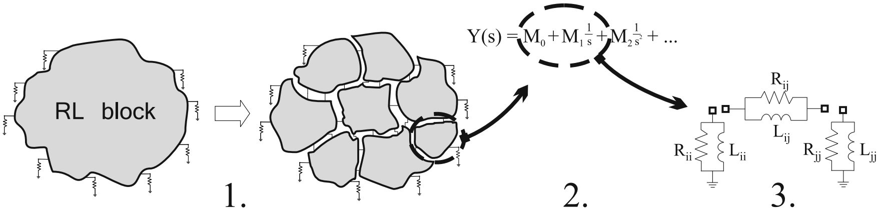


**Fig. 1:** The RL MOR concept: (1) The circuit is partitioned into subcircuits. (2) For each subcircuit, the *y*-parameter moments are calculated. (3) The macromodels for each partition are synthesized using the first two moments. Afterwards, the macromodels are coupled back together

## 2 Circuit Partitioning

Since the RL MOR method is based on approximating interconnects between port nodes with low-order macromodels, it is necessary to perform a partition on the large RL(C) circuit prior to macromodel synthesis. The size of the subcircuits (measured in the number of elements) is critical: If the subcircuit is too large, the low-order macromodel used later is not accurate enough to model the partition, and precision is lost. On the other hand, if the subcircuit is too small, the replacing macromodel is of the same size as the original subcircuit, and no actual reduction takes place.

By using partitioning to match a small section of the original circuit with a low-order approximation, we avoid the possible ill-conditioning issues related to direct high-order macromodel matching approaches. On the other hand, when the partitions are combined together, the final approximation is, in a sense, of order $q = q_\mathrm{p} \times N_\mathrm{part}$, where $q_\mathrm{p}$ is the order of reduction of one partition and $N_\mathrm{part}$ is the number of partitions between two ports. Thus, despite the low-order approximation per partition, we can reach high accuracy for the total reduction, depending on the number of partitions used.

In general, it is assumed that the original circuit is used to model phenomena that are best described by a large number of circuit element blocks of relatively equal importance and complexity. If this is not the case, the partitioning should differ in the size of partitions, such that for sections that need finer precision, smaller partitions are used.

METIS [8] is an algorithm package for partitioning large irregular graphs, partitioning large meshes, and computing fill-in-reducing orderings of sparse matrices. The METIS algorithms are based on multilevel graph-partitioning algorithms, which first reduce the size of the graph by coarsening the graph's details. This takes the form of collapsing adjacent vertices and edges. The smaller graph is then partitioned and refined into the original graph. hMETIS is an extension of METIS that uses hypergraphs instead of graphs [9]. This paper considers the hMETIS algorithm as a partitioning method in the MOR flow [6].

As the partitioning algorithms operate with the reduced-size graph, they are extremely fast compared to traditional partitioning algorithms that compute a partition directly on the original graph. In [8], extensive testing showed that the partitions provided by METIS are consistently better (as measured by the sizes of the cut sets) than those produced by spectral partitioning algorithms.

The use of METIS and hMETIS algorithms especially in circuit partitioning was studied in [10], where it was noted that they both produced excellent partitionings of equal size. The criteria for generating the partitions is to obtain subcircuits of (nearly) equal size with the fewest possible number of external nodes.

## 3 Calculation of *y*-Parameter Moments

Once the RL(C) circuit is partitioned into RL subcircuits, the *y*-parameters are needed to calculate the corresponding macromodel. The Laplace-domain circuit equations for an RL circuit can be expressed as

$$\begin{cases} (\mathbf{G}+\frac{1}{s}\Gamma)\mathbf{x}(s)=\mathbf{B}\mathbf{u}(s) \\ \mathbf{i}(s)=\mathbf{L}^{\mathrm{T}}\mathbf{x}(s), \end{cases} \tag{1}$$

where $\mathbf{x}$ denotes the (internal and external) nodal voltages and port currents, $\mathbf{u}$ denotes the port voltages, and $\mathbf{i}$ denotes the port currents. Here, $\mathbf{B}=\mathbf{L}$ is a selector matrix consisting of ones, minus ones, and zeroes,

$$\mathbf{G}=\begin{bmatrix} \mathbf{G}_{11} & \mathbf{M}_{\mathrm{u}} \\ -\mathbf{M}_{\mathrm{u}}^{\mathrm{T}} & 0 \end{bmatrix}, \text{ and } \Gamma=\begin{bmatrix} \Gamma_{11} & 0 \\ 0 & 0 \end{bmatrix}. \tag{2}$$

Matrices $\mathbf{G}_{11}$ and $\Gamma_{11}$ are symmetric positive semidefinite, and contain the conductance and inverse inductance element stamps, while $\mathbf{M}_{\mathrm{u}}$ consists of the stamps for the port-voltage sources. The size of the $\mathbf{G}$ and $\Gamma$ matrices is thus $n\times n$, with $n=n_{\mathrm{i+e}}+N$, where $n_{\mathrm{i+e}}$ is the total number of nodes and $N$ is the number of ports in the circuit (also, the number of external nodes, $n_{\mathrm{e}}=N$). Solving now for the port currents and defining $\mathbf{A}\equiv-\mathbf{G}^{-1}\Gamma$, $\mathbf{R}\equiv\mathbf{G}^{-1}\mathbf{B}$, results in the following *y*-parameter matrix:

$$\mathbf{Y}(s)=\mathbf{L}^{\mathrm{T}}(\mathbf{I}-\frac{1}{s}\mathbf{A})^{-1}\mathbf{R}, \tag{3}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. Finally, the term $(\mathbf{I} - \frac{1}{s}\mathbf{A})^{-1}$ can be expanded into a Neumann series to obtain

$$\mathbf{Y}(s) = \mathbf{M}_0 + \mathbf{M}_1 \frac{1}{s} + \mathbf{M}_2 \frac{1}{s^2} + \cdots, \tag{4}$$

where $\mathbf{M}_i = \mathbf{L}^\mathrm{T}\mathbf{A}^i\mathbf{R}$. Note that the dimension $(N \times N)$ of the block moments $\mathbf{M}_i$ is the same as the number $N$ of ports in the (sub)circuit. For a typical interconnect-type circuit, $N$ is generally very small ($N \approx 2$). In the case of more complex circuits, $N$ is larger, depending on the connectivity of the topology.

## 4 Macromodel Synthesis

For an $N$-port RL circuit, the admittance between the $i$th port and ground is given by the sum of the $i$th row (or column) of its Y-matrix, $\mathbf{Y}(s)$. The admittance connecting port $i$ and port $j$ is $-y_{ij}$. Thus, the circuit synthesis problem amounts to synthesizing admittances between pairs of ports and between a port and ground with lumped R and L elements. Once $\mathbf{M}_0$ and $\mathbf{M}_1$ have been calculated, each element of $\mathbf{Y}(s)$ can be approximated as

$$y_{ij} \approx m_0^{ij} + m_1^{ij}\frac{1}{s}. \tag{5}$$

Using a direct synthesis, the first two moments are realized with parallel R and L elements. A subcircuit between two ports, $i$ and $j$, is then realized with the macromodel shown in Fig. 2a (included are also the port macromodels). For off-diagonal elements $y_{ij} (i \neq j)$,

$$R_{ij} = -\frac{1}{m_0^{ij}} \quad \text{and} \quad L_{ij} = -\frac{1}{m_1^{ij}}. \tag{6}$$

However, in some situations, $m_1^{ij}$ may be positive. In this case, the macromodel shown in Fig. 2b is used, i.e., the negative inductance is not realized. For diagonal elements $y_{ii}$, the values for port macromodels are

$$R_{ii} = \frac{1}{m_0^{i0}} \quad \text{and} \quad L_{ii} = \frac{1}{m_1^{i0}}, \quad \text{where} \quad m_k^{i0} = m_k^{ii} + \sum_{j \neq i}^{N} m_k^{ij}, \quad k = 0, 1. \tag{7}$$

If $L_{ii}$ is negative, we can set $1/L_{ii} = 0$ and scale down all the $1/L_{ij}$ to keep the total inductance unchanged similar to [4]. In practise, however, the approximation error of simply removing the negative inductance is of the same magnitude, and the latter approach is used in this paper.

Since the admittance between two ports is matched here with a macromodel using only the first two moments, the transfer function to be approximated (i.e., one partition) may not be very complex in terms of poles and zeros. As described in

Sect. 2, the circuit partitioning should ensure that the final partitions are of appropriate size for desired reduction-accuracy ratio.

It should be noted that $\mathbf{M}_0$ describes the circuit's DC characteristics with precise accuracy. This leads to the result that if the dominant elements in the circuit are mostly resistances, the reduction is more accurate. In the extreme case with a resistance-only circuit, the reduction is error-free. It is worth mentioning that the resistance-only MOR method presented in [11] is, conceptually, a special case of the proposed RL MOR method (and of the RC methods [4] and [5]).

As all the synthesized resistances and inductances in the reduced circuit are nonnegative, the reduced circuits are passive and thus stable.

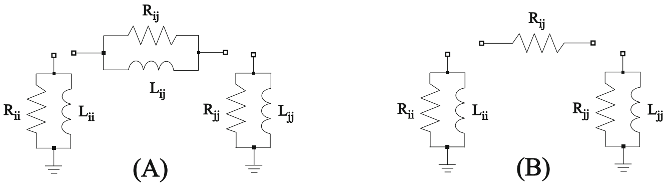


**Fig. 2:** The macromodels used for reduced circuit synthesis, **a** if $m_1^{ij} < 0$, and **b** if $m_1^{ij} > 0$

# 5 Simulation Results

The RL MOR algorithm was verified and simulated with several interconnect RL circuits, of which four representative cases are shown in Table 1. Here, $N_{\mathrm{p}}$, p.size, $n$, $n_e$, R, L, rr, $E_{\mathrm{tr}}$, $T_{\mathrm{tr}}$, and $t_{\mathrm{tr}}$ stand for the number of partitions, approximate size of one partition (in number of elements), the total number of nodes, number of external nodes (ports), resistances, inductances, the element reduction ratio, the normalized transient analysis error, transient analysis CPU time, and relative transient analysis CPU time, respectively. For example: The first row in the table shows the statistics for the original circuit `rlchain1` before reduction. The following three rows show the statistics for the reduced `rlchain1` with various number of partitions. Here, the circuit `rlchain1` is partitioned into 11, 31, and 101 partitions with the partitions modelled with macromodels. Figure 3 shows the transient simulation of the reduced `rlchain31` and the normalized error with $N_{\mathrm{p}} = 17$.

The circuits `rlchain1`–`rlchain31` are of a ladder circuit-type construction with series and/or parallel R and L elements in turns. The circuit `rlclock21` consists of a ladder circuit — with R and L in series and resistances to the ground — forming four connected loops, i.e., a four-leaved clover.

The table shows that the algorithm achieves good reduction of CPU time with only a minimal error in transient simulation compared to the original circuit. Depending on the number and size of the partitions, a trade-off between simulation speed and accuracy can be obtained. With larger (and fewer) partitions, greater reduction is achieved, but typically at the cost of a larger error.

The partitioning-based RL-in–RL-out MOR method was implemented in C and Matlab using SPICE netlists for circuit description. All the simulations were done on a HP RX5670/1.3 GHz computer. The netlists were first reduced with the MOR method, and then transient analysis was carried out on the reduced netlists using APLAC [12].

**Table 1:** Transient simulation results after RL MOR

| Circuit | $N_\mathrm{p}$ | p.size | $n$ | $n_\mathrm{e}$ | R | L | rr/% | $E_\mathrm{tr}$/% | $T_\mathrm{tr}$/s | $t_\mathrm{tr}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| `rlchain1` | Original | - | 2000 | 2 | 1998 | 999 | - | - | 1.40 | 1.000 |
| | 11 | 300 | 33 | 2 | 21 | 21 | 98.0 | 0.33 | 0.11 | 0.079 |
| | 31 | 100 | 93 | 2 | 61 | 61 | 94.2 | 0.06 | 0.13 | 0.093 |
| | 101 | 30 | 303 | 2 | 201 | 201 | 80.8 | 0.05 | 0.23 | 0.164 |
| `rlchain7` | Original | - | 7998 | 2 | 5997 | 3998 | - | - | 12.21 | 1.000 |
| | 41 | 250 | 123 | 2 | 81 | 81 | 98.4 | 1.04 | 0.15 | 0.012 |
| | 101 | 100 | 303 | 2 | 201 | 201 | 96.0 | 0.21 | 0.23 | 0.019 |
| | 335 | 30 | 1005 | 2 | 669 | 669 | 86.6 | 0.13 | 0.76 | 0.064 |
| `rlchain31` | Original | - | 23000 | 2 | 11499 | 11499 | - | - | 150 | 1.000 |
| | 17 | 1500 | 51 | 2 | 28 | 17 | 99.7 | 0.09 | 0.11 | 0.001 |
| | 78 | 300 | 234 | 2 | 153 | 78 | 98.7 | 0.03 | 0.16 | 0.001 |
| | 461 | 50 | 1383 | 2 | 790 | 461 | 94.6 | 0.01 | 0.49 | 0.003 |
| `rlclock21` | Original | - | 23998 | 3 | 24000 | 12000 | - | - | 145 | 1.000 |
| | 182 | 200 | 573 | 3 | 384 | 384 | 97.9 | 2.92 | 0.56 | 0.004 |
| | 482 | 75 | 1472 | 3 | 983 | 984 | 94.5 | 0.42 | 1.28 | 0.009 |
| | 1202 | 30 | 3627 | 3 | 2419 | 2412 | 86.6 | 0.10 | 20.06 | 0.138 |

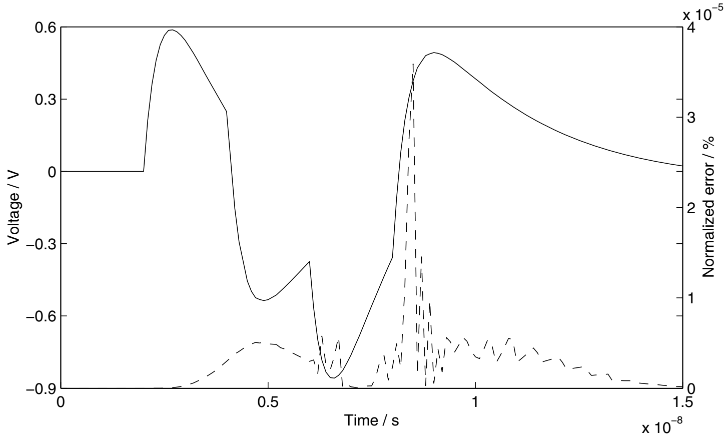


**Fig. 3:** Transient simulation of reduced circuit `rlchain31` (*solid line*) and the normalized error (*dashed line*), $N_\mathrm{p} = 17$

## *5.1 Comparison to PRIMA*

Table 2 shows the results of a reduction with transient simulation using PRIMA (with diagonalization in macromodel synthesis and without partitioning). Note that here, the system equations are formulated in a different manner than described in

Sec. 3. Rather, the MNA formulation presented in [7] is used. This also results in significantly larger system matrices in general, e.g. the size of **G** (and the number of equations) is $n_{\mathrm{i+e}} + n_{\mathrm{L}} + N$, where $n_{\mathrm{L}}$ is the number of inductances, compared to $n_{\mathrm{i+e}} + N$.

Here, $q$, C, and VCCS stand for the order of reduction, the number of capacitances, and voltage-controlled current sources (SPICE G element), respectively. As can be seen comparing the results in the two tables, Table 1 and Table 2, the RL MOR reduction results seem to be of the same order of magnitude as those reached with PRIMA.

**Table 2:** Transient simulation results after PRIMA reduction

| Circuit | $q$ | $n$ | $n_{\mathrm{e}}$ | R | L | C | VCCS | $E_{\mathrm{tr}}/\%$ | $T_{\mathrm{tr}}/\mathrm{s}$ | $t_{\mathrm{tr}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| `rlchain1` | Original | 2000 | 2 | 1998 | 999 | 0 | 0 | - | 1.40 | 1.000 |
| | 10 | 14 | 2 | 10 | 0 | 10 | 60 | 0.21 | 0.11 | 0.079 |
| | 30 | 34 | 2 | 30 | 0 | 34 | 180 | 0.19 | 0.14 | 0.100 |
| `rlchain7` | Original | 7998 | 2 | 5997 | 3998 | 0 | 0 | - | 12.21 | 1.000 |
| | 10 | 14 | 2 | 10 | 0 | 10 | 60 | 0.69 | 0.12 | 0.010 |
| | 30 | 34 | 2 | 30 | 0 | 30 | 180 | 0.12 | 0.16 | 0.013 |

Although the main interest in MOR is the end result — a reduced netlist and its properties —, it is worth mentioning that the reduction process using PRIMA took significantly more time and computer resources than with RL MOR, so that only the first two circuits could be analyzed in reasonable time with the limited computational resources available. Depending on algorithm implementation, code optimization, and available CPU hardware, PRIMA has been successfully applied to systems of much higher order. However, at some point, methods that manipulate the circuit's system matrices as a whole, inherently run into problems regarding available memory. As the presented RL MOR method first partitions the circuit into partitions, the problem is naturally divided into smaller sections, which even a low-end computer can manage with ease. The partitioning also presents an in-built possibility for parallel computing.

## 6 Limitations of the RL MOR Method

The optimum number of the partitions is generally unknown a priori. This leads to a situation, where a short iterative process must be performed for each new type of circuit to obtain the range of usable values for partition size. Of course, with careful study of the circuit's characteristics and element values, this process may be bypassed by manual estimation, but typically a short iteration of different subcircuit sizes is the fastest resort.

As described earlier, the RL MOR method uses only the first two moments to describe a connection between two subcircuit ports. In a typical case, the circuit can be divided into subsections that are simple enough to model with sufficient accuracy. It is possible, however, that the connection between two ports appears earliest at the moment $\mathbf{M}_2$, or even at higher moments, and thus the connection is left as an open circuit. If this connection is topologically critical to the circuit's behavior, the reduction fails.

## 7 Conclusions

In this paper, a new RL MOR method capable of efficient reduction of very large RL circuits was proposed. Using partitioning, the method generates a positive-element RL netlist from the first two moments of the *y*-parameters, preserving passivity and stability of the original circuit. Simulation results showing excellent reduction were presented along with a comparison to PRIMA reduction. Also, the limitations of the partitioning-based RL-in–RL-out MOR method, derived from theory or discovered during test simulations, were reported.

**Acknowledgements** This research was funded by the Academy of Finland.

## References

1. Krah, J. H.: Optimum Discretization of a Physical Cauer Circuit. IEEE Trans. Magnetics, **41**, no. 5, 1444–1447 (2005)
2. Mei, S., Ismail, Y.I.: Modeling Skin and Proximity Effects with Reduced Realizable RL Circuits. IEEE Trans. Very Large Scale Integration (VLSI) Systems, **12**, no. 4, 437 – 447 (2004)
3. Yu, Q., Kuh, E.S.: Accurate Reduced RL Model for Frequency Dependent Transmission Lines. In: Proc. ICECS 2002, **2**, pp. 761–764, Dubrovnik (2002)
4. Liao, H., Dai, W.W.-M.: Partitioning and Reduction of RC Interconnect Networks Based on Scattering Parameter Macromodels. In: Proc. ICCAD 1995, pp. 704–709 (1995)
5. Miettinen, P., Honkala, M., Roos, J.: Study and Development of an Efficient RC-in–RC-out MOR Method. In: Proc. ICECS 2008, pp. 1277–1280, Malta (2008)
6. Miettinen, P.: Hierarchical Model-Order Reduction Tool for RLC Circuits. Master's thesis, Helsinki University of Technology, 2007. URL http://www.ct.tkk.fi/publications/dt-pekka/main.html. Cited Oct 10 2008.
7. Odabasioglu, A., Celik, M., Pileggi, L.T.: PRIMA: Passive Reduced-Order Interconnect Macromodeling Algorithm. IEEE Trans. Computer-Aided design of Integrated Circuits and Systems, **17**, 645–654 (1998)
8. Karypis, G., Kumar, V.: METIS, A software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices (Version 4.0). URL http://glaros.dtc.umn.edu/gkhome/views/metis. Cited Oct 10 2008.
9. Karypis, G., Kumar, V.: hMETIS, A Hypergraph Partitioning Package (Version 1.5.3). URL http://glaros.dtc.umn.edu/gkhome/metis/hmetis/overview. Cited Oct 10 2008.
10. Miettinen, P., Honkala, M., Roos, J.: Using METIS and hMETIS Algorithms in Circuit Partitioning. Report CT-49, Circuit Theory Laboratory, Helsinki University of Technology (2006). URL http://www.ct.tkk.fi/publications/ct-49/main.html. Cited Oct 10 2008.
11. Rommes, J., Lenaers, P., Schilders, W.H.A.: Model order reduction for large resistance networks, In: J. Roos and L.R.J. Costa (eds.) SCEE 2008 Book of Abstracts, TKK Radio Science and Engineering Publications, Report R4, pp. 27–28. Picaset Oy, Helsinki (2008)
12. APLAC — Circuit Simulation and Design Tool, Version 8.4 Manuals, AWR–APLAC Corporation, Finland, 2008.

# Reduction of Large Resistor Networks

Joost Rommes, Peter Lenaers, and Wil H.A. Schilders

**Abstract** Electro Static Discharge (ESD) analysis is of vital importance during the design of large-scale integrated circuits, since it gives insight in how well the interconnect can handle unintended peak charges. Due to the increasing amount of interconnect and metal layers, ESD analysis may become very time consuming or even unfeasible. We propose an algorithm for the reduction of large resistor networks, that typically arise during ESD, to much smaller equivalent networks. Experiments show reduction and speed-ups up to a factor 10.

## 1 Introduction

Electro Static Discharge (ESD) analysis is of vital importance during the design of large-scale integrated circuits and derived products. A human touch charged by walking across a carpet, for instance, can affect or destroy a device containing electric components. The costs involved may vary from a few cents to millions if, due to interconnect failures, a respin of the chip is needed. An example of a damaged piece of interconnect that was too small to conduct the amount of current is shown in Figure 1.

ESD analysis [1, 2] requires knowledge on how fast electrical charge on the pins of a package can be discharged. In many cases, the discharge is done through the power network, the interconnect and the substrate, which are resistive. Diodes are used to protect transistors on a chip against peak charges. The discharge paths, that

Joost Rommes, Wil H.A. Schilders
NXP Semiconductors, HTC 37, 5656 AE Eindhoven, The Netherlands, e-mail: joost.rommes@nxp.com, wil.schilders@nxp.com

Peter Lenaers
Mathematics for Industry, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, e-mail: plenaers@gmail.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_68,

consist of very large resistor networks connected through diodes, must be of low resistance to allow for sufficient discharge.

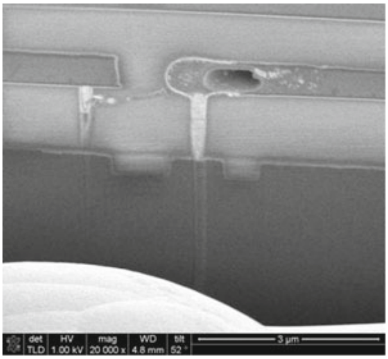


**Fig. 1:** Example of a piece of interconnect that was damaged because it was too small to conduct the amount of current caused by a peak charge

In practice, one is only interested in the path resistances from the output of one device to the input of another. But since one device can serve as driver to multiple other devices, the network that needs to be analyzed can be regarded as a tree with one root and many leaves. To complicate matters each branch (path from one internal node to another) can consist of multiple parallel paths, thus complicating the computation of the correct resistance.

The interconnect and resistance network are typically modeled by resistors, and diodes are used to connect different parts of the network. The resulting resistive network may contain up to millions of resistors, hundreds of thousands of internal nodes, and thousands of external nodes (nodes with connections to diodes). Simulation of such large networks within reasonable time is often not possible, and including such networks in full system simulations may be even unfeasible. Hence, there is need for much smaller networks that accurately or even exactly describe the resistive behavior of the original network, but allow for fast analysis.

In this paper we describe a new approach for the reduction of large resistor networks. We show how insights from graph theory, numerical linear algebra, and matrix reordering algorithms can be used to construct an equivalent network with the same number of external nodes, but much less internal nodes and resistors. This equivalent reduced network exactly describes the behavior of the original network, i.e., no approximation error is made. The approach is illustrated by numerical results.

The paper is organized as follows. In section 2 we describe the relevant properties of resistor networks and formulate the network reduction problem. An overview of existing approaches to deal with large resistor networks is given in section 3. In section 4 we describe a new approach to reduce resistor networks. Results of the new approach are shown in section 5. Section 6 concludes.

## 2 Properties of Resistor Networks

A resistor network consists of internal nodes, external nodes (or terminals), and resistors. Figure 2 shows a simple resistor network with external nodes $Z$, $A$, $B$, and $C$, and internal nodes $X$ and $Y$ (there are five resistors). Of interest are the path resistances from $Z$ to $A$, $B$, and $C$. This small example is purely for illustrational purposes; in real-life applications the number of nodes and resistors is much larger: typical networks consist of millions of resistors and nodes, of which (tens of) thousands are external nodes. In the following it will be assumed that the network has $n > 0$ internal nodes, $m > 0$ external nodes, and $r > 0$ resistors.

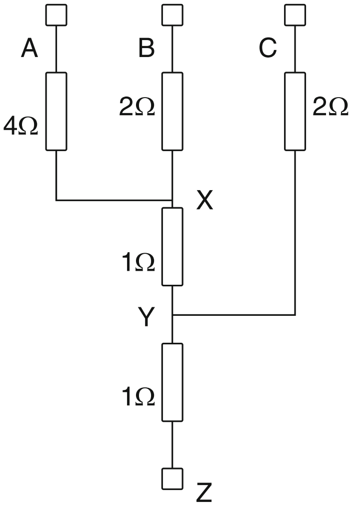


**Fig. 2:** Simple resistor network with external nodes $Z$, $A$, $B$, and $C$. Of interest are the path resistances between external nodes

### 2.1 Mathematical Formulation

Using Ohm's Law for resistors and Kirchhoff's Current Law [3], the electrical behavior of a resistance network can be described by

$$\mathbf{i} = Y \cdot \mathbf{v}, \tag{1}$$

where $\mathbf{i}, \mathbf{v} \in \mathbb{R}^N$ and $Y \in \mathbb{R}^{N \times N}$ (with $N = n + m$) contain the unknown inflowing currents, node voltages, and conductances, respectively.

We distinguish between internal and external nodes:

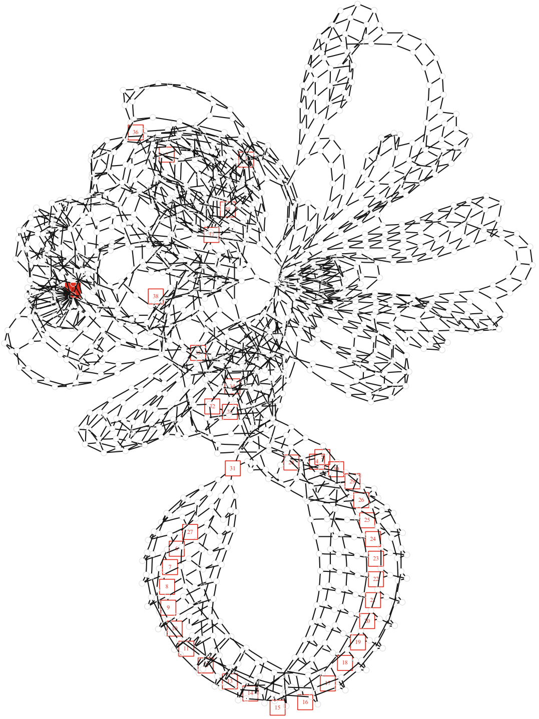


**Fig. 3:** Graph representation of a realistic resistor network. The *squares* are external nodes and need to be preserved in the reduced network. Of interest are the path resistances between external nodes

$$\begin{bmatrix} \mathbf{i}_e \\ \mathbf{i}_i \end{bmatrix} = \begin{bmatrix} Y_{ee} & Y_{ei} \\ Y_{ei}^T & Y_{ii} \end{bmatrix} \begin{bmatrix} \mathbf{v}_e \\ \mathbf{v}_i \end{bmatrix},$$

where $\mathbf{i}_e, \mathbf{v}_e \in \mathbb{R}^m$ and $\mathbf{i}_i, \mathbf{v}_i \in \mathbb{R}^n$ correspond to external and internal nodes, respectively, and $Y$ is partitioned accordingly. Note that $\mathbf{i}_i = 0$ since it is assumed that currents can only be injected in external nodes.

One node is chosen as reference (ground) node: this makes the $Y$ matrix nonsingular. All diagonal elements of $Y$ are strictly positive and all off-diagonal elements are negative or zero. The conductance matrix $Y = (y_{ij})$ is symmetric and (after grounding) positive-definite ($\mathbf{x}^T Y \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$). In most of the applications, the conductance matrix $Y$ is very sparse, typically having $O(1)$ nonzeros per row. Figure 3 shows a realistic resistor network.

The impedance matrix $Z$ can be obtained by inverting $Y$: $Z = Y^{-1}$. For large networks this is not possible due to memory and CPU limitations, and it is neither necessary since usually only specific elements are needed: the path resistance from the reference node (terminal) to another terminal $b$, for instance, is given by the diagonal element $z_{bb}$.

### 2.2 Problem Formulation

The problem is: given a very large resistor network described by (1), find an equivalent network with (a) the same external nodes, (b) exactly the same path resistances between external nodes, (c) $\hat{n} \ll n$ internal nodes, and (d) $\hat{r} \ll r$ resistors. Additionally, (e) the reduced network must be realizable as a netlist so that it can be (re)used in the design flow as subcircuit of large systems (see Figure 4 for an example use of a reduced netlist).

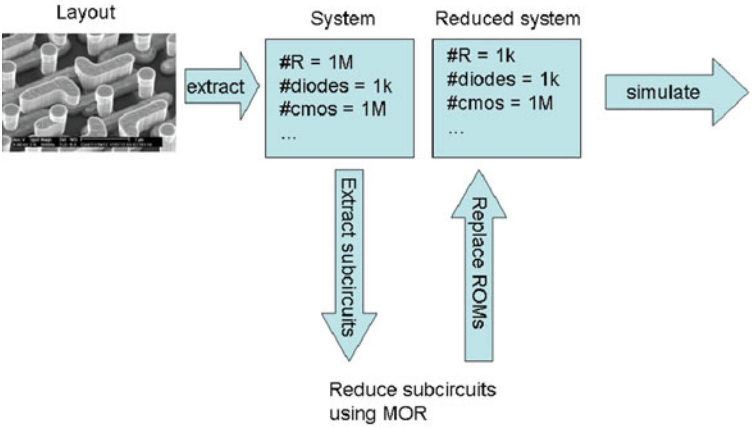


**Fig. 4:** Typical (re)use of reduced equivalent network in the design flow: the original network is reduced to a smaller network that replaces the original network in the complete system

Simply eliminating all internal nodes will lead to an equivalent network that satisfies conditions (a)–(c), but violates (d) and (e): for large numbers $m$ of external nodes, the number of resistors $\hat{r} = (m^2 - m)/2$ in the dense reduced network is in general much larger than the number of resistors in the sparse original network ($r$ of $O(n)$), leading to increased memory and CPU requirements.

## 3 Existing Approaches

There are several approaches to deal with large resistor networks. If the need for an equivalent reduced network can be circumvented in some way, this is usually the best to do. To see this, one has to take into account that due to sparsity of the original network, memory usage and computational complexity are *in principle* not an issue, even not for networks containing millions of resistors. Solving linear systems with the related conductance matrices is typically of complexity $O(n^\alpha)$, where $1 < \alpha \leq 2$, instead of the traditional $O(n^3)$ [4], and hence the path resistance problem can be solved directly. Of course, $\alpha$ depends on the sparsity and will rapidly increase as sparsity decreases. This also explains why eliminating all internal nodes does not work in practice: the large reduction in unknowns is easily undone by the enormous increase in number of resistors, mutually connecting all external nodes.

However, if we want to (re)use the network in full system simulations, a reduced equivalent network is needed to limit simulation times or make simulation possible at all. There is software [5, 6] available for the reduction of parasitic reduction networks, but this software produces approximate reduced networks while in many cases an exact reduced network is needed. In [7] approaches based on large-scale graph partitioning packages such as (h)METIS [8] are described, but only applied to small networks. Structure preserving projection methods for model reduction [9,10], finally, have the disadvantage that they lead to dense reduced-order models if the number of terminals is large.

## 4 Improved Approach

Knowing that eliminating all internal nodes is not an option and that projection methods lead to dense reduced-order models, we use concepts from matrix reordering algorithms such as AMD [11] and BBBD [12], usually used as preprocessing step for (parallel) LU- or Cholesky-factorization, to determine which nodes to eliminate. The fill-in reducing properties of these methods also guarantee sparsity of the reduced network. Similar ideas have also been used in [7, 13].

Our main motivation for this approach is that large resistor networks in ESD typically are extracted networks with a structure that is related to the underlying (interconnect) layout. Unfortunately, the extracted networks are usually produced by extraction software of which the algorithms are unknown, and hence the structure of the extracted network is difficult to recover. Standard tools from graph theory, however, can be used to recover at least part of the structure.

Note that in the context of this paper, with structure we refer to the topological structure of the network. This is in contrast with structure preserving model order reduction methods [9], where structure usually refers to the mathematical structure of the dynamical system. In our applications, the reduced network should have approximately the same sparsity and topology as the original network.

Our approach can be summarized as follows:

1. The first step is to bring the conductance matrix $Y$ into Balanced Border Block Diagonal (BBBD) form using techniques of [11, 12, 14], see Figure 5. In this form, the matrix consists of two parts: the main body $A_{11}$ and border blocks $A_{12}$, $A_{21} = A_{12}^T$ and $A_{22}$. The main body is partitioned into subblocks, where each block $B_{ii}$ represents a cluster in the network. Block $B_{ii}$ has a nonzero entry when two nodes in cluster $i$ are connected. Internal nodes that connect different clusters are in the border. Borderblock $A_{22}$ contains information on the connections between bordernodes, while borderblocks $A_{12}$ and $A_{21}$ contain information on the connections between bordernodes and the different clusters. The clusters contain both external and internal nodes, while all nodes in the border are internal.

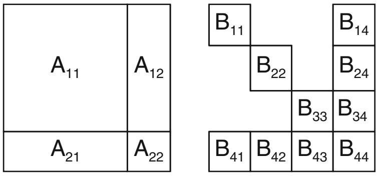


**Fig. 5:** Matrix in BBBD-form (*left*) with subblocks

2. The second step is to eliminate the internal nodes in block $A_{11}$. This is done using the Schur complement [15]. Since the ordering is chosen to minimize fill-in, the resulting reduced matrix is sparse. Note that all operations are exact, i.e., we do not make any approximations. As a result, the path resistances between external nodes remain equal to the path resistances in the original network.
3. Finally, the reduced conductance matrix can be realized as an reduced resistor network that is equivalent to the original network. Since the number of resistors (and number of nodes) is smaller than in the original network, also the resulting netlist is smaller in size.

An additional reduction could be obtained by removing relatively large resistors from the resulting reduced network. However, this will introduce an approximation error that might be hard to control a priori, since no sharp upper bounds on the error are available [16]. Another issue that is subject to further research is that the optimal ratio of number of (internal) nodes to resistors (sparsity) may also depend on the ratio of number of external to internal nodes, and on the type of simulation that will be done with the network.

## 5 Numerical Results

Table 1 shows results for three resistor networks of realistic interconnect layouts. The number of nodes is reduced by a factor $> 10$ and the number of resistors by a factor $> 3$. As a result, the computing time for calculating path resistances in the original network (including nonlinear elements such as diodes) is 10 times smaller.

**Table 1:** Results of reduction algorithm

| | Network I | | Network II | | Network III | |
|---|---|---|---|---|---|---|
| | Original | Reduced | Original | Reduced | Original | Reduced |
| #external nodes | 274 | | 3399 | | 1978 | |
| #internal nodes | 5558 | 516 | 99112 | 6012 | 101571 | 1902 |
| #resistors | 8997 | 1505 | 161183 | 62685 | 164213 | 39011 |
| CPU time | 10 s | 1 s | 67 hrs | 7 hrs | 20 hrs | 2 hrs |
| Speed up | 10x | | 9.5x | | 10x | |

# 6 Conclusions

Electro Static Discharge analysis is of crucial importance for present chip design. Because the resulting resistor networks may contain millions of nodes and resistors, full system simulation becomes too expensive or unfeasible, leading to delay in the design cycle. Hence, there is need for reduced networks that are much smaller but exactly reproduce the behavior of the original networks. We propose an algorithm based on concepts from graph and matrix reordering theory. The new method can reduce large resistor networks to small equivalent networks. Since the reduced network exactly matches the behavior of the original network, it can replace the original network in the design flow for Electro Static Discharge analysis. Speedups of up to a factor 10 are obtained for industrial circuits.

**Acknowledgements** Supported by O-MOORENICE! (MCA FP6 MTKI-CT-2006-042477).

# References

1. Kolyer, J.M., Watson, D.: ESD: From A To Z. Springer (1996)
2. Electrostatic discharge association. http://www.esda.org
3. Chua, L.O., Lin, P.: Computer aided analysis of electric circuits: algorithms and computational techniques, first edn. Prentice Hall (1975)
4. Phillips, J.R., Silveira, L.M.: Poor man's tbr: A simple model reduction scheme. IEEE Trans. CAD Circ. Syst. **24**(1), 283–288 (2005)
5. Edxact: Jivaro. http://www.edxact.com
6. Cadence: AssuraRCX. http://www.cadence.com
7. Miettinen, P., Honkala, M., Roos, J.: Using metis and hmetis algorithms in circuit partitioning. Circuit Theory Laboratory Report Series CT-49, Helsinki University of Technology (2006)
8. Karypis, G., Kumar, V.: METIS, A software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices. http://glaros.dtc.umn.edu/gkhome/metis/
9. Freund, R.W.: SPRIM: Structure-preserving reduced-order interconnect macromodeling. In: Technical Digest of the 2004 IEEE/ACM International Conference on CAD, pp. 80–87 (2004)
10. Schilders, W.H.A., van der Vorst, H.A., Rommes, J. (eds.): Model Order Reduction: Theory, Research Aspects and Applications, *Mathematics in Industry*, vol. 13. Springer (2008)
11. Amestoy, P.R., Davis, T.A., Duff, I.S.: An Approximate Minimum Degree Ordering Algorithm. SIAM J. Matrix Anal. Appl. **17**(4), 886–905 (1996)
12. Zečević, A.I., Šiljak, D.D.: Balanced decompositions of sparse systems for multilevel prallel processing. IEEE Trans. Circ. Syst.–I:Fund. Theory and Appl. **41**(3), 220–233 (1994)
13. Zhou, Q., Sun, K., Mohanram, K., Sorensen, D.C.: Large power grid analysis using domain decomposition. In: Proc. Design Automation and Test in Europe, pp. 27–32 (2006)
14. Duff, I.S., Erisman, A.M., Reid, J.K.: Direct methods for sparse matrices. Oxford: Clarendon Press (1986)
15. Golub, G.H., van Loan, C.F.: Matrix Computations, third edn. John Hopkins University Press (1996)
16. Yang, F., Zeng, X., Su, Y., Zhou, D.: RLC equivalent circuit synthesis method for structure-preserved reduced-order model of interconnect in VLSI. Communications in computational physics **3**(2), 376–396 (2008)

# Nonlinear Model Order Reduction Based on Trajectory Piecewise Linear Approach: Comparing Different Linear Cores

Kasra Mohaghegh, Michael Striebel, E. Jan W. ter Maten, and Roland Pulch

**Abstract** Refined models for MOS-devices and increasing complexity of circuit designs cause the need for Model Order Reduction (MOR) techniques that are capable of treating nonlinear problems. In time-domain simulation the Trajectory PieceWise Linear (TPWL) approach is promising as it is designed to use MOR methodologies for linear problems as the core of the reduction process. We compare different linear approaches with respect to their performance when used as kernel for TPWL.

## 1 Introduction

The tendency to analyze and design systems of ever increasing complexity is becoming more and more a dominating factor in progress of chip design. Along with this tendency, the complexity of the mathematical models increases both in structure and dimension. Complex models are more difficult to analyze, and due to this it is also harder to develop control algorithms. Therefore Model Order Reduction (MOR) is of utmost importance. For linear systems, quite a number of approaches are well-established and have proved to be very useful [1]. However, accurate models for MOS-devices introduce highly nonlinear equations. And, as the packing density in circuit design is growing, very large nonlinear systems arise. Hence, there is a growing request for reduced order modeling of nonlinear problems. In transient analysis the Trajectory PieceWise Linear (TPWL) approach [2, 3] is a promising technique

Kasra Mohaghegh, Roland Pulch
Bergische Universität Wuppertal, Wuppertal, Germany, e-mail: mohaghegh@math.uni-wuppertal.de, pulch@math.uni-wuppertal.de

Michael Striebel
Technische Universität Chemnitz, Chemnitz, Germany, e-mail: michael.striebel@mathematik.tu-chemnitz.de

E. Jan W. ter Maten
NXP Semiconductors, Eindhoven, The Netherlands, e-mail: jan.ter.maten@nxp.com

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_69,

as it makes use of linear MOR methods. A brief introduction to TPWL is given below. Analyzing the TPWL approach, we are interested in how different linear MOR techniques perform when used as a linear kernel, how robust the reduced models are and how they behave when combined to more complex systems.

## 2 MOR for Linear Problems

A continuous time-invariant (lumped) multi-input multi-output linear dynamical system is of the form:

$$\begin{cases} C\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -Gx(t)+Bu(t), \\ \quad y(t) = Lx(t)+Du(t), \quad x(0)=x_0, \end{cases} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the inner state, $u(t) \in \mathbb{R}^m$ is the input, $y(t) \in \mathbb{R}^p$ is the output. The dimension $n$ of the state vector is called the order of the system. $C$, $G$, $B$, $L$ and $D$ are the state space matrices. The dimension $n$ of the system exhibits the order of elements contained in the circuit. As VLSI systems exhibit a large density of elements, $n$ can easily reach a million.

Basically, MOR techniques aim to derive a system:

$$\begin{cases} \tilde{C}\frac{\mathrm{d}\tilde{x}(t)}{\mathrm{d}t} = -\tilde{G}\tilde{x}(t)+\tilde{B}u(t), \quad \tilde{x}(t) \in \mathbb{R}^q, \\ \quad \tilde{y}(t) = \tilde{L}\tilde{x}(t)+\tilde{D}u(t), \quad \tilde{x}(0)=\tilde{x}_0, \;\; \tilde{y}(t) \in \mathbb{R}^p, \end{cases} \tag{2}$$

of order $q$ with $q \ll n$ that can replace the original high-order system (1) in the sense, that the input-output behavior, described by the transfer function in the frequency domain, of both systems agrees. A common way is to identify a subspace of dimension $q \ll n$, that captures the dominant information of the dynamics and project (1) onto this subspace, spanned by some basis vectors $\{v_1,\dots,v_q\}$.

The reduction can be carried out by means of different techniques. Approaches like PRIMA [4], SPRIM [5], and PMTBR [6] project the full problem (1) onto a subspace of dimension $q$. The first two rely on Krylov subspace methods. The latter one exploits the direct relation between the multipoint rational projection framework and the Truncated Balanced Realization (TBR). This approach can take advantage of some a-priori knowledge of the system properties, and is based on a statistical interpretation of the system Gramians. We give a brief review on these techniques and analyze their behavior when used as linear kernels in TPWL.

### *2.1 Krylov Projection Techniques and Poor Man's TBR*

In recent years, MOR techniques based on Krylov subspaces have become the methods of choice for generating macromodels of large multi-port RLC circuits. Krylov subspace methods provide numerically robust algorithms for generating a basis of the reduced space, such that a certain number of moments of the transfer function of the original system is matched. Consequently, the transfer function of the reduced system approximates the original transfer functions around a specified frequency, or a collection of frequency points [7]. Owing to their robustness and low computational cost, Krylov subspace algorithms proved suitable for the reduction of large-scale systems, and gained considerable popularity, especially in electrical engineering. A number of Krylov-based MOR algorithms have been developed, including techniques based on the Lanczos method [8, 9] and the Arnoldi algorithm [4, 10]. The main drawbacks of these methods are, in general, lack of provable error bounds for the extracted reduced models, and no guarantees for preserving stability and passivity. Nevertheless, it has been demonstrated that if the original system has a specific structure, both stability and passivity can be preserved in the reduced system, by exploiting the fact that congruence transformations preserve the definiteness of a matrix. PRIMA [4] combines the moment matching approach with projection to arrive at a reduced system of type (2). Its main feature is that it produces provably passive reduced models.

However, PRIMA does not preserve the structure of the system matrices which is of an interest when trying to realize the reduced model. SPRIM [5], an adaption of this method, preserves block structures of the circuit matrices and generates provably passive and reciprocal macromodels of multiport RLC circuits. The SPRIM models match twice as many moments as the corresponding PRIMA models obtained with the same amount of computational work. Also SPRIM is less restrictive to matrices $C$ and $G$ in system (1), see [11].

Poor Man's TBR (PMTBR) [6] is a projection MOR technique that exploits the direct relation between the multipoint rational projection framework and the Truncated Balanced Realization (TBR). More details on PMTBR can be found in [6]. In the following simulation we assume that $C = I$ and $D = 0$ in (1).

## 2.2 Examples

We consider the RLC ladder networks, illustrated in Figure 1.

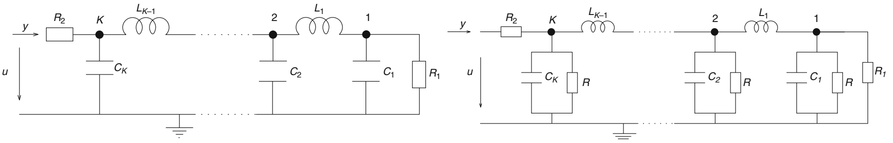


**Fig. 1:** *Left*: RLC circuit example 1; *Right*: RLC circuit example 2

The state variable $x \in \mathbb{R}^{2K-1}$ consists of the voltages of the $K$ nodes and the currents traversing the inductors $\{L_1, \ldots, L_{K-1}\}$. The voltage $u$ and the current $y$ represent input and output, respectively. Note that when the number of nodes is $K$ the order of the system becomes $n = 2K - 1$.

*Example 1.* We choose an RLC ladder network shown in Figure 1 (left). We set all the capacitances and inductances to the same value 1 while $R_1 = \frac{1}{2}$ and $R_2 = \frac{1}{5}$, see [12]. We arrange 51 nodes which gives us the order 101 for the circuit.

*Example 2.* We use an RLC ladder network given in Figure 1 (right). We set all the capacitances and inductances to the same value 1 while $R_1 = \frac{1}{2}$, $R_2 = \frac{1}{5}$ and $R = 1$, we choose 51 nodes which results in order 101 for the circuit.

The main reason for choosing these two examples is the behavior of Hankel singular values, see [1]. The Hankel singular values for the first example do not show any significant decay while in the second example we observe a rapid decay in the values. The model is reduced by three linear techniques (PRIMA, SPRIM and PMTBR) from order 101 to order 34 for both examples. Figure 2 shows the absolute error between the transfer function of the full system and the transfer function of the reduced system.

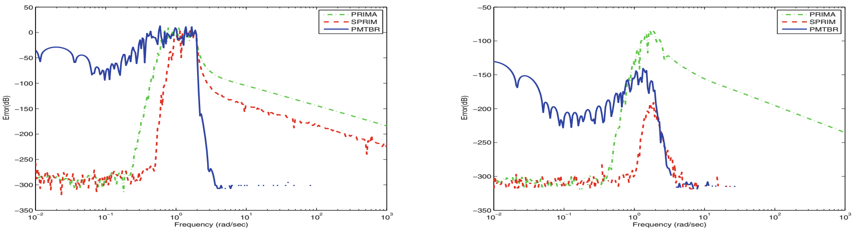


**Fig. 2:** *Left*: Error plot for the Example 1; *Right*: Error plot for the Example 2

As we expected the SPRIM produces a better approximation than PRIMA since it matches twice as much moments. Although both methods have a good match around the expansion point 0, the error increases as we are far from the expansion point. As the Hankel singular values for the first example do not decay, the PMTBR cannot produce an accurate model for low frequency in that case. This shows that we can not stick to one method for reduction in general and the method should be chosen depending on the circuit behavior.

## 3 MOR for Nonlinear Problems

Large linear problems most frequently arise from modeling parasitic effects introduced by the layout, i.e., the wiring. As structure sizes decrease and packing densities increase the growing complexity of the nominal circuitry that is build up from transistors showing highly nonlinear behavior generates the need of MOR for nonlinear problems as well. In general an electric circuit can be described by a system of differential-algebraic equations (DAEs) of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}[q(x(t))] + j(x(t)) + Bu(t) = 0, \tag{3}$$

where $x(t) \in \mathbb{R}^n$ represents the unknown vector of circuit variables at time $t \in [t_0, t_e]$; the nonlinear functions $q, j : \mathbb{R}^n \to \mathbb{R}^n$ describe the contribution of reactive and nonreactive elements, respectively, and the matrix $B$ distributes the input excitation $u : [t_0, t_e] \to \mathbb{R}^m$. Note that we concentrate on the state $x$ only and omit the output stage $y$ in our consideration.

MOR techniques developed for linear problems (1) cannot be applied directly to nonlinear models (3) as the transfer to a lower dimensional problem does not guarantee a reduction in the computational effort from evaluating the nonlinear model.

## 3.1 Trajectory Piecewise Linearization

The idea of TPWL [2, 3] is to represent the full nonlinear system (3) by a bunch of order reduced linear models that can reproduce the typical behavior of the system.

For this purpose a training input $\bar{u}(t)$ for $t \in [t_0, t_e]$ is chosen and a transient simulation is run in order to get a trajectory, i.e., a collection of points $\bar{x}_0, \ldots, \bar{x}_N$ approximating $x(t_i)$ at time-points $t_0 < t_1 < \cdots < t_N = t_e$, that reflect typical states of the system. On the trajectory, points $\{x_1^{\text{lin}}, \ldots, x_s^{\text{lin}}\} \subset \{\bar{x}_0, \ldots, \bar{x}_N\}$ are selected around which the nonlinear functions $q$ and $j$ are linearized. To the linear models, that are all of dimension $n$, any MOR for linear problems can be applied. This delivers local reduced subspaces $V_1, \ldots, V_s$ of possibly different dimensions $k_1, \ldots, k_s$. One common subspace $V$ of dimension $k \ll n$ is constructed that describes the primary information of all local subspaces and on which all linear models are projected. Finally a weighting $w_i(Vz) \in [0, 1]$ for $i = 1, \ldots, s$ with $\sum_{i=1}^{s} w_i(Vz) = 1$ is introduced to decide which linear submodels are valid in a certain situation. The full system shall be replaced by the reduced one given by

$$\sum_{i=1}^{s} w_i(Vz) \left[ V^T C_i V \frac{\mathrm{d}}{\mathrm{d}t} z + V^T G_i V z + V^T (j(x_i^{\text{lin}}) - G_i x_i^{\text{lin}}) \right] + V^T B u(t) = 0 \tag{4}$$

with $C_i = \left.\frac{\partial q}{\partial x}\right|_{x = x_i^{\text{lin}}}$ and $G_i = \left.\frac{\partial j}{\partial x}\right|_{x = x_i^{\text{lin}}}$

Besides the freedom in choosing which linear MOR technique to use there are also different strategies reported for determining the linearization points along the trajectory. In our considerations we stick to the strategy described in [3]. There at each time-point $t_i$ both the full nonlinear system and the currently responsible reduced linear model are discretized with the same stepsize leading to two different approximations $\bar{x}_i$ and $\hat{x}_i = V z_i$. Whenever the difference $\bar{x}_i - \hat{x}_i$ becomes too large, a new linearization point is arranged.

### 3.2 Example

We apply only PRIMA and PMTBR as a linear core for TPWL. In all simulation below the PMTBR is used unless stated otherwise. One of the partitions which is used inside the SPRIM algorithm is always of size 2 by 2 and the other part becomes larger as there is no inductor in the structure of the inverter chain. Therefore SPRIM is not reasonable to apply in this test case. The inverter chain constitutes a special class of circuit problems. Here a signal passes through the system, activating at each time-slot just a few elements and leaving the others latent. However, as the signal passes through, each element is active at some time and sleeping at some others. As in [13], the training of the inverter chain during the TPWL model extraction was done with a single piecewise linear input voltage at $\bar{u}(t)$ (see also Figure 3), defined by

$$\bar{u}(0) = 0,\ \ \bar{u}(5\text{ns}) = 0,\ \ \bar{u}(10\text{ns}) = 5,\ \ \bar{u}(15\text{ns}) = 5,\ \ \bar{u}(17\text{ns}) = 0.$$

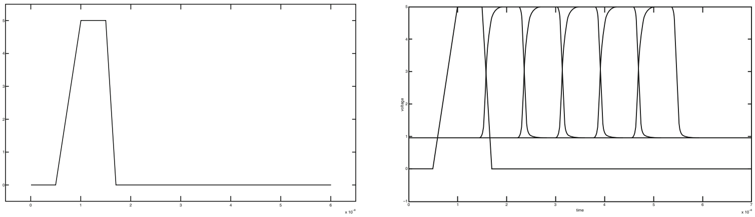


**Fig. 3:** Inverter chain: training input (*left*) and state response (*right*, all stages)

In Figure 4 we see the danger of defining distances to linearization points not in the full space but in the reduced space. Both plots are showing the signal at inverter 24. In Figure 4 in the right plot the second impulse is just not recognized where this seems to be no problem in the left plot. However, something else seems to be missing, even if we take the distance in the full space. In Figure 5 the voltage at inverters 68 and 92 is given. In both cases, the signal cannot be recovered correctly. In the latter one it is even not recognized at all. At the moment we cannot state reasons for that. Obviously this is not caused by the reduction but by the linearization or the weighting procedure as we get similar results when turning off the reduction step.

The impact of broadening the input signal $u$ can be seen in Figure 6, which displays the voltage at inverters 18 and 68. The signals are far away from the expected behavior. However, there seems to be a trend towards the situation that was encountered during the training. And indeed in Figure 6 (right), at inverter 68 we find a time shifted version of the training signal instead of the wide input signal that has been applied now.

Finally, in Figure 7 the result of using the reduced model that arises from training input $\bar{u}$ of given pulse width with a slightly tighter input signal $u$ is given for the inverters 6 and 68, respectively. In the former the characteristic is reflected quite well. However, in the latter the output signal seems to be just a time shifted version of the situation during the training. Having a closer look at how the inverter chain is

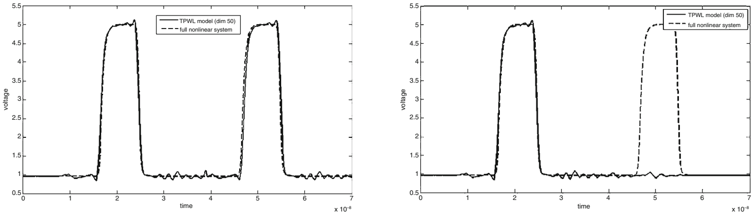


**Fig. 4:** Inverter chain: TPWL-resimulation, reduction to order 50, repeated pulse, inverter 24, *Left*: distance defined in full space; *Right*: distance defined in reduced space

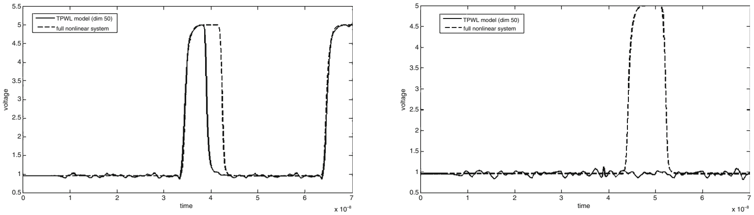


**Fig. 5:** Inverter chain: TPWL-resimulation, reduction to order 50, repeated pulse, distance in full space, *Left*: inverter 68; *Right*: inverter 92

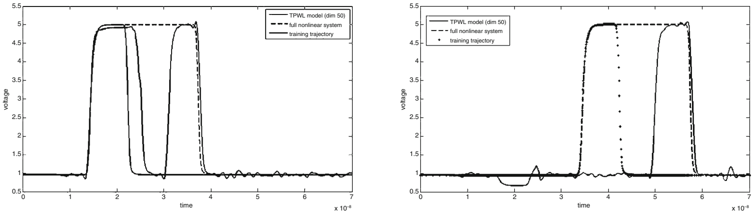


**Fig. 6:** Inverter chain: TPWL-resimulation, reduction to order 50, wider pulse, distance in full space, *Left*: inverter 18; *Right*: inverter 68

modeled we see that the input voltage is applied at a floating node. This could give reasoning for the behavior encountered. However, also the backward and forward validity of the linear models could be the reasons.

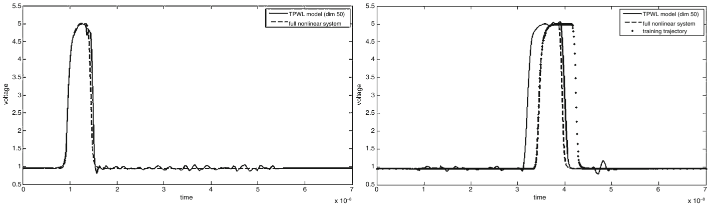


**Fig. 7:** Inverter chain: TPWL-resimulation, reduction to order 50, tighter impulse, distance in full space, *Left*: inverter 6; *Right*: inverter 68

The error in Figure 8 is an overall error for all nodes. This total error shows that PMTBR yields better approximations than PRIMA. As changing from one linear method to the other the problems stay the same. Thus the reduction steps do not cause them.

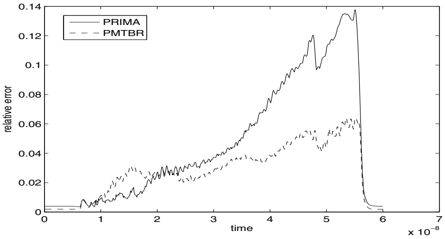


**Fig. 8:** Overall error for PRIMA and PMTBR used inside TPWL

**Acknowledgements** The work presented is supported by the Marie Curie RTN COMSON and ToK project O-MOORE-NICE!.

## References

1. A.C. Antoulas.: Approximation of large-scale Dynamical Systems, advance in design and control, SIAM, 2005.
2. M.J. Rewieński.: A Trajectory Piecewise-Linear Approach to Model Order Reduction of Nonlinear Dynamical Systems. Ph.D. thesis, MIT, USA, 2003.
3. T. Voß, R. Pulch, J. ter Maten, A. El Guennouni.: Trajectory piecewise linear approach for nonlinear differential-algebraic equations in circuit simulation. In: G. Ciuprina, D. Ioan (eds.): *Proc. SCEE. Mathematics in Industry,* vol. 11, Springer, pp. 167–174, 2007.
4. A. Odabasioglu, M. Celik, L.T. Paganini.: PRIMA: Passive reduced-order interconnect macromodeling algorithm. *IEEE TCAD of Integ. Circuits and Systems,* vol. 17(8), pp. 645–654, 1998.
5. R.W. Freund.: SPRIM: structure preserving reduced-order interconnect macromodeling. *Proc. ICCAD,* pp. 80–87, 2004.
6. J. Phillips, L.M. Silveira.: Poor man's TBR: a simple model reduction scheme. *Proc. DATE,* vol. 2, pp. 938–943, 2004.
7. E. J. Grimme.: Krylov Projection Methods for Model Reduction. Ph.D. thesis, University of Illinois at Urbana-Champaign, IL, 1997.
8. P. Feldmann, R.W. Freund.: Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and System* vol. 14(5), pp. 639–649, 1995.
9. K. Gallivan, E. Grimme, P. Van Dooren.: Asymptotic waveform evaluation via a Lanczos method. *Proceedings of the 33rd Conference on Decision and Control* vol. 1, pp. 443–448, 1994.
10. K. Willcox, J. Peraire, J. White.: An Arnoldi approach for generation of reduced-order models for turbomachinery. *Computers and Fluids,* vol. 31(3), pp. 369–389, 2002.
11. R.W. Freund.: Structure-Preserving Model Order Reduction of RCL Circuit Equations. In: H.A. van der Vorst, W.H.A. Schilders & J. Rommes (eds.): Model Order Reduction: Theory, Research Aspects and Applications, Springer 2008.
12. D.C. Sorensen.: Passivity Preserving Model reduction via interpolation of spectral zeros. *System and Control Letters,* vol. 54(4), pp. 347–360, 2005.
13. T. Voß.: Model reduction for nonlinear differential algebraic equations. M.Sc. thesis, University of Wuppertal, Germany, 2005.

# Model Order Reduction for Nonlinear IC Models with POD

Arie Verhoeven, Michael Striebel, and E. Jan W. ter Maten

**Abstract** Due to refined modelling of semiconductor devices and increasing packing densities, reduced order modelling of large nonlinear systems is of great importance in the design of integrated circuits (ICs). Despite the linear case, methodologies for nonlinear problems are only beginning to develop. The most practical approaches rely either on linearisation, making techniques from linear model order reduction applicable, or on proper orthogonal decomposition (POD), preserving the nonlinear characteristic. In this paper we focus on POD. We demonstrate the missing point estimation and propose a new adaption of POD to reduce both dimension of the problem under consideration and cost for evaluating the full nonlinear system.

## 1 Introduction

The dynamics of electrical circuits at time $t$ can be generally described by a nonlinear, first order, differential-algebraic equation (DAE) system of the form:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}[\mathbf{q}(\mathbf{x}(t))] + \mathbf{j}(\mathbf{x}(t)) + \mathbf{B}\mathbf{u}(t) & = \mathbf{0}, \\ \mathbf{y}(t) & = \mathbf{C}^T\mathbf{x}(t), \end{cases} \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ represents the unknown vector of circuit variables at time $t \in \mathbb{R}$; $\mathbf{q}, \mathbf{j} : \mathbb{R}^n \to \mathbb{R}^n$ describe the contribution of reactive and nonreactive elements, respectively; $\mathbf{B} \in \mathbb{R}^{n\times m}$ distributes the input excitation $u : \mathbb{R} \to \mathbb{R}^m$ and $\mathbf{C} \in \mathbb{R}^{n\times q}$

Arie Verhoeven
VORtech Computing, Delft, The Netherlands, e-mail: arie.verhoeven@na-net.ornl.gov

E. Jan W. ter Maten
NXP Semiconductors, Eindhoven, The Netherlands, e-mail: jan.ter.maten@nxp.com

Michael Striebel
Chemnitz University of Technology, Chemnitz, Germany, e-mail: michael.striebel@mathematik.tu-chemnitz.de

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_70,

maps the state $\mathbf{x}$ to the system response $\mathbf{y}(t) \in \mathbb{R}^q$. In circuit design the input $\mathbf{u}$ and the output $\mathbf{y}$ are terminal voltages and terminal currents, respectively, or vice versa. Therefore, we assume that they are linearly injected and extracted, respectively.

The dimension $n$ of the unknown vector $\mathbf{x}(t)$ is of the order of the number of elements in the circuit, which can easily reach hundreds of millions. Therefore, one may solve the network equations (1) by means of computer algebra in an unreasonable amount of time only.

Model order reduction (MOR) aims to replace the original model (1) by a system

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}[\tilde{\mathbf{q}}(\mathbf{z}(t))] + \tilde{\mathbf{j}}(\mathbf{z}(t)) + \tilde{\mathbf{B}}\mathbf{u}(t) &= \mathbf{0}, \\ \tilde{\mathbf{y}}(t) &= \tilde{\mathbf{C}}^T \tilde{\mathbf{x}}(t), \end{cases} \tag{2}$$

with $\mathbf{z}(t) \in \mathbb{R}^r$; $\tilde{\mathbf{q}}, \tilde{\mathbf{j}} : \mathbb{R}^r \to \mathbb{R}^r$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{r\times m}$ and $\tilde{\mathbf{C}} \in \mathbb{R}^{r\times q}$, which can compute a system response $\tilde{\mathbf{y}}(t) \in \mathbb{R}^q$ that is sufficiently close to $\mathbf{y}(t)$ given the same input signal $\mathbf{u}(t)$, but in much less time.

## 2 Linear Versus Nonlinear Model Order Reduction

So far most research effort was spent on developing and analysing MOR techniques suitable for linear problems. For an overview on these methods we refer to [1].

Research on and applications of MOR for nonlinear problems can still be found less frequent. Some approaches like balanced truncation for nonlinear problems [2, 3] are accurate but yet hard to be applied in an industrial context. Others are only feasible for weakly nonlinear dependencies. Then again, when trying to transfer approaches from linear MOR, especially projection based methods, fundamental differences emerge.

To see this, first consider a linear problem of the form

$$\mathbf{E}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) + \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) = 0, \quad \text{with } \mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}. \tag{3}$$

Usually the state $\mathbf{x}(t)$ is approximated in a lower dimensional space of dimension $r \ll n$, spanned by basis vectors which we subsume in $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_r) \in \mathbb{R}^{n\times r}$:

$$\mathbf{x}(t) \approx \mathbf{V}\mathbf{z}(t), \quad \text{with } \mathbf{z}(t) \in \mathbb{R}^r. \tag{4}$$

The reduced state $\mathbf{z}$, i.e., the coefficients of the expansion in the reduced space, is defined by a reduced dynamical system that arises from projecting (3) on a test space spanned by the columns of $\mathbf{W}$. There, $\mathbf{W}$ and $\mathbf{V}$ are chosen, such that their columns are biorthonormal, i.e., $\mathbf{W}^T\mathbf{V} = \mathbf{I}_{r\times r}$. The Galerkin projection[1] yields

$$\tilde{\mathbf{E}}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{z}(t) + \tilde{\mathbf{A}}\mathbf{z}(t) + \tilde{\mathbf{B}}\mathbf{u}(t) = 0, \tag{5}$$

[1] Most frequently $\mathbf{V}$ is constructed to be orthogonal, such that $\mathbf{W} = \mathbf{V}$ can be chosen.

with $\tilde{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{V}$, $\tilde{\mathbf{A}} = \mathbf{W}^T\mathbf{A}\mathbf{V} \in \mathbb{R}^{r\times r}$ and $\tilde{\mathbf{B}} = \mathbf{W}^T\mathbf{B} \in \mathbb{R}^{r\times m}$. The system matrices $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}$ of this reduced substitute model are of smaller dimension and fixed, i.e., need to be computed only once. However, $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}$ are usually dense whereas the system matrices $\mathbf{E}$ and $\mathbf{A}$ are usually very sparse.

Applying the same technique directly to the nonlinear system means obtaining the reduced formulation (2) by defining $\tilde{\mathbf{q}}(\mathbf{z}) = \mathbf{W}^T\mathbf{q}(\mathbf{V}\mathbf{z})$ and $\tilde{\mathbf{j}}(\mathbf{z}) = \mathbf{W}^T\mathbf{j}(\mathbf{V}\mathbf{z})$. Clearly, $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{j}}$ map from $\mathbb{R}^r$ to $\mathbb{R}^r$.

To solve network problems of type (2) numerically, usually multistep methods are used. This means that at each timepoint $t_l$ a nonlinear equation

$$\alpha\tilde{\mathbf{q}}(\mathbf{z}_l) + \tilde{\beta} + \tilde{\mathbf{j}}(\mathbf{z}_l) + \tilde{\mathbf{B}}\mathbf{u}(t_l) = \mathbf{0}, \tag{6}$$

has to be solved for $\mathbf{z}_l$ which is the approximation of $\mathbf{z}(t_l)$. In the above equation $\alpha$ is the integration coefficient of the method and $\tilde{\beta} \in \mathbb{R}^r$ contains history from previous timesteps. Newton techniques that are used to solve (6) usually require an update of the system's Jacobian matrix in each iterations $\nu$:

$$\tilde{\mathbf{J}}_l^{(\nu)} = \left(\alpha\frac{\partial\tilde{\mathbf{q}}}{\partial\mathbf{z}} + \frac{\partial\tilde{\mathbf{j}}}{\partial\mathbf{z}}\right)\Big|_{\mathbf{z}=\mathbf{z}_l^{(\nu)}} = \mathbf{W}^T\left[\alpha\frac{\partial\mathbf{q}}{\partial\mathbf{x}} + \frac{\partial\mathbf{j}}{\partial\mathbf{x}}\right]\Big|_{\mathbf{x}^{(\nu)}=\mathbf{V}\mathbf{z}_l^{(\nu)}}\mathbf{V}. \tag{7}$$

The evaluation of the reduced system, i.e., $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{j}}$, necessitates in each step the back projection of the argument $\mathbf{z}$ to its counterpart $\mathbf{V}\mathbf{z}$ followed by the evaluation of the full system $\mathbf{q}$ and $\mathbf{j}$ and the projection to the reduced space with $\mathbf{W}$ and $\mathbf{V}$.

Consequently, with respect to computation time no reduction will be obtained unless additional measures are taken or other strategies are pursued.

Up to now, approaches based on linearisation, especially the approach of trajectory piecewise linearisation (TPWL) [4, 5], and projection methods based on the Proper Orthogonal Decomposition (POD) are popular. In the following we concentrate on POD and discuss adaptions.

## 3 Proper Orthogonal Decomposition and Adaptions

The POD method, also known as the principal component analysis and Karhunen–Loève expansion, provides a technique for analysing multidimensional data [6–8].

POD sets work on data extracted from a benchmark simulation. In a finite dimensional setup like it is given by (1), $K$ snapshots of the state $\mathbf{x}(t)$, the system is in during the training interval $[t_0, t_e]$, are collected in a snapshot matrix

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K) \ \in \mathbb{R}^{n\times K}. \tag{8}$$

The snapshots, i.e., the columns of $\mathbf{X}$, span a space of dimension $k \le K$. We search for an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of this space that is optimal in the sense that the time-averaged error that is made when the snapshots are expanded in the space spanned by just $r < k$ basis vectors to $\tilde{\mathbf{x}}_{r,i}$,

$$\langle \|\mathbf{x} - \tilde{\mathbf{x}}_r\|_2^2 \rangle \quad \text{with the averaging operator} \quad \langle \mathbf{f} \rangle = \frac{1}{K} \sum_{i=1}^{K} \mathbf{f}_i \tag{9}$$

is minimised. This least squares problem is solved by computing the eigenvalue decomposition of the state covariance matrix $\frac{1}{K}\mathbf{X}\mathbf{X}^T$ or, equivalently by the singular value decomposition (SVD) of the snapshot matrix (assuming $K > n$)

$$\mathbf{X} = \mathbf{UST} \quad \text{with} \quad \mathbf{U} \in \mathbb{R}^{n \times n}, \mathbf{T} \in \mathbb{R}^{K \times K} \text{ and } \mathbf{S} = \left( \begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0}_{n \times (K-n)} \\ & & \sigma_n & \end{array} \right), \tag{10}$$

where $\mathbf{U}$ and $\mathbf{T}$ are orthogonal and the singular values satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_n \geq 0$. The matrix $\mathbf{V} \in \mathbb{R}^{n \times r}$ whose columns span the reduced subspace is now build from the first $r$ columns of $\mathbf{U}$, where the truncation $r$ is chosen such that

$$1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2} \leq \text{tol}. \tag{11}$$

For the, in this way constructed matrix, it holds $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{r \times r}$. Therefore, Galerkin projection as described above can be applied to create a reduced system (2).

For a more detailed introduction to POD in MOR we refer to [9]. For further studies we point to [8] which addresses error analysis for the MOR with POD and [10] where the connection of POD to balanced model reduction can be found.

In the following we reflect two adaptions of POD to overcome the problems that occur in MOR for nonlinear problems and where described in Sec. 2.

### 3.1 Missing Point Estimation

The missing point estimation (MPE) was proposed in [11] to reduce the cost of updating system information in the solution process of time varying systems arising in computational fluid dynamics. In [12] the MPE approach was brought forward to circuit simulation.

Here, once a POD basis is found, such that (4) holds, there is no Galerkin projection applied. Instead a numerical integration scheme is applied which in general leads to system of $n$ nonlinear equations, analogue to (6), for the $r$ dimensional unknown $\mathbf{z}$. In MPE this system is reduced to dimension $g$ with $r \leq g < n$ by discarding $n - g$ equations. Formally this can be described by multiplying the system with a selection matrix[2] $\mathbf{P}_g \in \{0,1\}^{g \times n}$, stating a $g$-dimensional overdetermined problem

$$\alpha \mathbf{P}_g \mathbf{q}(\mathbf{V}\mathbf{z}_l) + \mathbf{P}_g \beta + \mathbf{P}_g \mathbf{j}(\mathbf{V}\mathbf{z}_l) + \mathbf{P}_g \mathbf{B}\mathbf{u}(t_l) = \mathbf{0}, \tag{12}$$

[2] This means, the matrix has exactly one non-zero entry per row.

which is solved at each timepoint $t_l$ for $\mathbf{z}_l$ in the least-squares sense [12]. The benefit is that due to the structure of $\mathbf{P}_g$ not the full nonlinear functions $\mathbf{q},\mathbf{j}$ have to be evaluated but just $g$ components.

The choice of $\mathbf{P}_g$ is motivated by identifying the $g$ most dominant state variables, i.e., components of $\mathbf{x}$. In terms of the POD basis this is connected to restricting the orthogonal $\mathbf{V}$ to $\tilde{\mathbf{V}} = \mathbf{P}_g\mathbf{V} \in \mathbb{R}^{g\times r}$ in an optimal way. This in turn goes down to minimising

$$\| \left(\tilde{\mathbf{V}}^T\tilde{\mathbf{V}}\right)^{-1} - \mathbf{I}_{r\times r}\|. \tag{13}$$

Details on reasoning and solving (13) can be found in [13, 14]

## 3.2 *Adapted POD*

We put a new approach up for discussion that combines the Galerkin projection with the MPE method. Like described in Sec. 3 we collect snapshots in $\mathbf{X}$ on which we apply an SVD (10). Then we define the matrix $\mathbf{L} = \mathbf{U}\Sigma \in \mathbb{R}^{n\times n}$, with $\Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_n)$, i.e., we first scale the left-singular vectors with the corresponding singular values. Next we transform the original system (1) by writing $\mathbf{x}(t) = \mathbf{L}\mathbf{w}(t)$ and using Galerkin projection:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathbf{L}^T\mathbf{q}(\mathbf{L}\mathbf{w}(t))\right] + \mathbf{L}^T\mathbf{j}(\mathbf{L}\mathbf{w}(t)) + \mathbf{L}^T\mathbf{B}\mathbf{u}(t) = \mathbf{0}. \tag{14}$$

Now, we identify separately the $r$ and $g$ most dominant columns of $\mathbf{L}$ and $\mathbf{L}^T$, respectively, where the predominance of a column vector $v \in \mathbb{R}^n$ is determined by its 2-norm $\|v\|_2$. Note that this selection is directly connected to the singular values, i.e., if they decrease rapidly we can expect $r$ and $g$ to be small. We use this information to approximate $\mathbf{L}$ and $\mathbf{L}^T$ by matrices that agree with the respective matrix in the selected $r$ and $g$ selected columns but have the $n-r$ and $n-g$ remaining columns set to $\mathbf{0} \in \mathbb{R}^n$, respectively. Again, formally this can be expressed with the help of selection matrices $\mathbf{P}_r \in \{0,1\}^{r\times n}$ and $\mathbf{P}_g \in \{0,1\}^{g\times n}$, respectively:

$$\mathbf{L} \approx \mathbf{L}\mathbf{P}_r^T\mathbf{P}_r \quad \text{and} \quad \mathbf{L}^T \approx \mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g. \tag{15}$$

From this we conclude $\mathbf{L}^T \approx \mathbf{P}_r^T\mathbf{P}_r\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g$. We insert these approximations in (14) and multiply with $\mathbf{P}_r$, bearing in mind that $\mathbf{P}_r\mathbf{P}_r^T = \mathbf{I}_{r\times r}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathbf{P}_r\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g\mathbf{q}(\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r\tilde{\mathbf{w}})\right] + \mathbf{P}_r\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g\mathbf{j}(\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r\tilde{\mathbf{w}}) + \mathbf{P}_r^T\mathbf{L}^T\mathbf{B}\mathbf{u} = \mathbf{0}. \tag{16}$$

Note that due to the approximations to $\mathbf{L}$ and $\mathbf{L}^T$ in the above equation $\mathbf{w}$ has changed to $\tilde{\mathbf{w}}$ which can merely be an approximation to the former. We introduce $\mathbf{S}_r = \mathrm{diag}(\sigma_1,\dots,\sigma_r)$ and keep the first $r$ columns of $\mathbf{U}$ in $\mathbf{V} \in \mathbb{R}^{n\times r}$. Therewith we express $\mathbf{L}\mathbf{P}_r^T = \mathbf{V}\mathbf{S}_r$. Finally we scale (16) with $\mathbf{S}_r^{-1}$ and introduce a new unknown

$\mathbf{z} = \mathbf{S}_r \mathbf{P}_r \tilde{\mathbf{w}} \in \mathbb{R}^r$ from which we can reconstruct the full state by approximation $\mathbf{x} \approx \mathbf{V}\mathbf{z}$. We end up with

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathbf{W}_{r,g}\mathbf{P}_g\mathbf{q}(\mathbf{V}\mathbf{z})\right] + \mathbf{W}_{r,g}\mathbf{P}_g\mathbf{j}(\mathbf{V}\mathbf{z}) + \tilde{\mathbf{B}}\mathbf{u}(t) = \mathbf{0}, \tag{17}$$

with $\mathbf{W}_{r,g} = \mathbf{V}^T\mathbf{P}_g^T \in \mathbb{R}^{r\times g}$ and $\tilde{\mathbf{B}} = \mathbf{V}^T\mathbf{B}$. Like in the MPE approach just $g$ components of the nonlinear function $\mathbf{q}$ and $\mathbf{j}$ have to be evaluated.

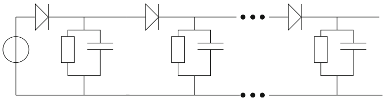
## 4 Numerical Results

We consider the academic diode chain model shown in Fig. 1 with 300 nodes. The current traversing a diode with potential $V_a$ and $V_b$ at the input- and output-node, respectively is described by the nonlinear equation

$$q(V_a, V_b) = \begin{cases} I_s(e^{\frac{V_a - V_b}{V_T}} - 1) & \text{if } V_a - V_b > 0.5, \\ 0 & \text{otherwise,} \end{cases}$$

with threshold voltage $V_T = 0.0256\,\mathrm{V}$ and static current $I_s = 10^{-14}\,\mathrm{A}$. The resistors and capacitors have uniform size $R = 10\,\mathrm{k}\Omega$ and $C = 1\,\mathrm{pF}$.

**Fig. 1** Diode chain



The voltage source defines the input $u(t)$. For the model extraction we choose the step given by

$$u(t) = \begin{cases} 20 & \text{if } t \leq 10\,\mathrm{ns}, \\ 170 - 15\cdot 10^9 \cdot t & \text{if } 10\,\mathrm{ns} < t \leq 11\,\mathrm{ns}, \\ 5 & \text{if } t > 11\mathrm{ns}. \end{cases}$$

As Fig. 2 shows, the signal dies out very quickly and just the first 30 diodes operate. This reflects also in the singular values which drop very rapidly. Therefore, for extracting a reduced order model we start the algorithm with the parameters $r = 30$ and $g = 35$, i.e., the state space is reduced to dimension 30 and the nonlinear functions are downsized to dimension 35.

Of special interest is how a reduced substitute model behaves when signals different to the training signal are applied. For testing purposes we choose

$$\bar{u}_1(t) = 7.5\cos\left(\frac{2\pi t}{60\cdot 10^{-9}}\right) + 12.5 \quad \text{and} \quad \bar{u}_2(t) = 9.5\cos\left(\frac{2\pi t}{60\cdot 10^{-9}}\right) + 12.5.$$

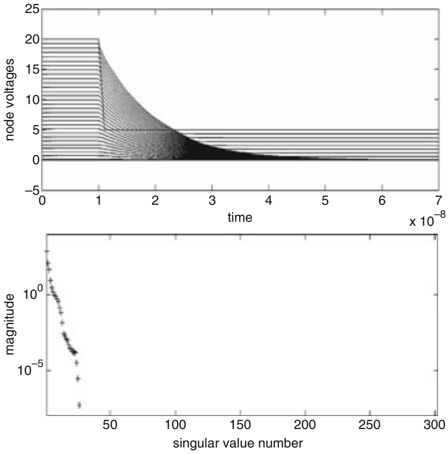


**Fig. 2:** Diode chain: system's response (*left*) and singular values (*right*)

Note that the maximum of $\bar{u}_1(t)$ is less than the maximum of the signal $u(t)$ applied for training, whereas $\bar{u}_2$ exceeds $u(t)$.

Figure 3 shows the voltages of different nodes as they were produced by solving both the full and the reduced nonlinear system. With the reduced model we were able to accurately reproduce the behaviour of the full system when $\bar{u}_1(t)$ was taken as the input. From Table 1 we see that we also achieved a high speedup. Here we also see that the classical POD, i.e, the combination with direct Galerkin projection may even cause more computational work. But, considering the trajectory that was produced with $\bar{u}_2(t)$, we see one of the limitations. An explanation might be that the energy in the system during resimulation was higher than during training and extraction. Similar statements can be found in [15] with respect to TPWL.

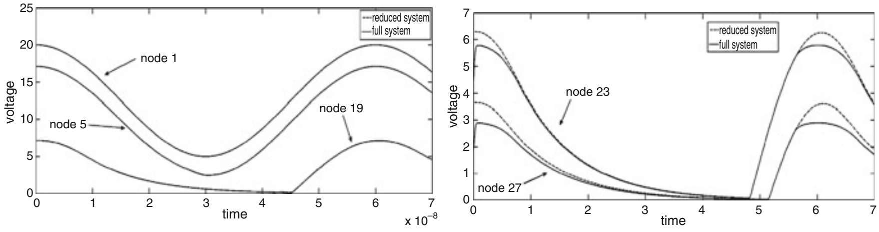


**Fig. 3:** Resimulation with differing input signal $\bar{u}_1(t)$ and $\bar{u}_2(t)$

**Table 1:** Comparison of cpu time [s]

| Input | Full | Classical POD | Adapted POD |
|---|---|---|---|
| Like training | 42.01 | 35.51 | 5.12 |
| $7.5\cos\ldots$ | 40.22 | 45.34 | 6.28 |

## 5 Conclusion and Outlook

In this paper we study reduced order modelling of nonlinear IC models. We review the problems that show up when MOR techniques for linear problems are applied to nonlinear systems. These problems arise from the necessity to still evaluate the full nonlinear system. To this point ways to overcome the problem are to either linearise the nonlinear system and apply MOR to the arising linear systems, like done in TPWL, or to adapt projection methods, like done in MPE in connection with POD. We introduce a new adaption of the latter approach. Put to test with an academic example it shows nice results, especially with input signals that differ from training signals. However, the new approach has to be studied more carefully regarding its general applicability.

**Acknowledgements** The work presented is funded by the Marie-Curie Transfer-of-Knowledge project O-MOORE-NICE!

## References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM (2005)
2. Scherpen, J.M.A.: Balancing for nonlinear systems. Ph.D. thesis, University of Twente (1994)
3. Ionescu, T.C., Scherpen, J.M.A.: Positive Real Balancing for Nonlinear Systems. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering – SCEE 2006, *Mathematics in Industry*, vol. 11, pp. 153–159. The European Consortium for Mathematics in Industry, Springer-Verlag Berlin Heidelberg (2007)
4. Rewieński, M.J., White, J.: A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. IEEE Trans. CAD Int. Circ. Syst. **22**(2), 155–170 (2003)
5. Voß, T., Pulch, R., ter Maten, J., El Guennouni, A.: Trajector piecewise linear aproach for nonlinear differential-algebraic equations in circuit simulation. In: G. Ciuprina, D. Ioan (eds.) Scientific Computing in Electrical Engineering – SCEE 2006, pp. 167–173. Springer (2007)
6. Holmes, P., Lumley, J., Berkooz, G.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambrige University Press, Cambrige, UK (1996)
7. Loève, M.: Probability Theory. Van Nostrand (1955)
8. Rathinam, M., Petzold, L.R.: A new look at proper orthogonal decomposition. SIAM J. Numer. Anal. **41**(5), 1893–1925 (2003)
9. Pinnau, R.: Model reduction via proper orthogonal decomposition. In: W. Schilders, H. van der Vorst, J. Rommes (eds.) Model order reduction: theory, applications, and research aspects, pp. 95–109. Springer (2008)
10. Willcox, K., Peraire, J.: Balanced model reduction via the proper orthogonal decomposition. AIAA Journal **40**(11), 2323–2330 (2002)
11. Astrid, P.: Reduction of process simulation models: a proper orthogonal decomposition approach. Ph.D. thesis, Technische Universiteit Eindhoven (2004)
12. Astrid, P., Verhoeven, A.: Application of least squares mpe technique in the reduced order modeling of electrical circuits. In: Proceedings of the 17th Int. Symp. MTNS, pp. 1980–1986 (2006)
13. Astrid, P., Weiland, S.: On the construction of pod models from partial observations. In: Proceedings of the 44rd IEEE Conference on Decision and Control, pp. 2272–2277 (2005)
14. Verhoeven, A.: Redundancy reduction of ic models by multirate time-integration and model order reduction. Ph.D. thesis, Technische Universiteit Eindhoven (2008)
15. Rewieński, M.J.: A trajectory piecewise-linear approach to model order reduction of nonlinear dynamical systems. Ph.D. thesis, Massachusetts Institute of Technology (2003)

# On Model Order Reduction of Perturbed Nonlinear Neural Networks with Feedback

Marissa Condon and Georgi G. Grahovski

**Abstract** The paper addresses the dynamical properties of large-scale perturbed nonlinear systems of the Hopfield type with feedback. In particular, it focuses on the hyperstability of the equilibria of the system. It proceeds to examine the effect of the empirical balanced truncation model reduction technique on the hyperstability properties. Finally, estimates of the additional conditions for preserving hyperstability when perturbations are present are derived.

## 1 Introduction

Neural networks have attracted the attention of the scientific community for several decades [2, 6]. One of the most important nonlinear neural networks is the Hopfield model [4, 5] which was introduced by J. J. Hopfield in the 1980s. It has been extensively studied (see, e.g., [7] and the references therein) and has found many important applications such as pattern recognition, associative memory and combinatorial optimisation.

The study of the stability of the equilibrium points of dynamical systems is an important area which has been the focus of study over the past number of years. One of the reasons for its importance is that if an equilibrium of a Hopfield neural network is globally asymptotically stable, then the domain of attraction of this point is the entire state space [7].

Marissa Condon, Georgi G. Grahovski
School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland, e-mail: marissa.condon@dcu.ie, grah@eeng.dcu.ie

Georgi G. Grahovski
Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, 72 Tsarigradsko chaussée, 1784 Sofia, Bulgaria

J. Roos and Luis R.J. Costa (eds.), *Scientific Computing in Electrical Engineering SCEE 2008*, Mathematics in Industry 14, DOI 10.1007/978-3-642-12294-1_71,

A special subclass of Hopfield neural networks are Hopfield networks models with feedback. In this case, there is an additional functional dependence imposed between the inputs and the corresponding outputs.

In most of the applications involving neural networks, the model equations form a large-scale system (see e.g. [7] and the references therein). For example, there are approximately $10^{12}$ neurons in the human brain [2]. As a rule, this leads to costly and inefficient computations. Therefore, model reduction is of paramount importance. The reduced model must mirror the properties of the original system if it is to be of practical utility.

Nonlinear model reduction has increasingly become a focus of research as in general, linear models are inadequate to describe real-world processes. While numerous approaches for linear model reduction have been proposed [1], there is a dearth of effective nonlinear model reduction techniques. Balanced truncation, as pioneered by Moore [9], is a very effective linear model reduction technique and consequently, it has been extended by several authors for nonlinear systems. For example, Scherpen introduced the notion of controllability and observability functions to generalise the controllability and observability gramians which characterise linear systems [10]. However, their calculation is computationally expensive and their use is hence, restricted [10, 11]. To counteract this, empirical gramians have been proposed by several authors [12], [13] and [3]. It is the technique in [3] that is adopted in this work.

In the present article, the effect of model reduction on the hyper/stability properties of the nonlinear Hopfield model with feedback is studied.

The structure of the paper is as follows: In Section 2, the perturbed Hopfield neural network model with feedback is described briefly. In Section 3, empirical balanced truncation as a form of model reduction technique is reviewed. The hyperstability criteria (in Popov's sense) and their modification for perturbed nonlinear systems are outlined in Section 4. The effects of model reduction (balanced truncation style) on the hyperstability of nonlinear systems of Hopfield type are studied in Section 5.

## 2 Perturbed Hopfield Model with Feedback

Consider the following system of non-linear ODE's (known as Hopfield models [4, 5]) of the form

$$\dot{x}_i = -b_i x_i + \sum_{j=1}^{N} A_{ij} G_j(x_j) + U_i(t), \tag{1}$$

where $i = 1, \dots, n$, $b_i$'s are constants, $A_{ij}$ form a constant matrix and the external inputs $U_i(t)$ are functions of the time variable $t$. The functions $G_j$ are, in general, nonlinear with respect to the state variables $x_j$, $j = 1, \dots N$ (here $N$ is the number of the neurons in the network). The second term in (1) gives the interconnection between the neurons.

The corresponding perturbed version of the Hopfield model (1) takes the form

$$\dot{x}_i = -\tilde{b}_i x_i + \sum_{j=1}^{N} \tilde{A}_{ij} \tilde{G}_j(x_j) + U_i(t), \tag{2}$$

where $\tilde{b}_i = b_i + \Delta b_i$, $\tilde{A}_{ij} = A_{ij} + \Delta A_{ij}$ (note, that $\Delta A_{ij}$ does not need to be symmetric) and $\tilde{G}_j(u_j) = G_j(u_j) + \Delta G_j(u_j)$, $1 \le i \le n$. Here $\Delta b_i$, $\Delta A_{ij}$ and $\Delta G_j(u_j)$ are considered as (small) perturbations of the system (1).

Model order reduction (Empirical Balance truncation) is applied to the model equations (1), and the paper studies the qualitative behaviour of the solutions of the reduced perturbed model. Special attention is paid to the Popov hyper-stability properties.

Feedback $\mathbf{v}(t)$ in the Hopfield model (1) is introduced by defining a (vector) function $\mathbf{F}$ (nonlinear, in general) of the corresponding outputs $y_i(t)$: $v_i(t) = F(y_i)$. The function $\mathbf{v}(t)$ is, in fact, the output of the corresponding feedback block. This will be discussed in Section 3.

## 3 Model Reduction

The Hopfield model corresponds to a nonlinear system of the generic form

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t)) + \mathbf{B}(t)\mathbf{u}(t), \\ \mathbf{y}(t) &= \mathbf{h}(t, \mathbf{x}(t)), \end{aligned} \tag{3}$$

where $\mathbf{f} : \mathrm{I\!R}^n \to \mathrm{I\!R}^n$ and $\mathbf{h} : \mathrm{I\!R}^n \to \mathrm{I\!R}^q$ are non-linear functions, the vector $\mathbf{u}(t) \in \mathrm{I\!R}^n$ is the input to the system (3), while the vector $\mathbf{y}(t) \in \mathrm{I\!R}^q$ is regarded as an output. For the model equations (1), we identify

$$\mathbf{f}(t, \mathbf{x}(t)) = -b_i x_i + \sum_{j=1}^{n} A_{ij} G_j(x_j)$$

and the feedback $\mathbf{v}(t)$ is given by a nonlinear function $\mathbf{F}(\mathbf{y})$ relating the output $\mathbf{y}(t)$ of the system (3) and the output of the corresponding feedback block $\mathbf{v}(t)$: $\mathbf{v}(t) = -\mathbf{F}(\mathbf{y})$. The block diagram of such a model is depicted on Fig. 1.

Suppose that the equilibrium point is reached when $\mathbf{u}(t) = 0$. Consider the vicinity of an isolated asymptotically stable equilibrium point (steady–state solution) which is supposed to be a constant solution and is chosen for simplicity at $\mathbf{x} = 0$, i.e. $\mathbf{f}(t, 0) \equiv 0$. It is also assumed that the system does not leave the region of attraction of this equilibrium point when the input is applied. If the system exhibits multiple steady–state solutions, then the analysis may be applied separately in the vicinity of each solution provided that extra care is taken to ensure that the system

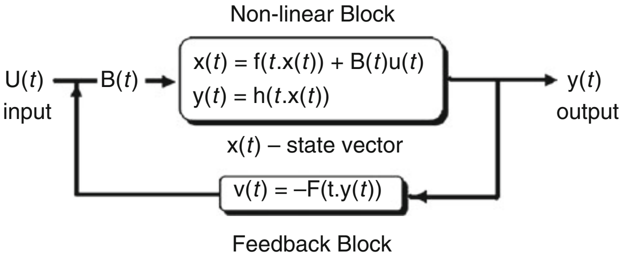


**Fig. 1:** Block diagram of a nonlinear system (3) with feedback

does not leave the region of attraction of the corresponding (asymptotically stable) equilibrium point.

Let also $\mathbf{x}^{ilm}(t)$ be the solution of (3) with $\mathbf{u} \equiv 0$:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)), \qquad \mathbf{x}^{ilm}(0) = c_m T_l e_i. \tag{4}$$

It is assumed that the initial condition in (4) does not take the system outside the region of attraction of the equilibrium point $\mathbf{x} = 0$. Then the ‘state-space average’ of the ‘nonlinear’ fundamental solution may be defined as

$$\langle \Theta(t) \rangle = \frac{1}{rs} \sum_{m=1}^{s} \sum_{l=1}^{r} \sum_{i=1}^{n} \frac{1}{c_m} x^{ilm}(t) e_i^T T_l^T, \tag{5}$$

where $\mathbf{M} \equiv \{c_1, c_2, \ldots, c_s\}$ is the set of $s$ positive constants, $\mathbf{T}^n \equiv \{T_1, T_2, \ldots, T_r\}$ is the set of $r$ orthogonal $n \times n$ matrices and $\mathbf{E}^n \equiv \{e_1, e_2, \ldots, e_n\}$ is the set of standard unit vectors in $\mathbb{R}^n$. Here, also, the superscript “$T$” denotes transposition of a matrix. So for the system in (3), the *nonlinear* controllability gramian is defined as

$$P = \int_0^{\infty} \langle \Theta(-\tau) \rangle^{-1} B(-\tau) B^T(-\tau) \langle \Theta(-\tau) \rangle^{-1T} \mathrm{d}\tau, \tag{6}$$

where $\langle \Theta(t) \rangle$ is as described in (5) and the *nonlinear* observability gramian is defined as [3]

$$Q = \int_0^{\infty} z^T(\tau) z(\tau) \mathrm{d}\tau, \qquad z(t) = \frac{1}{rs} \sum_{i,l,m} \frac{1}{c_m} y^{ilm}(t) e_i^T T_l^T. \tag{7}$$

$y^{ilm}(t)$ is the output which corresponds to an initial state $x^{ilm}(0) = c_m T_l e_i$ and a zero source term.

Let $T$ be the matrix that transforms both $P$ and $Q$ into diagonal form $S$ as follows:

$$TPT^* = S, \quad (T^{-1})^* Q T^{-1} = S, \quad (TPQT^{-1} = S^2).$$

The states of the system are then ordered according to decreasing values of the diagonal entries in $S$. Once balanced, a Galerkin projection $\Pi = [I,0]$, where $\Pi$ is $k \times n$ projection matrix and $I$ is $k \times k$ unit matrix, is then employed to project the transformed system onto the states corresponding to the $k$ largest singular values (i.e. the $k$ largest values of the diagonal matrix $S$ where $k$ is the desired dimension of the reduced-order model).

The reduced model (via empirical balanced truncation) that corresponds to (3) has the form

$$\begin{aligned} \dot{\mathbf{z}}(t) &= \Pi T \mathbf{f}(t, T^{-1}\Pi^{*}\mathbf{z}(t)) + \Pi T \mathbf{B}(t)\mathbf{u}(t), \\ \mathbf{y}(t) &= \mathbf{h}(t, T^{-1}\Pi^{*}\mathbf{z}(t)), \end{aligned} \tag{8}$$

where $T$ is the transformation matrix which casts into a diagonal form both the empirical *controllability* and *observability* gramians, associated with the nonlinear system (3), and $\Pi$ is a Galerkin projection [3].

## 4 Hyperstability of Nonlinear Neural Networks with Feedback

The hyperstability property of dynamical systems is a generalisation of Lyapunov stability. It gives the most general conditions to be imposed on the system in (3) in order to ensure that the solutions are bounded. V. M. Popov introduced the concept of hyperstability in 1973. He introduced it as a generalization of absolute stability for nonlinear systems.

Consider a linear, time-invariant, completely controllable and completely observable system:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)u(t), \qquad \mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t). \tag{9}$$

It is said to be hyperstable [8], if there exists a positive definite symmetric matrix $\mathbf{P}$, a regular matrix $\mathbf{L}$ and an arbitrary matrix $\mathbf{V}$ satisfying the so-called *Kalman-Yakubovich-Equations (KYEs)*:

$$\mathbf{A}^T\mathbf{P} + \mathbf{PA} = -\mathbf{LL}^T, \qquad \mathbf{C}^T = \mathbf{PB} + \mathbf{LV}, \qquad \mathbf{D} + \mathbf{D}^T = \mathbf{V}^T\mathbf{V}, \tag{10}$$

and the *Popov integral inequality*

$$\int_0^t \mathbf{v}^T(\tau)\mathbf{y}(\tau)\,\mathrm{d}\tau \geq -\beta_0^2 \tag{11}$$

holds for all $t \geq 0$ and for some positive constant $\beta_0$. Here, $\mathbf{v}(t)$ is the output of the nonlinear feedback block (Fig. 1). Note that Lyapunov stability is governed by the first Kalman-Yakubovich equation [8].

For nonlinear systems of the form (3), let $\mathbf{x}_R = 0$ be the equilibrium point and $\mathbf{f}(0) = 0$. The condition for *asymptotical hyperstability* of (3) is as follows: The system (3) is asymptotically hyperstable if there exists a continuous feedback control $\mathbf{v}(t) = -F(\mathbf{y})$ satisfying

$$\int_0^t \mathbf{v}^T(\tau) Q \mathbf{x}(\tau)\, \mathrm{d}\tau \leq \beta^2, \tag{12}$$

where $Q$ is the nonlinear observability gramian (7) (being a positive-definite and symmetric matrix) and for some positive constant $\beta < \infty$. This condition must hold true for arbitrary time $t \geq 0$ and does not depend on the initial conditions $\mathbf{x}(0)$.

In order to ensure that Popov's hyperstability criteria is satisfied for generic perturbed nonlinear systems, the perturbations of the nonlinear system must satisfy some additional relations.

In particular, for perturbed nonlinear models of Hopfield type (2), if the perturbations satisfy the estimates

$$\int_0^t |\Delta \tilde{b}_i|^2\, \mathrm{d}t < \infty$$

and

$$\begin{aligned} &\int_0^t |\sum_{i=1}^N \Delta A_{ii} G_i|^2\, \mathrm{d}t + \int_0^t |\sum_{i=1}^N A_{ii} \Delta G_i|^2\, \mathrm{d}t + \int_0^t |\sum_{i=1}^N \Delta A_{ii} \Delta G_i|^2\, \mathrm{d}t \\ &< \alpha \int_0^t \mathbf{v}^T(t) Q \mathbf{x}(t)\, \mathrm{d}t \end{aligned} \tag{13}$$

for some positive constant $\alpha$, then the perturbed Hopfield network (2) will be again asymptotically hyperstable at the origin.

## 5 Model Reduction and Hyperstability of Perturbed Neural Networks

Since most of the nonlinear neural networks with feedback that are of interest are large-scale systems, it is important to determine whether model order reduction for such systems will affect its hyperstability.

Application of empirical balanced truncation as outlined in Section 3, yields a reduced model of the form

$$\bar{\mathbf{x}} = \Pi T \mathbf{x},$$

$$\bar{\mathbf{A}} = \Pi T \mathbf{A} T^{-1} \Pi^* \qquad \bar{\mathbf{G}} = \Pi T \mathbf{G} T^{-1} \Pi^*.$$

Since the transformations $T$ do not depend on time $t$, applying model reduction to the Popov inequality does not affect it. Hence, the hyperstability condition also holds true for the reduced model.

For perturbed Hopfield models, one can derive the following integral estimate for perturbations that preserve the hyperstability property after model reduction:

$$\int_0^t |\sum_{i=1}^{N} \Delta A_{ii} G_i|^2 \,\mathrm{d}t + \int_0^t |\sum_{i=1}^{N} A_{ii} \Delta G_i|^2 \,\mathrm{d}t + \int_0^t |\sum_{i=1}^{N} \Delta A_{ii} \Delta G_i|^2 \,\mathrm{d}t \qquad (14)$$

$$< \alpha \int_0^t \mathbf{v}^T(t) Q \mathbf{x}(t) \,\mathrm{d}t / ||T||^2,$$

where $T$ is the matrix which casts both the gramians into diagonal form and $||\cdot||$ is the standard matrix norm in $\mathrm{I\!R}^n$. This is an explicit relation between the perturbations of the interconnection matrix and the nonlinearities of (2) in order to preserve the stability of the origin for the reduced model.

## 6 Conclusions

Hyperstability of nonlinear networks of the Hopfield type with feedback has been studied. It is shown that empirical balanced truncation preserves hyperstability. In addition, it is shown that if the perturbations of the model parameters satisfy certain additional conditions: (12) and (13), then the reduced perturbed nonlinear Hopfield network is also hyperstable for perturbations satisfying the estimate (14).

**Acknowledgements** One of us (GGG) would like to thank Dr. Neslihan S. Sengor for the valuable discussions during the SCEE 2008 Conference. The authors wish to acknowledge the financial support of Science Foundation Ireland for this research (PI grant 05/IN.1/I18 is acknowledged).

## References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. (Advances in Design and Control Series), SIAM, Philadelphia (2003)
2. Borisyuk, A., Friedman, A., Ermentrout, B., Terman, D.: Tutorials in Mathematical Biosciences I: Mathematical Neuroscience. Lecture Notes in Mathematics **1860**, Springer-Verlag, Berlin (2005)
3. Condon, M., Ivanov, R.: Nonlinear systems-algebraic gramians and model reduction. J. Nonl. Sci. **14**, 405–414 (2004)
4. Hopfield, J.J.: Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proc. Nat. Acad. Sci. USA **79**, 2554–2558 (1982);
   Neurons with graded response have collective computationa properties like those of two-state neurons. Proc. Nat. Acad. Sci. USA **81**, 3088–3092 (1984)
5. Hopfield, J.J., Tank, D.W.: Computing with neural circuits: A model. Science **233**, 625–633 (1986)
6. Edelstein-Keshet, L.: Mathematical Models in Biology. Classics in Applied Mathematics **46**, SIAM, Philadelphia (2005)
7. Michel, A.N., Liu, D.: Qualitative Analysis and Synthesis of Recurrent Neural Networks. Marcel Dekker Inc., New York (2004)
8. Popov, V.M.: Hyperstability of Control Systems. Die Grundlehren der Mathematischen Wissenschaften **245**, Springer-Verlag, Berlin (1973)
9. Moore, B.: Principal component analysis in linear systems: Controllability, Observability and model reduction. IEEE Trans. on Automatic Control **AC-26**(1) (1981)

10. Scherpen, J.M.A.: Balancing of nonlinear systems. Systems and Control Letters **21**, 143–153 (1993)
11. Gray, W.S., Scherpen, J.M.A.: Hankel Operators and Gramians for Nonlinear Systems. Proceedings of the 37th IEEE Conference on Decision and Control (CDC'98), pp. 1416–1421, Tampa, Fl, USA (1998)
12. Lall, S., Marsden, J.E., Glavaski S.: A subspace approach to balanced truncation for model reduction of nonlinear control systems. International Journal of Robust and Nonlinear Control **12**, 519–535 (2002)
13. Hahn, J., Edgar, T.F.: An Improved Method for Nonlinear Model Reduction Using Balancing of Empirical Gramians. Computers and Chemical Engineering **16**, 1379–1397 (2002)

# Author Index