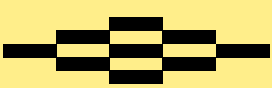
Mathematics in Industry 23
The European Consortium for Mathematics in Industry

Andreas Bartel
Markus Clemens
Michael Günther
E. Jan W. ter Maten *Editors*

Scientific Computing in Electrical Engineering

SCEE 2014, Wuppertal, Germany, July 2014



Springer

MATHEMATICS IN INDUSTRY **23**

THE EUROPEAN CONSORTIUM
FOR MATHEMATICS IN INDUSTRY

SUBSERIES

More information about this series at http://www.springer.com/series/4650

Andreas Bartel • Markus Clemens •
Michael Günther • E. Jan W. ter Maten
Editors

Scientific Computing in Electrical Engineering

SCEE 2014, Wuppertal, Germany, July 2014

Springer

Editors
Andreas Bartel
Applied Math. & Numerical Analysis
Bergische Universität Wuppertal
Wuppertal, Germany

Markus Clemens
Chair of Electromagnetic Theory
Bergische Universität Wuppertal
Wuppertal, Germany

Michael Günther
Applied Math. & Numerical Analysis
Bergische Universität Wuppertal
Wuppertal, Germany

E. Jan W. ter Maten
Applied Math. & Numerical Analysis
Bergische Universität Wuppertal
Wuppertal, Germany

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
ISBN 978-3-319-30398-7 ISBN 978-3-319-30399-4 (eBook)
DOI 10.1007/978-3-319-30399-4

Library of Congress Control Number: 2016939583

Mathematics Subject Classification (2010): 65-06, 65Lxx, 65Mxx, 65Nxx, 65L06, 65L12, 65L15, 65L60, 65L80, 65M06, 65M60, 78-06

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

From July 22 until July 25, 2014, the 10th International Conference on "Scientific Computing in Electrical Engineering" (SCEE) was held in Wuppertal, Germany. It was jointly organized by the Chair of Applied Mathematics and Numerical Analysis and the Chair of Electromagnetic Theory, Bergische Universität Wuppertal.

Due to a generous donation, we were able to use the beautiful *Historische Stadthalle Wuppertal* as our conference venue: a remarkable building in Wilhelminian style, which was inaugurated in 1900. There we welcomed our participants in the Offenbach Saal for all our talks, and we had registration, poster sessions, conference cafe, and personal meetings in the impressive Wandelhalle.

The tenth edition of the SCEE brought together more than 90 scientists from the fields of applied mathematics, electrical engineering, and the computer sciences as well as scientists from industry. Again, it created an excellent working atmosphere, especially due to its unique workshop character, where all talks and poster introductions were presented in the plenary. In addition, we had very clear talks and poster presentations, lively and fruitful discussions, and a great deal of personal networking.

We had a large variety of different talks from excellent invited scientists representing both academia and industry, including an inspiring opening talk by Stéphane Clénet. Our keynote speakers were (in alphabetical order):

Piergiorgio Alotto (Università di Padova, Italy), "Parallelization and Sparsification of a Surface-Volume Integral Code for Plasma-Antenna Interaction"
Stéphane Clénet (Arts & Métiers ParisTech, France), "Approximation Methods to Solve Stochastic Problems in Computational Electromagnetics"
Andreas Frommer (University of Wuppertal, Germany), "Computing $f(A)\mathbf{b}$: The Action of a Matrix Function on a Vector"
Daniel Klagges (Kostal GmbH & Co. KG, Germany), "Simulation of Power Electronics in Automotive Product Development"
Antonino La Magna (CNR Catania, Italy), "Graphene Nano-device Design from First Principles Calculations"

Markus Pistauer (CISC Semiconductor GmbH, Austria), "High-Level Simulation of Cyber-Physical Systems"
Joost Rommes (Mentor Graphics, France), "Different Views on Model-Order Reduction for the Electronics Industry"
Sebastian Schöps (TU Darmstadt, Germany), "Iterative Schemes for Coupled Multiphysical Problems in Electrical Engineering"

Participants of the SCEE 2014 at Schloss Burg, Germany

The topics above are representative of the conference's range. From Tuesday to Friday, we had a total of 30 oral presentations. And in two sessions, 24 posters were presented and discussed.

A special highlight of the SCEE 2014 was our conference excursion to the nearby Wupper valley. Starting at the "Müngstener Brücke" bridge, we went on a small hike, following the river to "Schloss Burg." Visiting the charming residence of the "Counts of Berg," we were told the history of the region "Bergisches Land" and enjoyed a joint dinner, where many ideas and new research directions were discussed.

The book in your hands collects the conference outcomes as proceeding papers. All these papers have successfully passed a standard peer review process. The contributions are divided into five parts, which reflect the main focus areas of the SCEE 2014:

I Device Modeling, Electric Circuits, and Simulation
II Computational Electromagnetics
III Coupled Problems
IV Model-Order Reduction
V Uncertainty Quantification

In the end, we feel we have compiled a very useful and interesting collection. We wish to thank all the participants for their valued contributions to the SCEE 2014 and to this book.

Wuppertal, Germany
November 2015

Andreas Bartel
Markus Clemens
Michael Günther
E. Jan W. ter Maten

Contents

Part I Device Modelling, Electric Circuits and Simulation

Electron Quantum Transport in Disordered Graphene 3
I. Deretzis, V. Romano, and A. La Magna

Latency Exploitation in Wavelet-Based Multirate Circuit Simulation 13
Kai Bittner and Hans Georg Brachtendorf

Turning Points of Nonlinear Circuits 21
Ignacio García de la Vega and Ricardo Riaza

Mixed Domain Macromodels for RF MEMS Capacitive Switches 31
Gabriela Ciuprina, Aurel-Sorin Lup, Bogdan Diţă, Daniel Ioan, Ştefan Sorohan, Dragoş Isvoranu, and Sebastian Kula

Part II Computational Electromagnetics

Systematic Determination of Eigenfields in Frequency Domain 43
Todorka Banova, Wolfgang Ackermann, and Thomas Weiland

DG Treatment of Non-conforming Interfaces in 3D Curl-Curl Problems 53
Raffael Casagrande, Christoph Winkelmann, Ralf Hiptmair, and Joerg Ostrowski

A Symmetric and Low-Frequency Stable Potential Formulation for the Finite-Element Simulation of Electromagnetic Fields 63
Martin Jochum, Ortwin Farle, and Romanus Dyczij-Edlinger

Local Multiple Traces Formulation for High-Frequency Scattering Problems by Spectral Elements 73
Carlos Jerez-Hanckes, José Pinto, and Simon Tournier

Multi-GPU Acceleration of Algebraic Multigrid Preconditioners 83
Christian Richter, Sebastian Schöps, and Markus Clemens

On Several Green's Function Methods for Fast Poisson Solver in Free Space 91
Dawei Zheng and Ursula van Rienen

Part III Coupled Problems

Thermal Simulations for Optimization of Dry Transformers Cooling System 103
Andrea Cremasco, Paolo Di Barba, Bogdan Cranganu-Cretu, Wei Wu, and Andreas Blaszczyk

Multirate GARK Schemes for Multiphysics Problems 115
Michael Günther, Christoph Hachtel, and Adrian Sandu

Iterative Software Agent Based Solution of Multiphysics Problems 123
Matthias Jüttner, André Buchau, Desirée Vögeli, Wolfgang M. Rucker, and Peter Göhner

Simulation of Thermomechanical Behavior Subjected to Induction Hardening 133
Qingzhe Liu, Thomas Petzold, Dawid Nadolski, and Roland Pulch

Tools for Aiding the Design of Photovoltaic Systems 143
Timo Rahkonen and Christian Schuss

Part IV Model Order Reduction

Parametric and Reduced-Order Modeling for the Thermal Analysis of Nanoelectronic Structures 155
Lihong Feng, Peter Meuris, Wim Schoenmaker, and Peter Benner

On Tuning Passive Black-Box Macromodels of LTI Systems via Adaptive Weighting 165
Stefano Grivet-Talocia, Andrea Ubolli, Alessandro Chinea, and Michelangelo Bandinu

Multipoint Model Order Reduction Using Reflective Exploration 175
Elizabeth Rita Samuel, Luc Knockaert, and Tom Dhaene

Interface Reduction for Multirate ODE-Solver 185
Christoph Hachtel, Andreas Bartel, and Michael Günther

Part V Uncertainty Quantification

Approximation Methods to Solve Stochastic Problems in Computational Electromagnetics 199
Stéphane Clénet

Reduced Basis Modeling for Uncertainty Quantification of Electromagnetic Problems in Stochastically Varying Domains 215
Peter Benner and Martin W. Hess

Model Order Reduction for Stochastic Expansions of Electric Circuits ... 223
Roland Pulch

Robust Topology Optimization of a Permanent Magnet Synchronous Machine Using Multi-Level Set and Stochastic Collocation Methods 233
Piotr Putek, Kai Gausling, Andreas Bartel, Konstanty M. Gawrylczyk, E. Jan W. ter Maten, Roland Pulch, and Michael Günther

First Results for Uncertainty Quantification in Co-Simulation of Coupled Electrical Circuits 243
Kai Gausling and Andreas Bartel

Index 253

Authors Index 255

Part I
Device Modelling, Electric Circuits and Simulation

Today's electric and electronic industries rely heavily on computer aided engineering tools. The high complexity of devices and the increasing speed of innovation cycles necessitate virtual prototyping. This allows such production at a competitive time to market because virtual experiments are faster and cheaper than their physical ancestors. Thus numerical tools for those simulations play a key role in the electrical engineering industry. The research focuses in particular on (a) improving the general efficiency and robustness of simulations and (b) the interaction/coupling of multiphysical problems.

The former focus is addressed by Bittner and Brachtendorf in *'Latency Exploitation in Wavelet-based Multirate Circuit Simulation'* in the case of Design Automation of radio frequency (RF) circuits, where the information signal or envelope is modulated by a carrier signal with a center frequency typically in the GHz range. To overcome the prohibitively small time steps in transient simulation demanded by Nyquist's sampling theorem, multirate schemes can be used that transform the DAE network equations into a system of partial DAEs, for short PDAEs. To even speed up classical multirate strategies, wavelet techniques are used combined with subcircuit partitioning strategies to exploit the latencies in different parts of the circuit.

The latter focus is addressed by Ciuprina et al. in *'Mixed Domain Macromodels for RF MEMS Capacitive Switches'* for RF applications again, which follows the companion model approach for a multiphysical device, i.e., instead of simulating the coupled PDE models directly one replaces the device by a subcircuit model, which can be used within circuit simulation packages directly. Here a method is discussed to extract macromodels for radio frequency micromechanical switches, for short RF MEMS switches, which preserve both the multiphysical and the RF behaviour of the device. The outcome is a Spice model, which is controlled by the MEMS actuation voltage and is excited with the RF signal. The modelling errors obtained range from 1 % for the mechanical characteristics to less than 3 % for the RF characteristics.

With decreasing dimensions, new materials attract more attention such as graphene. Deretzis, Romano, and La Magna discuss in *'Electron quantum transport in disordered graphene'* computational strategies for the calculation of quantum

transport in disordered graphene systems from the quasi-one-dimensional to the two-dimensional limit. Usally these strategies suffer from cubic computational costs. Different versions of the non-equilibrium Green's function formalism along with acceleration algorithms can overcome these computational limitations when dealing with two-terminal devices of dimensions that range from the nano- to the micro-scale.

Despite the focus on more computational aspects, some problems in nonlinear circuit theory are still unsolved, which are important from both an analytical and a numerical point of view. A problem in bifurcation theory is discussed by de la Vega and Riaza in *'Turning points of nonlinear circuits'*, which focuses on quadratic turning points. These points may yield saddle-node bifurcations, describing qualitative changes in the solutions. Existence conditions for these points are generlized from the ODE case to the case of semi-explicit DAEs, leading to a characterization in terms of the underlying circuit digraph and the devices' characteristics.

Electron Quantum Transport in Disordered Graphene

I. Deretzis, V. Romano, and A. La Magna

Abstract We discuss the strategies for the calculation of quantum transport in disordered graphene systems from the quasi-one-dimensional to the two-dimensional limit. To this end, we employ real- and momentum-space versions of the non-equilibrium Green's function formalism along with acceleration algorithms that can overcome computational limitations when dealing with two-terminal devices of dimensions that range from the nano- to the micro-scale. We apply this formalism for the case of rectangular graphene samples with a finite concentration of single-vacancy defects and discuss the resulting localization regimes.

1 Introduction

Methodological approaches for the calculation of quantum transport in non-ideal systems are often compromised by computational restrictions, as complex arithmetics and matrix operations that can involve N^3 processes (where $N \times N$ are matrix dimensions) may be necessary. The problem increases when simulations are used for the interpretation of experimental results, as sample dimensions of real devices often range from nanometers to micrometers. It is therefore important to define strategies for the calculation of the conduction characteristics of two-terminal systems with comparable structural characteristics as in real-world experiments. A particular case within this context is the calculation of quantum transport in two-dimensional systems like graphene [1], where disorder is inherently found due to the membrane-like structure of the material, in the form of defects [2] or due to interaction with external parameters like the metallic contacts [3] or the substrate [4]. Additionally, disorder can be engineered through ion-beam processes [5], nano-pattering and nano-lithography [6] or through chemical funtionalization [7]. In all cases, the

I. Deretzis • A. La Magna (✉)
Istituto per la Microelettronica e Microsistemi (CNR-IMM), Z.I. VIII strada 5, 95121 Catania, Italy
e-mail: ioannis.deretzis@imm.cnr.it; antonino.lamagna@imm.cnr.it

V. Romano
Dipartimento di Matematica e Informatica, Università di Catania, Via A. Doria 6, 95125 Catania, Italy

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_1

intrinsic conduction characteristics of the graphene system get significantly altered, with alterations being strongly related to the defect-type or interaction. Such direct relationship between the resulting conduction alteration and its defect origin, make the quantum treatment of transport in graphene-based systems indispensable.

In this paper we discuss strategies for the statistical calculation of quantum transport in disordered graphene systems based on a multiscale approach for the description of the electronic structure and the non-equilibrium Green's function formalism [8]. We pay particular attention to the computational techniques that can allow for the calculation of the conduction variations when gradually passing from the quasi-one-dimensional limit (graphene nanoribbons) to the two-dimensional case (graphene). We finally discuss localization phenomena, the formation of conduction gaps, transport length scales and conductance characteristics for single-vacancy defected graphene.

2 Methodology

Quantum transport is calculated in two-terminal graphene devices, i.e. devices that comprise of a single graphene channel of finite dimensions in contact with two semi-infinite leads. For the sake of simplicity here we consider ideal contacts, i.e. contacts made of graphene with the same lateral width as the channel material. We start from the single-particle retarded Green's function matrix

$$\mathscr{G}^r(\varepsilon) = [\varepsilon S - H - \Sigma_L - \Sigma_R]^{-1}, \tag{1}$$

where ε is the energy, H the real-space Hamiltonian and S the overlap matrix, which in the case of an orthonormal basis set is identical with the unitary matrix I. $\Sigma_{L,R}$ are self-energies that account for the effect of ideal semi-infinite contacts, which can be calculated as:

$$\Sigma_{L(R)} = \tau^{\dagger}_{L(R)} g_{L(R)} \tau_{L(R)} \tag{2}$$

Here $\tau_{L,R}$ are interaction Hamiltonians that describe the coupling between the contacts and the device and $g_{L,R}$ the surface Green functions of the contacts, which can be computed through optimized iterative techniques [9]. The transmission probability of an incident Bloch state with energy ε can be thereon computed as the trace of the following matrix product:

$$T(\varepsilon) = Tr\{\Gamma_L \mathscr{G}^r \Gamma_R [\mathscr{G}^r]^{\dagger}\}, \tag{3}$$

where

$$\Gamma_{L(R)} = i\{\Sigma_{L(R)} - [\Sigma_{L(R)}]^{\dagger}\} \tag{4}$$

are the spectral functions of the two contacts. The reflection coefficient of a single quantum channel can be defined as $R = 1 - T$. According to the Landauer-Buttiker theory [8], conductance can be calculated as:

$$G = \frac{2e^2}{h}T, \tag{5}$$

where $G_0 = 2e^2/h \approx 77.5\mu S$ is the conductance quantum.

The electronic structure of graphene can be easily calculated within a next-neighbor tight-binding (TB) model. Such a description accounts only for the linear combination of π atomic orbitals of graphene, which is however sufficient for the low-energy spectrum of the material. Hence, the next-neighbor TB Hamiltonian can be written as

$$H = -t \sum_{<i,j>,\sigma} c^{\dagger}_{i,\sigma} c_{j,\sigma} + H.c., \tag{6}$$

where $c_i(c_i^{\dagger})$ is the annihilation (creation) operator for an electron with spin σ at site i, and t is the hopping integral with a typical value $t = 2.7\,\text{eV}$. As the objective of the study is to calculate the transport properties of disordered graphene, here we consider the presence of a single type of defect, i.e. carbon vacancies. The simplest and most common method to include a vacancy in a site i of the graphene lattice is to remove its π electron from the model by switching to infinite the related on-site energy term ε_i in the Hamiltonian, or equivalently, by switching to zero the hopping t_{ij} terms between the defected and the neighboring sites. However, a more accurate treatment of the resulting defect states within the electronic spectrum has to take into account the structural reconstruction around the defected site. A method to incorporate such information within the TB model is to perform calculations with methods of higher accuracy (e.g. the density functional theory) and calibrate the TB Hamiltonian in order to reproduce the *ab initio* results. Here, based on density functional theory calculations of defected graphene quantum dots [10], the tuned values of the on-site energy of the defect site and the hopping integrals between this and neighboring sites have been set to $\varepsilon_i = 10\,\text{eV}$ and $t_{ij} = 1.9\,\text{eV}$, respectively. This example is a typical paradigm of the multiscale approach often used for conductance calculations in doped and defected graphene systems.

The previous formalism can be considered as the base-formalism for the calculation of quantum transport in laterally confined graphene systems, as the device Hamiltonian H is written in real space. Considering that direct matrix inversions needed for the calculation of the Green function require N^3 operations, it becomes obvious that such an approach can be solely applied for rather short graphene nanoribbons, i.e. laterally confined stripes of graphene. Notwithstanding the computational power offered by modern processing units, it is difficult for this non-optimized approach to reach dimensions higher than the nanoscale. This aspect introduces a non-negligible problem, especially when direct comparisons between theory and experiments are needed, as graphene samples used for electrical

measurements often have μm dimensions. It is then obvious that new methodologies as well as optimized numerical approaches are crucial for the calculation of quantum transport in such systems. Scaling as a function of the device length can be achieved by taking advantage of the sparsity in the matrices used within the transport formalism (e.g. Hamiltonian and overlap matrices), in order to reduce the required computations. A linear scaling of matrix operations with the system size can be reached through $O(N)$ techniques [11, 12] by creating tridiagonal blocks within the device Hamiltonian. However, even in this case, scalability is limited to the device length, whereas the lateral confinement remains a problem.

A way to overcome the lateral scalability problem in the quantum transport calculation of disordered graphene structures is to consider systems with lateral periodicity and use the discretization of the wave vector perpendicular to the transport direction in order to define the width of the device. Hence, in this case the electronic structure description starts with the k-space Hamiltonian matrix

$$H(\mathbf{k}_\perp) = \sum_m H_{nm} e^{i\mathbf{k}_\perp \cdot (\mathbf{d}_m - \mathbf{d}_n)}, \tag{7}$$

where $\mathbf{k}_\perp$ is the Bloch wave vector within the first Brillouin zone and matrices H_{nm} are written in real space on the previously discussed TB basis set, noting that for $n \neq m$, H_{nm} are interaction matrices between neighboring unit cells, whereas in the case of $n = m$, H_{nn} refers to the Hamiltonian matrix of unit cell n. The single-particle retarded Green's function matrix then becomes

$$\mathcal{G}^r_{k_\perp}(\varepsilon) = [\varepsilon S_{k_\perp} - H_{k_\perp} - \Sigma_{L,k_\perp} - \Sigma_{R,k_\perp}]^{-1}, \tag{8}$$

whereas all equations of the non-equilibrium Green's function formalism maintain the same form. The total conductance of the system in this case is the sum of the single conductances calculated at each sampled k-point and the total width of the device is $W = T_\perp \times n_k$, where $T_\perp$ is the translation vector and n_k the total number of k-points for the $\Gamma \rightarrow X$ path of the rectangular Brillouin zone. We note here that the device unit cell can be any rectangular graphene ribbon that can be periodically repeated along the direction which is perpendicular to transport, with the smallest possible cell being adequate for calculations in ideal (non-defected) systems or systems with line defects parallel to the contacts. However, when random defectiveness is the case, the use of rectangular supercells is mandatory. In this case the bigger the periodic supercell used for each k-point calculation, the smaller the error due to periodicity will be. Finally, particular attention has to be paid when performing statistical calculations in disordered graphene systems. Here statistical variations (e.g. the fluctuations of the conductance) can only be correctly evaluated by the real-space formalism, whereas statistical means (e.g. the total conductance) can be correctly evaluated by both real- and momentum space formalisms.

3 Results

3.1 Ideal Graphene

The conductance of ideal graphene systems is characterized by significant qualitative variations when quantum confinement becomes important, i.e. in the case of narrow graphene nanoribbons. In this case the formation of sub-bands in the electronic structure [13] gives rise to integer plateaus in the calculated conductance. Figure 1 shows the ideal conductance of graphene nanoribbons with armchair-type edges and variable widths. It is important to note that within the nearest-neighbor TB picture, narrow armchair ribbons can be either metallic or semiconducting, strictly based on the number of dimer lines that define their width. A simple geometric rule deriving from such calculation shows that $\forall p \in N$, ribbons with $N_a = 3p + 2$ dimer lines are metallic while the rest are semiconducting. The main differences observed in the calculated conductance when gradually increasing the lateral width W of the ribbons are: (a) the total conductance proportionally increases with W, as new conduction channels are added to the device, (b) the conductance plateaus progressively become smaller and (c) the band gaps (when they exist) also follow a decreasing trend. From Fig. 1 it is also clear to see that after a transition range when $W \approx 50 - 100$ nm, both conductance plateaus and bandgaps become extremely small and the V-shape of two-dimensional graphene conductance appears. A further increase of W only gives rise to quantitative differences whereas the conductivity of the system remains the same.

Considering calculations in ideal systems, the transport signatures of both one- and two-dimensional graphene should be clearly identifiable in experiments, as in the former case conductance plateaus should appear, whereas in the latter the conductance should reveal a V-shape. In practice, experimentally it is very

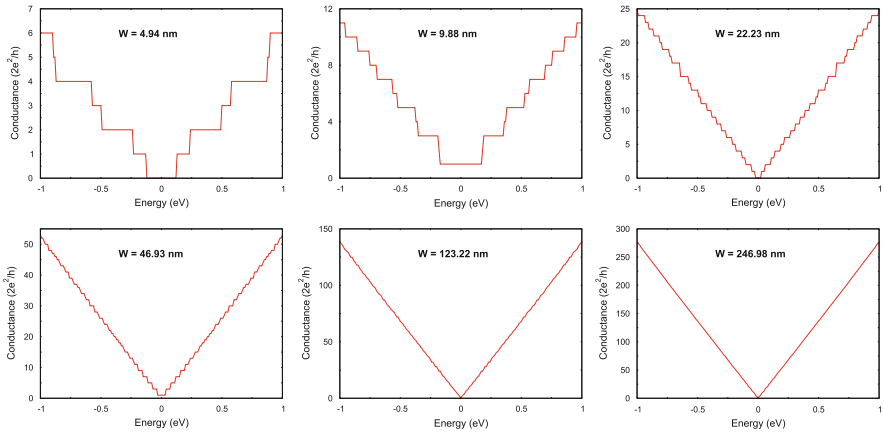


Fig. 1 Ideal conductance of graphene ribbons with armchair-type edges and variable widths

difficult to demonstrate a quantized graphene conductance even for very narrow graphene ribbons [14]. The origin of this discrepancy is often identified in the presence of disorder, in the form of either structural defects or in the interaction between graphene and its substrate or contacts. It therefore becomes clear that the calculation of quantum transport in graphene considering plausible sources of disorder is fundamental for the correct assessment of the experimental results by the simulations.

3.2 *Defected Graphene*

Structural defects are very common in graphene samples as they can be generated during the mechanical, chemical or epitaxial growth process. Apart from the local transformation of the hexagonal graphene lattice, such defects give rise to quasi-localized states within the eigenspectrum [15] with resonances that are characteristic of the defect type [16]. The simplest structural defect in graphene is the single vacancy, whose defect states have resonances which impact more heavily on the valence band of the low-energy spectrum rather than on the conduction band, resulting in a conductance asymmetry [17]. Considering a disordered graphene system with just this type of defect, it is very interesting to visualize the alterations of the conductance characteristics as well as the transition of the various localization regimes when altering the geometrical characteristics of the devices.

Figure 2 shows conductance means (solid lines) and fluctuations (points) for a statistical calculation of 100 replicas of a graphene system with fixed $W = 9.88\,\text{nm}$ and defect concentration 0.5 %, while scaling the device length L from 20 to 212 nm. Starting from $L = 20\,\text{nm}$, the conductance distribution has the following characteristics: (a) the ideal symmetry of the graphene conductance around the charge neutrality point breaks, as a result of the resonances of the single-vacancy states that have a higher density at the valence band of the system rather than the

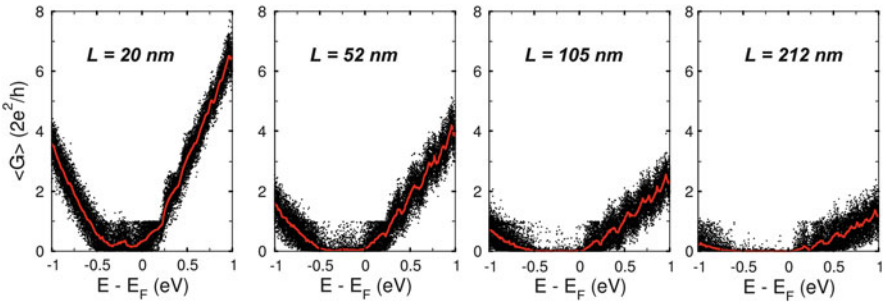


Fig. 2 Scaling of the conductance distribution as a function of length for a graphene nanoribbon with $W = 9.88\,\text{nm}$ and a fixed vacancy concentration of 0.5 %. *Solid lines* represent the mean conductance obtained from 100 different configurations of the system (shown as points)

conduction band. In particular a defect state at $\varepsilon \approx -0.35\,\text{eV}$ gives rise to a local dip of the conductance. (b) Notwithstanding the ideal conductance of this system is characterized by plateaus due to its relatively small width (see Fig. 1), the presence of defects suppresses such feature. It can be therefore argued that even for relatively low defect concentrations it is very difficult to recover quantization features in the conductance. (c) Conductance fluctuations are present throughout the calculated energy spectrum with the conductance variation being $\delta G \sim 2e^2/h$, implying a weak localization regime with universal conductance fluctuations regardless of the number of conduction channels. By gradually increasing the length of the device the following conductance alterations can be seen: (a) the total conductance of the system follows a decreasing trend as a result of the increase of the scattering processes within the device. (b) A transport gap opens at an energy range $-0.5\,\text{eV} \leq \varepsilon \leq 0\,\text{eV}$ due to the total scattering of the electron waves at such energies from the defect states. It is also very important to see that at this energy region the opening of such a transport gap is accompanied by a strong suppression of the conductance fluctuations δG, which also defines a change in the localization regime. A method for defining if a disordered system operates within the weak or strong localization regimes is by calculating its characteristic localization length ξ from:

$$< G > \propto e^{-\frac{2L}{\xi}} \tag{9}$$

Then, a system with fixed W and defect concentration can be characterized as being in the weak localization regime if the device length $L \ll \xi$, and similarly, being in the strong localization regime if $L \gg \xi$. For the case of the ribbon of Fig. 2 the calculated value of ξ for $\varepsilon = -0.15\,\text{eV}$ (i.e. an energy value within the transport gap) is found to be 39.24 nm, implying that for the two configurations with $L = 105\,\text{nm}$ and $L = 212\,\text{nm}$, the system is within the strong localization regime at this energy. It is important to note that the localization length strongly depends on the density of defects. Table 1 shows the calculated ξ for the graphene ribbon of $W = 9.88\,\text{nm}$ and single-vacancy concentrations of 0.2, 0.5, 1, and 2 % at energy $\varepsilon = -0.15\,\text{eV}$. It is clear that ξ becomes smaller as the defect concentration increases.

The influence of the device geometry on the conductance distribution of a disordered graphene system is also important when scaling concerns its lateral width. Figure 3 shows conductance means (solid lines) and fluctuations (points) for a statistical calculation of 100 replicas of a graphene system with fixed $L = 20\,\text{nm}$ and defect concentration (0.5 %),while varying the device width W from 4.3 to 21 nm. Apart from a gradual increase of the conductance due to the insertion of new

Table 1 Localization length ξ for a graphene nanoribbon with width $W = 9.88\,\text{nm}$ and single-vacancy concentrations of 0.2, 0.5, 1, and 2 % at energy $\varepsilon = -0.15\,\text{eV}$

Vac. (%)	ξ (nm)
0.2	50.40
0.5	39.24
1.0	28.39
2.0	25.64

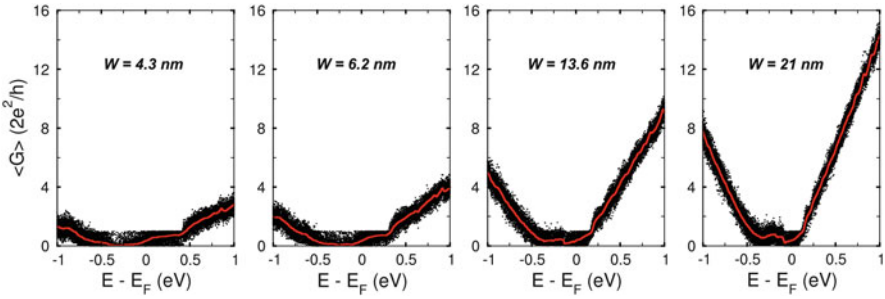


Fig. 3 Scaling of the conductance distribution as a function of width for a graphene nanoribbon with $L = 20\,\text{nm}$ and a fixed vacancy concentration of 0.5 %. *Solid lines* represent the mean conductance obtained from 100 different configurations of the system (shown as points)

conductance channels as the device becomes larger, there are also some qualitative aspects that denote transitions between localization regimes. In particular, it is clear that in the narrower graphene ribbons ($W = 4.3\,\text{nm}$ and $W = 6.2\,\text{nm}$) the defect concentration is high enough for the creation of transport gaps at energy resonances relative to single-vacancy defect states, where also the conductance fluctuations are partially suppressed. Such characteristics are typical of the strong localization regime for disordered graphene nanoribbons. On the contrary, the same defect concentration fails in opening a transport gap for wider ribbons, and similarly, the conductance fluctuations recover characteristics which can be attributed to the weak localization regime. The key issue that emerges here is that for the same level of disorder, strong localization can be achieved easier for narrower graphene samples rather than for wider ones. Another important issue regards the conductance fluctuations δG in the weak localization regime, which appear to be independent from the width of the device, maintaining a fixed value around the conductance mean.

A further increase of the width for the disordered graphene ribbon shows that above a certain value of W, differences become only quantitative, as the total conductance increases proportionally with W. Figure 4 shows the mean conductance calculated for a graphene system with $L = 20\,\text{nm}$, $W = 39.4\,\text{nm}$ and the same defect concentration as before. For this calculation the k-space formalism has been employed, using a rectangular graphene supercell of $L = 20\,\text{nm}$ and $W = 9.88\,\text{nm}$, while sampling the Brillouin zone $\Gamma \rightarrow X$ path at $n_k = 4$ k-points (we note that the k-point sampling is not arbitrary, but considers an equidistant separation of the entire $\Gamma \rightarrow X \rightarrow \Gamma$ path in $[2 \times n_k + 1]$ regions). Further increases of the width practically give rise to the same conductivity. This aspect brings to discussion the definition of the transition range between quasi-one-dimensional and two-dimensional transport. Our calculations show that such transition depends on the concentration of the defects, with systems having higher concentrations transiting faster towards the two-dimensional limit. In all cases, the presence of disorder should facilitate the

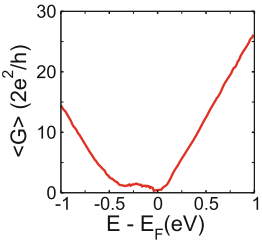


Fig. 4 Mean conductance as a function of energy for a graphene ribbon with $W = 39.4$ nm, $L = 20$ nm and a fixed vacancy concentration of 0.5 %. The mean value has been calculated over 100 replicas of equivalent systems with variable random positions of the defect sites

transition of a graphene system to the two-dimensional situation with respect to the ideal case (see Fig. 1).

4 Discussion

The objective of this paper has been to discuss the computational strategies for the calculation of quantum transport in disordered graphene systems, in a way that these can be relevant to experimental conductance measurements. To this end, both real- and momentum-space formulations of the non-equilibrium Green's function formalism have been employed along with acceleration algorithms that can make calculations computationally more affordable. Within this context, the paper has tried to evidence that the dimensions of the graphene channels as well as their level of disorder can be fundamental for the manifestation of different transport features and localization regimes. As a general picture, our results have evidenced that narrow graphene nanoribbons are more influenced by defectiveness with respect to wider ones, as the strong localization regime can be reached easier. We have moreover seen that disorder facilitates the transition of the conduction characteristics from the one-dimensional to the two-dimensional limit.

Although the level of knowledge regarding quantum transport calculations in graphene-based systems is by now well consolidated, there are still plenty of challenges and open issues that have to be affronted in the forthcoming years. A first aspect has to do with the level of complexity in the modeling of defectiveness, as often calculations consider single-type defects, contrary to the intrinsically more complex experimental scenario. A second issue regards the correct assignment of statistical distributions in quantities like the conductance fluctuations [17, 18], especially in the presence of asymmetric disorder (e.g. when defect states are not symmetric with respect to the charge neutrality point). Finally an important issue remains the calculation of disorder effects on the transport characteristics of two-dimensional materials beyond graphene, as in most cases the nearest-neighbor TB Hamiltonian is not adequate for these systems.

References

1. Geim, A.K., Novoselov, K.S.: The rise of graphene. Nat. Mater. **6**, 183–191 (2007)
2. Banhart, F., Kotakoski, J., Krasheninnikov, A.V.: Structural defects in graphene. ACS Nano **5**, 26–41 (2010)
3. Deretzis, I., Fiori, G., Iannaccone, G., La Magna, A.: Atomistic quantum transport modeling of metal-graphene nanoribbon heterojunctions. Phys. Rev. B **82**, 161413 (2010)
4. Nicotra, G., Ramasse, Q.M., Deretzis, I., La Magna, A., Spinella, C., Giannazzo, F.: Delaminated graphene at silicon carbide facets: atomic scale imaging and spectroscopy. ACS Nano **7**, 3045–3052 (2013)
5. Compagnini, G., Giannazzo, F., Sonde, S., Raineri, V., Rimini, E.: Ion irradiation and defect formation in single layer graphene. Carbon **47**, 3201–3207 (2009)
6. Ci, L., Xu, Z., Wang, L., Gao, W., Ding, F., Kelly, K.F., Yakobson, B.I., Ajayan, P.M.: Controlled nanocutting of graphene. Nano Res. **1**, 116–122 (2008)
7. Georgakilas, V., Otyepka, M., Bourlinos, A.B., Chandra, V., Kim, N., Kemp, K.C., Hobza P., Zboril R., Kim, K.S.: Functionalization of graphene: covalent and non-covalent approaches, derivatives and applications. Chem. Rev. **112**, 6156–6214 (2012)
8. Datta, S.: Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge (1997)
9. Sancho, M.L., Sancho, J.L., Sancho, J.L., Rubio, J.: Highly convergent schemes for the calculation of bulk and surface Green functions. J. Phys. F: Met. Phys. **15**, 851 (1985)
10. Deretzis, I., Forte, G., Grassi, A., La Magna, A., Piccitto, G., Pucci, R.: A multiscale study of electronic structure and quantum transport in $C_{6n^2}H_{6n}$-based graphene quantum dots. J. Phys. Condens. Matter **22**, 095504 (2010)
11. Anantram, M.P., Govindan, T.R.: Conductance of carbon nanotubes with disorder: A numerical study. Phys. Rev. B, **58**, 4882 (1998)
12. Petersen, D.E., Li, S., Stokbro, K., Sørensen, H.H.B., Hansen, P.C., Skelboe, S., Darve, E.: A hybrid method for the parallel computation of Green's functions. J. Comput. Phys. **228**, 5020–5039 (2009)
13. Deretzis, I., La Magna, A.: Coherent electron transport in quasi one-dimensional carbon-based systems. Eur. Phys. J. B **81**, 15–36 (2011)
14. Lin, Y.M., Perebeinos, V., Chen, Z., Avouris, P.: Electrical observation of subband formation in graphene nanoribbons. Phys. Rev. B **78**, 161409 (2008)
15. Deretzis, I., Fiori, G., Iannaccone, G., La Magna, A.: Effects due to backscattering and pseudogap features in graphene nanoribbons with single vacancies. Phys. Rev. B **81**, 085427 (2010)
16. Deretzis, I., Piccitto, G., La Magna, A.: Electronic transport signatures of common defects in irradiated graphene-based systems. Nucl. Instrum. Methods Phys. Res., Sect. B Beam Interactions Mater. Atoms **282**, 108–111 (2012)
17. La Magna, A., Deretzis, I., Forte, G., Pucci, R.: Conductance distribution in doped and defected graphene nanoribbons. Phys. Rev. B **80**, 195413 (2009)
18. La Magna, A., Deretzis, I., Forte, G., Pucci, R.: Violation of the single-parameter scaling hypothesis in disordered graphene nanoribbons. Phys. Rev. B **78**, 153405 (2008)

Latency Exploitation in Wavelet-Based Multirate Circuit Simulation

Kai Bittner and Hans Georg Brachtendorf

Abstract The simulation of radio frequency (RF) circuits is one of the severest problems in Design Automation: the information signal or envelope is modulated by a carrier signal with a center frequency typically in the GHz range. Due to Nyquist's sampling theorem the time steps in conventional transient analysis are prohibitively small. A technique to overcome Nyquist's bottleneck is the multirate method which reformulates the ordinary circuit's differential algebraic equations (DAEs) as a system of partial DAEs (PDAEs). In this paper further improvements of the wavelet multirate circuit simulation technique are presented. In the new algorithm we use different grids for the approximation of the solution on different circuit parts, exploiting latency. In particular, for circuits comprising latent parts the grids can be much sparser, which results in the reduction of the overall problem size and leads to a faster simulation.

1 Introduction

In simulation of RF circuits one faces waveforms with a spectrum centered around a center frequency, which is typically in the GHz range for modern communication standards. Due to the Nyquist's theorem the waveforms must be discretized with a sampling rate, which is at least twice as high as the highest relevant frequency in the spectrum. Classical transient solvers which solve the initial value problem (IVP) show unacceptably long run times. To overcome this bottleneck envelope methods based on a reformulation of the ordinary DAEs by partial DAEs, known as multirate PDAE have been developed [1–8]. However, despite this tremendous progress, the run time is often prohibitively long for circuits such as PLLs. In this paper improvements based on latency exploitation are proposed, which utilize some specific properties of (sub-) circuits of RF circuitry and properties of the multirate PDAE.

K. Bittner (✉) • H.G. Brachtendorf
University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria
e-mail: Kai.Bittner@fh-hagenberg.at; Hans-Georg.Brachtendorf@fh-hagenberg.at

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_2

For speeding up conventional transient analysis, several attempts have been made for exploiting latency and other specific properties of circuit DAEs. An excellent overview of these methods can be found, e.g., in [9]. In [10] a relaxation method denoted as timing analysis has been presented based essentially on one Gauss-Seidel (GS) iteration per time step. This method has an emphasis on CMOS circuits where the diagonal part of the Jacobian matrix is dominant. Alternatively, the waveform relaxation (WR), e.g., [11, 12], has attracted attention for several decades. Here the circuit is divided into sub-circuits wherein the coupling of these sub-circuits is relatively week. Each sub-circuit is simulated for a time period while the remaining sub-circuits are idle or latent. The method is repeated until convergence is achieved. WR may be interpreted as a block Gauss-Seidel for a time period. It has been developed for CMOS circuits, too. In [13] the latency insertion method (LIM) has been proposed, which has its origin in electromagnetic field simulation. Essentially, the discretization grid for the unknown currents and voltages are shifted. The technique is advantageous when delays of interconnects are dominant. The time steps however are limited similar to the CFL condition. Node tearing, often with latency exploitation, has been reported, e.g., in [14–17]. In [15, 17] the sub-circuits are allowed to have separate integration step sizes reflecting their activity level.

For all the cited methods a careful partitioning and/or time step control is required to achieve convergence. The method for the latency exploitation considered in this paper, which is based both on the multirate PDAE and spline-wavelet technique, has none of these restrictions.

2 The Multirate Circuit Simulation Problem

We consider circuit equations in the charge/flux oriented modified nodal analysis (MNA) formulation, which yields a mathematical model in the form of a system of differential-algebraic equations (DAEs):

$$\frac{d}{dt}q\big(x(t)\big) + g\big(x(t)\big) = s(t). \tag{1}$$

For RF circuits the circuit DAE (1) exhibits multirate behavior, i.e., (most) of the signal waveforms have a bandpass spectrum, where the spectrum is centered around a center frequency, which is typically in the GHz range for state of the art mobile phone standards. The time steps employing conventional solvers for ordinary DAEs must be kept sufficiently small to avoid aliasing of the numerical solution. The run time is therefore prohibitively long. One method to overcome this bottleneck reformulates the underlying ordinary DAEs by a system of partial DAEs [1, 2]. Several modifications of this method have been proposed [3–7, 18]. To separate different time scales the problem is reformulated as a multirate PDAE, i.e.,

$$\Big(\tfrac{\partial}{\partial\tau} + \omega(\tau)\,\tfrac{\partial}{\partial t}\Big)q\big(\hat{x}(\tau,t)\big) + g\big(\hat{x}(\tau,t)\big) = \hat{s}\big(\tau,t\big) \tag{2}$$

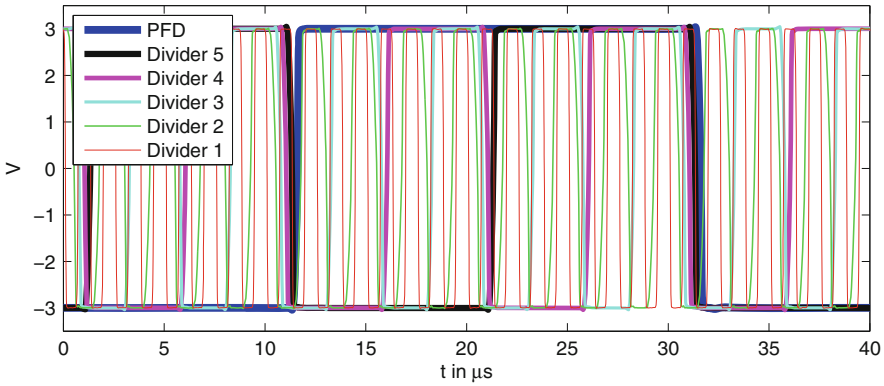


Fig. 1 Several signals in a frequency divider chain as part of a PLL

with mixed initial-boundary conditions $x(0,t) = X_0(t)$ and $x(\tau,t) = x(\tau,t+P)$. A solution of the original circuit equations can be found along certain characteristic lines [8].

Discretization with respect to τ (Rothe's method) using a linear multistep method results in a periodic boundary value problem in t of the form

$$\begin{aligned} \omega_k \tfrac{d}{dt} q_k\big(X_k(t)\big) + f_k(X_k,t) &= 0, \\ X_k(t) &= X_k(t+P), \end{aligned} \tag{3}$$

where $X_k(t)$ is the approximation of $\hat{x}(\tau_k,t)$ for the k-th time step τ_k (cf. [8, 19]). The periodic boundary value problem (3) can be solved by several methods, as Shooting, Finite Differences, Harmonic Balance, etc. Here, we consider the spline wavelet based method introduced by the authors in [19], following ideas from [20, 21]. One problem of traditional methods is that all signals in the circuit are discretized over the same grid. This can pose a problem if different signal shapes are present in the circuit, which may be approximated more efficiently if individual grids are used for each of the signals. As an example we consider a chain of 5 frequency dividers (as part of a PLL). In each step the frequency is reduced by a factor 2 as one can see in Fig. 1, where the solution for a fixed τ is shown. Obviously, for the low frequency signals towards the end of the divider chain a much sparser grid would be sufficient for an accurate representation, in comparison to the high frequency input signal.

3 Division into Subcircuits

Although the representation of each signal over its own individual grid seems to give maximal flexibility, this approach leads to several problems, which make the simulation inefficient. One problem is that the evaluation of the circuit depends on

the evaluation of device models, which is usually very costly, e.g., for transistor models used nowadays. Usually a device model has to be evaluated for every grid point. For a four terminal transistor we need therefore to evaluate the device model for four different grids. In many cases this will be more costly than the evaluations for one optimized single grid, which is against our intention to reduce the computational effort. This effort might be reduced if one has a strategy to "synchronize" the grids, but that does not seem to be a trivial task. On the other side, the signals show often similar behavior, at least on parts of the circuit, such that the same grid might be (nearly) optimal for many signals. Thus, it might be a better idea to collect signals of similar shape and use the same grid for all of these signals. Then we have to store only a few grids, which makes it also easier to design an effective grid adaptation strategy. Therefore, we consider groups of signals with similar shape appearing in a part of the circuit. In the current implementation a priori knowledge of the circuit design is required.

The circuit is divided into N subcircuits which are connected at terminal nodes. To facilitate different expansions of signals on the subcircuits we replace each common node by a pair of nodes connected by a perfect conductor, which is referred to as node tearing. Namely, we introduce the "connection" $C^{k,\ell}_{\mu,\nu}$, if the μ-th node of subcircuit k is identified with the ν-th node in subcircuit ℓ, as one can see in Fig. 2. Thus, the circuit equations from the modified nodal analysis (MNA) of the subcircuits have to be supplemented by additional conditions for the connections. The perfect conductor for the connection is modeled as voltage source of voltage zero. That is, we need the current through the connection $C^{k,\ell}_{\mu,\nu}$, as additional unknowns i^k_μ and i^ℓ_ν for each of the two involved subcircuits.

In addition to the resulting circuit equations

$$\frac{d}{dt}q^k\big(x^k(t)\big) + g^k\big(x^k(t), t\big) = 0, \qquad k = 1, \ldots, N \tag{4}$$

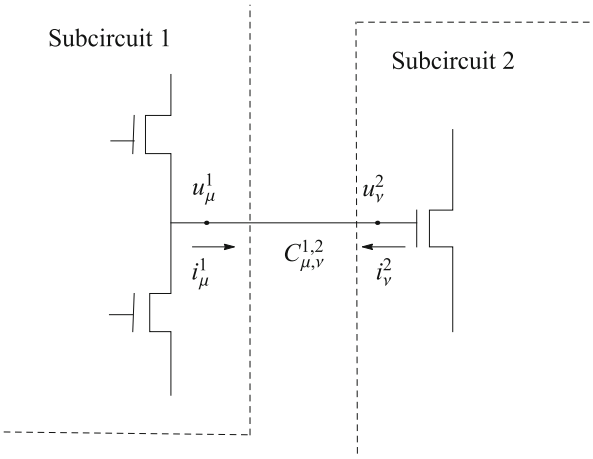


Fig. 2 Splitting of a circuit into subcircuits witch connections

of the N subcircuits, we need for each connection $C_{\mu,\nu}^{k,\ell}$ that voltages and currents coincide, that is we include the equations

$$u_\mu^k(t) - u_\nu^\ell(t) = 0 \quad \text{and} \quad i_\mu^k(t) + i_\nu^\ell(t) = 0. \tag{5}$$

For the correct understanding of the above formulation one needs to recall that $u_\mu^k(t)$ and $i_\mu^k(t)$ are components of the vector $x^k(t)$ of unknowns, which contains *all node voltages* (except ground) and currents through voltage sources, inductors, and *connections*.

The splitting into subcircuits introduces the additional Eq. (5), which will increase the problem size. This may lead to a loss of performance if the splitting is done poorly. For a successful use of our method the splitting, either done by hand or automatically, should follow some rules. First, a splitting should only be done if the signal shapes in the resulting subcircuits differ enough to justify the use of different discretization grids, which are significantly coarser than the grid for the (sub)circuit, which is splitted. Furthermore, the splitting should only generate few new *connections*. We expect that the second requirement is often fulfilled if the first requirement is satisfied

4 Spline Galerkin Discretization and Wavelet Based Adaptivity

Our goal is to approximate the solution of the Eqs. (4) and (5) by spline functions as it was done in [19]. However, we want to use an adapted spline representation for each subcircuit, i.e.,

$$x^i(t) = \sum_{k=1}^{n_i} c_k^i \varphi_k^i(t), \quad i = 1, \ldots, N,$$

where the families $\{\phi_k^i : k = 1, \ldots, n_i\}$ are periodic B-spline bases for spline spaces over grids of spline knots $T^i := \{t_k^i \in (0, P] : k = 1, \ldots, n_i\}$, which may be mutually different. We use a Petrov-Galerkin discretization to obtain a system of nonlinear equations, which determines the coefficients c_k^i. In particular, we integrate the Eqs. (4) and (5) over subintervals, i.e.,

$$\int_{\tau_{\ell-1}^i}^{\tau_\ell^i} \tfrac{d}{dt} q^i\big(x^i(t)\big) + g^i\big(x^i(t), t\big)\, dt = 0, \qquad \ell = 1, \ldots, n_i,$$

for each subcircuit and

$$\int_{\tau^i_{\ell-1}}^{\tau^i_\ell} u^i_\mu(t) - u^j_\nu(t)\,dt = 0, \qquad \ell = 1,\ldots,n_i \tag{6}$$

$$\int_{\tau^j_{\ell-1}}^{\tau^j_\ell} i^i_\mu(t) + i^j_\nu(t)\,dt = 0, \qquad \ell = 1,\ldots,n_j \tag{7}$$

for each connection $C^{i,j}_{\mu,\nu}$ between two subcircuits. The splitting points τ^i_ℓ are chosen in close relation to the spline grid, namely such the $t^i_\ell \in (\tau^i_{\ell-1}, \tau^i_\ell)$. By using the grid T^i in (6) but T^j in (7), we assure that the number of unknowns and equations coincide.

The wavelet based coarsening and refinement methods described in [19, 22] are used to generate adaptive grids for an efficient signal representation. An advantage of this approach is that grid and solution from the previous envelope time step are used to generate an initial guess for Newton's method. Since the waveforms change only slowly with τ, we have usually a very good initial guess on a nearly optimal grid and the solution is obtained with only few iteration steps.

5 Numerical Test

The algorithm was implemented in C++ and tested on a PLL with frequency divider. The solutions for a fixed τ_k are shown in Fig. 1.

For comparison we show in Fig. 3 the spline grid generated by the classical spline wavelet method (see [19]) using the same grid for all signals. We have plotted the grid points t_i against their index i, which allows to recognize the local density of the grid.

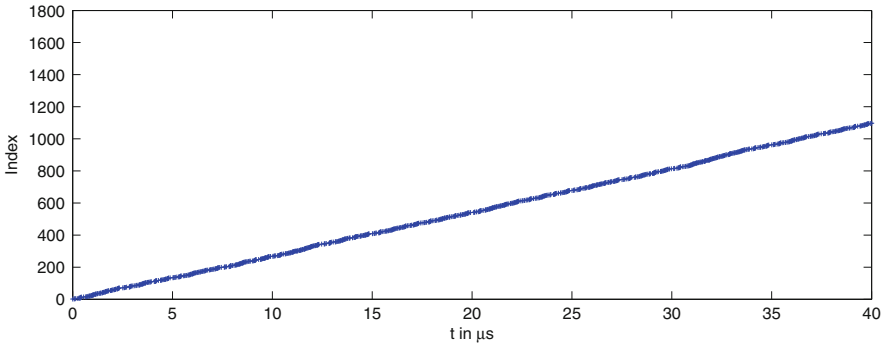


Fig. 3 Grid of the single grid method

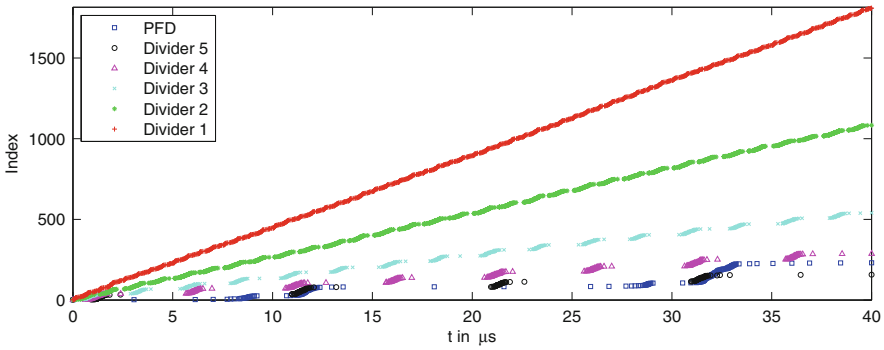


Fig. 4 Several signals in a frequency divider chain as part of a PLL

The grids used in our new multiple grid method can be seen in Fig. 4. Obviously, one gets much better adapted, smaller grids for the lower frequency signals. This leads to a reduction of the total number of equations from roughly 130,000 to 85,000. The number of nonzeros in the Jacobian for Newton's method is reduced from 5,000,000 to 2,500,000. Consequently the time for assembling resp. solving the linear system was reduced from 4 to 2 s respectively 8 to 4 s.

A further effect is that the larger the nonlinear system, the harder it is to solve by Newton's method, which results in more Newton iterations as well as shorter envelope time steps. Thus, an envelope simulation with a frequency modulated signal over 0.3 s worked well for the multiple grid method and was done in 37 min. A similar simulation with the single grid method needed almost 5 h.

6 Conclusion

An improvement of the spline wavelet based envelope method from [19] has been developed. It uses different spline grids for different parts of the circuit. This leads to a more efficient representation of the solution, which results in a significant reduction of computation time.

Acknowledgements This work has been partly supported by the ENIAC research project ARTEMOS under grant 829397 and the FWF under grant P22549.

References

1. Brachtendorf, H.G.: Simulation des eingeschwungenen Verhaltens elektronischer Schaltungen. Shaker, Aachen (1994)
2. Brachtendorf, H.G., Welsch, G., Laur, R., Bunse-Gerstner, A.: Numerical steady state analysis of electronic circuits driven by multi-tone signals. Electr. Eng. **79**(2), 103–112 (1996)

3. Ngoya, E., Larchevèque, R.: Envelope transient analysis: a new method for the transient and steady state analysis of microwave communication circuit and systems. In: Proceedings of the IEEE MTT-S International Microwave Symposium, pp. 1365–1368. San Francisco (1996)
4. Roychowdhury, J.: Efficient methods for simulating highly nonlinear multi-rate circuits. In: Proceedings of the IEEE Design Automation Conference, pp. 269–274 (1997)
5. Pulch, R., Günther, M.: A method of characteristics for solving multirate partial differential equations in radio frequency application. Appl. Numer. Math. **42**, 399–409 (2002)
6. Brachtendorf, H.G.: On the relation of certain classes of ordinary differential algebraic equations with partial differential algebraic equations. Technical Report 1131G0-971114-19TM, Bell-Laboratories (1997)
7. Brachtendorf, H.G.: Theorie und Analyse von autonomen und quasiperiodisch angeregten elektrischen Netzwerken. Eine algorithmisch orientierte Betrachtung. Universität Bremen (2001). Habilitationsschrift
8. Bittner, K., Brachtendorf, H.G.: Optimal frequency sweep method in multi-rate circuit simulation. COMPEL **33**(4), 1189–1197 (2014)
9. Ogrodzki, J.: Circuit simulation methods and algorithms. Electronic Engineering Systems. CRC, Boca Raton (1994)
10. Newton, A.: Techniques for the simulation of large-scale integrated circuits. IEEE Trans. Circuits Syst. **26**(9), 741–749 (1979)
11. Lelarasmee, E., Ruehli, A.E., Sangiovanni-Vincentelli, A.: The waveform relaxation method for time-domain analysis of large-scale integrated circuits. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **1**, 131–145 (1982)
12. Fang, W., Mokari, M., Smart, D.: Robust VLSI circuit simulation techniques based on overlapped waveform relaxation. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **14**(4), 510–518 (1995)
13. Schutt-Aine, J.: Latency insertion method (LIM) for the fast transient simulation of large networks. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **48**(1), 81–89 (2001)
14. Cox, P., Burch, R., Hocevar, D., Ping Yang, B., Epler, B.: Direct circuit simulation algorithms for parallel processing [VLSI]. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **10**(6), 714–725 (1991)
15. Sakallah, K., Director, S.: SAMSON2: an event driven VLSI circuit simulator. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **4**(4), 668–684 (1985)
16. Rabbat, N., Sangiovanni-Vincentelli, A., Hsieh, H.: A multilevel Newton algorithm with macromodeling and latency for the analysis of large-scale nonlinear circuits in the time domain. IEEE Trans. Circuits Syst. **26**(9), 733–741 (1979)
17. Rabbat, N., Hsieh, H.: A latent macromodular approach to large-scale sparse networks. IEEE Trans. Circuits Syst. **23**(12), 745–752 (1976)
18. Bittner, K., Brachtendorf, H.G.: Trigonometric splines for oscillator simulation. In: 22nd International Conference Radioelektronika, pp. 1–4 (2012)
19. Bittner, K., Brachtendorf, H.G.: Adaptive multi-rate wavelet method for circuit simulation. Radioengineering **23**(1), 300–307 (2014)
20. Bittner, K., Dautbegovic, E.: Wavelets algorithm for circuit simulation. In: Günther, M., Bartel, A., Brunk, M., Schöps, S., Striebel, M. (eds.) Progress in Industrial Mathematics at ECMI 2010. Mathematics in Industry, pp. 5–11. Springer, Berlin/Heidelberg (2012)
21. Bittner, K., Dautbegovic, E.: Adaptive wavelet-based method for simulation of electronic circuits. In: Michielsen, B., Poirier, J.R. (eds.) Scientific Computing in Electrical Engineering 2010. Mathematics in Industry, pp. 321–328. Springer, Berlin/Heidelberg (2012)
22. Bittner, K., Brachtendorf, H.G.: Fast algorithms for grid adaptation using non-uniform biorthogonal spline wavelets. SIAM J. Sci. Comput. **37**(2), B283–B304 (2015)

Turning Points of Nonlinear Circuits

Ignacio García de la Vega and Ricardo Riaza

Abstract Bifurcation theory plays a key role in the qualitative analysis of dynamical systems. In nonlinear circuit theory, bifurcations of equilibria describe qualitative changes in the local phase portrait near an operating point, and are important from both an analytical and a numerical point of view. This work is focused on quadratic turning points, which, in certain circumstances, yield saddle-node bifurcations. Algebraic conditions guaranteeing the existence of this kind of points are well-known in the context of explicit ordinary differential equations (ODEs). We transfer these conditions to semiexplicit differential-algebraic equations (DAEs), in order to impose them to branch-oriented models of nonlinear circuits. This way, we obtain a description of the conditions characterizing these turning points in terms of the underlying circuit digraph and the devices' characteristics.

1 Introduction

The context of the present work is the study of bifurcation phenomena in nonlinear circuits. We have focused on quadratic turning points, which are related to certain local bifurcations in dynamical systems, in particular to the saddle-node bifurcation. With terminological abuse, we will often use the expression "turning point" to mean a "quadratic turning point". We are interested in the analysis of turning points in the equations governing nonlinear circuits, which have the structure of a semiexplicit DAE. Therefore, our first efforts are directed to adequate the classical conditions characterizing turning points in ODEs to a semiexplicit index-one DAE context (Sect. 2). Afterwards, in Sect. 3, we will analyze these reformulated conditions in terms of the circuit topology and the devices' characteristics. Finally, Sect. 4 briefly compiles some concluding remarks.

I.G. de la Vega (✉) • R. Riaza
Depto. de Matemática Aplicada a las Tecnologías de la Información y las Comunicaciones, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n., 28040 Madrid, Spain
e-mail: ignacio.garciadelavega@gmail.com; ricardo.riaza@upm.es

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_3

Turning points in explicit ODEs Let us recall the algebraic conditions defining quadratic turning points in ODEs. Consider the ordinary differential equation

$$x' = f(x, \mu), \tag{1}$$

with $x \in \mathbb{R}^n$, and f sufficiently smooth and depending on a parameter $\mu \in \mathbb{R}$. Provided that $f(x^*, \mu^*) = 0$, (x^*, μ^*) is called a *quadratic turning point* of (1) if the conditions 1–3 below are satisfied [4].

1. $\mathrm{rk} f_x(x^*, \mu^*) = n - 1$;
2. $w^T f_\mu(x^*, \mu^*) \neq 0$;
3. $w^T f_{xx}(x^*, \mu^*)(v, v) \neq 0$.

Here v (resp. w) denotes a right (resp. left) eigenvector of the zero eigenvalue of the matrix of partial derivatives $f_x(x^*, \mu^*)$. Such turning points are important e.g. in numerical continuation theory [1]. If, additionally,

4. the algebraic multiplicity of the null eigenvalue of $f_x(x^*, \mu^*)$ is one; and
5. the remaining eigenvalues of $f_x(x^*, \mu^*)$ have non-zero real parts,

then (x^*, μ^*) is called a *saddle-node bifurcation point*, because the system undergoes a saddle-node bifurcation as μ crosses μ^* [5, 7, 10]. Near (x^*, μ^*) we will observe that when $\mu < \mu^*$ (resp. when $\mu > \mu^*$) there are no equilibria, whereas for $\mu > \mu^*$ (resp. $\mu < \mu^*$) there are two hyperbolic equilibrium points. These two equilibria differ in the sign of one real eigenvalue, being in particular a saddle and a node when $x \in \mathbb{R}^2$.

2 Turning Points in Semiexplicit DAEs

Our purpose is to characterize the existence of turning points and saddle-node bifurcations in electrical circuit models and, specifically, in branch-oriented models. These models have the structure of a semiexplicit DAE [3, 8], that is,

$$y' = h(y, z, \mu) \tag{2a}$$

$$0 = g(y, z, \mu), \tag{2b}$$

where $y \in \mathbb{R}^r$, $z \in \mathbb{R}^p$, $\mu \in \mathbb{R}$, and h and g are sufficiently smooth. We will group together y and z into a single variable $x = (y, z) \in \mathbb{R}^n$, with $n = r + p$. For later use let us also define the matrices

$$M = \begin{pmatrix} h_y & h_z \\ g_y & g_z \end{pmatrix}, \quad \tilde{M} = \begin{pmatrix} M \\ (\det M)_x \end{pmatrix}. \tag{3}$$

Specifically, we will work in a local index-one context [3, 8]; this means that the matrix of partial derivatives $g_z(y^*, z^*, \mu^*)$ is non-singular. By the implicit function theorem this implies that there is a local map $\psi(y, \mu)$ such that $0 = g(y, z, \mu)$ if and only if $z = \psi(y, \mu)$, with $\psi_y = -(g_z)^{-1} g_y$. This, together with (2a), enables one to express the local dynamics of the DAE (2) in terms of the reduced ODE

$$y' = \eta(y, \mu), \tag{4}$$

with $\eta(y, \mu) = h(y, \psi(y, \mu), \mu)$. In turn, this makes it possible to define an equilibrium (y^*, z^*, μ^*) of the semiexplicit index-one DAE (2) as a (quadratic) turning point (resp. a saddle-node bifurcation point) if the reduction (4) satisfies the conditions 1–3 (respectively 1–5) stated in Sect. 1.

Theorem 1 provides conditions for system (2) to have a turning point. Additional conditions for the existence of a saddle-node point will be formulated in terms of the reduction (4); this point of view will be exploited in Sect. 3.

Theorem 1 *Consider the semiexplicit DAE (2) and assume there exists a point (x^*, μ^*) such that $h(x^*, \mu^*) = 0$ and $g(x^*, \mu^*) = 0$, with $g_z(x^*, \mu^*)$ non-singular. Then (x^*, μ^*) is a quadratic turning point if the following conditions are satisfied:*

1. $\operatorname{rk} M(x^*, \mu^*) = n - 1$;
2. $\begin{pmatrix} h_\mu \\ g_\mu \end{pmatrix}(x^*, \mu^*) \notin \operatorname{im} M\,(x^*, \mu^*)$;
3. $\operatorname{rk} \tilde{M}(x^*, \mu^*) = n$.

Proof Write $x^* = (y^*, z^*)$ and note that (y^*, μ^*) is an equilibrium point of (4), because $\eta(y^*, \mu^*) = h(y^*, z^*, \mu^*) = 0$. We check below that conditions 1–3 in Sect. 1 hold for the reduction (4) at (y^*, μ^*).

1. If we compute η_y in terms of the maps h and g, we obtain

$$\eta_y = \begin{pmatrix} h_y & h_z \end{pmatrix} \begin{pmatrix} I \\ -(g_z)^{-1} g_y \end{pmatrix} = h_y - h_z (g_z)^{-1} g_y,$$

which is the Schur complement of g_z in M [6]. The corank of a matrix and the corank of its Schur reduction are equal, therefore $\operatorname{rk} M(x^*, \mu^*) = n - 1$ implies $\operatorname{rk} \eta_y(y^*, \mu^*) = r - 1$.

2. The second condition is $w^T \eta_\mu(y^*, \mu^*) \neq 0$, where w is an eigenvector associated to the zero eigenvalue of the matrix A^T with $A = \eta_y(y^*, \mu^*)$; note that $w^T A = 0 \Leftrightarrow w^T \perp \operatorname{im} A$. Therefore, $w^T \eta_\mu(y^*, \mu^*) \neq 0 \Leftrightarrow \eta_\mu(y^*, \mu^*) \notin \operatorname{im} A$, that is, $(h_\mu - h_z g_z^{-1} g_\mu)(x^*, \mu^*) \notin \operatorname{im} (h_y - h_z (g_z)^{-1} g_y)(x^*, \mu^*)$ which is equivalent to

$$\begin{pmatrix} h_\mu \\ g_\mu \end{pmatrix}(x^*, \mu^*) \notin \operatorname{im} \begin{pmatrix} h_y & h_z \\ g_y & g_z \end{pmatrix}(x^*, \mu^*).$$

3. Equation $w^T \eta_{yy}(y^*, \mu^*)(v, v) \neq 0$ can be recast as $\eta_{yy}(y^*, \mu^*)(v, v) \notin \operatorname{im} \eta_y(y^*, \mu^*)$. The fact that for a C^2 map $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfying $\operatorname{cork} f'(x^*) = 1$ we have $(\det f'(x))' v \neq 0 \Leftrightarrow f''(x)(v, v) \notin \operatorname{im} f'(x)$, where v is a non-null vector belonging to $\operatorname{Ker} f'(x)$, allows us to transform this condition into $(\det(\eta_y))_y(y^*, \mu^*)v \neq 0$, where $v \in \operatorname{Ker} \eta_y(y^*, \mu^*)$. Additionally, because η_y is the Schur complement of g_z in M, we have $\det \eta_y = \det\left(g_z^{-1}\right) \det M$, and then

$$(\det(\eta_y))_y = (\det\left(g_z^{-1}\right))_x \det M \begin{pmatrix} I \\ -(g_z)^{-1} g_y \end{pmatrix} + \det\left(g_z^{-1}\right) (\det M)_x \begin{pmatrix} I \\ -(g_z)^{-1} g_y \end{pmatrix}.$$

Condition 1 states that $\operatorname{rk} M(x^*, \mu^*) = n - 1$, thus $\det(M(x^*, \mu^*)) = 0$. Additionally, $\det\left(g_z^{-1}\right)(x^*, \mu^*) \neq 0$; therefore condition 3 is satisfied if and only if

$$\left((\det M)_y \;\; (\det M)_z\right) \begin{pmatrix} I \\ -(g_z)^{-1} g_y \end{pmatrix} (x^*, \mu^*) v \neq 0,$$

for some (hence any) non-vanishing vector v belonging to $\operatorname{Ker} \eta_y(y^*, \mu^*)$. Because of the identity $\operatorname{Ker} \eta_y(y^*, \mu^*) = \operatorname{Ker}(h_y - h_z(g_z)^{-1} g_y)(x^*, \mu^*)$, condition 3 is then equivalent to the requirement that the system

$$(h_y - h_z(g_z)^{-1} g_y)(x^*, \mu^*) v = 0 \tag{5a}$$

$$((\det M)_y - (\det M)_z g_z^{-1} g_y)(x^*, \mu^*) v = 0 \tag{5b}$$

only possesses the trivial solution. Equivalently, the matrix of coefficients of (5),

$$M_1 = \begin{pmatrix} h_y - h_z(g_z)^{-1} g_y \\ \det M_y - \det M_z g_z^{-1} g_y \end{pmatrix} (x^*, \mu^*),$$

must have maximum column rank. But M_1 is the Schur complement of g_z in the matrix $\tilde{M}(x^*, \mu^*)$ arising in the statement of condition 3 of Theorem 1; hence, the maximum column rank condition on M_1 is transferred to $\tilde{M}(x^*, \mu^*)$. This means that condition 3 in Sect. 1 holds for (4) at (y^*, μ^*) and the proof is complete.

3 Nonlinear Circuits Exhibiting Turning Points

In this section, we characterize the existence of turning points and saddle-node bifurcations for nonlinear circuits, under certain restrictions to be specified later. For this purpose, we use branch-oriented circuit models [8] defined by:

$$C(v_c) v_c' = i_c \tag{6a}$$

$$L(i_l) i_l' = v_l \tag{6b}$$

$$0 = B_c v_c + B_l v_l + B_g v_g + B_n v_n + B_j v_j + B_v V \tag{6c}$$

$$0 = Q_c i_c + Q_l i_l + Q_g \gamma_1(v_g) + Q_n \gamma_2(v_n) + Q_j \mu + Q_v i_v, \tag{6d}$$

where we denote the branch voltages by v, the currents by i, and use the subscripts c, l, g, n, j and v to denote capacitors, inductors, passive resistors, non-passive resistors, current sources and voltage sources, respectively. All devices may be nonlinear, often without explicit mention. We assume that there exists only one non-passive resistor and a unique DC current source, whose current $i_j = \mu$ is the parameter of the system. The reader may think of a tunnel diode as an example of a (locally) non-passive resistor. We also assume that there exists an equilibrium point that we will denote by $(x^*, \mu^*) = (v_c^*, i_l^*, i_c^*, v_l^*, v_g^*, v_n^*, v_j^*, i_v^*, \mu^*)$. The incremental capacitance and inductance matrices, C and L, are both non-singular at (x^*, μ^*) and, finally, V is the vector of voltages in the DC voltage sources.

System (6) has the semiexplicit DAE structure displayed in (2) with $y = (v_c, i_l)$ and $z = (i_c, v_l, v_g, v_n, v_j, i_v)$. Note that Eqs. (6a) and (6b) stand for the constitutive relations of capacitors and inductors, whereas Eqs. (6c) and (6d) are the expression of Kirchhoff laws. In (6d) we have eliminated the resistors currents using the constitutive relations γ_1 and γ_2. In the formulation of Kirchhoff laws we have made use of the so-called loop and cutset matrices B, Q, which are well-known in digraph theory and whose main properties are compiled in Lemma 1 [2, 9].

Lemma 1 *The loop and cutset matrices B, Q of a digraph verify the following.*

1. *B_K (resp. Q_K) has full column rank if and only if the branches specified by K do not contain any cutset (resp. loop).*
2. *The loop and cutset spaces are orthogonal to each other, that is, if columns of Q and B are arranged in the same order, then $QB^T = 0$.*
3. *Suppose the branches of a given digraph are split in four disjoint sets K_1, K_2, K_3 and K_4, and denote by B_i and Q_i the submatrices of the loop and cutset matrices defined by K_i; assume additionally that P is a positive definite matrix. Then*

$$\mathrm{Ker} \begin{pmatrix} B_1 & 0 & B_3 \\ 0 & Q_2 & Q_3 P \end{pmatrix} = \mathrm{Ker} B_1 \times \mathrm{Ker} Q_2 \times \{0\}.$$

These properties allow us to prove Theorem 2, which characterizes turning points and saddle-node bifurcations for the circuit model (6). By a K-loop (resp. K-cutset) we mean a loop (resp. cutset) defined only by elements of K; this way, for instance a JCN-cutset is a cutset defined only by current sources, capacitors and/or non-passive resistors. JLN-cutsets, VC-loops, etc. are defined analogously.

Theorem 2 *In the setting defined above, assume that $\gamma_2'(v_n^*) = 0$, $\gamma_2''(v_n^*) \neq 0$ at the equilibrium point (x^*, μ^*). This equilibrium is then a turning point of (6) if*

- *there is a unique JCN-cutset, which includes the current source, the non-passive resistor and at least one capacitor; and*
- *there are no JLN-cutsets, VC-loops or JVL-loops.*

If, additionally, L and C are symmetric positive definite and there are no VCL-loops, then the turning point yields a saddle-node bifurcation.

Proof The matrices $g_z(x^*,\mu^*)$ and $M(x^*,\mu^*)$ read for system (6) as:

$$g_z(x^*,\mu^*)=\begin{pmatrix} 0 & B_l & B_g & B_n & B_j & 0 \\ Q_c & 0 & Q_gG & 0 & 0 & Q_v \end{pmatrix},\ M(x^*,\mu^*)=\begin{pmatrix} 0 & 0 & C^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & L^{-1} & 0 & 0 & 0 & 0 \\ B_c & 0 & 0 & B_l & B_g & B_n & B_j & 0 \\ 0 & Q_l & Q_c & 0 & Q_gG & 0 & 0 & Q_v \end{pmatrix},$$

where $G=\gamma_1'(v_g^*)$ is the incremental conductance matrix of passive resistors, which is positive definite. In light of item 3 in Lemma 1, non-trivial entries in $\mathrm{Ker} g_z(x^*,\mu^*)$ must come either from $\mathrm{Ker}\begin{pmatrix} B_l & B_n & B_j \end{pmatrix}$ or from $\mathrm{Ker}\begin{pmatrix} Q_c & Q_v \end{pmatrix}$. Since there are neither JLN-cutsets nor VC-loops, we conclude that $g_z(x^*,\mu^*)$ is non-singular.

1. The non-singularity of C, L allows us to study the rank of the matrix $M(x^*,\mu^*)$ in terms of

$$\begin{pmatrix} B_c & 0 & B_g & B_n & B_j & 0 \\ 0 & Q_l & Q_gG & 0 & 0 & Q_v \end{pmatrix}.$$

By applying item 3 of Lemma 1, non-zero entries of $\mathrm{Ker} M(x^*,\mu^*)$ must come either from $\mathrm{Ker}\begin{pmatrix} B_c & B_n & B_j \end{pmatrix}$ or from $\mathrm{Ker}\begin{pmatrix} Q_l & Q_v \end{pmatrix}$. Since there is a unique JCN-cutset and no JVL-loops, we have $\mathrm{null}\begin{pmatrix} B_c & B_n & B_j \end{pmatrix}=1$, where null stands for the nullity, that is, the dimension of the kernel, $\mathrm{null}\begin{pmatrix} Q_l & Q_v \end{pmatrix}=0$ and therefore $\mathrm{null}\, M(x^*,\mu^*)=1$, that is, $\mathrm{rk}\, M(x^*,\mu^*)=n-1$, which is condition 1 in Theorem 1.

2. The 2nd condition in Theorem 1 may be restated as $\mathrm{null}\, M(x^*,\mu^*)=\mathrm{null}\,\hat{M}(x^*,\mu^*)$, with

$$\hat{M}(x^*,\mu^*)=\begin{pmatrix} 0 & 0 & C^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & L^{-1} & 0 & 0 & 0 & 0 & 0 \\ B_c & 0 & 0 & B_l & B_g & B_n & B_j & 0 & 0 \\ 0 & Q_l & Q_c & 0 & Q_gG & 0 & 0 & Q_v & Q_j \end{pmatrix}.$$

Proceeding as above, we observe that non-trivial entries in $\mathrm{Ker}\hat{M}(x^*,\mu^*)$ must be due to those in $\mathrm{Ker}\begin{pmatrix} B_c & B_n & B_j \end{pmatrix}$ or in $\mathrm{Ker}\begin{pmatrix} Q_l & Q_v & Q_j \end{pmatrix}$. The absence of JVL-loops implies $\mathrm{null}\,\hat{M}(x^*,\mu^*)=\mathrm{null}\begin{pmatrix} B_c & B_n & B_j \end{pmatrix}$ and therefore $\mathrm{null}\, M=\mathrm{null}\,\hat{M}$.

3. The third condition in Theorem 1 says that the matrix $\tilde{M}$ (cf. (3)) has full column rank or, equivalently, $\mathrm{rk}\,\tilde{M}=n$. Provided that $\mathrm{null}\, M=1$, requiring $\tilde{M}$ to have full column rank is equivalent to $(\det M)_x v\neq 0$, where v is any vector that spans $\mathrm{Ker} M$. For any point $\hat{x}=(v_c,i_l,i_c,v_l,v_g,v_n^*,v_j,i_v)$, $M(\hat{x},\mu)$ is a singular matrix

because $\gamma_2'(v_n^*) = 0$. Thus, $(\det M)_x = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & a & 0 & 0 \end{pmatrix}$ and $a \neq 0$ because $\gamma_2''(v_n^*) \neq 0$.

The absence of VL-loops and the existence of a JCN-cutset imply that vectors belonging to Ker$M$ have the form of v where $v^T = (v_1, 0, 0, 0, 0, v_6, v_7, 0)$. Additionally, the fact that there are no JC-cutsets implies $v_6 \neq 0$; it follows that the multiplication of $(\det M)_x$ by vectors of Ker$M$ does not vanish.

4. To complete the proof it remains to show that the absence of VCL-loops leads to a saddle-node bifurcation. To do this we make use of conditions 4 and 5 in Sect. 1.

In order to prove that the zero eigenvalue is simple, we will show that the intersection of the kernel and the image of $\eta_y = (h_y - h_z(g_z)^{-1} g_y)$ at (x^*, μ^*) only contains the null vector. First, a vector u belongs to $\operatorname{im} \eta_y$ if and only if $\hat{u}$ belongs to $\operatorname{im} M$, with $\hat{u}^T = \begin{pmatrix} u^T & 0 \end{pmatrix}$, that is, if and only if there exists a vector v satisfying

$$u_1 = C^{-1} v_3 \tag{7a}$$

$$u_2 = L^{-1} v_4 \tag{7b}$$

$$0 = B_c v_1 + B_l v_4 + B_g v_5 + B_n v_6 + B_j v_7 \tag{7c}$$

$$0 = Q_l v_2 + Q_c v_3 + Q_g G v_5 + Q_v v_8. \tag{7d}$$

On the other hand, a vector u belongs to Kerη_y if and only if

$$\begin{pmatrix} C^{-1} & 0 & 0\,0\,0\,0 \\ 0 & L^{-1} & 0\,0\,0\,0 \end{pmatrix} (g_z^{-1}) \begin{pmatrix} B_c & 0 \\ 0 & Q_l \end{pmatrix} u = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In order to satisfy this relation there must be a vector y such that

$$(g_z^{-1}) \begin{pmatrix} B_c u_1 \\ Q_l u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \Leftrightarrow \begin{pmatrix} B_c u_1 \\ Q_l u_2 \end{pmatrix} = g_z \begin{pmatrix} 0 \\ 0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix},$$

that is,

$$B_c u_1 = B_g y_1 + B_n y_2 + B_j y_3 \tag{8a}$$

$$Q_l u_2 = Q_g G y_1 + Q_v y_4. \tag{8b}$$

Using the orthogonality of the cutset and loop spaces, namely, the fact that Ker$B$ and Ker$Q$ are orthogonal to one another (cf. [2]), it is not difficult to obtain from (8) the relation $y_1^T G y_1 = 0$; y_1 must then vanish because G is positive

definite. Making use of (7a) and (7b), Eqs. (8a) and (8b) then read as

$$0 = B_c C^{-1} v_3 - B_n y_2 - B_j y_3 \tag{9a}$$

$$0 = Q_l L^{-1} v_4 - Q_v y_4. \tag{9b}$$

Therefore if $u \in \mathrm{Ker}\eta_y \cap \mathrm{im}\,\eta_y$, then (7c), (7d) and (9) must hold. Applying the aforementioned orthogonality property to (7c) and (9b) we obtain that $v_4^T L^{-1} v_4 = 0$ and from Eqs. (7d) and (9a), $v_3^T C^{-1} v_3 = 0$. Altogether this yields $u = 0$.

5. It remains to be proved that if there are no VCL-loops then $\eta_y(x^*, \mu^*)$ has no purely imaginary eigenvalues. A complex number λ is an eigenvalue of $\eta_y(x^*, \mu^*)$ if and only if

$$\lambda \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} h_y & h_z \\ g_y & g_z \end{pmatrix} (x^*, \mu^*)$$

is singular or, equivalently, there exists non-trivial solutions to

$$0 = \lambda^{-1} B_c C^{-1} i_c + \lambda B_l L i_l + B_g v_g + B_n v_n + B_i v_i \tag{10a}$$

$$0 = Q_c i_c + Q_l i_l + Q_g G v_g + Q_v i_v. \tag{10b}$$

The orthogonality of the cutset and cycle spaces implies that if $Qp = 0$ and $Bq = 0$ then $p^T q = 0$. Applying this result to the conjugate of (10b) in conjunction with (10a), we obtain

$$0 = \lambda^{-1} i_c^* C^{-1} i_c + \lambda i_l^* L i_l + v_g^* G v_g, \tag{11}$$

where * stands for the Hermitian (conjugate transpose). If we take the sum of (11) and its Hermitian, we obtain:

$$0 = 2\mathrm{Re}(\lambda^{-1}) i_c^* C^{-1} i_c + 2\mathrm{Re}(\lambda) i_l^* L i_l + v_g^* (G + G^T) v_g. \tag{12}$$

For purely imaginary eigenvalues, $\mathrm{Re}(\lambda^{-1}) = \mathrm{Re}(\lambda) = 0$ and therefore we must have $v_g = 0$ for (12) to hold. System (10) can be then simplified to:

$$0 = \lambda^{-1} B_c C^{-1} i_c + \lambda B_l L i_l + B_n v_n + B_i v_i \tag{13a}$$

$$0 = Q_c i_c + Q_l i_l + Q_v i_v. \tag{13b}$$

Since there are no VCL-loops, $\begin{pmatrix} Q_c & Q_l & Q_v \end{pmatrix}$ has full column rank and, consequently, $i_c = i_l = i_v = 0$ must hold to satisfy (13b). The absence of JLN-cutsets then yields $v_n = v_i = 0$ from (13a). This means that (10) only has the trivial solution, and this rules out purely imaginary eigenvalues. The proof of Theorem 2 is then complete.

4 Concluding Remarks

We have performed a circuit-theoretic analysis of the existence of turning points and saddle-node bifurcations in nonlinear circuits. The analysis of these phenomena in broader contexts, including e.g. other non-passive devices, higher-index configurations or parameters with other roles, as well as the study of other related bifurcations in similar terms, are in the scope of future research.

Acknowledgements Research supported by Project MTM2010-15102 of Ministerio de Ciencia e Innovación, Spain.

References

1. Allgower, E.L., Georg, K.: Introduction to Numerical Continuation Methods. SIAM, Philadelphia (2003)
2. Bollobás, B.: Modern Graph Theory. Springer, New York (1998)
3. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential Algebraic Equations. SIAM, Philadelphia (1996)
4. Govaerts, W.J.F.: Numerical Methods for Bifurcations of Dynamical Equilibria. SIAM, Philadelphia (2000)
5. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, New York (1983)
6. Horn, R.A., Johnson, C.R., Matrix Analysis. Cambridge University Press, New York (1985)
7. Perko, L.: Differential Equations and Dynamical Systems. Springer, New York (1991)
8. Riaza, R.: Differential-Algebraic Systems. World Scientific, Singapore (2008)
9. Riaza, R.: Manifolds of equilibria and bifurcations without parameters in memristive circuits. SIAM J. Appl. Math. **72** (2012) 877–896
10. Sotomayor, J.: Generic bifurcations of dynamical systems. In: Peixoto, M.M. (ed.) Dynamical Systems, pp. 561–582. Academic, New York (1973)

Mixed Domain Macromodels for RF MEMS Capacitive Switches

Gabriela Ciuprina, Aurel-Sorin Lup, Bogdan Diţă, Daniel Ioan, Ştefan Sorohan, Dragoş Isvoranu, and Sebastian Kula

Abstract A method to extract macromodels for RF MEMS switches is proposed. The macromodels include both the coupled structural-electric behavior of the switch as well as its RF behavior. The device with distributed parameters is subject to several analyses from which the parameters of the macromodel are extracted, by model reduction.

From the coupled structural-electrostatic analysis the parametric capacitance and the effective stiffness coefficients of the switch are extracted. From the RF characteristics in the up stable state, the transmission line parameters are extracted. Finally, all parameters are combined in a Spice circuit model, which is controlled by the MEMS actuation voltage and is excited with the RF signal.

The procedure is applied to a capacitive switch. Relative modeling errors with respect to the non-reduced model, considered as reference, of less than 3 % for the RF characteristics and less than 1 % for the mechanical characteristics are obtained.

1 Introduction

RF MEMS switches are devices containing electrostatic actuated movable parts with two stable states (up and down), used to allow or block the propagation of RF signals in various applications. They are based on micromachining technologies, being more suitable than solid electronic switching devices [1]. A typical capacitive RF switch contains an elastic bridge over a coplanar waveguide line (Fig. 1). The capacitance between the grounded bridge and the signal line, isolated with a dielectric layer is strongly dependent on the bridge position.

The design of this device focuses not only on the RF performances (S parameters at the RF ports) in its stable states, but also on other relevant quantities

G. Ciuprina (✉) • A.-S. Lup • B. Diţă • D. Ioan • Ş. Sorohan • D. Isvoranu
Politehnica University of Bucharest, Spl. Independenţei 313, 060042 Bucharest, Romania
e-mail: gabriela.ciuprina@upb.ro; sorin@lmn.pub.ro; bogdan@lmn.pub.ro; daniel.ioan@upb.ro; stefan.sorohan@upb.ro; dragos.isvoranu@upb.ro

S. Kula
Kazimierz Wielki University, Bydgoszczy, Poland
e-mail: skula@ukw.edu.pl

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_4

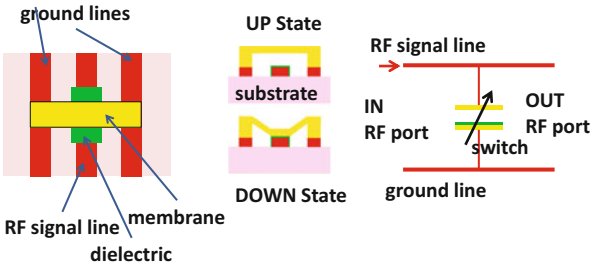


Fig. 1 Typical capacitive RF switch of bridge type. *Left*—the movable part (membrane, or bridge) is placed transversely with respect to a coplanar wave guide RF signal line. *Middle*—the switch has two stable states: up and down. *Right*—the switch is used in an RF circuit, being able to allow the signal to pass (if it is in the up state), or block it (in down state)

(pull-in/out voltages at the actuation terminals, commutation time between the stable states) related to its switching from one stable state to the other. The investigation of the latter aspects needs multiphysics simulations since several physical effects (mechanical motion, air damping, electrostatic actuation) come together. Even since the early development of these devices, the computational challenges identified are the multiphysics modeling, required for the estimation of the switching properties, and the nonlinear macromodeling or the nonlinear order reduction, which is very important for the designers who need dynamical device level models. The effective macromodels should be accurate enough and have few degrees of freedom, and they have to be correlated to design parameters such as dimensions and material properties, with the aim of being embedded in system-level models [2]. The multiphysics modeling is still a difficult challenge [3]. A common approach for design is to use separate macromodels for the physical domains involved, depending on the investigated properties. The RF macromodels, consist of short sections of transmission lines (TLs) and R, L, C elements, and they are used to model the S-parameters of the switch in its stable states. The values of the capacitance are different for the down and up states. They are computed with simple formulas based on an uniform electrostatic field assumption as in [4], whereas R and L are computed from down-state simulations with an EM field solver and fitting of the obtained S parameters.

Circuit macromodels are also proposed for the multiphysics domain, as in [5], where large signal dynamic circuit simulation models for MEMS devices using controlled current sources are proposed and implemented in APLAC. The importance of device models at global level is that they can be combined and integrated into existing design environments [6]. Aspects related to the mixed-domain electromechanical and electromagnetic simulation of RF-MEMS devices and network are reported in [7] . The author develops and use lumped component models for elementary components such as the flexible beams and the rigid plates. The elements are implemented in the VerilogA programming language, within the Cadence IC development environment, the simulations being completed in Spectre. This strategy is discussed also in [8] where macromodels are derived using a hierarchical modeling approach that use the generalized Kirchhoff network

theory. Combined techniques that derive both lumped and distributed components are used to obtain a fully coupled model described in a hardware description language. A MEMS component model library is offered by this team at http://rfmems.sourceforge.net/.

The goal of this paper is to obtain a combined macromodel, that includes both the multiphysics behavior and the RF behavior of the switch. For this, the device with distributed parameters is subject to several analyses from which the parameters of the macromodel are extracted, by model reduction. Finally, all parameters are combined in a Spice circuit model. The test used is the capacitive bridge-type switch proposed by Qian (Fig. 1) and its detailed description can be found in [9].

2 Multiphysics Macromodel

In order to change the stable state of the switch (e.g., from up to down), an electric voltage has to be applied between the central line and the membrane. The electric force that appears moves the mobile part until the mechanical contact is achieved; when the voltage is zeroed, the system moves back to the initial position due to the elastic forces in the membrane. During the movement, there is also a damping force due to the relative moment of the mobile plate with respect to the gas that surrounds it. It is obvious that the movement is non-uniform: the velocity is not constant, the acceleration is non-zero, so when writing equilibrium equations in a reference system attached to the mobile plate, an inertial force has to be considered. The most simple reduced order model appropriate for this coupled structural-electrostatic-fluid formulation corresponds to the equation of motion of a mobile plate of a parallel plate capacitor, suspended by a spring, when an actuation voltage is applied between its plates [1]:

$$m\frac{\mathrm{d}^2 z}{\mathrm{d}\,t^2} + b\frac{\mathrm{d}\,z}{\mathrm{d}\,t} + kz + k_s z^3 = F_{\mathrm{ES}}(u, z), \tag{1}$$

where m is the effective mass, b is the effective damping coefficient, k is the linear elasticity constant, k_s is the nonlinear elasticity (spring) constant, F_{ES} is the electrostatic force, which depends on the applied voltage u and the displacement of the membrane z. If the applied actuation voltage is not high enough, the electrostatic force might be not high enough to ensure the contact, but only to change the gap between the armatures. If the actuation voltage is higher than a certain value called *pull-in voltage* V_{pi} then the mobile part collapses on the fixed part. The pull-in voltage is an important characteristic of a switch and therefore, it has to be caught by a multiphysics macromodel. When solving a set of static multiphysics coupled simulations, corresponding to increasing values of the applied actuation voltage, an instability occurs when the pull-in voltage is reached. Figure 2 shows the computational domain of a 2D model for the Qian switch [9] and the Dirichlet boundary conditions used by the multiphysics formulation.

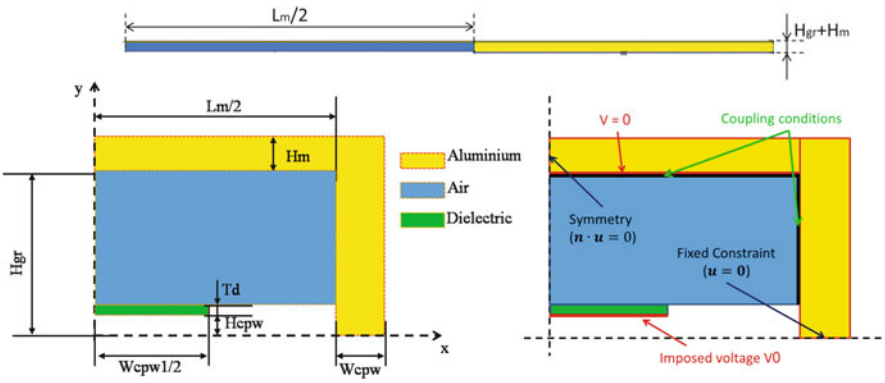


Fig. 2 2D Multiphysics domain: *up*—drawing at scale; *down left*—drawing not at scale, showing the geometric parameters: beam length $L_m = 280\,\mu\text{m}$; beam height $H_m = 0.4\,\mu\text{m}$; dielectric thickness $T_d = 0.1\,\mu\text{m}$; height of the RF signal line $H_{\text{cpw}} = 0.4\,\mu\text{m}$; height of the RF signal line $H_{\text{gr}} = 4\,\mu\text{m}$; $W_{\text{cpw}} = W_{\text{cpw1}} = 120\,\mu\text{m}$. For some postprocessing, the beam width $W_m = 280\,\mu\text{m}$ is needed; *down right*—computational multiphysics domain and boundary/interface conditions

In order to extract the lumped effective parameters k and k_s, it is enough that a set of static coupled (structural-electrostatic) finite element analysis simulations for several applied voltages $u = V0$, be carried out for the model described above. Equation (1) written for the static case suggests the following *extraction algorithm for the effective elastic coefficients*:

1. Do coupled static numerical simulations (e.g. FEM) for increasing values of the actuation voltage u. Record position $z(u)$ and electrostatic energy $W_{\text{ES}}(u)$;
2. Compute the dependence of the switch capacitance $C(z) = 2W_{\text{ES}}/u^2$, where $u = u(z)$, on the membrane displacement. Approximate the dependence $1/C(z)$ with a first order least square approximation $c_1 z + c_2$. The result of this step is shown in Fig. 3—left.
3. Compute the dependence of the electrostatic force $F_{\text{ES}}(z)$ on the displacement by using the generalized force theorem $F_{\text{ES}}(z) = (u^2/2)\mathrm{d}\,C(z)/\mathrm{d}z$. Since the simulations at step 1 were static, this electrostatic force is equal to the elastic force that acts on the membrane.
4. Do a cubic least square approximation of the dependence found at step 3 in order to find k and k_s. A less accurate model can be obtained if the least square approximation is of order 1, meaning that k_s is neglected. The result of this step is shown in Fig. 3—right.

The SPICE circuit that synthesizes Eq. (1) in which the damping term is not considered is shown in Fig. 4. The actuation voltage is modeled by the independent voltage source V1. The behavioral current source B1 models the electrostatic force. The behavioral current source B2 models the elastic force. The current flowing through the mass capacitor is the inertial force. All the important mechanical and electric characteristics—displacement $z(t)$, velocity $v(t)$, capacitance $C(z)$ and its

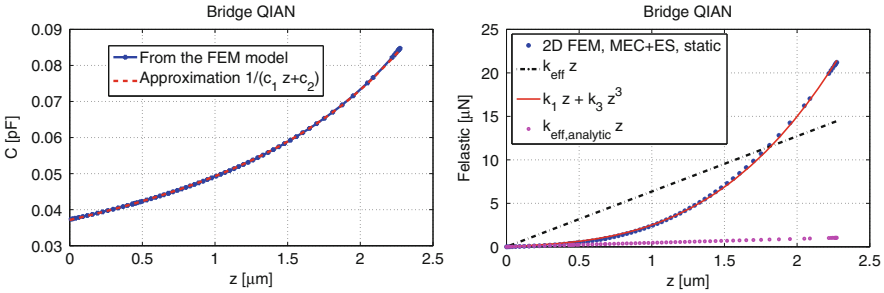


Fig. 3 Extraction of effective elastic coefficients from the multiphysics simulation: *left*—rational approximation of the capacitance; *right*—various possible approximations: linear or cubic least square; analytical evaluation of the elasticity coefficient is valid only for very small displacements

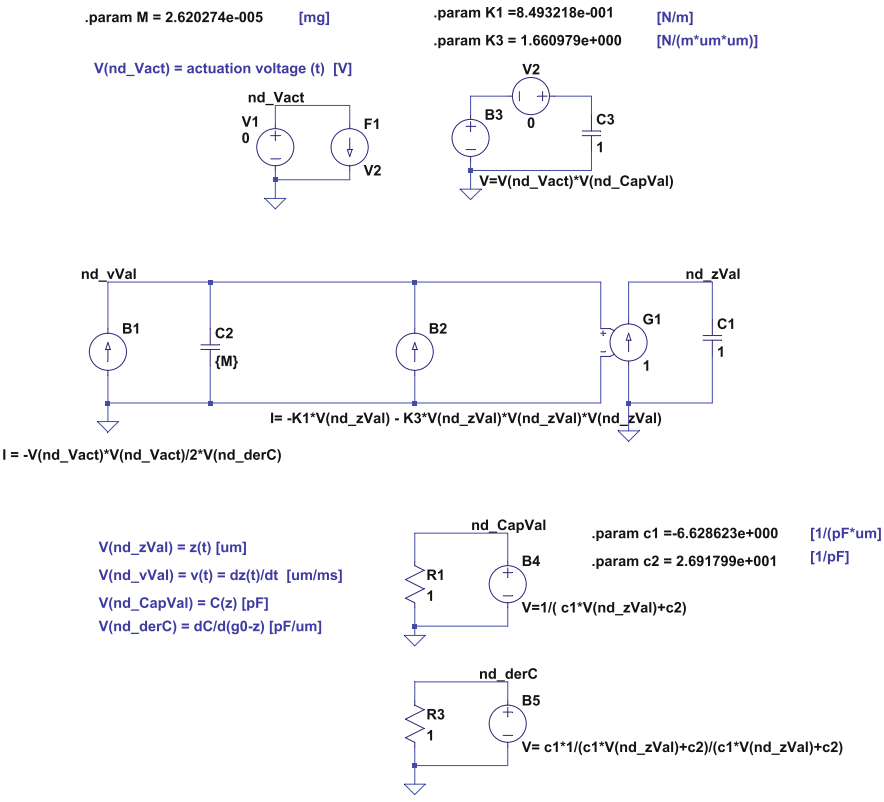


Fig. 4 Equivalent SPICE multiphysics macromodel. The "currents" flowing through this model are forces. The displacement in µm is the voltage at node `nd_sVal` and is used by the source B2 to provide the elastic force. The velocity in µm/ms is the voltage at node `nd_vVal`. The capacitance of the switch is the voltage at node `nd_CapVal`. The derivative of the capacitance with respect to the gap is the voltage at node `nd_derC` and is used by B1 to provide the electrostatic force

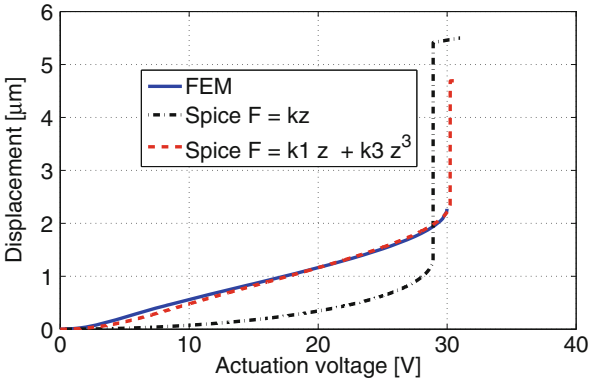


Fig. 5 Static simulations: FEM vs. SPICE equivalent macromodel

derivative with respect to the displacement $\mathrm{d}C/\mathrm{d}z$ are voltages in this schematic. Scaled values have been used.

The set of static simulations of this circuit are shown in Fig. 5 and reveal a relative error of the pull-in voltage with respect to its value from the FEM multiphysics model, of 3.55 % if a linear approximation of the elastic force is used, and a relative error of 0.82 % if a cubic approximation of the elastic force is used. The cubic approximation is not only very accurate for the pull in voltage, but also for all the dependence $z(u)$.

3 Mixed RF-Multiphysics Macromodel

To allow the coupling with the rest of the RF circuit, the macromodel of the switch has to include both a model for the RF signal lines and a model for the switch itself. The signal lines are best described by transmission lines (TL) models, whereas for the switching part lumped components are used (Fig. 6). Thus, the resulting RF macromodel includes both distributed and lumped parameters. The transmission lines placed on both sides are considered identical, of length l, complex impedance $\underline{Z}_c$ and complex propagation constant $\underline{\gamma}$.

In order to extract the line parameters of the TLs, an EM full-wave simulation of the switch in the up position has been done, as described in [10]. Electromagnetic circuit element (EMCE) boundary conditions were applied to surfaces on the boundary of the domain that correspond to the RF terminals, and the mathematical model thus obtained was numerically discretized with the finite integration technique. From the frequency simulation of this numerical model the impedance transfer matrix $\mathbf{Z}$ is obtained. On the other hand, the analysis of the schematic in Fig. 6 in which TLs relationships are used leads to the following analytical expressions for

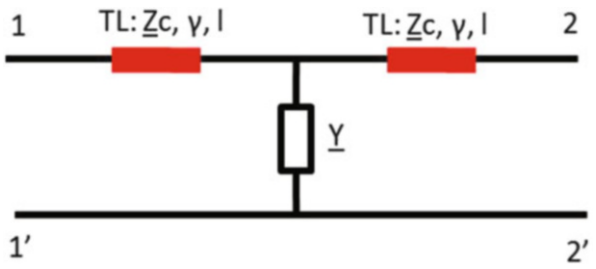


Fig. 6 Typical RF macromodel. The switch model is represented by an admittance $\underline{Y}$ is synthesized by using lumped R, L, C series connected components

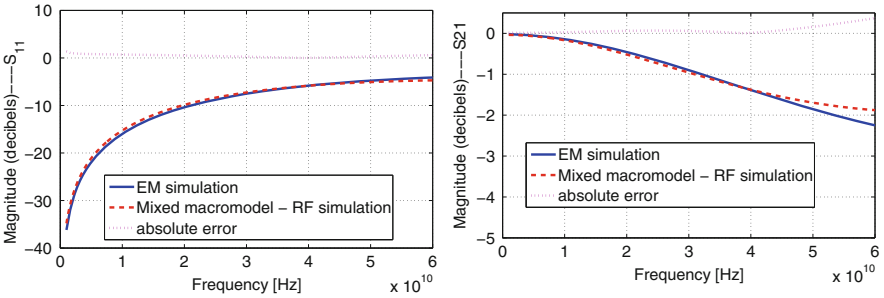


Fig. 7 EM simulation vs. mixed macromodel RF simulation: *left*—return loss (S_{11} signal pass), *right*—insertion loss (S_{21}, signal pass)

the transfer impedance components:

$$\underline{Z}_{12} = \underline{Z}_{21} = \frac{1}{\Delta}, \qquad \text{where} \qquad \Delta = 2\frac{\cosh(\underline{\gamma} l)\sinh(\underline{\gamma} l)}{\underline{Z}_c} + \underline{Y}\cosh^2(\underline{\gamma} l), \quad (2)$$

$$\underline{Z}_{11} = \underline{Z}_{22} = \frac{1}{\Delta}\left(\sinh^2(\underline{\gamma} l) + \cosh^2(\underline{\gamma} l) + \underline{Z}_c\underline{Y}\sinh(\underline{\gamma} l)\cosh(\underline{\gamma} l)\right). \quad (3)$$

From the multiphysics simulation discussed in the previous section, the dependence on z of the switch capacitance C was extracted. Assuming for the moment that we neglect R and L, the admittance needed in the formulas above is $\underline{Y} = j\omega C$ and it follows that the line parameters can be deduced quite straightforward from the formulas above. Values for the line parameters are obtained for every frequency, and an average value was computed for the frequency range of interest (Fig. 7).

The mixed macromodel is obtained by replacing the switch capacitance in the RF schematic by a model that connects it with the multiphysics macromodel, as in Fig. 8. A fixed capacitance has been added in parallel with the parametric one. It corresponds to the electric field lines that close through the substrate, and it has been computed by a separate electrostatic problem for the substrate. The validation of the model built so far is done by comparing the RF results (S parameters,

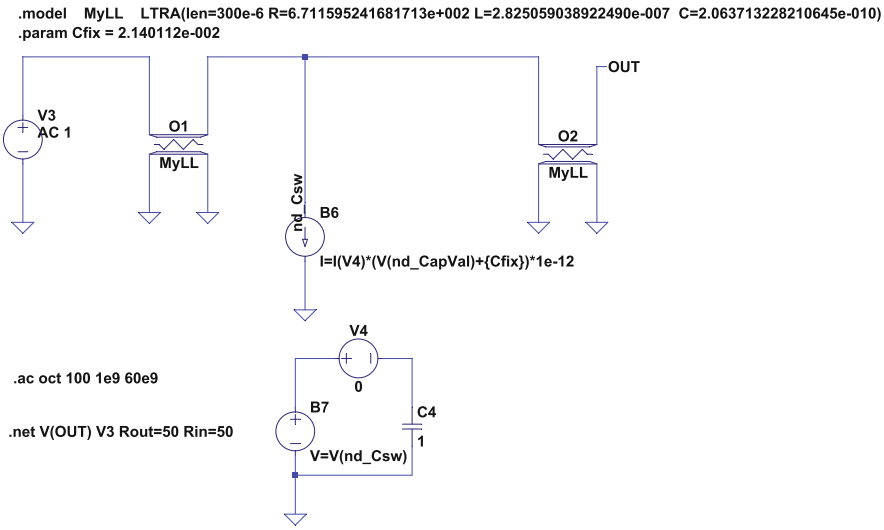


Fig. 8 Mixed macromodel: the RF part, the switch model is a current source controlled by the capacitance value that is taken from the multiphysics part (voltage at node `n_CapVal` in the multiphysics part of the schematic)

where $\mathbf{S} = (\mathbf{Z} - Z_0\mathbf{I})(\mathbf{Z} + Z_0\mathbf{I})^{-1}$, Z_0 being a reference impedance, according to http://eceweb1.rutgers.edu/~orfanidi/ewa/ch14.pdf) of the mixed schematic with the results from the EM simulation (Fig. 7). A relative error of 2.5 % in Frobenius norm is obtained.

4 Conclusions

A mixed macromodel of a RF-MEMS switch, with few degrees of freedom, was extracted from several analyses of the device with distributed parameters. All parameters are combined in a single Spice circuit model, which is controlled by the MEMS actuation voltage and is excited with the RF signal. A relative error less than 3 % in the S parameters and less than 1 % in the pull-in voltage is obtained, which is very satisfactory given the low order imposed for the reduced model. Our future studies will continue, the next step being to improve the multiphysics part, by including damping and contact phenomena, in order that the macromodel be able to carry out RF simulations up to the down position as well as transient simulations needed for the extraction of switching time and pull-out voltage.

Acknowledgements The financial support of the Romanian Government program PN-II-PT-PCCA-2011-3, no. 5/2012 and of the Sectoral Operational Programme HRD 2007–2013 of the Ministry of European Funds through POSDRU/159/1.5/S/132395 is acknowledged.

References

1. Rebeiz, G.M.: RF MEMS: Theory, Design, and Technology. Wiley, New York (2003)
2. Senturia, S.D., Aluru, N., White, J.: Simulating the behavior of MEMS devices: computational methods and needs. IEEE Comput. Sci. Eng. **16**(10), 30–43 (1997)
3. Hannot, S.: Modeling strategies for electro–mechanical microsystems with uncertainty quantification. Ph.D. Thesis, Delft University of Technology (2010)
4. Muldavin, J., Rebeiz, G.: High-isolation CPW MEMS shunt switches—part 1: modeling. IEEE Trans. Microw. Theory Tech. **48**(6), 1045–1052 (2000)
5. Veijola, T.: Nonlinear circuit simulation of MEMS components: controlled current source approach. In: ECCTD'01 - European Conference on Circuit Theory and Design, August 28–31, 2001, Espoo, Finland, vol. III, pp. 377–380 (2001)
6. Shafique, M., Virk, K., Menon, A., Madsen, J.: System-level modeling and simulation of MEMS-based sensors. In: 9th International Multitopic Conference, IEEE INMIC 2005, pp. 1–6 (2005)
7. Iannacci, J.: Mixed-Domain Fast Simulation of RF and Microwave MEMS-based Complex Networks within Standard IC Development Frameworks (2001). www.intechopen.com
8. Niessner, M., Schrag, G., Wachutka, G., Iannacci, J.: Modeling and fast simulation of RF-MEMS switches within standard IC design frameworks. In: International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2010, pp. 317–320 (2010)
9. Qian, J.Y., Li, G.P., De Flaviis, F.: A parametric model of low-loss RF MEMS capacitive switches. Asia-Pacific Microwave Conference, APMC 2001, Taipei, Taiwan (2001)
10. Ioan, D., Ciuprina, G.: Reduced order models of on-chip passive components and interconnects, workbench and test structures. In: Schilders, W.H.A., van der Vorst, H.A., Rommes, J. (eds.) Model Order Reduction: Theory, Research Aspects and Applications, vol. 13, pp. 447–467. Springer, Heidelberg (2008)

Part II
Computational Electromagnetics

Maxwell's equations of Electromagnetism were first published in 1865—now in 2015 it is exactly 150 years later as I write this text. These coupled partial differential equations provide a complete mathematical model to describe any macroscopic electromagnetic phenomenon—ranging from electric, magnetic or electromagnetic fields that are almost completely undetectable to human senses. (Well, except for a small range in the electromagnetic spectrum related to visible light and infrared heat radiation and for some second level effects to the human body exposed to strong electric and magnetic fields. And then you might also become seriously ill or even die when having been exposed to ionizing electromagnetic fields such as x-ray or gamma radiation for too long…)

The ability to describe electromagnetic fields by mathematically solving these Maxwell equations (or derived variants thereof) has had a strong impact on our technological society within the last 150 years. Essentially they are being instrumental to the way how we communicate today and power our homes or the majority of gadgets in our life that make use of electromagnetic energy.

More than half a decade ago computer-based methods were added to the toolbox of mathematical solution methods available for the electromagnetic field theory. It is worth to note at this point, that the famous Finite Difference Time Domain Method (FDTD method), by now a commonly used standard method for the solution of the Maxwell equations for transient electromagnetic wave phenomena published by K. Yee in 1966, will become half a century old by next year. By now, the computer aided solution of Maxwell's equations—summarized as *Computational Electromagnetics (CEM)*—is a long established branch of Scientific Computing research.

In the following chapter of this book we find six paper contributions from various authors presenting their results to the numerical calculation of electric, magnetic and electromagnetic fields. These papers each report on progress achieved either in the development of novel field formulations, for different discretization schemes or they present advances in sophisticated mathematical solution techniques and in high performance computing.

T. Banova et al. describe in the paper “Systematic determination of eigenfields in Frequency domain” a large scale parallel eigenvalue solver for a billiard resonator problem.

R. Casagrande et al. prove in their paper “DG Treatment of Non-Conforming Interfaces in 3D Curl-curl equations” that the use of these interfaces combined with interior penalty methods will result in the loss of one order of convergence that is critical when using lowest order edge element formulations.

M. Jochum and crew reformulate the time-harmonic Maxwell’s equations in terms of potentials in “ A symmetric and low-frequency stable potential formulation for the finite-element simulation of electromagnetic fields” which allows to calculate electro-and magneto-quasistatic problems combined in one formulation.

C. Jerez-Hankes et al. provide in their paper “Local Multiple Traces Formulation for High-Frequency Scattering Problems by Spectral Elements” a novel ready-to-precondition boundary integral formulation to solve 2D Helmholtz scattering problems including objects with high-contrast ratio.

C. Richter et al. develop an computationally efficient solver for large scale 3D electrostatic FEM problems in electric insulator design by using multiple GPUs in “Multi-GPU Acceleration of Algebraic Multigrid Preconditioners”.

D. Zheng et al. compare three different Green’s function methods for the fast solution of electric space charge studies in accelerator physics in “On several Green’s function methods for fast Poisson solver in free space”

This short collection of papers showcases Computational Electromagnetics to be a very active, interdisciplinary research field combining novel methods and advanced techniques in applied mathematics and computer science with sophisticated models in electromagnetic field theory—aiming at improvements in the computer aided simulation of complex application problems in physics and in electrical engineering.

Systematic Determination of Eigenfields in Frequency Domain

Todorka Banova, Wolfgang Ackermann, and Thomas Weiland

Abstract This paper addresses numerical procedures utilized to the accurate and robust calculation of thousands of eigenpairs for the Dirac billiard resonator. The main challenges posed by the present work are: first, the capability of the approaches to tackle the large-scale eigenvalue problem, second, the ability to accurately extract many, i.e. order of thousands, interior eigenfrequencies for the considered problem, and third, the efficient implementation of the underlying approaches. The eigenfield calculations are accomplished in two steps. Initially, the finite integration technique or the finite element method with higher order curvilinear elements is applied, and further, the (B-)Lanczos method with its variations is exploited for the eigenpair determination. The comparative assessment of the numerical results to the complementary measurements confirms the applicability of the approaches and points out the significant reductions of computational costs. Finally, all of the results indicate that the suggested techniques can be used for precise determination of many eigenfrequencies.

1 Introduction

Over the last years, the increasing number of applications has stimulated the development of new methods and software for the numerical solution of large-scale eigenvalue problems. At the same time, the realistic applications frequently challenge the limit of both computer hardware and numerical algorithms, as one might possibly need large number of eigenpairs for matrices with dimension in

T. Banova (✉)
Institut für Theorie Elektromagnetischer Felder (TEMF), Technische Universität Darmstadt, Schlossgartenstraße 8, D-64289 Darmstadt, Germany

Graduate School of Computational Engineering, Technische Universität Darmstadt, Dolivostraße 15, D-64293 Darmstadt, Germany
e-mail: banova@temf.tu-darmstadt.de

W. Ackermann • T. Weiland
Institut für Theorie Elektromagnetischer Felder (TEMF), Technische Universität Darmstadt, Schlossgartenstraße 8, D-64289 Darmstadt, Germany
e-mail: ackermann@temf.tu-darmstadt.de; thomas.weiland@temf.tu-darmstadt.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_5

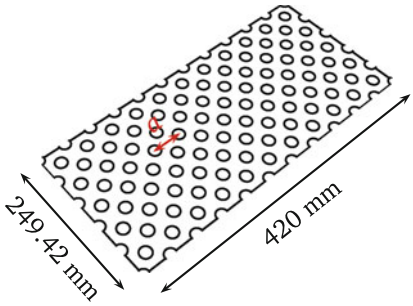


Fig. 1 Superconducting microwave Dirac billiard cavity containing 888 metal cylinders with radius $r = 4\,\text{mm}$ and height $h = 3\,\text{mm}$, squeezed between two metallic plates. The resonator is constructed from brass and coated with lead. As the lattice constant equals $a = 12\,\text{mm}$, the resulting photonic crystal has a total size of $249.42\,\text{mm} \times 420\,\text{mm} \times 3\,\text{mm}$

excess of several millions. In the present work, the investigations of the properties of a graphene using a microwave photonic crystal (Dirac billiard resonator) [1, 2] also emphasize the necessity for calculation of thousands of interior eigenfrequencies. Graphene is a monoatomic layer of carbon atoms arranged in a regular hexagonal pattern. Additionally, it can be described as one-atom thick layer of the mineral graphite [3, 4]. High-quality graphene is very strong, light, nearly transparent, and an excellent conductor of heat and electricity. Its interaction with other materials and light, and its inherently two-dimensional nature produce unique properties. Due to its peculiar electronic properties, the carbon allotrope attracted a lot of attention over the last years, which culminated in a Nobel Prize in 2010.

It is worth mentioning that the band structure of the photonic crystal, which is displayed in Fig. 1, possesses similar properties. More precisely, the photonic crystal considered in the present work is given by a three-dimensional Dirac billiard resonator composed of 888 metallic cylinders, which are arranged on a triangular lattice and squeezed between two metal plates. The cylinders are characterized with radius $r = 4\,\text{mm}$ and height $h = 3\,\text{mm}$. As the lattice constant equals $a = 12\,\text{mm}$, the resulting photonic crystal has a total size of $249.42\,\text{mm} \times 420\,\text{mm} \times 3\,\text{mm}$. Each cylinder is screwed to the top and the bottom brass plate to ensure a proper mechanical stability and thus, reproducibility of the measurements. Finally, both the lid and the body are leaded in order to reach superconductivity by cooling with liquid helium at low temperatures (4.2 K). Herewith, precise statistics for the superconducting Dirac billiard cavity can be generated only if thousands of eigenfrequencies are calculated. The problem to compute a large number of eigenfrequencies along with their associated eigenvectors is given by the condition that they are located inside the spectrum.

Reflecting the state that an analytical solution for the electromagnetic problem of a Dirac billiard is not available, this work resorts to a numerical solution. Namely, if the finite element method [5] is utilized to solve the electromagnetic problem of a superconducting cavity, the numerical solution of a generalized large-scale

eigenvalue problem

$$A\,\mathbf{x} = \lambda\, B\,\mathbf{x} \tag{1}$$

for given real symmetric matrices A and B is considered at the end. Thereon, the algebraic eigenvalue problem is solved with the B-Lanczos solvers [6]. Supposing that the numerical solution of the same problem is treated by the finite integration technique [7], finally it yields to a standard eigenvalue problem

$$A\,\mathbf{x} = \lambda\,\mathbf{x} \tag{2}$$

for a given symmetric sparse matrix $A \in \mathbb{R}^{n\times n}$. The novelty of this work is towards efficient, robust, and accurate determination of many interior eigenpairs for (2). For validation purposes, the numerical results are compared to the measurements.

Despite the fact that various types of numerical methods for eigenvalue determination (Krylov-Schur, Jacobi-Davidson, Arnoldi) are available in different software packages, not as many are specifically adapted for computing thousands of eigenpairs. The Lanczos method [8] with its variations is very attractive for the aforementioned project necessities, as it reduces the original eigenvalue problem to a tridiagonal one and takes a significant advantage over its competitors, which concentrate on individual frequency samples per iteration. Among the basic implementations of the Lanczos algorithm, a combination with a filtering method is used as a valuable tool to enable the computation of interior eigenpairs. Moreover, the solvers exploit all parallelism from a multi-threaded and multi-process implementation of the used libraries. Analogously, this facilitates a higher mesh resolution to be considered, whereby the computational costs will be kept on an acceptable level.

2 Eigenvalue Determination in Frequency Domain

Within this work, the excited electromagnetic fields inside closed resonators are considered under the assumption of perfectly electric conducting walls. Prior to frequency-domain simulations, the related geometry is modeled and decomposed into hexahedral or tetrahedral elements with the CST Microwave Studio® [9]. Afterward, the corresponding mesh information is passed to the CEM3D solver [5] in order to produce the sparse matrices that are used as input for the eigenmode solvers. Here, the CEM3D program solves the electromagnetic problem either with the finite integration technique or with a higher order finite element method. Respectively, the outcome is either a standard or a generalized eigenvalue problem, derived from the Maxwell's equations for a loss-free and source-free bounded domain with perfectly electric conducting walls on its surface.

2.1 Lanczos Method with Polynomial Filtering

The Lanczos algorithm with polynomial filtering (cf. Fig. 2) replaces the matrix-vector product $A\,\mathbf{v}_j$ in the Lanczos algorithm [6] by $\rho(A)\,\mathbf{v}_j$, where ρ is a polynomial being determined from the knowledge on the distribution of the sought eigenvalues. The main goal of the polynomial filtering is to enhance the Lanczos projection scheme by processing the vectors $\mathbf{v}_j$, such that their components in the unwanted parts of the spectrum are relatively reduced to those in the wanted parts. By means of a three-term recurrence formula

$$\beta_{j+1}\,\mathbf{v}_{j+1} = \rho(A)\,\mathbf{v}_j - \alpha_j\,\mathbf{v}_j - \beta_j\,\mathbf{v}_{j-1}, \tag{3}$$

the Lanczos recursion with polynomial filtering generates a highly-structured (in fact tridiagonal) real symmetric matrix T, which is defined as

$$T = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_j \\ & & & \beta_j & \alpha_j \end{bmatrix}. \tag{4}$$

1: {determine a polynomial filter $\rho(\lambda)$ with γ such that $\rho(\lambda) \geq \gamma$, $\lambda \in [\xi, \eta]$}
2: $\mathbf{v}_0 \leftarrow 0$, $\mathbf{v}_1 \leftarrow$ random vector, $\beta_1 \leftarrow 0$
3: **for all** $j = 1, 2, \ldots$ **do**
4: $\quad \mathbf{v}_{j+1} \leftarrow \rho(A)\,\mathbf{v}_j - \beta_j\,\mathbf{v}_{j-1}$
5: $\quad \alpha_j \leftarrow \langle \mathbf{v}_j, \mathbf{v}_{j+1} \rangle$
6: $\quad$ {calculate Ritz pairs}
7: $\quad$ {check convergence every tenth iteration}
8: $\quad \mathbf{v}_{j+1} \leftarrow \mathbf{v}_{j+1} - \alpha_j\,\mathbf{v}_j$
9: $\quad$ {apply reorthogonalization}
10: $\quad \beta_{j+1} \leftarrow \left|\left|\mathbf{v}_{j+1}\right|\right|$
11: $\quad$ **if** $\beta_{j+1} == 0$ **then**
12: $\quad\quad$ **break**
13: $\quad$ **end if**
14: $\quad \mathbf{v}_{j+1} \leftarrow \mathbf{v}_{j+1} / \beta_{j+1}$
15: **end for**
16: {compute the Ritz values θ_j of T and the corresponding Ritz vectors $\mathbf{y}_j$}
17: {compute the approximate eigenvalues $\lambda_j = \langle \mathbf{y}_j, A\,\mathbf{y}_j \rangle$}

Fig. 2 Lanczos algorithm with polynomial filtering for the solution of the standard eigenvalue problem (2). The polynomial filter $\rho(\lambda)$ is expanded in the proper scaled and shifted basis of the Chebyshev polynomials

Herein, the major practical advantage of this method is the tridiagonal reduction of the eigenvalue problem that yields minimal storage requirements, as do the associated algorithms for its eigenvalue and eigenvector computations.

According to the descriptive view in Fig. 2, the iterative process implements the modified Gram-Schmidt process, where in every Lanczos iteration the newest Lanczos vector $\mathbf{v}_{j+1}$ is determined by orthogonalizing the vector $\rho(A)\,\mathbf{v}_j$ with respect to $\alpha_j\,\mathbf{v}_j$ and $\beta_j\,\mathbf{v}_{j-1}$. Due to the roundoff errors, which result from the finite computer arithmetic, and the convergence of the eigenvalues of the matrix T to the eigenvalues of the original matrix $\rho(A)$, spurious eigenvalues are attributed to the losses in the orthogonality of the Lanczos vectors. Therefore, the Lanczos vectors are reorthogonalized in line 9 of Fig. 2. Various reorthogonalization schemes have been proposed in the literature to correct the loss of orthogonality of the Lanczos vectors. Within this work, the implemented Lanczos method with polynomial filtering uses a full reorthogonalization for simplicity [10]. That is, the orthogonality of the current Lanczos vector $\mathbf{v}_j$ against all the previous vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$ is reinstated at each step j.

A fundamental problem lies in computing an appropriate polynomial ρ in order to approximate a step function that covers the interval of the desired eigenvalues $[\xi, \eta]$. It should be noted that the matrices A and $\rho(A)$ share the same eigenvectors, and the matrix $\rho(A)$ has eigenvalues $\rho(\lambda_1), \ldots, \rho(\lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix A. If the polynomial $\rho(\lambda)$ is chosen such that $\rho([\xi, \eta])$ is in an extreme region of the spectrum, the eigenvalues of the matrix $\rho(A)$ in $\rho([\xi, \eta])$ will be approximated first. Afterwards, the corresponding eigenvectors can be used to extract the eigenvalues of the matrix A in $[\xi, \eta]$. However, a high-degree polynomial approximation to a discontinuous step function exhibits parasitic oscillations. Therefore, a two-stage process is adapted [11]. First, a smooth function similar to the step function is selected and then a polynomial approximation $\rho(\lambda)$ to this function is applied in the least-squares sense (see line 1 of Fig. 2).

A variant, known as the filtered conjugate residual polynomial algorithm is proposed in [11]. Here, the functions are expanded in the proper scaled and shifted basis of the Chebyshev polynomials. Thus, all inner product operations as well as the adding and the scaling operations of two expanded polynomials can be easily performed with the expansion coefficients. As a consequence of the 3-term recurrence of the Chebyshev polynomials, the polynomial multiplication by λ can be also easily implemented. The details are omitted here and can be found in [11, 12]. Due to the fact that the procedure is performed in a polynomial space, for the standard eigenvalue problem the matrix is never invoked and therefore, the resulting computing costs are negligible.

The convergence of the algorithm is checked in line 7 of Fig. 2. With a given tolerance ε, the desired eigenvalues are deemed to have converged at the j iteration if the number of sought eigenvalues of T_j is the same as the number of eigenvalues of T_{j-1} and the error of the sought eigenvalues, measured in the relative and the average sense, is below the tolerance ε.

3 Application Example: Dirac Billiard Resonator

The dedicated eigenmode solvers are implemented in C++ and based on PETSc data structures [13]. At this point, it should be pointed out that the parallel vectors and the sparse matrices are easily and efficiently assembled through the mechanisms provided by PETSc. Additionally, the PETSc library enables parallel computing by employing the MPI standard for all message-passing communication. Moreover, the implemented solvers employ the Intel MKL 10.2 library with LAPACK [14]. In case of the standard eigenvalue problem (2), the algorithm presented in Sect. 2.1 performs repeated computations of matrix-vector products, which are the only large-scale operations included within this approach. On the other hand, the solution of the problem (1) introduces a factorization of the matrix B. The details are omitted and can be found in [6].

The frequency spectrum from 19 to 31 GHz is numerically calculated (1656 eigenpairs in total) and then, compared with the measurements.[1] During the measurements, the analyzed structure is cooled down to a temperature of 4.2 K, which is naturally accompanied with a geometrical shrinkage. Thus, the raw measurement data are scaled with a factor that compensates for the difference in the dimensions of the measured and the simulated structure. In the numerical studies, the eigenfrequencies are determined for the cases when the Dirac billiard is discretized with 4,515,840 hexahedrons and 630,348 tetrahedrons by using the Lanczos solver with polynomial filtering and the B-Lanczos solver with shift-and-invert, respectively. The Lanczos solver with polynomial filtering is set to calculate the above mentioned eigenfrequencies in two simulations, whereas the B-Lanczos solver with shift-and-invert computed the sought eigenfrequencies in eight simulations.

Prior to comparing the level-density analyses based on the eigenvalues determined with the different approaches, additional information about the accounted computational resources is presented. In the simulation studies, it was experienced that 15 cluster nodes are suited for problems with more than 10^6 mesh cells when using the Lanczos solver with polynomial filtering. To be precise, each node has two six-core Intel Xeon X5650 3.0 GHz processors along with 24 GB of available working memory. On the other side, the B-Lanczos solver with shift-and-invert is run on a powerful computer with 256 GB of RAM memory and two quad-core Intel Xeon E5-2643 processors, clocked at 3.3 GHz. The computational time as well as the memory consumptions for the eigenpair determination are summarized in Table 1.

The results for the level-density analysis are compared in Fig. 3. On the abscissa are given the frequencies in GHz, whereas the ordinate presents the occurrences, i.e. obtained with the help of a Lorentz function [15], which belong to a specific

[1]The measurements are kindly provided from the Institute for Nuclear Physics at Technical University of Darmstadt [2, 15].

Table 1 Computational time and memory consumptions for the determination of 1656 eigenpairs with the Lanczos solver with polynomial filtering and the B-Lanczos solver with shift-and-invert

	Lanczos with polynomial filtering	B-Lanczos with shift-and-invert
Eigenfrequencies	1656	1656
Time (days)	0.4	1.6
Memory/eig (MB)	201.3	295.2

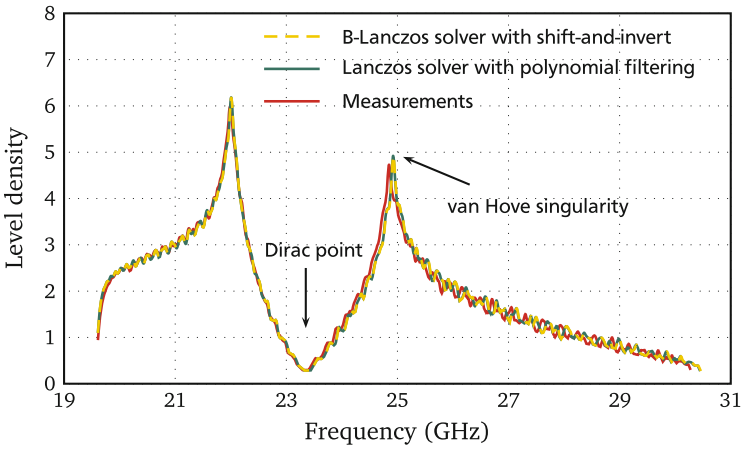


Fig. 3 Comparison of the numerical results to the complementary measurements [2, 15] for a superconducting Dirac billard cavity. The B-Lanczos solver with shift-and-invert and the Lanczos solver with polynomial filtering are exploited for the computation of the eigenpairs

frequency. The red line shows the reflection spectrum measured with antennas placed at different positions inside of the photonic crystal. The locations of the antennas are chosen in the center of three cylinders, forming a triangle, in order to minimize the disturbance of the propagating mode at the Dirac frequency.

In the considered frequency spectrum, only one band with a Dirac point around 23.5 GHz is present. Below 19 GHz and above 31 GHz band gaps can be noticed. As displayed in this figure, it is clear that the number of resonances in the range of 23.5 GHz decreases greatly. This behavior reflects the vanishing density of states at the Dirac point. The experimental reflection spectrum also has a clearly pronounced minimum around 23.5 GHz, i.e. within the frequency range where the Dirac point is expected, and shows the characteristic cusp structure. The sharp resonances at the edges of the bands are related to the so-called van Hove singularities. Evidently, the measured spectrum closely resembles those obtained by the numerical simulations.

4 Conclusions

In this paper, the statistical properties of a Dirac billiard resonator are investigated via an employment of different numerical approaches to calculate thousands of its eigenpairs. The numerical approaches are initially based on the finite integration technique or the finite element method with higher order curvilinear elements, and afterwards, the (B-)Lanczos method with its variations takes advantage over its competitors for the solution of the (generalized) eigenvalue problem.

In addition to the need to ensure high precision of the proposed approaches, the numerically calculated eigenfrequencies are compared side by side with the reference data, which are determined by the measurements. Hereby, the findings show that the proposed approaches deliver solutions, which agree well with the reference data, gaining high accuracy and efficiency in eigenfield determination. Beside the accuracy, the robustness of the underlying approaches is also investigated throughout this work. Finally, all of the results indicate that the suggested techniques can be applicable in different areas of applications, where a precise determination of plenty of eigenfrequencies takes a crucial role.

Acknowledgements This work was supported by the "Excellence Initiative" of the German Federal and State Governments and the Graduate School of Computational Engineering at TU Darmstadt.

References

1. Bittner, S., Dietz, B., Miski-Oglu, M., Richter, A.: Extremal transmission through a microwave photonic crystal and the observation of edge states in a rectangular Dirac billiard. Phys. Rev. B **85**(6), 064301 (2012)
2. Bittner, S., Dietz, B., Miski-Oglu, M., Oria Iriarte, P., Richter, A., Schäfer, F.: Observation of a Dirac point in microwave experiments with a photonic crystal modeling graphene. Phys. Rev. B **82**(1), 014301 (2010)
3. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V., Firsov, A.A.: Electric field effect in atomically thin carbon films. Science **306**, 666–669 (2004)
4. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Katsnelson, M.I., Grigorieva, I.V., Dubonos, S.V., Firsov, A.A.: Two-dimensional gas of massless Dirac fermions in graphene. Nature **438**, 197–200 (2005)
5. Ackermann, W., Benderskaya, G., Weiland, T.: State of the art in the simulation of electromagnetic fields based on large scale finite element eigenanalysis. ICS Newslett. **17**(2), 3–12 (2010)
6. Banova, T., Ackermann, W., Weiland, T.: Accurate determination of thousands of eigenvalues for large-scale eigenvalue problems. IEEE Trans. Magn. **50(2)**, 481–484 (2014)
7. Weiland, T.: A discretization method for the solution of Maxwell's equations for six-component fields. Electr. Commun. (AEÜ) **31**(3), 116–120 (1977)
8. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Natl. Bur. Stand. **45**(4), 255–282 (1950)

9. CST - Computer Simulation Technology AG: CST Microwave Studio®. Darmstadt (2012). http://www.cst.com.CitedAug112014
10. Saad, Y.: Numerical Methods for Large Eigenvalue Problems. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2011)
11. Saad, Y.: Filtered conjugate residual-type algorithms with applications. SIAM J. Matrix Anal. Appl. **28**(3), 845–870 (2006)
12. Fang, H., Saad, Y.: A filtered Lanczos procedure for extreme and interior eigenvalue problems. SIAM J. Sci. Comput. **34**(4), A2220–A2246 (2012)
13. Balay, D., Brown, J., Buschelman, K., Eijkhout, V., Gropp, W., Kaushik, D., Knepley, M., McInnes, L.C., Smith, B., Zhang, H.: PETSc users manual. Argonne National Laboratory (2011)
14. Intel: Intel ©Math Kernel Library. (2010). http://www.intel.com.CitedAug112014
15. Cuno, C.: Randzustände in einem supraleitenden Mikrowellen-Diracbillard. Bachelor's thesis, Technische Universität Darmstadt, Darmstadt (2012)

DG Treatment of Non-conforming Interfaces in 3D Curl-Curl Problems

Raffael Casagrande, Christoph Winkelmann, Ralf Hiptmair, and Joerg Ostrowski

Abstract We consider 3D Curl-Curl type of problems in the presence of arbitrary, non-conforming mesh-interfaces. The Interior Penalty/Nitsche's Method (Stenberg, Computational mechanics, 1998) is extended to these problems for edge functions of the first kind. We present an a priori error estimate which indicates that one order of convergence is lost in comparison to conforming meshes due to insufficient approximation properties of edge functions. This estimate is sharp for first order edge functions: In a numerical experiment the method does not converge as the mesh is refined.

1 Introduction

The Curl-Curl equation,

$$\nabla \times (\mu^{-1} \nabla \times \mathbf{A}) = \mathbf{j}^i, \tag{1}$$

can be used to calculate the magnetic field that originates from a stationary current $\mathbf{j}^i$. Herein μ denotes the magnetic permeability, $\mathbf{A}$ is the magnetic vector potential and the magnetic field is $\mathbf{B} = \nabla \times \mathbf{A}$. The Magnetostatic model (1) is a special case of the temporally gauged Eddy Current model (note that (2) reduces to (1) in static cases as well as in regions where the electric conductivity σ vanishes) ,

$$\sigma \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}t} + \nabla \times (\mu^{-1} \nabla \times \mathbf{A}) = \mathbf{j}^i. \tag{2}$$

R. Casagrande (✉) • C. Winkelmann • R. Hiptmair
Seminar for Applied Mathematics, ETH Zürich, Rämistr. 101, CH-8092 Zürich, Switzerland
e-mail: raffael.casagrande@sam.math.ethz.ch; christoph.winkelmann@sam.math.ethz.ch; hiptmair@sam.math.ethz.ch

J. Ostrowski
ABB Switzerland Ltd., Corporate Research, Segelhofstrasse 30–34, CH-5405 Baden, Switzerland
e-mail: joerg.ostrowski@ch.abb.com

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_6

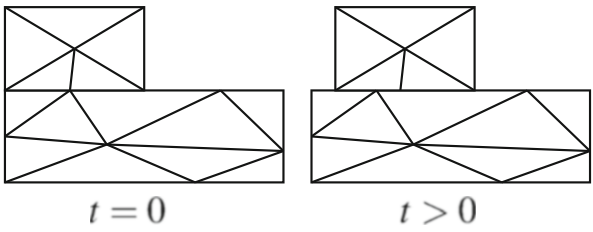


Fig. 1 Initially conforming sub-meshes become non-conforming when the upper sub-mesh starts moving

In some applications like the simulation of electric machines or magnetic actuators, magnetic fields have to be computed in the presence of moving, rigid parts. Then one may use separate, moving sub-meshes for them in order to avoid remeshing. However, this leads to so-called "sliding interfaces", i.e. meshes with hanging nodes (cf. Fig. 1).

Our aim is to construct a method which solves (1) such that the solution is not affected by the "non-conformity" of the sub-meshes at the common interface. This problem has been studied in depth in the framework of so called Mortar Methods where the continuity requirements are incorporated directly into the trial-space [1] or they are enforced by additional Lagrange Multipliers [2]. These approaches have been proven to be successful, but they require the inversion of a full matrix respectively additional unknowns. A related approach uses a primal/dual formulation and couples the systems in a weak sense across the interface [3].

We pursue a different approach that fits into the framework of Discontinuous Galerkin (DG) methods which support non-conforming meshes naturally. A Locally Discontinuous Galerkin scheme for sliding meshes has already been proposed and analyzed in [4]. We will study a simpler method which has it's origins in Nitsche's Method. The idea is to penalize tangential discontinuities along the non-conforming interface, but not in the interior of the subdomains where we use a standard FEM discretization. The method has been analyzed for the Poisson Equation in [5] where it was shown that a symmetric positive definite system matrix results. We aim to extend this idea to (1).

It is important to realize that (1) and (2) (if $\sigma = 0$ anywhere) don't have a unique solution (in the L^2-norm). In this work, we will therefore study the regularized problem,

$$\nabla \times (\mu^{-1} \nabla \times \mathbf{A}) + \varepsilon \mathbf{A} = \mathbf{j}^i, \quad \text{in } \Omega \tag{3}$$

$$\mathbf{n} \times \mathbf{A} = \mathbf{g}_D \quad \text{on } \partial\Omega. \tag{4}$$

Here $\varepsilon > 0$ is the regularization parameter that renders the solution unique. We discuss the influence and necessity of the regularization term in [6]. Finally we want to point out that the boundary condition (4) implies $(\nabla \times \mathbf{A}) \cdot \mathbf{n} = \mathbf{B} \cdot \mathbf{n} = 0$ on $\partial\Omega$ which reflects the decay of the fields away from the source.

We start our discussion by introducing DG notations (Sect. 2) that we need in order to introduce the interior penalty formulation in Sect. 3. Section 3 also analyzes

the behavior of the discrete solution as the mesh is refined (h-refinement) and the role of the approximation space is studied. The theoretical results are compared to numerical experiments in Sect. 4. We finish our discussion by a short conclusion and outlook (Sect. 5).

2 Preliminaries

Before we can introduce the Symmetric Weighted Interior penalty (SWIP) formulation of (3)–(4) we introduce some definitions and notations:

Subdomains and Sub Meshes Let us assume that the domain Ω, on which (3)–(4) is posed, is a simply connected polyhedron with Lipschitz boundary. Furthermore we assume Ω to be split into two non-overlapping subdomains $\overline{\Omega_1} \cup \overline{\Omega_2} = \overline{\Omega}$. On each subdomain we introduce a sequence of shape regular, simplical meshes in the sense of Ciarlet such that the union $\mathscr{T}_{\mathscr{H}} = \mathscr{T}_{\mathscr{H},1} \cup \mathscr{T}_{\mathscr{H},2}$ is *quasi-uniform* at the non-conforming interface $\Gamma = \overline{\Omega}_1 \cap \overline{\Omega}_2$ (cf. [6], Definition 1). It is easy to check that meshes created by the motion of individual sub-meshes (cf. Fig. 1) fit this definition.

Magnetic Permeability We assume there exists a partition $P_\Omega = \{\Omega_{i,\mu}\}$ such that each $\Omega_{i,\mu}$ is a polyhedron and such that the permeability $\mu > 0$ is constant on each $\Omega_{i,\mu}$. Furthermore the mesh sequence $\mathscr{T}_{\mathscr{H}}$ is *compatible* with the partition P_Ω: For each $\mathscr{T}_h \in \mathscr{T}_{\mathscr{H}}$, each element $T \in \mathscr{T}_h$ belongs to exactly one $\Omega_{i,\mu} \in P_\Omega$. I.e. the magnetic permeability is allowed to jump over element boundaries, and in particular over the non-conforming interface Γ.

Polynomial Approximation Later on we will seek our discrete solution in the piecewise polynomial space (cf. [7]),

$$P_3^k(\mathscr{T}_h) := \left\{ p \in L^2(\Omega) \,\middle|\, \forall T \in \mathscr{T}_h,\, p|_T \in P_3^k(T) \right\} \tag{5}$$

where $\mathscr{T}_h \in \mathscr{T}_{\mathscr{H}}$ and $P_3^k(T)$ is the usual space of polynomials up to degree k on mesh element T. Note that functions of $P_3^k(\mathscr{T}_h)$ are discontinuous across element boundaries.

Mesh Faces, Jump and Average Operators We denote by $\mathscr{F}_h = \mathscr{F}_h^b \cup \mathscr{F}_h^i$ the set of all boundary and inner faces of a given mesh $\mathscr{T}_h$. $\mathscr{F}_T$ stands for all faces of the mesh element $T \in \mathscr{T}_h$. For each mesh face F, *vector valued* function $\mathbf{A}_h \in P_3^k(\mathscr{T}_h)^3$, we define

$$\begin{aligned}
&\text{if } F \in \mathscr{F}_h^i,\ F = \partial T_i \cap \partial T_j: \quad [\mathbf{A}_h]_T = \mathbf{n}_F \times \left(\mathbf{A}_h|_{T_i} - \mathbf{A}_h|_{T_j} \right), \\
&\text{if } F \in \mathscr{F}_h^b,\ F = \partial T_i \cap \partial \Omega: \quad [\mathbf{A}_h]_T = \mathbf{n}_F \times \mathbf{A}_h|_{T_i}, \\
&\text{if } F \in \mathscr{F}_h^i,\ F = \partial T_i \cap \partial T_j: \quad \{\mathbf{A}_h\}_\omega = \omega_1\, \mathbf{A}_h|_{T_i} + \omega_2\, \mathbf{A}_h|_{T_j}, \\
&\text{if } F \in \mathscr{F}_h^b,\ F = \partial T_i \cap \partial \Omega: \quad \{\mathbf{A}_h\}_\omega = \mathbf{A}_h|_{T_i}.
\end{aligned} \qquad \begin{matrix} \text{(jump)} \\ \\ \text{(average)} \\ \\ \end{matrix}$$

Here $\mathbf{n}_F$ always points from T_i to T_j and $\omega_i \in [0, 1]$ such that $\omega_1 + \omega_2 = 1$.

3 Symmetric Weighted Interior Penalty (SWIP) Formulation

We chose an arbitrary subspace $V_h \subseteq P_3^k(\mathscr{T}_h)^3$ as discrete test and trial space, and use integration by parts (cf. [7, 8] for details) to arrive at the SWIP formulation of (3): Find $\mathbf{A}_h \in V_h$ such that

$$a_h^{\mathrm{SWIP}}(\mathbf{A}_h, \mathbf{A}'_h) + \varepsilon \int_\Omega \mathbf{A}_h \cdot \mathbf{A}'_h = \ell_h(\mathbf{A}'_h) \qquad \forall \mathbf{A}'_h \in V_h \tag{6}$$

with

$$\begin{aligned} a_h^{\mathrm{SWIP}}(\mathbf{A}_h, \mathbf{A}'_h) = &\int_\Omega \left(\mu^{-1}\nabla \times \mathbf{A}_h\right) \cdot \left(\nabla \times \mathbf{A}'_h\right) - \sum_{F \in \mathscr{F}_h} \int_F \left\{\mu^{-1}\nabla \times \mathbf{A}_h\right\}_\omega \cdot \left[\mathbf{A}'_h\right]_T \\ &- \sum_{F \in \mathscr{F}_h} \int_F \left\{\mu^{-1}\nabla \times \mathbf{A}'_h\right\}_\omega \cdot [\mathbf{A}_h]_T + \sum_{F \in \mathscr{F}_h} \frac{\eta \gamma_{\mu,F}}{a_F} \int_F [\mathbf{A}_h]_T \cdot \left[\mathbf{A}'_h\right]_T, \end{aligned} \tag{7}$$

$$\begin{aligned} \ell_h(\mathbf{A}'_h) = &\int_\Omega \mathbf{j}^i \cdot \mathbf{A}'_h - \sum_{F \in \mathscr{F}_h^b} \int_F \left\{\mu^{-1}\nabla \times \mathbf{A}'_h\right\}_\omega \cdot (\mathbf{n} \times \mathbf{g}_D) \\ &+ \sum_{F \in \mathscr{F}_h^b} \frac{\eta \gamma_{\mu,F}}{a_F} \int_F \left[\mathbf{A}'_h\right]_T \cdot (\mathbf{n} \times \mathbf{g}_D). \end{aligned} \tag{8}$$

where $a_F = \frac{1}{2}(h_{T_1} + h_{T_2})$ is the average diameter of the adjacent elements of face F and η is the penalty parameter. The weights are chosen as

$$\gamma_{\mu,F} := \frac{2}{\mu_1 + \mu_2}, \qquad \omega_1 := \frac{\mu_1}{\mu_1 + \mu_2}, \qquad \omega_2 := \frac{\mu_2}{\mu_1 + \mu_2}.$$

Remark If $V_h \subseteq \mathbf{H}(\mathbf{curl})$, then all inner tangential jumps in (7) will drop out and only jumps at the boundary remain. I.e. we are left with a standard FEM formulation where the inhomogeneous boundary conditions (4) are enforced in a weak sense.

3.1 A Priori Error Estimate

Using the theory of DG Methods one can derive the following error estimate [6]:

Theorem 1 *Let* $\mathbf{A} \in V^* := \mathbf{H}(\mathbf{curl}, \Omega) \cap H^2(P_\Omega)^3$ *be a solution of the strong formulation (3)–(4) (in the sense of distributions) and let* $\mathbf{A}_h \in V_h \subseteq P_3^k(\mathscr{T}_h)^3$ *solve the variational formulation (6). Then there exist constants* $C > 0$, $C_\eta > 0$, *both*

independent of h, μ, such that for $\eta > C_\eta$,

$$\|\mathbf{A} - \mathbf{A}_h\|_{SWIP} < C \inf_{\mathbf{v}_h \in V_h} \|\mathbf{A} - \mathbf{v}_h\|_{SWIP,*}, \tag{9}$$

and the discrete problem (6) is well-posed.

The associated function spaces and norms are defined by

$$\begin{aligned} H^2(P_\Omega)^3 &:= \left\{ \mathbf{A} \in L^2(\Omega)^3 \,\middle|\, \forall \Omega_{i,\mu} \in P_\Omega : \ \mathbf{A}|_{\Omega_{i,\mu}} \in H^2(\Omega_{i,\mu})^3 \right\}, \\ \|\mathbf{A}\|^2_{\mathrm{SWIP}} &:= \|\mu^{-1/2} \nabla \times \mathbf{A}\|^2_{L^2(\Omega)} + \|\varepsilon^{1/2} \mathbf{A}\|^2_{L^2(\Omega)} + \sum_{F \in \mathscr{F}_h} \frac{\gamma_{\mu,F}}{a_F} \| [\mathbf{A}]_T \|^2_{L^2(F)}, \\ \|\mathbf{A}\|^2_{\mathrm{SWIP},*} &:= \|\mathbf{A}\|^2_{\mathrm{SWIP}} + \sum_{T \in \mathscr{T}_h} h_T \|\mu^{-1/2} \ \nabla \times \mathbf{A}|_T \|^2_{L^2(\partial T)}. \end{aligned} \tag{10}$$

Theorem 1 tells us that the total error is bounded by the best approximation error. In the following we will bound the best approximation error in the $\|\cdot\|_{\mathrm{SWIP},*}$ norm for two concrete choices of V_h.

Edge Functions of the First Kind In this section we assume $V_h = R^k(\Omega_1) \oplus R^k(\Omega_2) \subset P_3^k(\mathscr{T}_h)$, where R^k is the space of k-th order edge functions (cf. [9], Sect. 5.5) of the first kind. The following polynomial approximation result gives a bound on the right-hand side of (9):

Theorem 2 *Assume the exact solution $\mathbf{A} \in V := \mathbf{H}(\mathbf{curl}, \Omega) \cap H^{s+1}(\Omega_1 \oplus \Omega_2)^3$ with integer $1 \leq s \leq k$, then there exists a projector $\pi_h : V \mapsto V_h$ such that*

$$\|\mathbf{A} - \pi_h \mathbf{A}\|_{SWIP,*} < C h^{s-1} \left(\|\mathbf{A}\|_{H^{s+1}(\Omega_1)^3} + \|\mathbf{A}\|_{H^{s+1}(\Omega_2)} \right).$$

Where C is independent of h.

Sketch of Proof The approximation space V_h consists of two standard edge element spaces in Ω_1, Ω_2. We can thus use the standard projection operator r_h, as it is defined in [9], for edge functions on Ω_1, Ω_2 to compose our global projection operator π_h:

$$\pi_h(\mathbf{A}) := \left(r_h \left(\mathbf{A}|_{\Omega_1} \right), r_h \left(\mathbf{A}|_{\Omega_2} \right) \right).$$

Next, we note that all the tangential jumps in (7) and (10) that lie on interior, conforming faces drop out and only jumps over $F \in \mathscr{F}_h^{b,\Gamma} := \mathscr{F}_h^b \cup$

$\left\{F \in \mathscr{F}_h^i \middle| F \subset \overline{\Omega_1} \cap \overline{\Omega_2}\right\}$ remain. Thus,

$$\begin{aligned}\|\mathbf{A} - \pi_h\mathbf{A}\|^2_{\mathrm{SWIP},*} &= \|\mu^{-1/2}\nabla \times (\mathbf{A} - \pi_h\mathbf{A})\|^2_{L^2(\Omega)} + \|\varepsilon^{1/2}(\mathbf{A} - \pi_h\mathbf{A})\|^2_{L^2(\Omega)} \\ &+ \sum_{F\in\mathscr{F}_h^{b,\Gamma}} \frac{\gamma_{\mu,F}}{a_F} \| [\mathbf{A} - \pi_h\mathbf{A}]_T \|^2_{L^2(F)} + \sum_{T\in\mathscr{T}_h} h_T \|\mu^{-1/2}\nabla \times (\mathbf{A} - \pi_h\mathbf{A})\|^2_{L^2(\partial T)}, \\ &= T_1 + T_2 + T_3 + T_4.\end{aligned}$$

The terms T_1, T_2, T_4 are easily bounded in terms of $O(h^{2s})$ by standard polynomial approximation results (cf. Theorem 5.41 in [9]). However, for T_3 we have to use Lemma 5.52 in [9] to achieve a rate $O(h^{2s-2})$ which unfortunately dominates the other terms. The fact that the error contribution T_3 is confined to a neighborhood of the interface, respectively the boundary, does not help, because the solution may be concentrated there as well.

Piecewise Polynomials For the sake of completeness we shortly present an approximation result for the case $V_h = P_3^k(\mathscr{T}_h)$:

Theorem 3 *Assume the exact solution* $\mathbf{A} \in V := \mathbf{H}(\mathbf{curl}, \Omega) \cap H^{s+1}(P_\Omega)^3$ *with integer* $1 \le s \le k$, *then there exists a projector* $\pi_h^P : V \mapsto V_h$ *such that*

$$\|\mathbf{A} - \pi_h^P\mathbf{A}\|_{SWIP,*} < Ch^s \|\mathbf{A}\|_{H^{s+1}(P_\Omega)^3}$$

where C is independent of h.

For the proof of this theorem we refer the reader to the proof of Theorem 3.21 in [8]. The important point is that piecewise polynomials $P_3^k(\mathscr{T}_h)$ yield the expected rate of convergence because they span the full polynomial space.

4 Numerical Results

We consider a 3D sphere with radius 1 that is split into two half-spheres which are then meshed individually (Fig. 2). We impose the analytical solution $\mathbf{A} = (\sin y, \cos z, \sin x)$ and choose $\mathbf{j}^i$, $\mathbf{g}_D$ correspondingly ($\mu = 1$, $\varepsilon = 10^{-6}$).

Figure 3 shows the error for a sequence of quasi-uniform meshes which approximate the boundary linearly. We can see that piecewise polynomials yield always the expected rate of converge, $O(h)$, but this does not hold for edge functions: For $V_h = R^1(\Omega_1) \oplus R^1(\Omega_2)$ the error fluctuates significantly depending on the angle (see Fig. 4) and for $\theta = 2.86$ (solid line) no convergence is observed. This shows that the estimate in Theorem 2 is sharp for $k = 1$. For $k = 2$ we would expect $O(h)$ convergence but for all configurations we observe a rate of order at least $O(h^{1.5})$ because in this experiment the solution is not concentrated at the interface/boundary: T_3 decays faster than in the worst case.

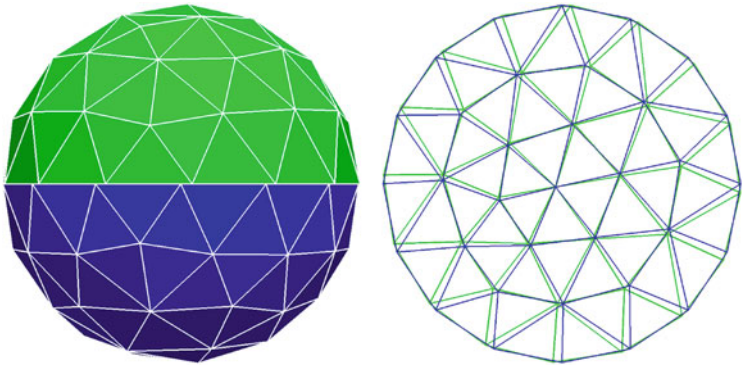


Fig. 2 The meshes for the two half spheres. The upper hemi-sphere is turned against the lower hemisphere by $\theta = 2.86$ degrees to create a non-conforming mesh

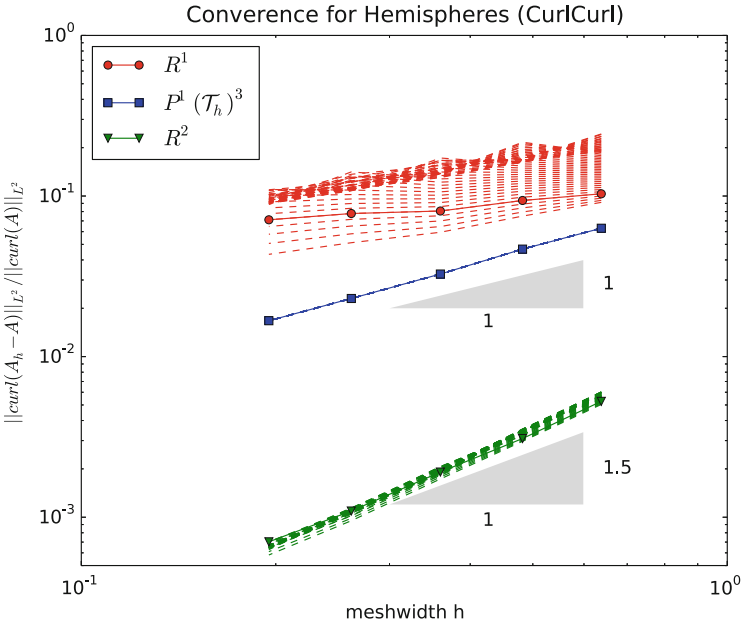


Fig. 3 The relative **H(curl)** error vs. mesh size h for rotation angle $\theta = 2.86$ degrees (*solid lines*). The *dashed gray lines* correspond to $\theta = 90n/(50\pi)$ degrees, $n \in \{0, 1, \ldots, 49\}$

Finally, Fig. 4 shows the relative error for different rotation angles for a fixed mesh width h. This confirms the previous result in that the error for $V_h = R^1(\Omega_1) \oplus R^1(\Omega_2)$ depends on the geometry of the overlapping meshes. For $V_h = P^1(\mathcal{T}_h)$, respectively $V_h = R^2(\Omega_1) \oplus R^2(\Omega_2)$ the error does not depend on θ.

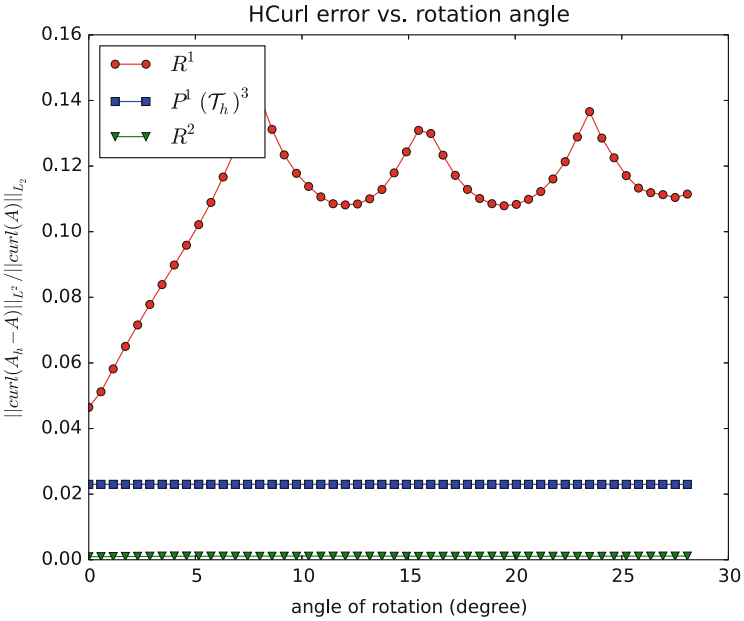


Fig. 4 The relative **H**(**curl**) error vs. the rotation angle for $h = 0.261$

5 Conclusion and Outlook

We have shown that a straightforward generalization of the Interior Penalty/Nitsche's Method to 3D Curl-Curl problems does not yield the expected rate of convergence if edge functions of the first kind are used. Generally one order of convergence is lost, i.e. for k-th order edge functions we observe $O(h^{k-1})$ convergence as the mesh is refined. The reason for this is that R^k doesn't span the full polynomial space P^k. Moreover, the result is sharp for $k = 1$, i.e. the method fails completely for first order edge functions. This problem can be cured by using either the full polynomial space P^1 or by using second order edge functions R^2.

The proposed SWIP scheme leads to a sparse, symmetric positive definite matrix which yields, together with the conjugate gradient method, a fast and robust solution scheme. Furthermore μ can be discontinuous across the non-conforming interface.

Outlook The proof of Theorem 2 suggests that it suffices to use 2nd order edge functions solely in elements adjacent to the non-conforming interface to achieve $O(h)$ convergence. This is easily implemented using hierarchical edge functions [10] and reduces the required number of degrees of freedom drastically. Another open question is whether the problem can still be solved using CG if the regularization

term in (3) is dropped because the system matrix then becomes positive semi-definite and the right-hand side is no longer in its range. Also, it is unclear whether the SWIP bilinear form offers a spectrally accurate discretization of the Curl-Curl operator [11] for $\varepsilon = 0$ and thus the convergence rate of CG may deteriorate as $h \to 0$.

References

1. Rapetti, F., Maday, Y., Bouillault, F., Razek, A.: Eddy-current calculations in three-dimensional moving structures. IEEE Trans. Magn. **38**(2), 613–616 (2002)
2. Wohlmuth, I.: Discretization Methods and Iterative Solvers Based on Domain Decomposition. Springer, Berlin (2001)
3. Rodriguez, A.A., Hiptmair, R., Valli, A.: A hybrid formulation of eddy current problems. Numer. Methods Partial Differ. Equ. **21**(4), 742–763 (2005)
4. Perugia, I., Schötzau, D.: On the coupling of local discontinuous Galerkin and conforming finite element methods. J. Sci. Comput. **16**(4), 411–433 (2001)
5. Stenberg, R.: Mortaring by a method of JA Nitsche. Computational mechanics (Buenos Aires, 1998), CD-ROM file, Centro Internac. Métodos Numér. Ing. Barcelona (1998)
6. Casagrande, R., Hiptmair, R.: An a priori error estimate for interior penalty discretizations of the Curl-Curl operator on non-conforming meshes. SAM Report, ETH Zürich, Zürich (2014)
7. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer, New York (2012)
8. Casagrande, R.: Sliding Interfaces for Eddy Current Simulations. Master's thesis, ETH Zürich, Zürich (2013)
9. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford University Press, New York (2003)
10. Bergot, M., Duruflé, M.: High-order optimal edge elements for pyramids, prisms and hexahedra. J. Comput. Phys. **232**, 189–213 (2013)
11. Buffa, A., Houston, P., Perugia, C.: Discontinuous Galerkin computation of the Maxwell eigenvalues on simplicial meshes. J. Comput. Appl. Math. **204**, 317–333 (2007)

A Symmetric and Low-Frequency Stable Potential Formulation for the Finite-Element Simulation of Electromagnetic Fields

Martin Jochum, Ortwin Farle, and Romanus Dyczij-Edlinger

Abstract A low-frequency stable potential formulation is presented. It covers lossy and lossless regions, results in symmetric finite-element matrices, and guarantees unique solutions. This contribution improves upon the authors' prior work by including general impressed currents and charge distributions. Moreover, it clarifies the interface condition to be imposed on the gauge on the common boundaries of the lossy and lossless regions.

1 Introduction

In recent years, the stability of finite-element (FE) formulations for the time-harmonic Maxwell equations in the low-frequency (LF) regime has gained a lot of attention. It is well known that the electric field formulation (EFF) breaks down in the static limit [1]. Various alternatives have been suggested [1–6], but none of them is completely satisfactory: The formulation of Dyczij-Edlinger et al. [1] does not consider ohmic losses, the method of Hiptmair et al. [2] leads to non-symmetric matrices and non-unique solutions, the purely algebraic approach of Ke et al. [3] relies on numerical break-down, and [4, 5] require an LF threshold and cannot recover magnetostatic fields. A promising approach is [6]; however, its matrices are non-symmetric.

In a recent publication [7], the authors presented an LF stable potential formulation that covers lossy and lossless domains and yields symmetric matrices and unique solutions. However, it requires all impressed currents to be solenoidal, and the lossless region to be free of charges. This contribution improves upon [7] by including general impressed currents and charge distributions. Moreover, a variational framework is provided that clarifies the interface condition to be imposed on the gauge on the common boundary of the lossy and lossless regions.

M. Jochum • O. Farle (✉) • R. Dyczij-Edlinger
Chair for Electromagnetic Theory, Saarland University, Saarland, Germany
e-mail: m.jochum@lte.uni-saarland.de; o.farle@lte.uni-saarland.de, edlinger@lte.uni-saarland.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_7

2 Time-Harmonic Boundary Value Problem

We write $\mathbf{E}$ and $\mathbf{H}$ for the electric and magnetic field strength, and $\mathbf{J}_i$ for the impressed current density, and ρ for the electric charge density. The wavenumber, characteristic impedance, and speed of light in free space are denoted by k_0, η_0, and c_0; the relative magnetic permeability and electric permittivity by μ_r and ε_r, respectively, and the electric conductivity by σ. We will also use the relative magnetic reluctivity $\nu_r = \mu_r^{-1}$. The indices C and N stand for "conducting" ($\sigma \neq 0$) and "non-conducting" ($\sigma = 0$), respectively.

Let Ω be a topologically simple domain which is partitioned into a conducting sub-domain Ω_C and a non-conducting region Ω_N. The interface of Ω_N and Ω_C is denoted by Γ and the unit surface normal by $\hat{\mathbf{n}}$.

We consider the Maxwell equations in the frequency domain,

$$\eta_0 \nabla\times \mathbf{H} = (\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\mathbf{E} + \eta_0\mathbf{J}_i \qquad \text{in } \Omega, \tag{1a}$$

$$\nabla\times \mathbf{E} = -\mathrm{i}k_0\eta_0\mu_r\mathbf{H} \qquad \text{in } \Omega, \tag{1b}$$

$$\nabla\cdot(\mu_r\mathbf{H}) = 0 \qquad \text{in } \Omega, \tag{1c}$$

$$\nabla\cdot(\varepsilon_r\mathbf{E}) = c_0\eta_0\rho \qquad \text{in } \Omega, \tag{1d}$$

subject to the boundary conditions (BC)

$$\mathbf{E}\times\hat{\mathbf{n}} = 0 \qquad \text{on } \partial\Omega, \tag{2a}$$

$$\hat{\mathbf{n}}\cdot(\mu_r\mathbf{H}) = 0 \qquad \text{on } \partial\Omega, \tag{2b}$$

and the interface conditions

$$(\mathbf{E}_C - \mathbf{E}_N)\times\hat{\mathbf{n}} = 0 \qquad \text{on } \Gamma, \tag{3a}$$

$$(\mathbf{H}_C - \mathbf{H}_N)\times\hat{\mathbf{n}} = 0 \qquad \text{on } \Gamma, \tag{3b}$$

$$[(\mu_r\mathbf{H})_C - (\mu_r\mathbf{H})_N]\cdot\hat{\mathbf{n}} = 0 \qquad \text{on } \Gamma, \tag{3c}$$

$$\{[(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\mathbf{E} + \eta_0\mathbf{J}_i]_C - (\mathrm{i}k_0\varepsilon_r\mathbf{E} + \eta_0\mathbf{J}_i)_N\}\cdot\hat{\mathbf{n}} = 0 \qquad \text{on } \Gamma. \tag{3d}$$

3 Source Modeling in the Lossless Domain

Taking the divergence of Ampère's Law (1a) leads to the continuity equation

$$\nabla\cdot[(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\mathbf{E}] = -\eta_0\,\nabla\cdot\mathbf{J}_i. \tag{4}$$

In the lossy domain Ω_C, the prescription of $\mathbf{J}_i$ fixes the sources of the electric field. Hence (1d) is not a governing equation for the electromagnetic fields. Rather, the

charge density ρ becomes a dependent quantity which is obtained from $\mathbf{E}$ in a post-processing step, via (1d). Moreover, there are no particular constraints on $\mathbf{J}_i$.

In the lossless domain Ω_N, (4) simplifies to

$$\mathrm{i}k_0 \nabla \cdot (\varepsilon_r \mathbf{E}) = -\eta_0 \nabla \cdot \mathbf{J}_i. \tag{5}$$

Substituting (1d) for the left-hand side of (5) shows that $\mathbf{J}_i$ and ρ are linked by

$$\nabla \cdot \mathbf{J}_i = -\mathrm{i}k_0 c_0 \rho. \tag{6}$$

For $k_0 = 0$, ρ becomes an independent excitation for the electrostatic problem in Ω_N:

$$\nabla \times \mathbf{E} = 0, \tag{7}$$

$$\nabla \cdot (\varepsilon_r \mathbf{E}) = c_0 \eta_0 \rho. \tag{8}$$

4 Electric Field Formulation and Low-Frequency Instability

A classical example of a formulation that breaks down in the static limit is given by the time-harmonic EFF. The corresponding boundary value problem (BVP) reads

$$\nabla \times (\nu_r \nabla \times \mathbf{E}) + \mathrm{i}k_0 \eta_0 \sigma \mathbf{E} - k_0^2 \varepsilon_r \mathbf{E} = -\mathrm{i}k_0 \eta_0 \mathbf{J}_i \qquad \text{in } \Omega, \tag{9a}$$

$$\mathbf{E} \times \hat{\mathbf{n}} = 0 \qquad \text{on } \partial\Omega, \tag{9b}$$

where $\mathbf{J}_i$ is given. As long as $k_0 > 0$ holds, (9a) incorporates the continuity equation (4) in lossy regions Ω_C and the electric flux balance (1d) in lossless regions Ω_N, respectively, as can be seen by taking the divergence of (9a):

$$\mathrm{i}k_0 \nabla \cdot [(\eta_0 \sigma + \mathrm{i}k_0 \varepsilon_r) \mathbf{E}] = -\mathrm{i}k_0 \eta_0 \nabla \cdot \mathbf{J}_i \qquad \text{in } \Omega_C, \tag{10}$$

$$-k_0^2 \nabla \cdot (\varepsilon_r \mathbf{E}) = -\mathrm{i}k_0 \eta_0 \nabla \cdot \mathbf{J}_i = -k_0^2 c_0 \eta_0 \rho \qquad \text{in } \Omega_N. \tag{11}$$

However, the two constraints are imposed in wavenumber-dependent form and vanish for $k_0 = 0$. Instability in the LF regime ($k_0 \ll 1$), and non-uniqueness in the static limit ($k_0 = 0$) follow.

5 Low-Frequency Stable Potential Formulations

To overcome the shortcomings of the EFF, the authors presented in [7] a gauged potential formulation that provides the basis for this work. We define a scaled magnetic vector potential $\mathbf{A} \in H_0^{\mathrm{curl}}(\Omega)$ and an electric scalar potential $V \in H_0^1(\Omega)$

by

$$\eta_0 \mu_r \mathbf{H} = \nabla \times \mathbf{A}, \tag{12}$$

$$\mathbf{E} = -\nabla V - \mathrm{i} k_0 \mathbf{A}. \tag{13}$$

We introduce some subspace $\tilde{H}_0^{\mathrm{curl}}(\Omega) \subset H_0^{\mathrm{curl}}(\Omega)$ via the inexact Helmholtz splitting

$$H_0^{\mathrm{curl}}(\Omega) = \tilde{H}_0^{\mathrm{curl}}(\Omega) \oplus \nabla H_0^1(\Omega). \tag{14}$$

In the discrete setting, (14) is realized by a tree-cotree splitting of the FE basis functions of lowest order [8]. If hierarchical FEs with an explicit basis for higher-order gradients [9, 10] are employed, that basis is discarded; see [1].

Equation (14) enables us to represent $\mathbf{A}$ in terms of a reduced potential $\mathbf{A}_c \in \tilde{H}_0^{\mathrm{curl}}(\Omega)$ and an auxiliary scalar potential $\psi \in H_0^1(\Omega)$:

$$\mathbf{A} = \mathbf{A}_c + \nabla \psi \qquad \text{with } \mathbf{A}_c \in \tilde{H}_0^{\mathrm{curl}}(\Omega),\ \psi \in H_0^1(\Omega). \tag{15}$$

Thus,

$$\eta_0 \mu_r \mathbf{H} = \nabla \times \mathbf{A}_c, \tag{16}$$

$$\mathbf{E} = -\nabla V - \mathrm{i} k_0 (\mathbf{A}_c + \nabla \psi). \tag{17}$$

5.1 Boundary Value Problem in Terms of Potentials

In the lossy sub-domain Ω_C we state

$$\nabla \times (\nu_r \nabla \times \mathbf{A}_c) + (\eta_0 \sigma + \mathrm{i} k_0 \varepsilon_r) [\mathrm{i} k_0 (\mathbf{A}_c + \nabla \psi) + \nabla V] = \eta_0 \mathbf{J}_i \quad \text{in } \Omega_C, \tag{18a}$$

$$-\nabla \cdot [(\eta_0 \sigma + \mathrm{i} k_0 \varepsilon_r) (\mathbf{A}_c + \nabla \psi)] = 0 \quad \text{in } \Omega_C. \tag{18b}$$

Therein (18a) represents Ampère's Law (1a), and (18b) is a gauge condition. In the lossless sub-domain Ω_N we employ Ampère's Law (1a), again, and the electric flux balance (1d):

$$\nabla \times (\nu_r \nabla \times \mathbf{A}_c) + \mathrm{i} k_0 \varepsilon_r [\mathrm{i} k_0 (\mathbf{A}_c + \nabla \psi) + \nabla V] = \eta_0 \mathbf{J}_i \quad \text{in } \Omega_N, \tag{19a}$$

$$-\nabla \cdot \{\varepsilon_r [\mathrm{i} k_0 (\mathbf{A}_c + \nabla \psi) + \nabla V]\} = \eta_0 c_0 \rho \quad \text{in } \Omega_N. \tag{19b}$$

A gauge will be imposed in the discrete setting. Note that (19a) implies (19b) for

$k_0 > 0$; see Sect. 5.3. BCs corresponding to (2) are given by

$$\mathbf{A}_c \times \hat{\mathbf{n}} = 0 \qquad \text{on } \partial\Omega, \tag{20a}$$

$$\psi = 0 \qquad \text{on } \partial\Omega, \tag{20b}$$

$$V = 0 \qquad \text{on } \partial\Omega. \tag{20c}$$

Interface conditions will be discussed in Sect. 5.4.

5.2 Weak Formulation in Lossy Sub-Domain

Testing (18a) by $\mathbf{w}_c \in \tilde{H}_0^{\text{curl}}$ and ∇N_ψ, with $N_\psi \in H_0^1$, and (18b) by $N_V \in H_0^1$ yields

$$\int_{\Omega_C} \nabla\times \mathbf{w}_c \cdot (\nu_r \nabla\times \mathbf{A}_c) + \mathbf{w}_c \cdot (\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\,[\mathrm{i}k_0\,(\mathbf{A}_c + \nabla\,\psi) + \nabla\, V]\;\mathrm{d}\,\Omega$$

$$= \eta_0 \int_{\Omega_C} \mathbf{w}_c \cdot \mathbf{J}_i\;\mathrm{d}\,\Omega + \eta_0 \int_{\Gamma} \mathbf{w}_c \cdot (\mathbf{H}\times \hat{\mathbf{n}}_{CN})\;\mathrm{d}\,\Gamma, \tag{21}$$

$$\int_{\Omega_C} \nabla N_\psi \cdot \{(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\,[\mathrm{i}k_0\,(\mathbf{A}_c + \nabla\,\psi) + \nabla\, V]\}\;\mathrm{d}\,\Omega$$

$$= \eta_0 \int_{\Omega_C} \nabla N_\psi \cdot \mathbf{J}_i\;\mathrm{d}\,\Omega - \int_{\Gamma} N_\psi[(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\mathbf{E} + \eta_0\mathbf{J}_i]\cdot\hat{\mathbf{n}}_{CN}\;\mathrm{d}\,\Gamma, \tag{22}$$

$$\int_{\Omega_C} \nabla N_V \cdot [(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\,(\mathbf{A}_c + \nabla\,\psi)]\;\mathrm{d}\,\Omega$$

$$= \int_{\Gamma} N_V\,[(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\,(\mathbf{A}_c + \nabla\,\psi)]\cdot\hat{\mathbf{n}}_{CN}\;\mathrm{d}\,\Gamma. \tag{23}$$

It can be shown that (22) is a weak form of the continuity equation

$$-\nabla\cdot\{(\eta_0\sigma + \mathrm{i}k_0\varepsilon_r)\,[\mathrm{i}k_0\,(\mathbf{A}_c + \nabla\psi) + \nabla V]\} = -\eta_0\nabla\cdot\mathbf{J}_i. \tag{24}$$

5.3 Weak Formulation in Lossless Sub-Domain

Testing (19a) by $\mathbf{w}_c \in \tilde{H}_0^{\text{curl}}$ and ∇N_ψ, with $N_\psi \in H_0^1$, and (19b) by $N_V \in H_0^1$ yields

$$\int_{\Omega_N} \nabla\times \mathbf{w}_c \cdot (\nu_r \nabla\times \mathbf{A}_c) + \mathbf{w}_c \cdot (\mathrm{i}k_0\varepsilon_r)\,[\mathrm{i}k_0\,(\mathbf{A}_c + \nabla\psi) + \nabla V]\;\mathrm{d}\,\Omega$$

$$= \eta_0 \int_{\Omega_N} \mathbf{w}_c \cdot \mathbf{J}_i\;\mathrm{d}\,\Omega + \eta_0 \int_{\Gamma} \mathbf{w}_c \cdot (\mathbf{H}\times \hat{\mathbf{n}}_{NC})\;\mathrm{d}\,\Gamma, \tag{25}$$

$$\mathrm{i}k_0 \int_{\Omega_N} \nabla N_\psi \cdot \varepsilon_r \left[\mathrm{i}k_0 \left(\mathbf{A}_c + \nabla \psi\right) + \nabla V\right] \mathrm{d}\,\Omega = \eta_0 \int_{\Omega_N} \nabla N_\psi \cdot \mathbf{J}_i \,\mathrm{d}\,\Omega - \int_\Gamma N_\psi \left(\mathrm{i}k_0 \varepsilon_r \mathbf{E} + \eta_0 \mathbf{J}_i\right) \cdot \hat{\mathbf{n}}_{NC} \,\mathrm{d}\,\Gamma, \tag{26}$$

$$\int_{\Omega_N} \nabla N_V \cdot \varepsilon_r \left[\mathrm{i}k_0 \left(\mathbf{A}_c + \nabla \psi\right) + \nabla V\right] \mathrm{d}\,\Omega = \eta_0 c_0 \int_{\Omega_N} N_V \rho \,\mathrm{d}\,\Omega - \int_\Gamma N_V (\varepsilon_r \mathbf{E}) \cdot \hat{\mathbf{n}}_{NC} \,\mathrm{d}\,\Gamma. \tag{27}$$

It can be shown that (26) is a weak form of the continuity equation

$$-\mathrm{i}k_0 \nabla \cdot \left\{\varepsilon_r \left[\mathrm{i}k_0 \left(\mathbf{A}_c + \nabla\psi\right) + \nabla V\right]\right\} = \mathrm{i}k_0 \eta_0 c_0 \rho = -\eta_0 \nabla \cdot \mathbf{J}_i. \tag{28}$$

It is apparent that (28) is a wavenumber-scaled version of (19b), in accordance with the lack of a gauge in Ω_N. The fact that (28) vanishes in the static case suggests to replace (26) by a suitable gauge condition, on the FE level.

5.4 Interface Conditions

The interface conditions (3a) and (3c) require that

$$(\mathbf{A}_{c,C} - \mathbf{A}_{c,N}) \times \hat{\mathbf{n}} = 0 \qquad \text{on } \Gamma, \tag{29}$$

$$\psi_C - \psi_N = 0 \qquad \text{on } \Gamma, \tag{30}$$

$$V_C - V_N = 0 \qquad \text{on } \Gamma, \tag{31}$$

which is to be imposed in strong form, by single-valued potentials on the interface. The interface conditions (3b) and (3d) are imposed in weak form, by requiring that the boundary integrals in (21) and (25) as well as in (22) and (26) cancel out.

Finally, we require that the boundary integrals in (23) and (27) cancel. This means, the gauge condition (18b) is supplemented by the constraint

$$\left[\left(\eta_0 \sigma + \mathrm{i}k_0 \varepsilon_r\right)_C \left(\mathbf{A}_c + \nabla\psi\right)_C + \left(\varepsilon_r \mathbf{E}\right)_N\right] \cdot \hat{\mathbf{n}} = 0 \qquad \text{on } \Gamma. \tag{32}$$

5.5 Finite-Element Representation and Gauge

The discrete formulation is obtained by restricting the spaces $\tilde{H}_0^{\mathrm{curl}}(\Omega)$ and $H_0^1(\Omega)$ to finite-dimensional FE spaces [9, 10]. Assuming (complex)-symmetric material

tensors, it can be seen from the weak forms of Sects. 5.2 and 5.3 that the resulting FE matrices will also be complex-symmetric, which can be exploited to reduce memory consumption and compute time. The computationally cheapest choice of gauge in Ω_N is to set all FE coefficients x_ψ associated with ψ basis functions in the interior of Ω_N to zero. In this case (26) still contributes to unknowns on Γ.

6 Numerical Examples

6.1 Partially Filled Cavity

Figure 1a shows a box-shaped cavity, which is half-filled by a lossy dielectric. To compare the LF properties of the EFF and the present approach, the frequency-dependence of the spectral condition number of the system matrix is shown in Fig. 1b. In the frequency range under consideration, the condition number remains almost constant for the new formulation, whereas that of the EFF grows rapidly as the frequency tends to zero. Saturation at $10^{21} \ldots 10^{25}$ is due to numerical noise.

6.2 RLC Circuit

The voltage-driven *RLC* circuit shown in Fig. 2 constitutes our second example. The wires and electrodes are taken to be lossy, whereas all other materials are assumed to be lossless. The field plots of Fig. 3 demonstrate that all relevant physical effects are represented correctly: In the static case, the structure serves as an ideal open circuit.

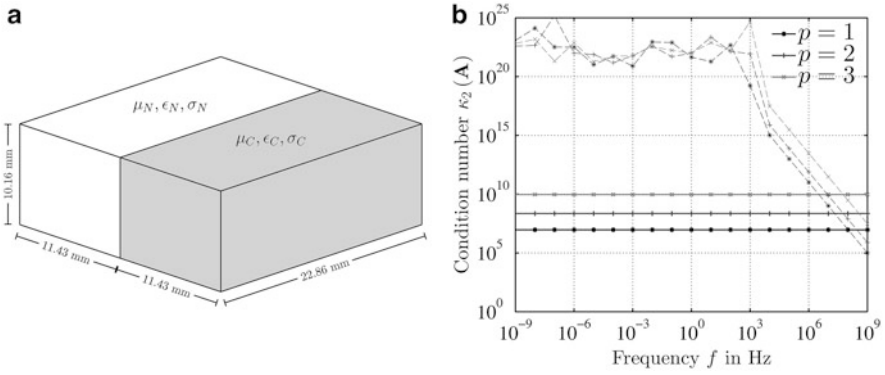


Fig. 1 Half-filled cavity: spectral condition number κ_2 of FE matrix vs. frequency for FE basis functions of different degree p. *dashed line*: **E** method; *solid line*: present approach. Materials: μ_N=μ_C=1, ε_N=ε_C=1, σ_N = 0 S/m, σ_C = 1 S/m. (**a**) Structure. (**b**) Spectral condition number κ_2

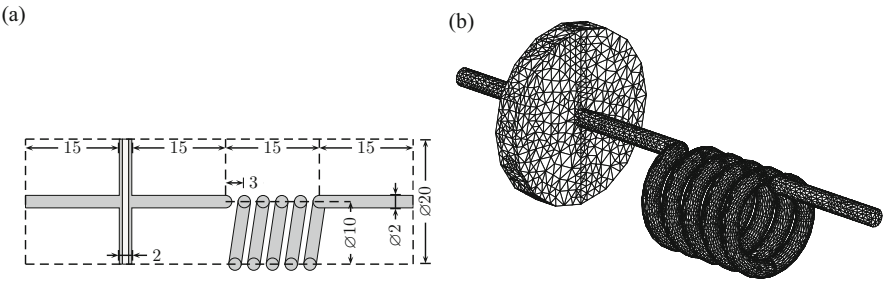


Fig. 2 RLC circuit. (**a**) Structure. Dimensions are in mm. (**b**) Mesh

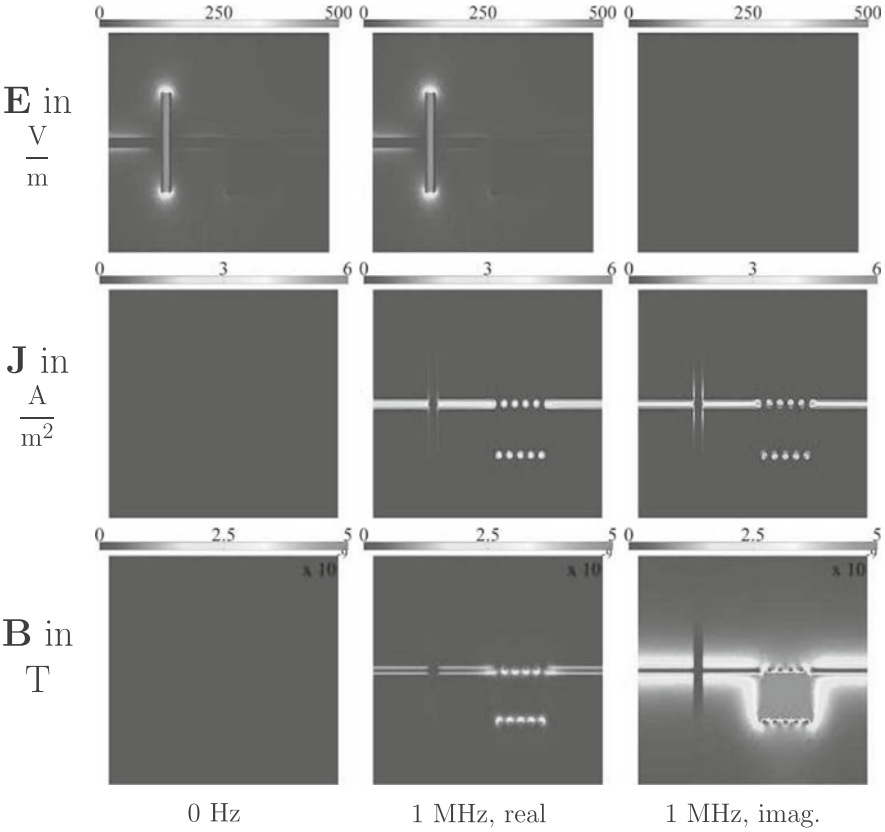


Fig. 3 RLC circuit: field patterns for different operating frequencies

As the frequency rises, significant currents and magnetic fields start to develop. In parallel, the skin and proximity effect become clearly visible in the wires.

References

1. Dyczij-Edlinger, R., Peng, G., Lee, J.-F.: Efficient finite-element solvers for the Maxwell equations in the frequency domain. Comput. Methods Appl. Mech. Eng. **169**(3), 297–309 (1999)
2. Hiptmair, R., Krämer, F., Ostrowski, J.: A robust Maxwell formulation for all frequencies. IEEE Trans. Magn. **44**(6), 682–685 (2008)
3. Ke, H., Hubing, T.H., Maradei, F.: Using the LU recombination method to extend the application of circuit-oriented finite-element methods to arbitrarily low frequencies. IEEE Trans. Microwave Theory Tech. **58**(5), 1189–1195 (2010)
4. Zhu, J., Jiao, D.: A rigorous solution to the low-frequency breakdown in full-wave finite-element-based analysis of general problems involving inhomogeneous lossless/lossy dielectrics and nonideal conductors. IEEE Trans. Microwave Theory Tech. **59**(12), 3294–3306 (2011)
5. Zhu, J., Jiao, D.: Fast full-wave solution that eliminates the low-frequency breakdown problem in a reduced system of order one. IEEE Trans. Compon. Packag. Manuf. Technol. **2**(11), 1871–1881 (2012)
6. Badics, Z., Pávó, J.: Full wave potential formulation with low-frequency stability including ohmic losses. IEEE Trans. Magn. **51**(3), 1–4 (2015). doi: 10.1109/TMAG.2014.2362114
7. Jochum, M., Farle, O., Dyczij-Edlinger, R.: A new low-frequency stable potential formulation for the finite-element simulation of electromagnetic fields. IEEE Trans. Magn. **51**(3), 1–4 (2015). doi: 10.1109/TMAG.2014.2360080
8. Albanese, R., Rubinacci, G.: Solution of three dimensional eddy current problems by integral and differential methods. IEEE Trans. Magn. **24**(1), 98–101 (1988)
9. Webb, J.P.: Hierarchal vector basis functions of arbitrary order for triangular and tetrahedral finite elements. IEEE Trans. Antennas Propag. **47**(8), 1244–1253 (1999)
10. Ingelström, P.: A new set H(curl)-conforming hierarchical basis functions for tetrahedral meshes. IEEE Trans. Microwave Theory Tech. **54**(1), 106–114 (2006)

Local Multiple Traces Formulation for High-Frequency Scattering Problems by Spectral Elements

Carlos Jerez-Hanckes, José Pinto, and Simon Tournier

Abstract We provide a novel ready-to-precondition boundary integral formulation to solve Helmholtz scattering problems by heterogenous penetrable objects in two dimensions exhibiting high-contrast ratios. By weakly imposing transmission conditions and integral representations per subdomain, we are able to devise a robust Galerkin-Petrov formulation employing weighted Chebyshev polynomials. Matrix entries are computed by fast Fourier transforms and regularization techniques. Computational results provided are consistent for large contrast scatterers and frequency sweep as well as efficient Calderón-type preconditioning.

1 Introduction

We consider the modeling of time-harmonic electromagnetic waves scattered by penetrable heterogeneous objects in $\mathbb{R}^2$ using Boundary Integral Equations (BIEs). Specifically, our interest lies in bounded scatterers composed of several subdomains, each of them characterized by constant but distinct wavenumbers exhibiting large ratios. Consequently, for a given excitation frequency, a subdomain may contain a large number of wavelengths, a situation referred to as *medium or medium-high frequency* regimes.[1] Such problems can be encountered in real life as, for instance, when designing phased-array antennae or analyzing electromagnetic compatibility.

From a computational perspective, this frequency regime renders standard low-order implementations either very expensive or simply impractical and one naturally adopts higher order approximations. For homogenous scatterers, several

[1] If the subdomain has a length L_i and waves propagating therein have a wavelength λ_i, we consider situations reaching $L_i/\lambda_i \cong 1000$ with $\max_{ij} \kappa_i/\kappa_j \preceq 100$. The case of very high frequencies will not be considered as it can only be practically resolved by asymptotic methods.

C. Jerez-Hanckes (✉) • J. Pinto • S. Tournier
School of Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile
e-mail: cjerez@ing.puc.cl, jspinto@uc.cl; simon.tournier@alumni.enseeiht.fr.

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_8

approaches based on Fourier-spectral Galerkin discretizations have been proposed with particular emphasis on non-penetrable objects [1, 2, 5, 9].

In the present work, we continue the analysis of the strategy devised in [8] to tackle such problems using the local *Multiple Traces Formulations* (MTFs) [3, 4, 6]. The MTF requires all unknown boundary traces to be defined independently per subdomain while transmission conditions are enforced weakly. In [8], we showed that in the continuous case, the resulting first-kind Fredholm equation possesses unique solutions with a block structure amenable to parallelization and operator preconditioning. Moreover, we provided numerical results for spectral trace discretization and confirmed the strong reduction in GMRes iterations.

We now proceed to recall a few definitions, assess computational costs and discuss the method's robustness for different wavenumbers and frequency sweep.

2 Generalized Local Multiple Traces Formulation

For the sake of brevity, we refer the reader to the notation, definitions and derivations introduced in [6, 8] and consider the geometry depicted in Fig. 1. Under this setting, κ_i are wavenumbers and $\boldsymbol{\gamma}$ denotes joint Dirichlet and Neumann trace operators. The partial differential problem can be cast as follows: to seek u such that for $\mathbf{g}$ being the traces of an incoming plane wave it holds,

$$-\Delta u - \kappa_i^2 u = 0, \qquad \forall\, \mathbf{x} \in \Omega_i, \quad i = 0, 1, 2, \tag{1a}$$

$$[\boldsymbol{\gamma} u] = \mathbf{g}, \qquad \forall\, \mathbf{x} \in \Gamma_{01} \cup \Gamma_{02}, \tag{1b}$$

$$[\boldsymbol{\gamma} u] = \mathbf{0}, \qquad \forall\, \mathbf{x} \in \Gamma_{12}, \tag{1c}$$

$$+ \text{radiation conditions} \qquad \text{when } \|\mathbf{x}\| \to \infty. \tag{1d}$$

where u on Ω_0 represents the scattered wave while inside $\Omega_1 \cup \Omega_2$ u is the total wave.

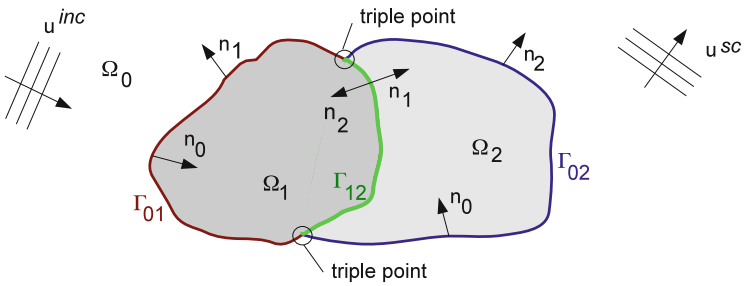


Fig. 1 Simple model geometry. Observe normal definitions

Transmission conditions will be weakly enforced across each interface Γ_{ij}. For this, one introduces *restriction-orientation-and-extension operators* $\widetilde{\mathsf{X}}_{ij}$ which act at each interface by restricting boundary unknowns, orienting Neumann traces and extending by zero onto the adjacent subdomain boundary. This is required to set up a single subdomain boundary equation. On each subdomain, volume Helmholtz equations are replaced by their integral representation counterparts. From here, one derives Calderón identities for each pair of Dirichlet and Neumann trace unknowns $\boldsymbol{\lambda}^i$ by testing with functions defined accordingly. More precisely,

$$\begin{aligned}\langle \boldsymbol{\lambda}^i \, , \, \boldsymbol{\varphi}^i \rangle_\times &= \left\langle \left(\frac{1}{2}\mathrm{Id} + \mathsf{A}_i \right) \boldsymbol{\lambda}^i \, , \, \boldsymbol{\varphi}^i \right\rangle_\times \\ &= \left\langle \begin{pmatrix} \frac{1}{2}\mathrm{Id} - \mathsf{K}_i & \mathsf{V}_i \\ \mathsf{W}_i & \frac{1}{2}\mathrm{Id} + \mathsf{K}_i' \end{pmatrix} \begin{pmatrix} \lambda_{\mathrm{D}}^i \\ \lambda_{\mathrm{N}}^i \end{pmatrix} \, , \, \begin{pmatrix} \varphi_{\mathrm{D}}^i \\ \varphi_{\mathrm{N}}^i \end{pmatrix} \right\rangle_\times , \end{aligned} \tag{2}$$

where the standard boundary integral operators (BIOs) show up and the subscript $\times$ denotes cross-duality products as defined in [6, Sect. 2], [3, Sects. 2.2 &2.3]. Please observe that this spectral form of the local MTF differs slightly from the versions described in [3, 4, 6], in that (2) must be taken over broken spaces for both *test and trial* functions as described in [8, Sect. 2]. With this, we can write down the MTF for two subdomains as follows: find $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2)$ such that for a suitable $\mathbf{g}$ it holds,

$$\langle \mathsf{M}\boldsymbol{\lambda} \, , \, \boldsymbol{\varphi} \rangle = \left\langle \mathsf{M} \begin{pmatrix} \boldsymbol{\lambda}^0 \\ \boldsymbol{\lambda}^1 \\ \boldsymbol{\lambda}^2 \end{pmatrix} \, , \, \begin{pmatrix} \boldsymbol{\varphi}^0 \\ \boldsymbol{\varphi}^1 \\ \boldsymbol{\varphi}^2 \end{pmatrix} \right\rangle_\times = \left\langle \begin{pmatrix} \mathbf{g}^0 \\ \mathbf{g}^1 \\ \mathbf{g}^2 \end{pmatrix} \, , \, \begin{pmatrix} \boldsymbol{\varphi}^0 \\ \boldsymbol{\varphi}^1 \\ \boldsymbol{\varphi}^2 \end{pmatrix} \right\rangle_\times , \tag{3}$$

where

$$\mathsf{M} := \begin{pmatrix} \mathsf{A}_0 & -\frac{1}{2}\widetilde{\mathsf{X}}_{01} & -\frac{1}{2}\widetilde{\mathsf{X}}_{02} \\ -\frac{1}{2}\widetilde{\mathsf{X}}_{10} & \mathsf{A}_1 & -\frac{1}{2}\widetilde{\mathsf{X}}_{12} \\ -\frac{1}{2}\widetilde{\mathsf{X}}_{20} & -\frac{1}{2}\widetilde{\mathsf{X}}_{21} & \mathsf{A}_2 \end{pmatrix} \tag{4}$$

reveals the block structure of the operator and one can show that the system (3) is well-posed [8, Theorem 1].

3 Discretization by Spectral Elements

As described in [8, Sect. 3], trial and test spaces for discretizing (3) are constructed as vectors of piecewise Dirichlet and Neumann functions per interface Γ_{ij}. Specifically, each component of these vector functions is defined as a pullback of first and second kind Chebyshev polynomials defined on the canonical segment $\hat{\Gamma} :=$

$[-1, 1]$, denoted by T_n, U_n, $n \in \mathbb{N}_0$, respectively. Test functions will be multiplied by the weight function $\omega(x) := \sqrt{1-x^2}$ to derive convenient orthogonality relations.

Thus, for the domain Ω_i with interface Γ_{ij} parametrized by the function $h_{ij} : \hat{\Gamma} \to \Gamma_{ij}$, the n-th trial and l-th test vector functions, $\boldsymbol{\lambda}_n^{ij}$ and $\mathbf{q}_l^{ik}$, respectively, are given by the following expressions:

$$\boldsymbol{\lambda}_n^{ij} := (T_n, U_n) \circ (h_{ij})^{-1}, \qquad \mathbf{q}_l^{ik} := (\omega U_l, \omega U_l) \circ (h_{ik})^{-1}.$$

With the above, and after rearranging $\times$-dual products, we can build the Galerkin-Petrov matrix originating from the MTF operator M. If L denotes any of the BIOs, matrix entries take the general structure:

$$I_{\mathsf{L}}^{ijk}[n, l] := \left\langle \mathsf{L}^{ijk} P_n \,,\, \omega U_l \right\rangle = \int_{\hat{\Gamma}} \int_{\hat{\Gamma}} F_{\mathsf{L}}^{ijk}(s,t) P_n(t) \omega(s) U_l(s) ds dt, \tag{5}$$

where F_{L}^{ijk} represents the associated BIO kernel including the mappings required to push the interfaces Γ_{ij} and Γ_{ik} onto $\hat{\Gamma}$, and P_n is either kind of Chebyshev polynomial pushed back to $\hat{\Gamma}$. As described in [8, Sect. 3.3], depending on whether the interfaces Γ_{ij} and Γ_{ik} coincide or not, singular behaviors show up in the integrands. We use standard regularization techniques to extract the singularities [7], i.e. the kernel $F_{\mathsf{L}}^{ijk}(s,t)$ is written as the sum of singular and continuous parts, where the singular part $S_{\mathsf{L}}(s,t)$ corresponds to the Laplacian kernel and has a known analytical decomposition:

$$S_{\mathsf{L}}(s,t) = \sum_{m=0}^{\infty} f_m^{\text{sing}}(t) U_m(s), \tag{6}$$

where the terms $f_m^{\text{sing}}(t)$ are polynomials of order m in t with order coefficients behaving according to the order of the operator, e.g. the coefficients decrease as $\mathcal{O}(1/m)$ and increase as $\mathcal{O}(m)$ for the weakly and hyper- singular operators, respectively. However, when the integration paths—interfaces Γ_{ij} and Γ_{jk}—do not coincide in (5), this term is unnecessary and we set $f_m^{\text{sing}} \equiv 0$ in (6). The continuous kernel $R_{\mathsf{L}}^{ijk} := F_{\mathsf{L}}^{ijk} - S_{\mathsf{L}}$ is approximated via Chebyshev polynomials for a fixed t as a degenerate kernel, i.e.

$$R_{\mathsf{L}}(s,t) \approx \sum_{m=0}^{M_c} f_m^{\text{reg}}(t) U_m(s), \tag{7}$$

using of the FFT to compute coefficients $f_m(t)$ for a suitable choice[2] of $M_c \in \mathbb{N}$. Then, if $f_l := f_l^{\text{sing}} + f_l^{\text{reg}}$, by applying the orthogonality properties of second kind

[2] In [8] we used M to account for the maximum polynomial order and N_c to denote the number of terms considered in the degenerate kernel approximation.

Chebyshev polynomials, i.e.

$$\langle U_l\,,\,\omega U_m\rangle_{\hat{\Gamma}} = \frac{\pi}{2}\delta_{lm},$$

one can quickly reduce the double integration to only one of the form as follows:

$$\begin{aligned} I_{\mathsf{L}}[n,l] &= \int_{\hat{\Gamma}}\int_{\hat{\Gamma}} \left(S_{\mathsf{L}}(s,t) + R_{\mathsf{L}}(s,t)\right) P_n(s)\omega(t)U_l(t)dsdt \\ &\approx \int_{\hat{\Gamma}}\int_{\hat{\Gamma}} \left(\sum_{m=0}^{\infty} f_m^{\text{sing}}(t)U_m(s) + \sum_{m=0}^{M_c} f_m^{\text{reg}}(t)U_m(s)\right) P_n(t)\omega(s)U_l(s)dsdt \\ &= \frac{\pi}{2}\int_{\hat{\Gamma}} \left(f_l^{\text{sing}}(t) + f_l^{\text{reg}}(t)\right) P_n(t)dt = \frac{\pi}{2}\int_{\hat{\Gamma}} f_l(t)P_n(t)dt, \end{aligned} \tag{8}$$

which is finally computed using Gauss-Legendre quadrature.

4 Computational Cost

It is possible to show that the previous scheme has a lower computational cost (in terms of number of operations) than a classical scheme based on pure quadrature for relative large problems.

Let us assume a maximum polynomial or modal order used for all discretized trace spaces with value $N \in \mathbb{N}$. Then, the number of trial and test functions considered is equal to $2N + 2$ per subdomain so that the full matrix size is $(2N + 2)^2 \cdot N_D$ as we have N_D subdomains. For short, the matrix size is then $\mathscr{O}(N_D \cdot N^2)$.

Let us first consider the case of computing the matrix entries via double quadrature. If $I(N)$ denotes the number of operations needed to compute a single integral with number of quadrature points determined by N, the computational cost of the full Petrov-Galerkin matrix is $\mathscr{O}\left(N^2 \cdot I(N)^2\right)$. Since the quadrature scheme will depend linearly on the order of the polynomials used, $I(N)$ is proportional to N, and the cost becomes $\mathscr{O}\left(N_D \cdot N^4\right)$.

Now, for the scheme described in Sect. 3, the main advantage is that all Fourier transforms can be pre-computed, so that for each quadrature point a Fourier transform with M_c points has to be taken. The cost of this pre-computation is $\mathscr{O}\left(N_D \cdot N \cdot M_c \log(M_c)\right)$. After that, each integral of the Petrov-Galerkin discretization is transformed into a single integral, generating a total cost:

$$\mathscr{O}\left(N_D \cdot N^2 \cdot (M_c \log(M_c) + N)\right),$$

which is significantly better than performing a double quadrature even for large M_c. Lastly, we observe that the values of M_c depend strongly on the wavenumbers of

the specific problem as discussed below but the accuracy obtained by the scheme is better than the one provided by full quadrature for similar computation costs.

5 Numerical Results

We now present numerical simulations for the unit circle divided in two halves so that three subdomains are created: $\Omega_0 := \{\mathbf{x} \in \mathbb{R}^2, \|\mathbf{x}\|_2 > 1\}$; $\Omega_1 := \{\mathbf{x} \in \mathbb{R}^2, \|\mathbf{x}\|_2 < 1, x_1 < 0\}$; and, $\Omega_2 := \{\mathbf{x} \in \mathbb{R}^2, \|\mathbf{x}\|_2 < 1, x_1 > 0\}$. This geometry contains all the difficulties portraying Lipschitz domains with sharp corners. As a rule of thumb, the free parameter used in the Chebyshev degenerate kernel expansion we will consider is set to $M_c = 2\text{ceil}(3N + \max_{i=0,1,2} \kappa_i) + 128$.

Depending on the wavenumber values for Ω_1 and Ω_2, we define two general scenarii called *symmetric* ($\kappa_1 = \kappa_2$) and *asymmetric* ($\kappa_1 \neq \kappa_2$). For the symmetric case, the L^2-error is computed using as reference the associated Mie series. For the asymmetric scenario, since no exact solutions are available, we check the L^2-norm of the interface jump relations and also the fulfillment of the weak Calderón identities, i.e. for each domain Ω_i, test functions $\mathbf{q}_l$ and discrete solution $\boldsymbol{\lambda}_N^i$, we measure the residual:

$$\left\langle 2\mathsf{A}_i \boldsymbol{\lambda}_N^i \,,\, \mathbf{q}_l \right\rangle - \left\langle \boldsymbol{\lambda}_N^i \,,\, \mathbf{q}_l \right\rangle .$$

Then, by sweeping in l and choosing the maximum value, we define a measure for the residual error in Calderón identities, referred to as *error in weak Calderón* in Figs. 2 and 3.

5.1 Convergence Analysis

We derive convergence results by assuming fixed frequency and permittivity values. From Mie series analysis and standard spectral methods, we should observe convergence starting at $N_i = 1.4\kappa_i$ with N_i being the number of modes describing the subdomain trace data. This hints at expecting different number of modes per subdomain required to represent physical wavelengths. However, as we will see, this only holds for the symmetric case. Figure 2a, b reveal consistent convergence behaviors for all three error measures, thereby validating the error proxies taken by measuring jump and Calderón conditions. Figure 2a, b show the convergence dependence on the contrast by changing the number of modes required to achieve similar error magnitudes.

Matters change for the asymmetric case (Fig. 2c, d), as the exterior number of modes N_0 required is now related to the highest value of the κ_i.

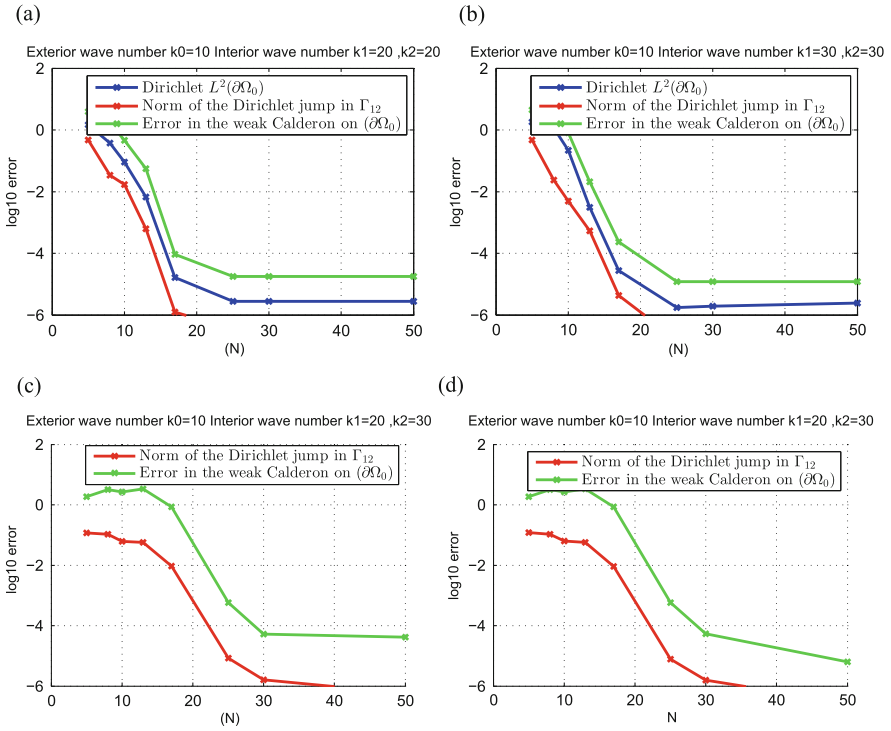


Fig. 2 Convergence analysis results for symmetric and asymmetric cases. (**a**) $k_0 = 10$,$\varepsilon_1 = 2, \varepsilon_2 = 2, N_{1,2} = 1.4\kappa_{1,2}$. (**b**) $k_0 = 10$,$\varepsilon_1 = 3, \varepsilon_2 = 3, N_{1,2} = 1.4\kappa_{1,2}$. (**c**) $k_0 = 10$, $\varepsilon_1 = 2, \varepsilon_2 = 3, N_{1,2} = 1.4\kappa_{1,2}$. (**d**) $k_0 = 10$, $\varepsilon_1 = 2, \varepsilon_2 = 3, N_{1,2} = 1.4 \max\{\kappa_1, \kappa_2\}$

5.2 *Frequency Range Analysis*

Figure 3a portrays errors for different exterior wavenumbers κ_0 and different contrasts using a fixed $N_i = 1.4k_i$ rule. This implies that for $\kappa_0 = 100$, the size of the linear system to solve is 1500×1500 ($\varepsilon_{1,2} = 2$). If the same rule is applied for the asymmetric case, the method fails to provide convergence. In this case, the strategy followed is to use the same maximum amount of modes originated by the largest κ_i. Figure 3b shows positive results for such a strategy. However, a robust rule of thumb is not clear and one can also observe an increase in errors as frequency moves up. We are currently trying to determine whether this is inherent to our method or related to the strategy used to set the number of modes.

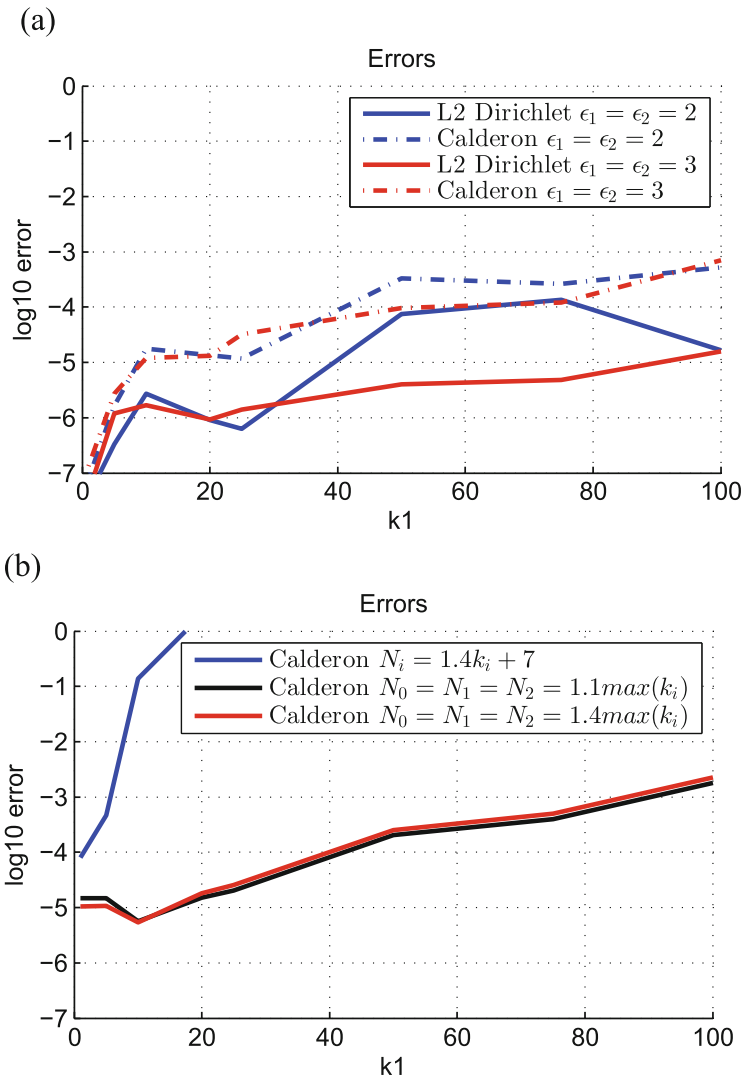


Fig. 3 Frequency sweep results. (**a**) Symmetric case $N_i = 1.4k_i + 7$. (**b**) Asymmetric case $\varepsilon_1 = 2, \varepsilon_2 = 3$

5.3 Krylov Iterative Solver: Built-In Preconditioner

When the wavenumber increases, the Helmholtz equation becomes more indefinite, deteriorating the convergence rate of Krylov subspace iterative solvers and, for our first kind formulation, quickly urging for preconditioning. Calderón identities used to establish the MTF also lead to a built-in preconditioner [6, Sect. 5.5], [3, Sect. 4],

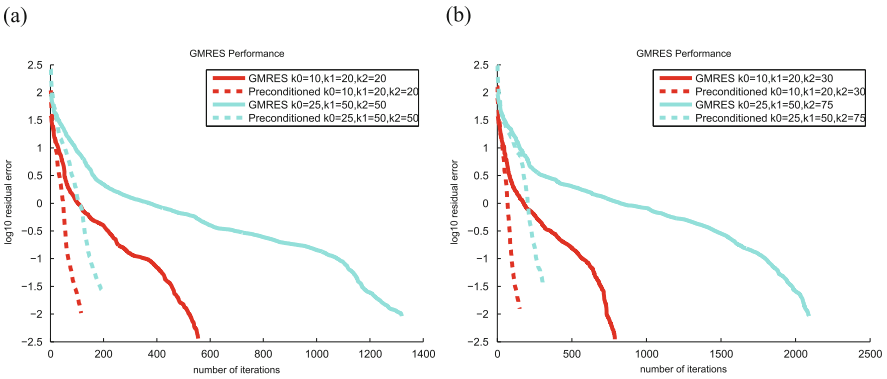


Fig. 4 GMRes residual error computations for symmetric and asymmetric cases. (**a**) Symmetric case. (**b**) Asymmetric case $\varepsilon_1 = 2, \varepsilon_2 = 3$

[8, Sect. 4.5]:

$$\mathbf{A} := \mathrm{diag}(\mathsf{A}_0, \mathsf{A}_1, \mathsf{A}_2), \tag{9}$$

whose discrete form is easily computed. The exact discrete version involves a duality pairing matrix which is block-diagonal with entries computed via orthogonality properties of Chebyshev polynomials. Specifically, we derive diagonal and bi-diagonal blocks according to the Neumann or Dirichlet dual pairing [8, Sect. 4.5]. The preconditioning matrix explicitly reads:

$$\mathsf{P} = \mathsf{M}_{ass}^{-1}\mathbf{A},$$

where $\mathbf{A}$ is the discretization of (9) and M_{ass}^{-1} is the duality pairing matrix. Since the structure of the duality pairing matrix is block diagonal, the numerical extra cost of its inverted matrix-vector product is negligible compared to the dense matrix vector product by $\mathbf{A}$. We use the naive strategy to only factorize by LU the relatively small bi-diagonal blocks, which in practice means a simple reordering, and then apply sparse substitutions at each iteration of GMRes [10] .

Figure 4 shows the convergence history for the homogeneous case ($\kappa_1 = \kappa_2 = 1$) and for the heterogeneous case ($\kappa_1 = 50$, $\kappa_2 = 1$), both comparing GMRes without and within the block diagonal preconditioner at moderate frequency ($\kappa_0 = 10$) and high frequency ($\kappa_0 = 100$).

6 Conclusions and Future Work

We have provided a formulation capable of dealing robustly with large contrasts and high-frequency problems. Numerical results for a simple configuration validate our claims. Further analysis should be carried out to determine optimal parametrization

for the number of modes to be used. Extensions to more general geometries and three-dimensional scatterers are under development.

Acknowledgements This project was partially funded by the Chilean National Science and Technology Commission CONICYT via projects FONDECYT Iniciación 11121166 and ACT 1118.

References

1. Boubendir, Y., Bruno O., Levadoux D., Turc, C.: Integral equations requiring small numbers of Krylov-subspace iterations for two-dimensional smooth penetrable scattering problems. Appl. Numer. Math. **95**, 82–98 (2015)
2. Cai, H.: A fast numerical solution for the first kind boundary integral equation for the Helmholtz equation. BIT Numer. Math. **52**(4), 851–875 (2012)
3. Claeys, X., Hiptmair, R., Jerez-Hanckes, Carlos.: Multitrace boundary integral equations. In: Direct and inverse problems in wave propagation and applications. Radon Series on Computational Applied Mathematics, vol. 14, pp. 51–100. De Gruyter, Berlin (2013)
4. Claeys, X., Hiptmair, R., Jerez-Hanckes, C., Pintarelli, S.: Novel multi-trace boundary integral equations for transmission boundary value problems. In: Fokas, A.S., Pelloni, B. (eds.) Unified Transform for Boundary Value Problems: Applications and Advances. SIAM, Philadelphia (2015)
5. Ganesh, M., Graham, I.G.: A high-order algorithm for obstacle scattering in three dimensions. J. Comput. Phys. **198**(1), 211–242 (2004)
6. Hiptmair, R., Jerez-Hanckes, C.: Multiple traces boundary integral formulation for Helmholtz transmission problems. Adv. Comput. Math. **37**(1), 39–91 (2012)
7. Hu, F.Q.: A spectral boundary integral equation method for the 2-D Helmholtz equation. Comput. Phys. **120**(2), 340–347 (1995)
8. Jerez-Hanckes, C., Pinto, J., Tournier, S.: Local multiple traces formulation for high-frequency scattering problems. J. Comput. Appl. Math. **289**, 306–321 (2015)
9. Nosich, A.I.: The method of analytical regularization in wave-scattering and eigenvalue problems: foundations and review of solutions. IEEE Antennas Propag. Mag. **41**(3), 34–49 (1999)
10. Saad, Y., Schultz, M.H.: GMRES : A generalized minimal residual algorithm for solving non symmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)

Multi-GPU Acceleration of Algebraic Multigrid Preconditioners

Christian Richter, Sebastian Schöps, and Markus Clemens

Abstract A multi-GPU implementation of Krylov subspace methods with an algebraic multigrid preconditioners is proposed. With this, large linear system are solved which result from electrostatic field problems after discretization with the Finite Element Method. As data is distributed across multiple GPUs the resulting impact on memory and execution time are discussed for a given problem solved with either first or second order ansatz functions.

1 Introduction

The solution of partial differential equations as they occur e.g. in electrostatics is of high importance in design and evaluation of virtual prototypes, e.g. of electric high-voltage system components. For this Finite Elements (FE) are very popular in electromagnetics, in particular for static and low frequency field simulations. After applying space and time discretization and possibly a nonlinear solver as e.g. the Newton-Raphson scheme, the resulting problem is a large symmetric positive definite linear algebraic system of equations. For solving these linear systems Krylov subspace method like conjugate gradients (CG) are a common approach [1]. The acceleration of the solution procedure with sophisticated preconditioners, i.e., the algebraic multigrid (AMG) method based on smoothed aggregation with graphic processing units is discussed in this paper.

As multicore systems are standard today, recent research focuses on GPUs as hardware accelerators. Sparse matrix vector (SpMV) operations can be implemented

C. Richter (✉)
University of Wuppertal, Chair of Electromagnetic Theory, 42119 Wuppertal, Germany
e-mail: christian.richter@uni-wuppertal.de

S. Schöps
Graduate School of Computational Engineering Institut für Theorie Elektromagnetischer Felder, Technische Universität Darmstadt, 64285 Darmstadt, Germany
e-mail: schoeps@gsc.ce.tu-darmstadt.de

M. Clemens
Chair of Electromagnetic Theory, Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: clemens@uni-wuppertal.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_9

efficiently on GPUs in general [2] and particularly in applications from electromagnetics [3]. The advantages of GPU acceleration have been demonstrated for Finite Differences [4] as well as FE [5, 6]. The major bottleneck of a GPU is its limited local memory that determines the maximum size of a problem that can be solved at once without swapping data between the GPU and the host memory which usually has a serious impact on the performance. Consequently, when it comes to problems exceeding the memory size of a single GPU the use of multiple GPUs becomes mandatory. But even for smaller problems the CG method can be accelerated by using multiple GPUs [7].

The results in this paper extend those reported in [8]. While previously the proposed addon of a multi-GPU-AMG solver to the CUSP library was presented for the first time, in this paper the results for second order ansatz functions and larger problems, exceeding the memory of one GPU, are discussed. Second order ansatz functions change the density of the matrix by increasing the number of non-zero matrix entries per degree of freedom. With the larger discrete problem exceeding one GPUs' memory it is shown that the code can solve large problems not only in theory, but in a real-world example.

The paper is structured as follows: first the problem formulation is introduced. In Sect. 2 the multi-GPU implementation is described in detail. A numerical example shows the effects when taking into account multiple GPUs for solving electromagnetic problems with either first or second order ansatz functions. In the end the work is summarized.

1.1 Problem Formulation

When solving an electrostatic problem, an elliptic boundary value problem has to be solved on a computational domain Ω, i.e.,

$$-\nabla \cdot (\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r})) = f(\mathbf{r}) \tag{1}$$

for $\mathbf{r} \in \Omega$ and where ε is the spatially distributed permittivity, f a given field source, ϕ the electric scalar potential with adequate boundary conditions, like a homogeneous Dirichlet constraint $\phi|_{\partial\Omega} = 0$. Discretizing the problem with FE results in a linear system of equations with a positive definite system matrix. The in-house simulation code 'MEQSICO' [9] is used that is capable of solving static and quasi-static electric and magnetic field problems and coupled multiphysical problems with high-order FEM ansatz functions [10]. An exemplary electrostatic problem is shown in Fig. 1 which is described by (1).

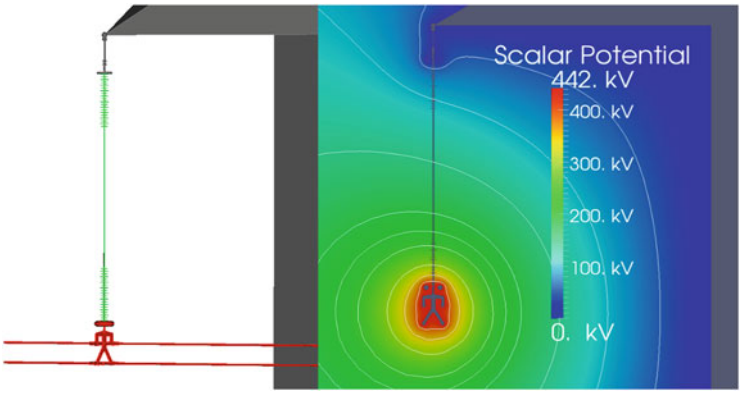


Fig. 1 CAD model and scalar potential of a high-voltage-insulator presented in [11]

1.2 Algebraic Multigrid

The AMG method is used as a preconditioner for a conjugate gradient solver [12]. Apart from classical AMG it can be based on smoothed aggregation [13] as employed in this paper. AMG consists of two parts: in the *setup-phase*, levels of increasing coarseness are assembled from the degrees of freedom. To enable the different grid levels to interact the prolongation and restriction operators are constructed, which connect two consecutive levels. With the so-called Galerkin product, a triple matrix product, the system matrix of the coarser level is constructed.

The multigrid preconditioner is applied in every iteration step within the *solve-phase* of the CG method. At first the given linear system is subject to a smoothing step and the afterwards computed residual is restricted to the next coarse level. Within this next level the described function is called recursively. After returning from the next coarse level, the result is corrected by an error calculated on the coarser grid level and a smoothing step is applied again. Instead of calling further recursive calls the system is solve on the coarsest level. The coarsest system is solved by direct or iterative solvers. Details of this V-cycle approach are given in [14, 15].

While the complex calculations of the setup phase are performed only once for a given system matrix, the solve phase is executed in every iteration step. Consisting only of some SpMVs the multigrid V-cycle in the solve phase is less time intense. These operations can be performed very efficiently on GPUs [2].

2 Algebraic Multigrid on Multiple GPUs

The CUSP library [16] is a well-established and fast library for solving linear equation systems on GPUs. Providing efficient implementations for matrix-vector operations and a set of solvers including an AMG-Preconditioner [17] it is well suited as a starting point for our multi-GPU approach. The AMG preconditioner [17] has a high memory demand due to the multiple grids, each storing its own system matrix as well as matrices for restriction and prolongation. To overcome the limitations we propose to distribute the data across multiple GPUs. As CUSP is an evolving software library the decision was made to implement a multi-GPU extension as an add-on to this library. The environment uses templated C++ classes and can interact with CUSP. Main parts of the addon are classes for multi-GPU vectors and matrices, communication routines for data exchange between the GPUs and a multi-GPU PCG solver with an AMG-preconditioner that solves the system on multiple GPUs in parallel.

2.1 Multi-GPU Datatypes

The major part of memory is spent on the storage of matrices. Therefore redundancy has to be avoided. To achieve this the matrices are split up in a row wise manner and the resulting parts are copied to the individual GPUs. During the splitting process the input matrix is converted into the Compressed Sparse Row (CSR) format, split up and the resulting parts are reconverted. Due to this the splitup is performed fast with almost no calculation effort and the resulting parts are load balanced as the entries per row remain unchanged supposing that each row has approximately the same number of non-zero elements. With respect to the construction of the multi-GPU matrix class and due to the fact that vectors have only low memory demand compared to a matrix, the vector class holds a full copy of every vector on each GPU. The vector part corresponding to the GPU is defined on the full vector via a vector view, i.e., a kind of pointer. When performing an operation it is executed simultaneously on all GPUs using OpenMP. During a vector-vector operation each GPU performs the operation only on the corresponding vector parts. A matrix-vector operation is performed on every GPU using the CUSP SpMV routine with the full vector as input and the corresponding vector view as output. Therefore it is important that the whole input vector is up to date on each GPU. This has to be ensured by communication routines.

2.2 Inter-GPU Communication

The exchange of data between the GPUs is the most critical part of the implementation. Sparse operations on a single GPU are already limited by the bandwidth of

the connection between the GPUs global memory and its processing unit. When it comes to inter-GPU communication there is a large gap between the GPUs internal bandwidth and the connecting Peripheral Component Interconnect Express (PCIe) bandwidth. A contemporary GPU like the Nvidia Tesla K20X as used in this work has a theoretical bandwidth of 250 GB/s. The PCIe bus which connects the GPUs has a bandwidth of 8 GB/s. Therefore sophisticated communication schemes have been developed to minimize the burden of communication. With the *copy-1n* routine a whole vector is distributed from one GPU to all GPUs involved in the calculation. With "direct access"[1] data can be copied from one GPU to another without going through the host. Due to this and by using asynchronous memory copy functions data can be copied between different GPUs in parallel. With these measures the bandwidth scales linearly with the number of parallel processes. As a result the vector is not copied to each GPU one after another but instead it is realized as follows: the first part of the vector is copied from the first GPU to the second one, from the second to the third and so on. In the meantime the second memory segment is transferred from the first to the second GPU in parallel. This can be realized because of the two copy engines of contemporary GPUs enabling them to send and receive data at the same time. The most important routine is the *gather-nn* routine. It is used when each GPU holds only its own piece of data and a vector has to be updated such that every vector holds a up-to-date version of the whole vector. This is the case before every SpMV. Here the same principle is used as described above. Each GPU copies its piece data to the next GPU in the cycle. When the transmission is finished the GPU sends the next piece of data it has just received to the next GPU. In this way each GPU sends and receives data during the whole procedure maximizing the parallel throughput.

2.3 Preconditioned Conjugate Gradients

The AMG preconditioner is set up by the CUSP library. As shown in [18] it should be set up on the host to overcome memory limitations of a single GPU. Within the multi-GPU splitup routine, the preconditioner is divided level by level and distributed across all GPUs involved. The multi-GPU CG routine then solves the problem by handing over the multi-GPU preconditioner and the original right hand side and solution vectors. The routine is build analogously to the CUSP AMG-CG, but uses multi-GPU routines. Due to the implementation only minimal changes are necessary in an existing code and the behavior in terms of residual reduction per step is almost identical.

[1]This technology was introduced with CUDA 4.0 within the framework of a "unified address space", i.e., a virtual address space for the host and all GPUs attached.

3 Numerical Example

As an example a real-world FE problem is solved using first and second order ansatz functions. The example, a high voltage insulator as is presented in [11], is shown in Fig. 1. The discrete model has 1.5×10^6 degrees of freedom and a linear system matrix consisting of 21×10^6 nonzero entries. When using second order ansatz functions the linear problem expands to 12×10^6 dof and 340×10^6 nonzero matrix entries. In both cases the problem is solved to a relative residual norm of 1×10^{-12}. Calculations are performed on a server running CentOS 6.5. It is equipped with two Intel Xeon E5-2670 CPUs and 128 GB RAM. Four Nvidia Tesla K20Xm GPUs are attached to the host. To ensure data integrity, error-correcting code (ECC) is enabled. Thus the effective bandwidth of each GPU is reduced from 250 to 200 GB/s. Host parallelization is done by OpenMP on the host. On the devices architecture model 3.5 is used. The code is compiled with CUDA 5.0, Thrust 1.8.0, CUSP 0.4.0 and GCC 4.4.7 with -O3. For comparison the problem is also solved on the host using the CUSP host version which has been shown to outperform [18] state of the art libraries like PetSc [19] or Trilinos ML [20]. The setup phase is performed on the host where it is stored and distributed to the GPUs. The preconditioner is only setup once and can be reused for multiple right hand sides, i.e., for several timesteps in a quasistatic simulation. The speedup of the GPU implementations over the host implementation is depicted in Fig. 2. It shows the individual speedup when solving the problem with first and second order ansatz functions for a varying number of GPUs. The problem cannot be solved on one GPU with second order ansatz functions due to memory limitations and therefore no results are presented. One can see that with first order ansatz functions a speedup of 7.7 times is achieved when using one GPU. It can be increased to a factor 10.8 with two GPUs but decreases to 9.5 when using four GPUs. This has two reasons: firstly the communication processes become so costly that the speedup in calculation cannot compensate for them. The data per GPU is not sufficient to keep the GPUs busy. Secondly every calculation and data movement operation needs a certain fixed time to be launched. This has a higher effect when the time to perform the operation is lower. The second order case differs from the first order one. The speedup is again more than doubled compared to the

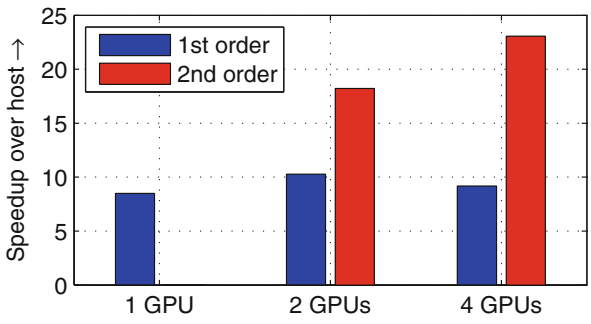


Fig. 2 Speedup of the solve-phase for the first- and second-order-problem on a varying number of GPUs

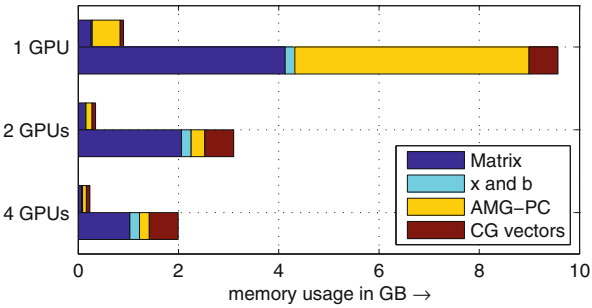


Fig. 3 Memory-usage per GPU for the first- (*upper bar*) and second-order-problem (*lower bar*) on a varying number of GPUs

first order case. It is increased to 18.1 on two GPUs and over 23 when using four GPUs. This speedup over the first order GPU calculation has two major reasons. At first the time the GPUs spend for calculations is larger because of the increased work each GPU has to do. This minimizes launch effects. Then the matrix itself is much denser with second order ansatz functions with an average of over 28.2 instead of 14.7 entries per row of the system matrix. This means that the calculations increase compared to the number of degrees of freedom, which are transferred at every individual communicational operation.

Figure 3 shows the memory consumption for the given problem. It is separated by the order of the ansatz functions and number of GPUs in use. Each bar is the sum of the individual parts of the CG-solver. The use of second order ansatz functions is shown to lead to a much higher memory demand. With the number of GPUs involved the memory demand per GPU of the matrices decreases linearly because they are split up and no information is saved redundantly. In contrast to this the memory demand for a vector remains unchanged because each GPU has to hold the full vector. Since the matrices memory demand is dominating the overall scaling remains almost linear. Another reduction of memory demand can be achieved by erasing redundancies between the matrix and the AMG preconditioner. In CUSP the system matrix is saved redundantly in the preconditioner as the system matrix for the finest level. As this is not needed in the proposed addon a further reduction is obtained.

4 Conclusion

The limitations of a single GPU can be overcome by using multiple GPUs for solving high dimensional discrete electric or magnetic field problems. This has been shown with an addon to the CUSP library that enables multi-GPU computing for FE simulations with first and second order ansatz functions. Memory consumption scales approximately linear with the number of used GPUs. Furthermore, significant speedups can be achieved by multiple GPUs, even though inter-GPU communi-

cation has to be taken into account. Especially higher order simulations can be accelerated significantly due to the higher density of their system matrices.

References

1. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM, Boston (2003)
2. Bell, N., Garland, M.: Efficient sparse matrix-vector multiplication on CUDA, NVIDIA Corporation, NVIDIA Technical Report NVR-2008-004 (2008)
3. Mehri Dehnavi, M., Fernández, D.M., Giannacopoulos, D.: Finite-Element sparse matrix vector multiplication on graphic processing units. IEEE Trans. Magn. **46**(8), 2982–2985 (2010)
4. Richter, C., Schöps, S., Clemens, M.: GPU acceleration of finite differences in coupled electromagnetic/thermal simulations. IEEE Trans. Magn. **49**(5), 1649–1652 (2013)
5. Mehri Dehnavi, M., Fernández, D.M., Giannacopoulos, D.: Enhancing the performance of conjugate gradient solvers on graphic processing units. IEEE Trans. Magn. **47**(5), 1162–1165 (2011)
6. Mehri Dehnavi, M., Fernández, D.M., Gaudiot, J.-L.: Parallel sparse approximate inverse preconditioning on graphic processing units. IEEE Trans. Parallel Distrib. Syst. **24**(9), 1852–1862 (2013)
7. Verschoor, M., Jalba, A.C.: Analysis and performance estimation of the conjugate gradient method on multiple GPUs. Parallel Comput. **38**(10–11), 552–575 (2012)
8. C. Richter; S. Schöps; M. Clemens Multi-GPU acceleration of algebraic multigrid preconditioners for elliptic field problems, IEEE Trans. Magn., **51**(3), 1–4 (2015). DOI: 10.1109/TMAG.2014.2357332
9. Steinmetz, T., Helias, M., Wimmer, G., et al.: Electro-quasistatic field simulations based on a discrete electromagnetism formulation. IEEE Trans. Magn. **42**(4), 755–758 (2006)
10. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford University Press, Oxford (2003)
11. Ye, H., Clemens, M., Seifert, J.: Electro-quasistatic field simulation for the layout optimization of outdoor insulation using microvaristor material. IEEE Trans. Magn. **49**(5), 1709–1712 (2013)
12. Stüben, K.: Algebraic multigrid (AMG): an introduction with applications, GMD, Report 53 (1999)
13. Vanek, P., Mandel, J., Bresina, M.: Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems. Computing **56**, 179–196 (1996)
14. Shapira, Y.: Matrix-Based Multigrid: Theory and Applications. Numerical Methods and Algorithms. Springer, New York (2008)
15. Trottenberg, U., Oosterlee, C., Schuller, A.: Multigrid. Academic, New York (2001)
16. Bell, N., Garland, M.: CUSP: generic parallel algorithms for sparse matrix and graph computations, version 0.4.0. (2012)
17. Bell, N., Dalton, S., Olson, L.N.: Exposing fine-grained parallelism in algebraic multigrid methods. SIAM J. Sci. Comput. **34**(4), C123–C152 (2012)
18. Richter, C., Schöps, S., Clemens, M.: GPU acceleration of algebraic multigrid preconditioners for discrete elliptic field problems. IEEE Trans. Magn. **50**(2), 461–464 (2014)
19. Balay, S., Brown, J., Buschelman, K., et al.: PETSc users manual, Argonne National Laboratory, Technical Report ANL-95/11 - Review 3.4, (2013)
20. Gee, M., Siefert, C., et al.: ML 5.0 smoothed aggregation user's guide, Sandia National Laboratories, Technical Report SAND2006-2649 (2006)

On Several Green's Function Methods for Fast Poisson Solver in Free Space

Dawei Zheng and Ursula van Rienen

Abstract We summarize four closely related numerical solution methods for Poisson's equation in free space: Green's function method, integrated Green's function method, reduced integrated Green's function method, and cutting integrated Green's function method. A new and final routine called cutting reduced Green's function method is carried out as well. These methods can be used for different practical problems to accelerate the calculation. Numerical examples are also given to compare the introduced methods.

1 Introduction

Poisson's equation is broadly used in many areas, such as electrostatics, mechanical engineering and theoretical physics—for instance in gravitational potential calculation and in beam dynamics simulations in particle accelerators. Particle accelerators have a long history. In fact, the first basics go back to Crookes who discovered cathode rays (1870), Thompson who showed that the cathode rays are composed of electrons (1896) and Röntgen who discovered X-rays (1895). The first milestone on the path to particle accelerators for high energy physics was Rutherford's scattering of alpha particles on a gold foil (1909). Modern accelerators for high energy physics still basically use the principle of scattering experiments. Since the energy of the electrons in cathodic ray tubes is limited, in the 1930s new types of generators for higher electric fields have been developed. Examples are the Van de Graff generator (1929) and the Cockcroft-Walton generator (1932) or the first cyclotron by Lawrence (1931). To overcome the limitations of these machines and achieve much higher energies of the electrons, radio-frequency (RF) cavities started to be used (and now are key elements of all accelerators) in which the energy of the high-frequency field is transferred to the passing electrons (or other elementary particles). Accelerators for high energy physics are either ring-like machines such as the Large Hadron Collider (LHC) at CERN or linear accelerators such as the

D. Zheng (✉) • U. van Rienen
Institute of General Electrical Engineering, University of Rostock, Albert-Einstein-Str. 2, 18059 Rostock, Germany
e-mail: dawei.zheng@uni-rostock.de; ursula.van-rienen@uni-rostock.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_10

design study for the International Linear Collider (ILC). The elementary particles are highly relativistic, i.e. have practically speed of light, and have very high energies. Synchrotron light sources, which are used for material studies, exploit the electromagnetic radiation which arises when an electron is forced (by magnets) on a curved path. Elaborating this principle more and more, new generations of brilliant light sources have been designed such as the European X-ray free-electron laser (XFEL) which is currently being built at DESY in Hamburg. Further, accelerators are used in medicine for cancer therapy.

No matter which type of accelerator is regarded, all of them start with a particle source where the elementary particles are produced (e.g. cathode, photocathode, ion source), some magnetic focusing elements and sections in which the stream of particles is bunched and a first acceleration takes place. Thus, a bunch is a large number of charged elementary particles. It achieves its relativistic speed and its higher and higher energy passing through RF cavities.

As long as the particles are non-relativistic, their self-electric space charge field is influencing the particles in the bunch while space charge fields don't play a role anymore for the highly relativistic particles. The space charge effect is of crucial importance for the next generation accelerators with their ultrashort, very dense bunches of high power, such as in the XFEL, since this naturally implies higher space charge effects. If one wouldn't do a careful design study, one possible effect could be e.g. that the bunch, which indeed should stay in tight dimensions, extremely grows due to the space charge effect and hits the wall of the vacuum chamber. The specific bunch characteristics of future accelerators makes simulation studies of space charge effects more challenging than before.

The most prominent, classical methodology for numerical space charge studies is known as the Particle-in-Cell (PIC) model [1]. The considered bunch is embedded in the computational domain Ω, which usually is a cubic or a cylindrical domain. The computational domain Ω is discretized and the charge of the particles inside each cell is assigned to neighboring grid points by algorithms like the Nearest Grid Point (NGP) or the Cloud in Cell (CIC) schemes. Note, that so-called macro-particles are introduced in order to achieve a computational load which is still manageable. Then, the space charge has to be calculated, applied to the (macro-)particles and the equation of motion has to be solved. A usual procedure is to use the Lorentz transformation in each time step to transfer between the laboratory system and the rest frame (of special relativity) and then compute the space charge fields in the rest frame. The self-electric field can be derived by solving Poisson's equation (in the rest frame). It is transferred back to the laboratory system by the Lorentz transformation with the Lorentz factor γ.

In this contribution, we concentrate on the efficient solution of Poisson's equation:

$$\left(\frac{\mathrm{d}^2\varphi(x,y,z)}{\mathrm{d}x^2}+\frac{\mathrm{d}^2\varphi(x,y,z)}{\mathrm{d}y^2}+\frac{\mathrm{d}^2\varphi(x,y,z)}{\mathrm{d}z^2}\right)=-\frac{\rho(x,y,z)}{\varepsilon_0},\quad (x,y,z)\in\Omega$$

where $\rho(x, y, z)$ is the charge density, ε_0 is the permittivity of vacuum and $\varphi(x, y, z)$ is the electrostatic potential, i.e. we study this problem in Cartesian coordinates in 3D. Free space boundary (or some say open boundary) conditions are regarded. Although this is not true in the real accelerator, this consideration is well-introduced and most common in the simulation of space charge effects as long as the bunch is far enough apart from the walls of the enclosing vacuum tube. The common way to solve this equation is to convolute the density of charged particles and the Green's function in free space, known as the Green's function method. However, in some cases such as a very long cigar-shape or short pancake-shape bunch the numerical calculation may suffer from errors. The so-called integrated Green's function (IGF) [2, 3] has especially been invented for such issues. It deals with an analytical integration rather than a numerical integration. However, the computation is rather involved and time-consuming and thus calls for an improvement to higher efficiency.

We present some appropriate methods, as accurate as the IGF method yet costing less CPU time for different practical problems. In general, the reduced integrated Green's function (RIGF) method is suitable for all problems applying the IGF method—for instance the near-bunch field calculation. In contrast, the cutting (integrated) Green's function (CIGF) method is only advantageous for far-bunch field calculation. A further new method, denoted as cutting reduced integrated Green's function (CRIGF) method can accelerate the calculations even more. This routine can also be used in other Poisson solver codes to improve efficiency.

2 GF, IGF and RIGF Integral for Poisson's Equation

The Green's function-type methods are often-used methods to solve Poisson's equation in free space, i.e. with open boundaries. The Green's function is given as:

$$G(x, x', y, y', z, z') = \frac{1}{\sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}}. \tag{1}$$

Using the Green's function, the solution of Poisson's equation in $\mathbf{R}^3$, i.e. the continuous electrostatic potential φ, reads as [1, 2]:

$$\varphi(x, y, z) = \frac{1}{4\pi\varepsilon_0} \cdot \int\int\int \rho(x', y', z') G(x, x', y, y', z, z') \mathrm{d}x' \mathrm{d}y' \mathrm{d}z'. \tag{2}$$

Now, regard a cubic computational domain Ω which is discretized by N_x, N_y and N_z steps, respectively, in each coordinate direction with equidistant step sizes h_x, h_y, h_z. Then, the discrete integral formula is given by

$$\varphi(x_i, y_j, z_k) \approx \frac{1}{4\pi\varepsilon_0} \cdot \sum_{i'=1}^{N_x} \sum_{j'=1}^{N_y} \sum_{k'=1}^{N_z} \rho(x_{i'}, y_{j'}, z_{k'}) \tilde{G}(x_i, x_{i'}, y_j, y_{j'}, z_k, z_{k'}), \tag{3}$$

where the grid points (x_i, y_j, z_k) are the center points of each integral. The integral cell is equal to the individual grid cells with side lengths h_x, h_y and h_z. Thus, the integral over one grid cell reads as:

$$\tilde{G}(x_i, x_{i'}, y_j, y_{j'}, z_k, z_{k'}) = \int_{x_i-h_x/2}^{x_i+h_x/2} \int_{y_j-h_y/2}^{y_j+h_y/2} \int_{z_k-h_z/2}^{z_k+h_z/2} G(x_i, x', y_j, y', z_k, z') \mathrm{d}x' \mathrm{d}y' \mathrm{d}z'. \tag{4}$$

In the following, for the different kinds of integrals, $\tilde{G}$ will be specified by different subscripts. Further, regarding the calculation of the Green's function values we will apply a coordinate translation, substituting $w - w'$ by w for w in $\{x, y, z\}$ and thus use w instead of $w - w'$. If we apply the midpoint rule, the numerical integral is known as GF integral:

$$\tilde{G}_{GF}(x_i, y_j, z_k) = h_x h_y h_z G(x_i, y_j, z_k). \tag{5}$$

In many applications, the midpoint rule can readily be used. Yet, often higher accuracy is needed. This can be achieved by higher order numerical integration rules or by the IGF integral:

$$\begin{aligned}
\tilde{G}_{IGF}(x_i, y_j, z_k) &= \int_{x_i-h_x/2}^{x_i+h_x/2} \int_{y_j-h_y/2}^{y_j+h_y/2} \int_{z_k-h_z/2}^{z_k+h_z/2} G(x', y', z') \mathrm{d}x' \mathrm{d}y' \mathrm{d}z' \\
&= IGF(x_i + \frac{h_x}{2}, y_j + \frac{h_y}{2}, z_k + \frac{h_z}{2}) - IGF(x_i + \frac{h_x}{2}, y_j + \frac{h_y}{2}, z_k - \frac{h_z}{2}) \\
&\quad - IGF(x_i + \frac{h_x}{2}, y_j - \frac{h_y}{2}, z_k + \frac{h_z}{2}) - IGF(x_i - \frac{h_x}{2}, y_j + \frac{h_y}{2}, z_k + \frac{h_z}{2}) \\
&\quad + IGF(x_i - \frac{h_x}{2}, y_j - \frac{h_y}{2}, z_k + \frac{h_z}{2}) + IGF(x_i + \frac{h_x}{2}, y_j - \frac{h_y}{2}, z_k - \frac{h_z}{2}) \\
&\quad + IGF(x_i - \frac{h_x}{2}, y_j + \frac{h_y}{2}, z_k - \frac{h_z}{2}) - IGF(x_i - \frac{h_x}{2}, y_j - \frac{h_y}{2}, z_k - \frac{h_z}{2}),
\end{aligned} \tag{6}$$

where the $IGF(x, y, z)$ function is the primitive function (antiderivative) of (1), which can be expressed as:

$$\begin{aligned}
IGF(x, y, z) \doteq \int\int\int \frac{1}{\sqrt{x^2 + y^2 + z^2}} \mathrm{d}x\mathrm{d}y\mathrm{d}z &= -\frac{z^2}{2} \arctan(\frac{xy}{z\sqrt{x^2 + y^2 + z^2}}) \\
-\frac{y^2}{2} \arctan(\frac{xz}{y\sqrt{x^2 + y^2 + z^2}}) - \frac{x^2}{2} \arctan(\frac{yz}{x\sqrt{x^2 + y^2 + z^2}}) &+ yz \ln(x \\
+ \sqrt{x^2 + y^2 + z^2}) + xz \ln(y + \sqrt{x^2 + y^2 + z^2}) + xy \ln(z + \sqrt{x^2 + y^2 + z^2}).&
\end{aligned} \tag{7}$$

Here, we present the simple form from [3].

RIGF integral: In order to figure out the improvement of the IGF integral compared to the GF integral, we define the local Green's function integral relative fraction as: $\eta_G(x_i, y_j, z_k) = \left|\delta\tilde{G}_{GF}(x_i, y_j, z_k)/\tilde{G}_{IGF}(x_i, y_j, z_k)\right|$, where $\delta\tilde{G}_k = \|\tilde{G}_{IGF_k} - \tilde{G}_{GF_k}\|$. To evaluate the variation of $\eta_G(x_i, y_j, z_k)$ visually and easily in the grid, we chose a computational domain with a large aspect ratio: $L_x : L_y : L_z = 1 : 1 : 30$, where L_x, L_y L_z are the edge lengths of the cubical domain Ω. It is discretized by $32 \times 32 \times 32 = 32{,}768$ grid points (In calculation, Green's function needs one more point on each axis, i.e. $33 \times 33 \times 33 = 35{,}937$ [2]). In Fig. 1 (left), we use a boxplot of $\eta G(:, :, z_k)$. Each column corresponds to one slice of index k. We can observe that the local relative errors exponentially decrease with an increasing value of k (z_k). Only in the very first slices, the errors are large and strongly varying. For increasing k, the errors inside a slice and compared with the neighbor slices errors coincide more and more.

The motivation of the RIGF integral is relatively natural and simple. In the calculation of $\tilde{G}(x_i, y_j, z_k)$, the IGF integral $\tilde{G}_{IGF}(x_i, y_j, z_k)$ has higher complexity than the numerical GF integral $\tilde{G}_{GF}(x_i, y_j, z_k)$, i.e. for each $\tilde{G}_{IGF}(x_i, y_j, z_k)$ we have to calculate eight terms in (6) and every term should be calculated by (7). Yet, $\tilde{G}_{GF}(x_i, y_j, z_k)$ has just one simple term, which is also faster to be evaluated. Thus, we calculate the $\tilde{G}_{IGF}(x_i, y_j, z_k)$ by the exact integral only over those grid cells where it is necessary and everywhere else we replace it by the numerical integral $\tilde{G}_{GF}(x_i, y_j, z_k)$. Practically, this means that only the near-origin parts, where the bunch is located, are treated by the IGF model. The remaining parts of the integral are calculated by the simpler standard GF model. We determine integer parameters (R_x, R_y, R_z) indicating at which grid line to switch from the IGF model to the GF model (see Fig. 1 (right) blue line between Ω_{IGF} and Ω_{GF}). For the following, we suppose that the large

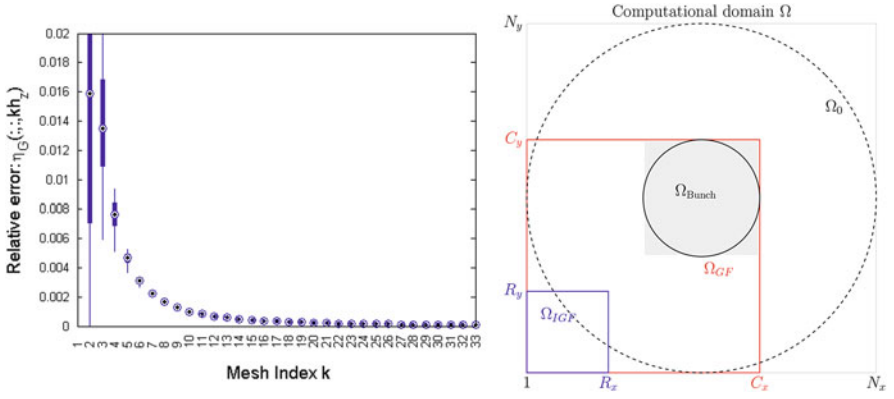


Fig. 1 (*Left*): The local relative error of the GF integral. (*Right*): A schematic plot of cutting Green's function domain for Cartesian coordinates

aspect ratio direction is on the z-axis. Then the new integral reads as follows:

$$\tilde{G}_{RIGF}(x_i, y_j, z_k) = \begin{cases} \tilde{G}_{IGF}(x_i, y_j, z_k), & (1,1,1) \le (i,j,k) \le (R_x, R_y, R_z); \\ \tilde{G}_{GF}(x_i, y_j, z_k), & \text{otherwise;} \end{cases}$$

It has been investigated how these parameters (R_x, R_y, R_z) should be chosen. There are two general key aspects which should be balanced in the chosen strategy: the computational time and the achieved accuracy.

With respect to the computational time, it would be an option to choose $R_w = (N_w + 1)/s_w$ for w in $\{x, y, z\}$. The larger value of s, the less computational time is needed by the IGF calculation. Regarding the cigar-shape bunch as an example, it is reasonable to choose $R_w = N_w + 1$ for w in $\{x, y\}$. Then, the computational time depends linearly on s which ranges from 1 (IGF routine) to $N_z + 1$ (GF routine):

$$t_{RIGF} = \frac{N_z + 1 - R_z}{N_z + 1} t_{GF} + \frac{R_z}{N_z + 1} t_{IGF} = \frac{R_z}{N_z + 1}(t_{IGF} - t_{GF}) + t_{GF}. \tag{8}$$

On the other hand, with respect to the computational errors of the numerical integral which imply errors of the final result as well, a different strategy would be appropriate. Since the $\tilde{G}_{IGF}$ is decreasing very fast with respect to the distance from the center of the bunch, the location where it starts to remain more or less stationary should be determined first. In practice, we use a reference function $f(N_z)$ to locate the stationary area. For example, we choose $1/\log_2 N_z$ as $f(N_z)$ to locate k by $\|\tilde{G}_{IGFk-1} - \tilde{G}_{IGFk}\|/\tilde{G}_{IGFk-1} < f(N_z)$. Secondly, we determine the accuracy tolerance: Choose the proper R_z given by the first k for which the magnitude of $\delta\tilde{G}_k/\delta\tilde{G}_{k\,\text{stable}}$ drops down to 10^{-s}, $s \ge 0$, where $\delta\tilde{G}_k = \|\tilde{G}_{IGF_k} - \tilde{G}_{GF_k}\|$. Note, s is the accuracy control integer for the RIGF. Of course, these parameters have to be determined individually for different problems under study.

3 CRIGF Method for Poisson's Equation

In many applications, the computational domain will be considerably larger than the domain occupied by the charged bunch. As shown in Fig. 1 (right), the bunch domain Ω_{Bunch} (the shadowed domain) lies in the center of the computational domain Ω. In this case, of course, all terms with zero charge density ρ (factor of the tilde Green's function) can be omitted in the summation of (3). Based on the convolution theory, the irrelevance of these terms should be still true if we take a Fourier transform and use it in the fast Poisson solver. Therefore, the CIGF [4] integral is recommended for high efficiency:

$$\tilde{G}_{CIGF}(x_i, y_j, z_k) = \begin{cases} \tilde{G}_{IGF}(x_i, y_j, z_k), & (1,1,1) \le (i,j,k) \le (C_x, C_y, C_z); \\ 0, & \text{otherwise;} \end{cases}$$

where (C_x, C_y, C_z) is determined by the domain-bunch ratio $\alpha_w = L_w\text{Bunch}/L_w\text{Domain}$, L_w is the length for w in $\{x, y, z\}$ and $C_w = [(2 + \alpha_w)/2\alpha_w]$. The CIGF is as accurate as the IGF. For far-bunch domain space charge simulation, the CIGF integral is efficient and does not waste calculations. When the near-bunch domain simulation takes place, the CIGF is not valid anymore. However, the RIGF can always be applied.

In total, we have the following CRIGF integral: The combination of RIGF and CIGF as the CRIGF should be more efficient than the pure CIGF for the same problem,

$$\tilde{G}_{CRIGF}(x_i, y_j, z_k) = \begin{cases} \tilde{G}_{IGF}(x_i, y_j, z_k), & (1,1,1) \le (i,j,k) \le (R_x, R_y, R_z); \\ \tilde{G}_{GF}(x_i, y_j, z_k), & (R_x, R_y, R_z) \le (i,j,k) \le (C_x, C_y, C_z); \\ 0, & \text{otherwise;} \end{cases}$$

where (C_x, C_y, C_z) and (R_x, R_y, R_z) are chosen as above.

In order to make the calculation of (3) more efficient, we should implement it as a cyclic convolution. The charge density ρ_{ex} is obtained by padding ρ with zeros in all expansion grid points, the tilde Green's function $\tilde{G}$ is expanded symmetrically as $\tilde{G}_{ex}$. Using 3D discrete Fourier transformation $\mathfrak{F}$ and convolution theory, the expanded potential expression is given by:

$$[\varphi_{ex}]_{i,j,k} = \frac{1}{4\pi\varepsilon_0} \mathfrak{F}^{-1}\{[\mathfrak{F}\tilde{G}_{ex}]_{i,j,k} \cdot [\mathfrak{F}\rho_{ex}]_{i,j,k}\}_{2N_x, 2N_y, 2N_z}. \tag{9}$$

The routine of (9) can be further improved with respect to both- less storage requirement and less time consumption [1]. We use a similar procedure. Then, the storage requirement of the convolution method is $2N_x \times N_y \times N_z$ plus two temporary 2D arrays sized $2N_y \times N_z$ and $2N_y \times 2N_z$. In fact, our algorithm uses a pruned Fourier transform, whose purpose is to save time while avoiding the wasteful transforms of zeros in each direction.

4 Discussions and Examples

We regard a uniformly charged ellipsoid to achieve an analytical validation. For the longitudinal-to-transverse ratio, we choose 30. The domain-bunch ratio is 2. The relative errors are defined as follows:

$$\eta_\varphi(i,j,k) := \frac{|\varphi_{i,j,k} - \varphi_{true_{i,j,k}}|}{\max_{i,j,k} |\varphi_{true_{i,j,k}}|}, \text{ and } \hat{\eta}_\varphi := \max_{i,j,k}(\eta_\varphi(i,j,k)).$$

Here, the notations are, $\eta_\varphi(i,j,k)$, $\hat{\eta}_\varphi$, $\varphi_{i,j,k}$ and $\varphi_{true_{i,j,k}}$ as the relative error of the potential at index (i,j,k), the global relative error of the potential, the computed

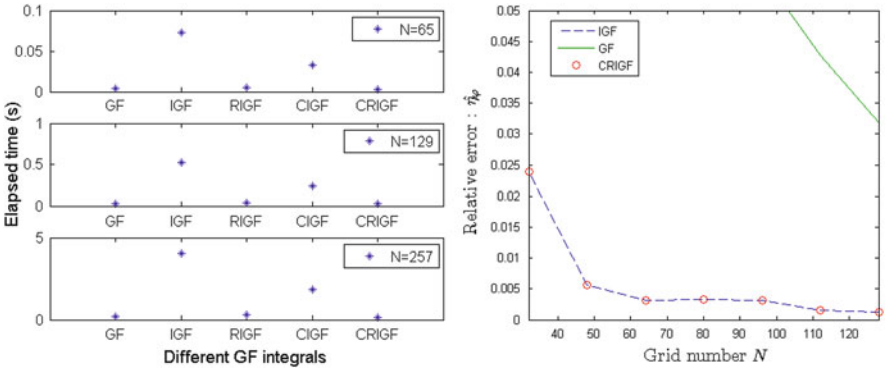


Fig. 2 (*Left*): Comparison of different Green's function integrals' elapsed time with increasing grid resolution. (*Right*): Convergence study of CRIGF, IGF and GF method

potential at index (i, j, k) and the true potential for the same index, respectively. The algorithm is implemented in C language on an Intel 2.6 GHz CPU.

Firstly, we compare the computation time of different GF integrals with increasing grid resolution as shown in Fig. 2 (left). Here $N = N_w, R_w = 8$ with w in $\{x, y, z\}$. RIGF is nearly 10 times faster than IGF, while CRIGF is more than 20 times faster for a domain-bunch ratio of 2. In Fig. 2 (right), we study the convergence of CRIGF with $s = 2$ for the accuracy control comparing to IGF and GF. As we can see, the CRIGF method agrees very well with the IGF method.

The whole implementation is carried out by using the FFTW [5] package. For the serial algorithm, the speed-up is around 15–25 % including the calculation of the discrete convolution. This needs to be further improved, since the convolution's calculation still takes most of the computational time. The efficiency results will be updated in our future studies, either by implementing a parallel algorithm or by applying a different discrete convolution routine.

5 Conclusion

In this paper, we introduced a 3D RIGF Poisson solver together with a routine called CRIGF method for beam dynamics simulations. We tested the new method with a model problem. On the practical side, RIGF is less time consuming, while it achieves almost the same accuracy as IGF for the electric potential. So we suggest to use the CRIGF routine rather than the IGF in order to speed up calculations.

References

1. Hockney, R.W., Eastwood, J.W.: Computer Simulation Using Particles. Institute of Physics Publishing, Bristol (1992)
2. Qiang, J., Lidia, S., Ryne, R.D., Limborg-Deprey, C.: Three-dimensional quasistatic model for high brightness beam dynamics simulation. Phys. Rev. ST Accel. Beams **9**:044204 (2006)
3. Qiang, J., Lidia, S., Ryne, R.D., Limborg-Deprey, C.: Erratum: three-dimensional quasistatic model for high brightness beam dynamics simulation. Phys. Rev. ST Accel. Beams **10**:129901 (2007)
4. Zheng, D., Markoviḱ, A., Pöplau, G., van Rienen, U.: Study of a fast convolution method for solving the space charge fields of charged particle bunches. Proc. IPAC 418–420 (2014)
5. Frigo, M., Johnson, S.G.: The design and implementation of FFTW3. Proc. IEEE **93**:216–231 (2005)

Part III
Coupled Problems

The chapter on *Coupled Problems* comprises five contributions. They range from adequate modelling via specialized integration techniques to automated integration by software agents. These papers underline the need for methods in multiphysical settings with possibly largely differing time scales.

A rapid heating and subsequent quenching of steel components induces a change of the microstructure of the material (phase change), which can be the basis of surface steel hardening. This process needs to be precisely controlled to avoid undesirable fatigue damage. Qingzhe Liu et al. describe in *Simulation of thermomechanical behavior subjected to induction hardening* a coupled model, which enables the numerical treatment of this problem. This system is given by the equations for electromagnetic field (Maxwell), the temperature evolution (heat equation), the mechanical deformations and stresses (momentum balances), and an equation for the steel phase transformations. The interplay of this coupled system includes Joule heat, thermal expansion, mechanical dissipation, transformation-induced plasticity, material/temperature dependent parameters. Eventually, the authors demonstrates the power of the coupled model by comparing a finite element simulation with physical experiments.

The paper by Andrea Cremasco et al. on *Thermal Simulations for Optimization of Dry Transformers Cooling System* addresses a coupled fluid-thermal system. In particular for optimization, one needs an efficient mathematical model for computation. To this end, the authors propose to employ an equivalent thermal/pressure network for their application. They could validated their model via ANSYS Fluent simulations and via measurements (of temperature in thermal steady state). Subsequently, a multi-objective optimization problem was formulated for the cooling system of a transformer based on the developed network model. A corresponding Pareto front was computed and optimal solutions identified.

Coupled systems will quite naturally involve multiple time scales. If these scales are largely differing, an overall time step oversamples the slowly varying subsystem, which is often the major part of the overall system. Therefore multirate schemes aim at employing different time steps for the subsystems to enable the use of an inherent time step and to avoid oversampling. The work by Michael Günther et al. on

Multirate GARK schemes for multiphysics problems provides a multirate extension for generalized additive Runge-Kutta schemes (GARK). The GARK-type methods enable different stage values for different components of the right hand side. This flexibility yields an opportunities to design multirate methods. In particular stability of corresponding multirate methods can be inherited from the underlying implicit schemes without further coupling constraint. First numerical results are shown for a benchmark from thermal-electric network analysis.

Co-Simulation is a term often used in applications to describes the coupling of various simulators to tackle complex real-life problems. A system of software agent in terms of the contribution by Matthias Jüttner et al. represents a software framework for coupled partial differential equations. In *Iterative Software Agent Based Solution of Multiphysics Problems*, they employ autonomous agents, which conduct the solving procedure of a specific subproblem and which may invoke commercial or in-house codes for the respective physical domain. This is based on a weak coupling, where convergence may be achieved by iteration. The idea is demonstrated numerically for a 3D waveguide system, where electromagnetism and heat transfer are coupled.

Due to the current policy changes in energy production, that is, the exit from nuclear and fossil-fuel energy in some countries, photovoltaic systems obtained more attention. The contribution by Timo Rahkonen and Christian Schuss on *Tools for Aiding the Design of Photovoltaic Systems* describes a set of simulation tools for the aiding the design of fixed and mobile photovoltaic energy harvesting systems in particular for moving panels. The main goal of the discussed methods is to estimate the available power and to include effects as self-heating, ambient temperature and bypassing to enable a fast and robust maximum power point tracking. To this end, the electric part of the photovoltaic panel is modelled via a nonlinear circuit using a pn-diode. This model is coupled to an equivalent circuit for the temperature.

These contributions give a small glimpse on the importance and richness of the field of coupled problems.

Thermal Simulations for Optimization of Dry Transformers Cooling System

Andrea Cremasco, Paolo Di Barba, Bogdan Cranganu-Cretu, Wei Wu, and Andreas Blaszczyk

Abstract An efficient computational model based on principles of thermo-fluid dynamics is crucial for thermal design and optimization of transformers. In this paper we propose a Thermal/Pressure Network (TPN) model of a dry transformer encapsulated in enclosure with natural or forced cooling. The network model has been validated by Computational Fluid Dynamics (CFD) simulations with ANSYS Fluent and then applied for the computation of real transformers, comparing results to thermal measurements. Finally, the parameterized transformer TPN model has been utilized in an optimization loop in order to improve the cooling system. In this respect, the use of a gradient-free optimization algorithm under a multi-objective frame is recommended to avoid local minima and smooth the dependency on the initial guess.

1 Introduction

For air-insulated (dry) transformers, the heat generated in the windings is transferred via convection to the bulk air above and then dissipated to the ambient air through the ventilation system, see Fig. 1. For a numerical simulation of such complex phenomena a very resource demanding CFD analysis is required [1], therefore, designers of transformers typically create their own simplified calculation procedures

A. Cremasco (✉) • P. Di Barba
Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
e-mail: andrea.cremasco01@ateneopv.it; paolo.dibarba@unipv.it

B. Cranganu-Cretu
ABB Transformers, Geneva, Switzerland
e-mail: bogdan.cranganu-cretu@ch.abb.com

W. Wu
ABB Transformers, Jefferson, MO, USA
e-mail: wei.wu@us.abb.com

A. Blaszczyk
ABB Schweiz AG, Corporate Research, 5405 Baden-Dättwil, Switzerland
e-mail: andreas.blaszczyk@ch.abb.com

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_11

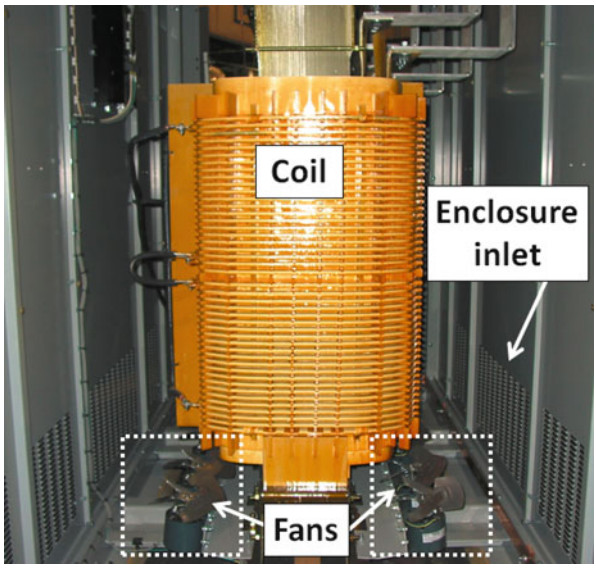


Fig. 1 Dry transformer with ventilation system including fan (which can be switched off) and the enclosure inlet/outlet openings, with grids and optional filters used to protect the transformers against designated conditions

based on empirical assessments and parameters of heat transfer phenomena that are valid for a specific transformer technology [2]. Such procedures are integrated into transformer design systems and used for optimization, thanks to very fast computation times.

In this paper we propose a new method for the thermal simulation of a dry transformer together with the cooling system. The new method is based on a coupled Pressure/Thermal Network (TPN) model, as described in [3]. The new method offers much better computational performance than the detailed CFD; since the TPN method is founded on thermo-fluid dynamics principles, it can be extended to all transformer technologies and cooling configurations including dry transformers as presented in [4].

The basic concept of the network approach is presented in Sect. 2. In this section we included an example for a CFD-based validation of a simple network element representing convection from a vertical wall.

The major new achievement reported in this paper is a network model for a dry-type transformer operated in an enclosure with ventilation openings. The transformer is cooled naturally or by means of fans installed inside the enclosure. The new model has been validated based on CFD computation for a simplified axial-symmetric transformer configuration. With the new network model we could reproduce with reasonable accuracy the fluid flow for different conditions: fans on/off, ventilation openings open/partially closed/closed. We applied the same TPN approach to the computation of real transformers in order to compare results with heat run test measurements. The result of CFD and experimental validations are presented in Sect. 3.

Finally, the fast speed of the network model calculation (less than a second for a typical design) made it possible to apply this method to design transformers in industrial design process and to optimize the cooling system based on multi-objective optimization [5], as shown in an example included in Sect. 4.

2 Network Concept and Network Elements

The TPN method is a lumped-parameter modelling approach based on substituting geometrical parts like windings, cooling ducts, enclosure walls, ventilation openings, fans, etc. by network elements in form of sources, resistors or sub-circuits representing thermo-fluid dynamic phenomena. The basic concepts of TPN, definition of network elements, coupling between the networks and the mathematical background of the solution method have been described in [3]. In Table 1 a short summary is presented.

In order to illustrate how the CFD validation of network elements has been performed, we present here a result for convection from the outer surface of the transformer coil to the bulk air. The height H of the cylindrical coil is variable in a typical range between 500 and 2000 mm. The heat flux $P/A = 150\,\mathrm{W/m^2}$ is dissipated through the cylindrical vertical wall, with radius $r = 350\,\mathrm{mm}$. The goal is calculation of the average temperature of the wall ϑ_{wall} assuming natural convection to the ambient air at temperature $\vartheta_{amb} = 20\,^{\circ}\mathrm{C}$ (radiation is not included).

Table 1 Characteristics of thermal and pressure networks and electrical analogy

Network type	Electric (analogy)	Thermal	Pressure
Quantities, units	Current I (A)	Power P (W)	Mass-flow rate $\dot{m}$ (kg/s)
	Voltage U (V)	Temperature $\Delta\vartheta$ (K)	Pressure Δp (Pa)
	Resistance R (Ω)	Thermal Res. R_t (K/W)	Flow res. S (1/(m · s))
Network principles	Current, voltage law	$\sum_i P_i = 0, \quad \sum_i \Delta\vartheta_i = 0$	$\sum_i \dot{m}_i = 0, \quad \sum_i \Delta p_i = 0$
	Ohm's law: $R = \mathrm{U}/I$	$R_t = \Delta\vartheta/P$	$S = \Delta p/\dot{m}$
Thermo-fluid dynamic principles for network elements evaluations[a, b, c]		Newton's law (convect.): $P = hA\Delta\vartheta$, Stefan-Boltzmann law (radiation), see [6]	Bernoulli's principle: Friction: $\Delta p = 1/2\,\xi\rho v^2$, Buoyancy: $\Delta p = gH_p(\rho_{ref} - \rho)$
Coupling equation		$P = \dot{m}c_p\Delta\vartheta$	

Symbols: h heat transfer coefficient, A heat transfer area, ξ friction factor, ρ fluid density, v fluid velocity (assumed uniform), g gravitation, H_p pressure height, c_p specific heat

[a] In the pressure network only relative pressure resulting from friction and buoyancy is included. Therefore, we can assume that the fluid properties are independent of Δp. The fluid density depends on static pressure according to ideal gas laws

[b] All fluid properties are temperature dependent, which results in non-linear resistors and sources of both pressure and thermal networks

[c] The flow resistance S depends for turbulent flows on velocity, which results in a strongly non-linear behavior of S, the thermal resistance R_t is also velocity dependent, in particular for convection in cooling ducts [4]

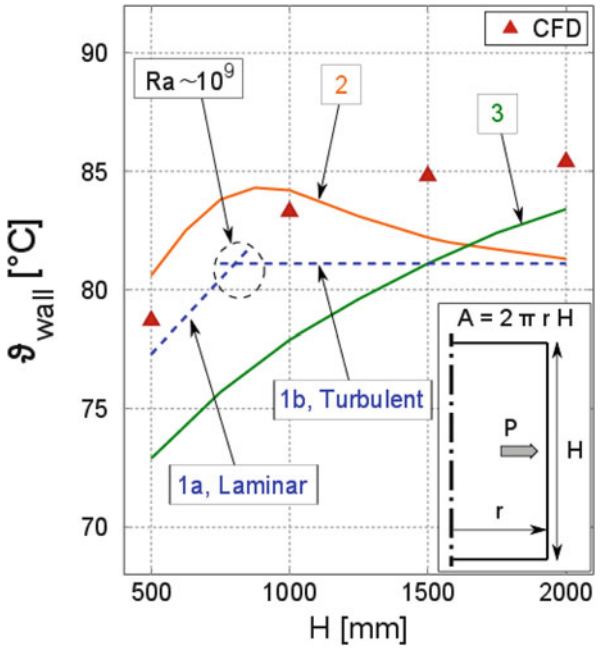


Fig. 2 CFD vs. network results for outer wall temperature of a transformer coil. Wall average temperature: $\vartheta_{wall} = \mathbf{R_t P} + \vartheta_{amb}$ with thermal resistance for convection $\mathbf{R_t} = 1/(\mathbf{h} \cdot \mathbf{A})$ and heat transfer coefficient $\mathbf{h} = (\mathbf{Nu} \cdot \mathbf{k_f})/\mathbf{l_{ch}}$, where k_f is thermal conductivity and $l_{ch} = H$ is characteristic length. The Nusselt number Nu is based on similarity theory, with the following correlations: $\mathbf{Nu} = \mathbf{c_1 Ra^{n1}}$ with $c_1 = 0.54$, $n_1 = 1/4$ for laminar flow, curve 1a (Rayleigh number $Ra < 10^9$); $c_1 = 0.1$, $n_1 = 1/3$ for turbulent flow, curve 1b ($Ra > 10^9$). Curves 2, 3 are based on correlations for constant temperature and constant heat flux models respectively, see Eq. 4.33–4.36 [8]

The results are presented in Fig. 2. The CFD solution is based on heat transfer coupled to Navier-Stokes equations, using $k\omega$-SST turbulence model [7]. The network result, including computation of convection resistor R_t as defined in Table 1, is based on thermodynamic correlations explained in Fig. 2. For all H variations, the difference between temperatures calculated by CFD and the TPN models is less than 5–6 °C, which is still acceptable for applying resistor R_t in the model of the cooling system (see Fig. 4 between HV winding and bypass-duct). Improvements and tuning of R_t can be a subject of future work.

3 Modelling of Cooling System

CFD Analysis and TPN Modelling For the CFD analysis we selected an equivalent axial-symmetric transformer model including a core leg, coil, fan and enclosure with bottom and top ventilation openings, see Fig. 3. The coil consists of a low voltage winding, LV, divided into two radially stacked segments, LV1 and LV2,

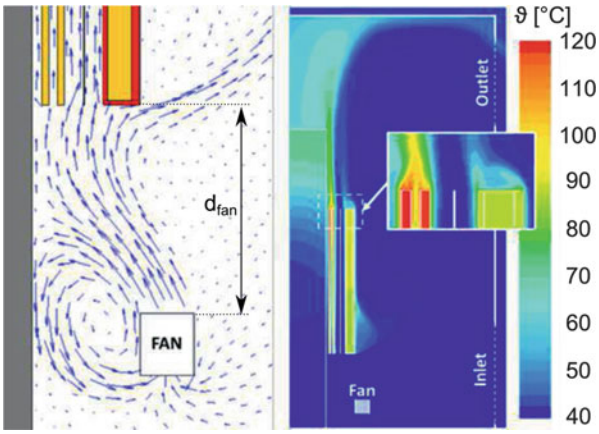


Fig. 3 *Left*: fluid flow recirculation of the fan. *Right*: temperature distribution

a barrier B and a high voltage winding HV casted in solid insulation; the region between the coil and the vertical wall of enclosure is called bypass-duct. The enclosure includes inlet and outlet ventilation openings at the bottom and the top respectively, whose friction was taken into account in CFD with pressure jump boundary conditions [7].

The cold air enters the enclosure from the inlet and flows through the cooling ducts between winding segments; for natural cooling (AN = air natural) the fluid is driven by buoyancy only, while for the forced cooling (AF = air forced) its major part is blown by fans. In both cases, there is air circulation from the bottom to the top of the coil, resulting in hot fluid flowing out from the enclosure through the outlet and taking heat away.

As the main extension of the standalone transformer model investigated in [4], we introduced the "Bypass Duct" as well as "Top" and "Bottom Fluids", see Fig. 4. Together with "Coil Ducts" these elements are responsible for controlling the temperature of the fluid according to the mass and power flow rates in each corresponding network branch (based on the coupling equation in Table 1). The fluid flow direction in the bypass-duct is reversible, see dashed lines in Fig. 4 and the $\dot{m}_{bypass}$ values in Fig. 5. Its direction depends on the performance of the fans and the ventilation grids. Due to reverse bypass flows and the recirculation of the hot air the temperature distribution inside the enclosure can be significantly influenced as shown in Fig. 5.

In the TPN the friction of the enclosure ventilation openings (called here vents) is modelled by a non-linear resistor whose characteristic is based on the equation $\Delta p_{grid} = 1/2\ \xi_{grid}\ \rho\ v^2$ in Table 1. The velocity v is calculated as $v = \dot{m}/(\rho A_{open})$, while A_{open} is the open surface area of the vents. ξ_{grid} is the friction factor of the vents, which depends on construction parameters such as the dimension and shape of the holes, the density of the grid and the presence of filters; these features are related to the Ingress Protection (IP) class of the enclosure [9]: for example dense grids with

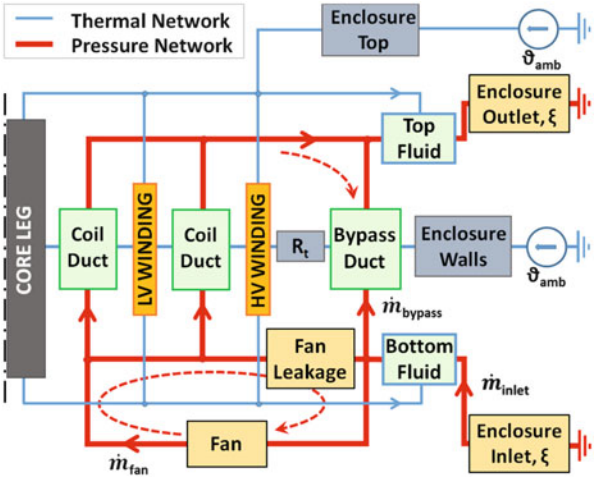
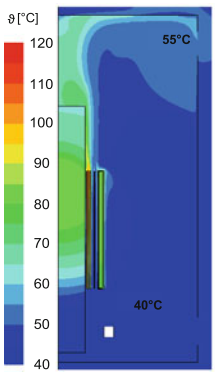
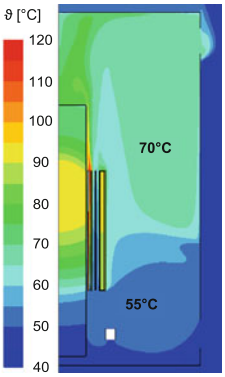


Fig. 4 Concept of the equivalent network model



$\xi^*_{grid} = 10$		
$\vartheta[°C]$	CFD	TPN
LV1	104.8	101.8
LV2	98.6	98.1
HV	81.2	80.0
Core	65.7	60.7
$\dot{m}[g/s]$	CFD	TPN
m_{inlet}	123.4	125.4
m_{bypass}	43.8	41.6
m_{duct1}	8.6	8.9
m_{duct2}	8.0	8.3
m_{duct3}	31.1	31.8
m_{duct4}	32.6	35.0



$\xi^*_{grid} = 200$		
$\vartheta[°C]$	CFD	TPN
LV1	114.3	109.7
LV2	108.5	106.4
HV	89.5	89.1
Core	72.4	67.7
$\dot{m}[g/s]$	CFD	TPN
m_{inlet}	51.2	48.2
m_{bypass}	-23.8	-31.6
m_{duct1}	7.9	8.4
m_{duct2}	7.4	7.9
m_{duct3}	29.0	30.2
m_{duct4}	30.7	33.3

Fig. 5 Temperature maps and results comparison between CFD and TPN model (AF)

filters provide higher IP class but reduce cooling because of the stronger friction. When comparing efficiency of different grids it is convenient to relate the pressure drop not to the open but to the total area A_{tot} of the vent (area that is occupied by the vent in the enclosure wall): $A_{open} = rA_{tot}$, with ratio $r < 1$. After applying this relation to calculation of the velocity and the pressure drop (see equations above) we define an equivalent friction factor $\xi^*_{grid} = \xi_{grid}/r^2$, which has been used for all computations in this paper. The value range for ξ^*_{grid} between 10 and 600 is typical for grids of transformer enclosures in a wide range of IP classes.

TPN Model Validation: CFD and Heat Run Tests In the Figs. 6, 7, 8 we show a comparison between CFD and TPN model results for the average temperature ϑ_{ave} of windings and enclosure walls as well as for the mass flow rate $\dot{m}$ through the enclosure vents and bypass-duct. All the results are referred to the same model

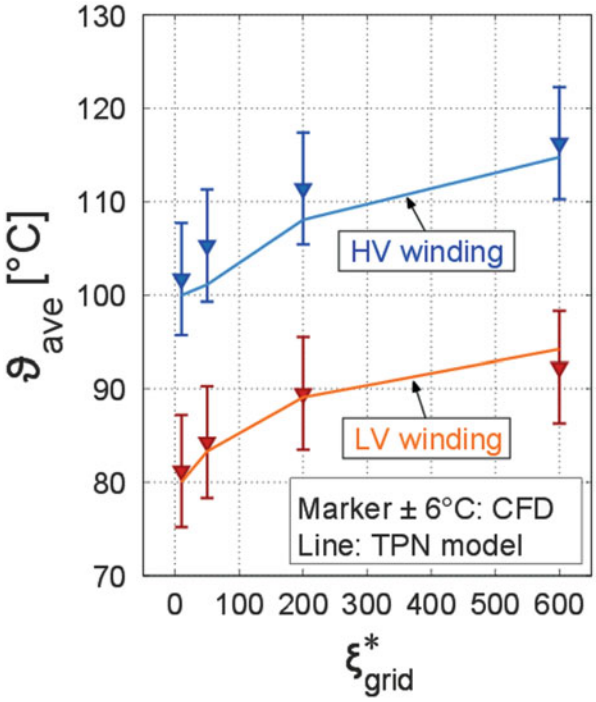


Fig. 6 LV and HV windings average temperature ϑ_{ave}

with load losses, varying only IP class to which different values of ξ^*_{grid} are related; radiation heat transfer has been included.

The temperature deviation is never beyond 6 °C and the CFD trends are followed by TPN, see Figs. 6 and 7. The mass flow rate deviation is always lower than 10 (g/s). The TPN model predicts the distribution of the flow inside the enclosure even when high ξ^*_{grid} limits the flow-rate of the outgoing fluid, causing a downward inversion of the flow in the bypass-duct (this corresponds to a negative value for $\dot{m}$, see Fig. 8).

We applied the TPN method to a real transformer tested with enclosure: the ϑ_{ave} of the windings was derived from electric resistance measurements after reaching the thermal steady state, see Table 2. The deviation from measurements falls into an applicable range of transformer designing.

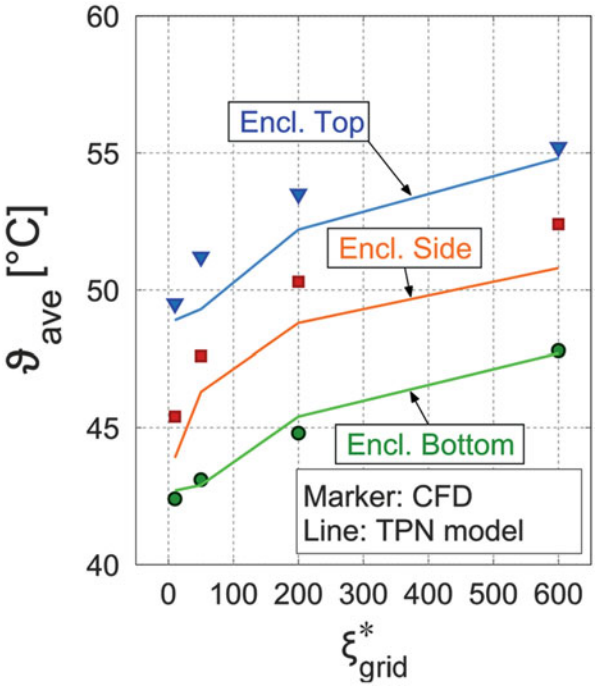


Fig. 7 Enclosure top, side and bottom walls ϑ_{ave}

4 Optimization of Cooling System

In this section we present a formulation of the multi-objective optimization problem applied to the network equivalent model of the dry transformer; the aim is to identify the Pareto front of the non dominated solutions, trading off three design criteria (objective functions). The objective functions to minimize and the design variables are described in Table 3.

In Fig. 9 the 2D projections of the 3D objective space are shown, with Pareto optimal solutions. Results have been obtained by means of the Non Dominated Sorting Algorithm NSGA-II [10]; finding the Pareto front lasted few hours on a standard processor for personal computing.

A posteriori, having identified the Pareto front, the designer can extract a single optimal solution taking into account extra preferences like e.g. the pressure vs. volume flow rate characteristics of a real fan.

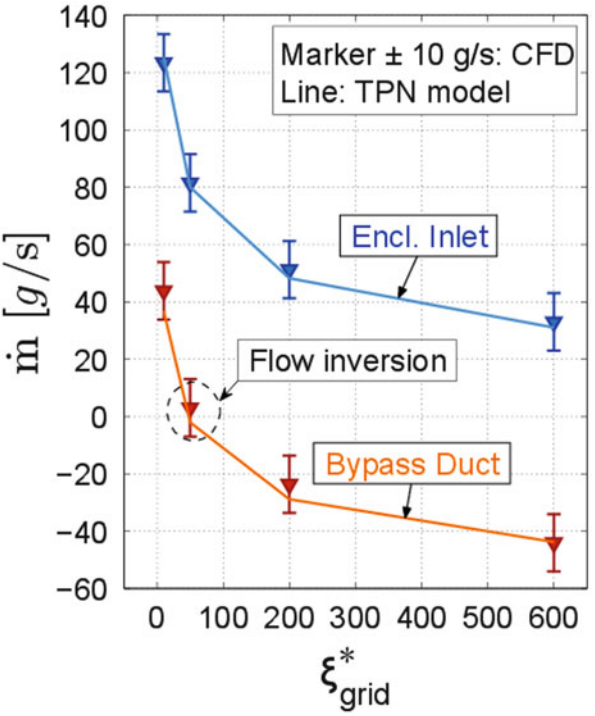


Fig. 8 Enclosure inlet and bypass-duct $\dot{m}$

Table 2 Temperature (°C) comparison between heat run test measurements and TPN model with natural (AN) and forced (AF) ventilation

Ventilation	LV winding			HV winding		
	Measured	TPN	Deviation	Measured	TPN	Deviation
AN	98.1	98.5	0.4	97.8	97.5	−0.3
AF	87.2	83.3	−3.9	95.8	93.5	−2.3

5 Conclusions. Next Steps

In this work we introduced the new model of equivalent Thermal/Pressure networks for dry transformer cooling systems. The CFD validation and comparison with heat run tests confirmed the applicability of the model to a wide range of enclosure Ingress Protection (IP) classes for natural and forced cooling. The new model has been integrated into a transformer design system (used in ABB) and will be a subject of tuning and statistical evaluations based on a large number of transformer designs. The presented application of finding the Pareto front from a multi-objective optimization will be considered as a possible extension of the transformer design system.

Table 3 Formulation of the optimization problem

Objective function to minimize		Description
$\vartheta_{ave} = \frac{\Sigma_k T_{wind,k} V_{wind,k}}{\Sigma_k V_{wind,k}}$ (°C)		Average temperature of the coil: the winding temperatures T_{wind} are weighted by the winding volumes V_{wind}
$\Delta p_{fan} = p_{fan,out} - p_{fan,in}$, (Pa)		Pressure jump provided by the fan [a]
$q_{rec,\%} = \left(1 - \frac{\dot{m}_{inlet}}{\dot{m}_{fan}}\right) \cdot 100$		Recirculation index: if $q_{rec,\%} = 0$, then $\dot{m}_{inlet} = \dot{m}_{fan}$ and there is no recirculation
Design variable	Bounds	Description
Q_{fan} (m^3/s)	(0.10, 0.35)	Fan volume flow rate. Note that there are two fans in parallel per coil, each one blowing the same Q_{fan}
d_{fan} (m)	(0.15, 0.40)	Axial distance of the fan out-take from the coil bottom, see Fig. 3
$k_v = \frac{A_{outlet}}{A_{inlet}}$	(0.7, 1.3)	Vent surface ratio, subject to the constraint on the enclosure design: $A_{tot} = A_{outlet} + A_{inlet} = 7\,m^2$. Vent surface is defined as $A_{inlet} = A_{tot} k_v/(1 + k_v)$, $A_{outlet} = A_{tot}/(1 + k_v)$

[a] The pressure jump Δp_{fan} provided by a fan blowing a certain volume flow rate Q_{fan} depends for example on the vent open surface or the distance d_{fan}. A real fan can supply higher Q_{fan} when lower Δp_{fan} is required, improving the cooling of the transformer

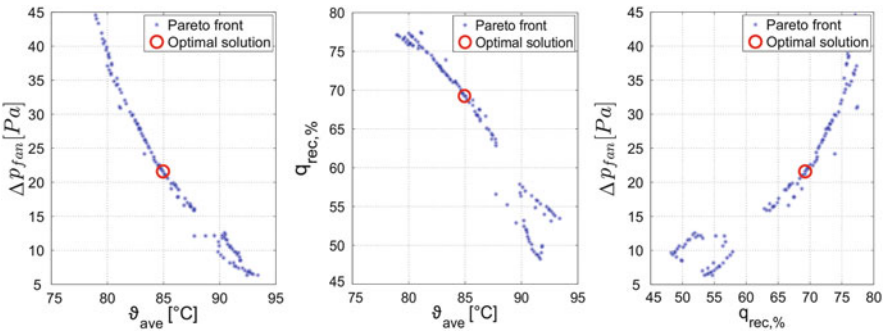


Fig. 9 2D projections of the 3D objective space. For the extracted optimal solution: $Q_{fan} = 0.25$ m^3/s, $d_{fan} = 0.21$ m, $k_v = 1.04$

Acknowledgements The authors sincerely thank Bernardo Galletti and Marcelo Buffoni, scientists from ABB Corporate Research, for their contribution to the CFD analysis.

References

1. Smolka, J., Nowak, A. J.: Experimental validation of the coupled fluid flow, heat transfer and electromagnetic numerical model of the medium-power dry-type electrical transformer. Int. J. Therm. Sci. **47**, 1393–1410 (2008)
2. Bockholt, M., Mönig, W., Weber, B., Patel, B., Cranganu-Cretu, B.: Thermal Design of VSD Dry-Type Transformer, SCEE 2012, Zurich

3. Blaszczyk, A., Flückiger, R., Müller, T., Olsson, C.-O.: Convergence behavior of coupled pressure and thermal networks. COMPEL J. **33**(4), 1233–1250 (2014)
4. Morelli, E., Di Barba, P., Cranganu-Cretu, B., Blaszczyk, A.: Network based cooling models for dry transformers. In: ARWtr Conference, Baiona, Spain (2013)
5. Di Barba, P., Dolezel, I., Karban, P., Kus, P., Mach, F., Mognaschi, M.E., Savini, A.: Multiphysics field analysis and multiobjective design optimization: a benchmark problem. Inverse Probl. Sci. Eng. **22**(7), 1214–1225 (2014)
6. Holman, J. P.: Heat transfer, Eq. (7–25), 6th edn. McGraw-Hill, NY (1999)
7. ANSYS Inc.: ANSYS Fluent 14.0 User's Guide, Boundary conditions, Porous Jump Boundary Conditions
8. Rohsenow, W.M., Hartnett, J.P., Cho, Y.I. (eds.): Handbook of Heat Transfer. McGraw-Hill, NY (1998)
9. IEC 60529. Degrees of protection provided by enclosures (IP Code)
10. Kalyanmoy, D., Amrit, P., Sameer, A., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)

Multirate GARK Schemes for Multiphysics Problems

Michael Günther, Christoph Hachtel, and Adrian Sandu

Abstract Multirate GARK schemes define a multirate extension of GARK schemes, generalized additive Runge-Kutta schemes. In contrast to additive schemes, GARK schemes allow for different stage values as arguments of different components of the right hand side. They introduce additional flexibility when compared to traditional partitioned Runge-Kutta methods, and therefore offer additional opportunities for the development of flexible solvers for systems with multiple scales, or driven by multiple physical processes.

Consequently, multirate GARK schemes allow for exploiting multirate behaviour in both the right-hand sides and in the components in a rather general setting, and are thus especially useful for coupled problems in a multiphysics setting. We apply MGARK schemes to a benchmark example from thermal-electrical coupling, characterized by a slow and fast part with a stiff and non-stiff characteristic, resp. We test two MGARK schemes: (a) an IMEX method, which completely utilizes the dynamics and differing stability properties of the coupled subsystem; and (b) a fully implicit schemes, which inherits the stability properties from both underlying schemes without any coupling constraint.

1 Introduction

Multiphysical systems are often characterized by a very different dynamical behavior in the subsystems, with time constants differing by orders of magnitude. To be efficient, numerical time integration schemes have to exploit this multirate behavior,

M. Günther (✉)
Applied Math. & Numerical Analysis, Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: guenther@math.uni-wuppertal.de

C. Hachtel
Bergische Universität Wuppertal, Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM), Gaußstraße 20, D-42119 Wuppertal, Germany
e-mail: hachtel@math.uni-wuppertal.de

A. Sandu
Department of Computer Science, Virginia Polytechnic Institute and State University, Computational Science Laboratory, 2202 Kraft Drive, Blacksburg, VA 22060, USA
e-mail: asandu@cs.vt.edu

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_12

which is physically given and allows for a static partitioning of the subsystems into slow and fast parts, resp.

Multirate time integration schemes aim at exploiting this property by applying different time step sizes to the subsystems, according to their different activity level. To get higher order schemes, these schemes have to fulfill additional order conditions, and at the same time preserve the stability properties of the respective subsystems.

This paper discusses the application of a new class of multirate schemes, multirate GARK [1] schemes based on a generalized view on additive Runge-Kutta schemes [3], to a multiphysical problem from electro-thermal coupling.

The paper is organized as follows: Sect. 2 gives a synopsis on multirate GARK schemes and their relation to GARK schemes. Section 3 introduces two multirate GARK schemes, based on an explicit-implicit and implicit-implicit pair of order-2 basis schemes. Section 4 discusses the numerical results obtained for both schemes. The last section concludes with final remarks and an outlook.

2 Multirate GARK Schemes

We consider a two-way partitioned system

$$y' = f(y) = f^{\{\mathfrak{s}\}}(y) + f^{\{\mathfrak{f}\}}, \quad y(t_0) = y_0\,, \tag{1}$$

with a slow component $\{\mathfrak{s}\}$, and an active (fast) component $\{\mathfrak{f}\}$. Note that this setting contains component-wise splitting as a special case:

$$y = \begin{pmatrix} y_{\mathfrak{s}} \\ y_{\mathfrak{f}} \end{pmatrix}, \quad f^{\mathfrak{s}} = \begin{pmatrix} f_{\mathfrak{s}} \\ 0 \end{pmatrix}, \quad f^{\mathfrak{f}} = \begin{pmatrix} 0 \\ f_{\mathfrak{f}} \end{pmatrix}. \tag{2}$$

The slow component is solved with a large step H, and the fast one with small steps $h = H/M$. We will consider the multirate generalization of GARK schemes [3] with M micro steps $h = H/M$, as given in the following

Definition 1 (Multirate GARK Method [1]) One macro-step of a generalized additive multirate Runge-Kutta method with M equal micro-steps reads

$$Y_i^{\{\mathfrak{s}\}} = y_n + H \sum_{j=1}^{s^{\{\mathfrak{s}\}}} a_{i,j}^{\{\mathfrak{s},\mathfrak{s}\}} f^{\{\mathfrak{s}\}}\left(Y_j^{\{\mathfrak{s}\}}\right) + h \sum_{\lambda=1}^{M} \sum_{j=1}^{s^{\{\mathfrak{f}\}}} a_{i,j}^{\{\mathfrak{s},\mathfrak{f},\lambda\}} f^{\{\mathfrak{f}\}}\left(Y_j^{\{\mathfrak{f},\lambda\}}\right),$$

$$Y_i^{\{\mathfrak{f},\lambda\}} = y_n + h \sum_{l=1}^{\lambda-1} \sum_{j=1}^{s^{\{\mathfrak{f}\}}} b_j^{\{\mathfrak{f}\}} f^{\{\mathfrak{f}\}}\left(Y_j^{\{\mathfrak{f},l\}}\right) + H \sum_{j=1}^{s^{\{\mathfrak{s}\}}} a_{i,j}^{\{\mathfrak{f},\mathfrak{s},\lambda\}} f^{\{\mathfrak{s}\}}\left(Y_j^{\{\mathfrak{s}\}}\right) +$$

$$+h\sum_{j=1}^{s^{\{\mathrm{f}\}}} a_{i,j}^{\{\mathrm{f},\mathrm{f}\}} f^{\{\mathrm{f}\}}\left(Y_j^{\{\mathrm{f},\lambda\}}\right), \quad \lambda = 1,\dots,M,$$

$$y_{n+1} = y_n + h\sum_{\lambda=1}^{M}\sum_{i=1}^{s^{\{\mathrm{f}\}}} b_i^{\{\mathrm{f}\}} f^{\{\mathrm{f}\}}\left(Y_i^{\{\mathrm{f},\lambda\}}\right) + H\sum_{j=1}^{s^{\{\mathrm{s}\}}} b_i^{\{\mathrm{s}\}} f^{\{\mathrm{s}\}}\left(Y_i^{\{\mathrm{s}\}}\right).$$

The base schemes are Runge-Kutta methods, $(A^{\{\mathrm{f},\mathrm{f}\}}, b^{\{\mathrm{f}\}})$ for the slow component and $(A^{\{\mathrm{s},\mathrm{s}\}}, b^{\{\mathrm{s}\}})$ for the fast component. The coefficients $A^{\{\mathrm{s},\mathrm{f},\lambda\}}$ and $A^{\{\mathrm{f},\mathrm{s},\lambda\}}$ for $\lambda = 1,\dots,M$ realize the coupling between the two components.

2.1 Order Conditions

The MGARK scheme can be written as a GARK scheme [3] over the macro-step H with the fast stage vectors $Y^{\{\mathrm{f}\}} := [Y^{\{\mathrm{f},1\}\,T},\dots,Y^{\{\mathrm{f},M\}\,T}]^T$. The corresponding Butcher tableau reads (with the vector $\mathbf{1} := (1,\dots,1)^T$ of ones)

$$\begin{array}{cccc|c}
\frac{1}{M}A^{\{\mathrm{f},\mathrm{f}\}} & 0 & \cdots & 0 & A^{\{\mathrm{f},\mathrm{s},1\}} \\
\frac{1}{M}\mathbf{1}b^{\{\mathrm{f}\}\,T} & \frac{1}{M}A^{\{\mathrm{f},\mathrm{f}\}} & \cdots & 0 & A^{\{\mathrm{f},\mathrm{s},2\}} \\
\vdots & & \ddots & & \vdots \\
\frac{1}{M}\mathbf{1}b^{\{\mathrm{f}\}\,T} & \frac{1}{M}\mathbf{1}b^{\{\mathrm{f}\}\,T} & \cdots & \frac{1}{M}A^{\{\mathrm{f},\mathrm{f}\}} & A^{\{\mathrm{f},\mathrm{s},M\}} \\
\hline
\frac{1}{M}A^{\{\mathrm{s},\mathrm{f},1\}} & \frac{1}{M}A^{\{\mathrm{s},\mathrm{f},2\}} & \cdots & \frac{1}{M}A^{\{\mathrm{s},\mathrm{f},M\}} & A^{\{\mathrm{s},\mathrm{s}\}} \\
\hline
\frac{1}{M}b^{\{\mathrm{f}\}\,T} & \frac{1}{M}b^{\{\mathrm{f}\}\,T} & \cdots & \frac{1}{M}b^{\{\mathrm{f}\}\,T} & b^{\{\mathrm{s}\}\,T}
\end{array}$$

Therefore the order conditions for MGARK schemes can be derived from the corresponding ones for GARK schemes [3]. Up to order two the order conditions given in Table 1 have to be fulfilled.

Table 1 Order conditions for MGARK schemes

p	Order condition
1	$b^{\{\mathrm{s}\}\,T}\mathbf{1} = 1$
	$b^{\{\mathrm{f}\}\,T}\mathbf{1} = 1$
2	$b^{\{\mathrm{s}\}\,T}A^{\{\mathrm{s},\mathrm{s}\}}\mathbf{1} = \frac{1}{2}$
	$b^{\{\mathrm{s}\}\,T}\left(\sum_{\lambda=1}^{M} A^{\{\mathrm{s},\mathrm{f},\lambda\}}\right)\mathbf{1} = \frac{M}{2}$
	$b^{\{\mathrm{f}\}\,T}A^{\{\mathrm{f},\mathrm{f}\}}\mathbf{1} = \frac{1}{2}$
	$b^{\{\mathrm{f}\}\,T}\left(\sum_{\lambda=1}^{M} A^{\{\mathrm{f},\mathrm{s},\lambda\}}\mathbf{1}\right) = \frac{M}{2}$

2.2 Stability

We consider systems (1) where each of the component functions is dispersive (with constants $\nu^{\{s\}} < 0,\ \nu^{\{f\}} < 0$):

$$\left\langle f^{\{s\}}(y) - f^{\{s\}}(z)\,,\, y - z \right\rangle \leq \nu^{\{s\}}\ \|y - z\|^2\,,$$
$$\left\langle f^{\{f\}}(y) - f^{\{f\}}(z)\,,\, y - z \right\rangle \leq \nu^{\{f\}}\ \|y - z\|^2\,,$$

with respect to the same scalar product $\langle \cdot, \cdot \rangle$. As for two solutions $y(t)$ and $\tilde{y}(t)$ of (1), each starting from a different initial condition, the norm of the solution difference $\Delta y(t) = \tilde{y}(t) - y(t)$ is non-increasing, we demand a similar property from the numerical approximations: the MGARK scheme is said to be nonlinearly stable, if the inequality

$$\| y_{n+1} - \tilde{y}_{n+1} \| \leq \| y_n - \tilde{y}_n \|$$

holds for any two numerical approximations y_{n+1} and $\tilde{y}_{n+1}$ obtained by applying the scheme to the ODE (1) with dispersive functions and with initial values y_n and $\tilde{y}_n$.

As a consequence of stability theory for GARK schemes, an MGARK scheme applied to a component-wise partitioned right-hand side (2) is nonlinearly stable, if both base schemes are algebraically stable [1].

3 Two Basic GARK Schemes for Multiphysics Application

In general, one is interested in a rough approximation of coupled multiphysics problems, which reflect the impact of the couplings of both systems. Hence we restrict to MGARK schemes of order 2. As we are interested in the nonlinear stability properties of MGARK schemes, and how the stability properties of both base schemes influence the stability of the overall scheme, we define two new IMEX and IMIM schemes as basic methods:

- MGARK-IMEX-2: The implicit-explicit version solves the fast, stiff part with an implicit base scheme, and the slow, non-stiff part with an explicit one. The coefficients are given by

$$b^{\{s\}} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \quad A^{\{s,s\}} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad A^{\{s,f,1\}} = \begin{pmatrix} 0 \\ M \end{pmatrix},$$
$$A^{\{s,f,\lambda\}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \forall \lambda = 2, \ldots, M,$$

$$b^{\{\mathrm{f}\}} = 1, \quad A^{\{\mathrm{f},\mathrm{f}\}} = \frac{1}{2}, \quad A^{\{\mathrm{f},\mathrm{s},\lambda\}} = \begin{pmatrix} \frac{1}{2} & 0 \end{pmatrix} \quad \forall \lambda = 1, \ldots, M.$$

The slow components are implicitly solved together with the fast components of the first micro step. The fast components of the remaining micro steps can be computed explicitly.

Note that only the fast part is algebraically stable, but neither the slow part nor the joint system.

- MGARK-IMIM-2: To get an overall stable scheme, both parts are solved by an implicit base scheme. The coefficients are given by

$$b^{\{\mathrm{s}\}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad A^{\{\mathrm{s},\mathrm{s}\}} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad A^{\{\mathrm{s},\mathrm{f},1\}} = \begin{pmatrix} 0 \\ \frac{M}{2} \end{pmatrix},$$

$$A^{\{\mathrm{s},\mathrm{f},\lambda\}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \forall \lambda = 2, \ldots, M,$$

$$b^{\{\mathrm{f}\}} = 1, \quad A^{\{\mathrm{f},\mathrm{f}\}} = \frac{1}{2}, \quad A^{\{\mathrm{f},\mathrm{s},\lambda\}} = \begin{pmatrix} \frac{1}{2} & 0 \end{pmatrix} \quad \forall \lambda = 1, \ldots, M.$$

Note that again the slow components are implicitly defined together with the fast components after the first micro step. The fast components of the remaining micro steps can be computed one after the other by solving nonlinear systems in the dimension of the active part only.

As both base schemes are algebraically stable, the MGARK method inherits this property for a component-wise partitioning.

4 Numerical Test Results for a Benchmark Example

We will test both MGARK implementations for a thermal–electrical multiphysics system, for specifications see [2]; its circuit diagram is given in Fig. 1 (left). The thermal component defines the slow (and non-stiff) part, the electrical component the fast (and stiff) part of the system.

The distributed temperature T of the resistor (wire) is described by the 1-D heat equation, which is semi-discretised using a finite volume approach, see Fig. 1 (right). Due to the electric current, the resistor is heated and so the resistance of this device changes: $R = R(T)$. The characteristic curve of the diode is also temperature dependent. The voltages are modeled by a nodal analysis using Kirchhoff's laws. Finally we get a partitioned system of ordinary differential equations like in (1). The vector of unknowns $y = (u_3, u_4, e, T)^T$ comprises the voltages u_3 and u_4 at nodes 3 and 4, resp., the dissipated energy e in the thermally dependent resistor and the vector of temperatures T in the semi-discretised resistor. The multirate behaviour of this system is given by the physical properties: the voltages and the dissipated energy change very fast (due to the source of the network), and the temperature in

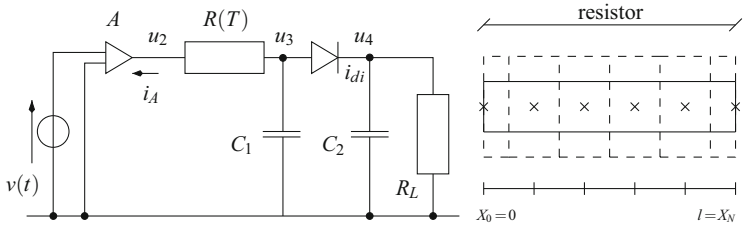


Fig. 1 Circuit and discretised resistor

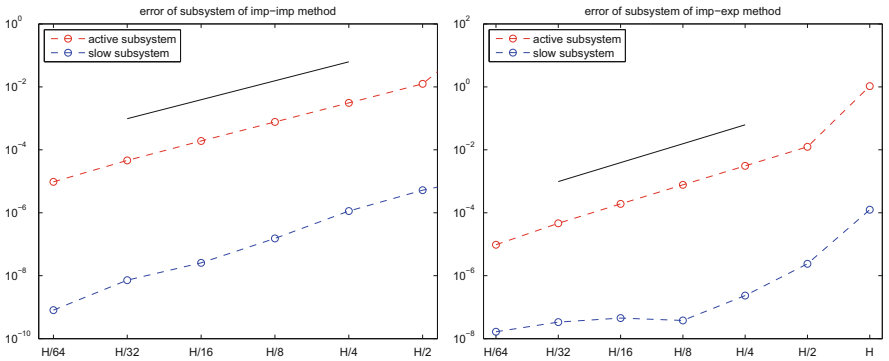


Fig. 2 Numerical results for the fast and slow subsystems (macro step size vs. achieved accuracy, measured in Euclidean norm): MGARK-IMIM-2 (*left*) and MGARK-IMEX-2 (*right*) with parameters $H = 10^{-3}$, $m = 5$. The *solid lines* represent the slope of order 2

the resistor changes much slower. Hence the partitioning according to the dynamical behaviour is quite natural:

$$y^{\{\mathfrak{f}\}} := \begin{pmatrix} u_3 \\ u_4 \\ e \end{pmatrix}, \qquad y^{\{\mathfrak{s}\}} := T.$$

The numerical results for both Multirate GARK schemes are given in Fig. 2. The IMIM scheme nicely shows in both fast and slow subsystems an order-2 behavior for all step sizes. The accuracy of the IMEX scheme in the slow part (which is not algebraically stable and computed explicitly), however, seems to be reduced for small step sizes.

5 Conclusion

By testing Multirate GARK schemes on a multiphysical test example from electro-thermal coupling, we have shown the feasibility of this multirate approach for both implicit-implicit and implicit-explicit pairing of basic schemes. Whereas the IMIM scheme shows an order-2 behavior for both subsystems at all step sizes, the IMEX schemes has a reduced accuracy in the slow system for small step sizes only. This behavior fits to the theoretical properties of both schemes: the IMIM scheme is algebraically stable in both subsystems, whereas the IMEX scheme is only stable in the fast (electric) part.

As next steps, we will follow three directions: (a) we will apply MGARK schemes to a range of multiphysical problems in a more realistic setting; (b) we will further analyze the stability of IMEX-MGARK schemes and its dependence on the coupling structure for both weak and slow coupling; (c) the excellent stability properties of IMIM-MGARK schemes suggest to use these schemes as basic schemes in a Multirate-MOR approach.

Acknowledgements The work of A. Sandu has been supported in part by NSF through awards NSF DMS—1419003, NSF CMMI—1130667, NSF CCF—1218454, AFOSR FA9550-12-1-0293-DEF, AFOSR 12-2640-06, and by the Computational Science Laboratory at Virginia Tech.

The work of M. Günther and C. Hachtel has been supported in part by the German BMBF Program, through grant 05M13PXA (BMBF Verbundprojekt KoSMOS, see http://scwww.math.uni-augsburg.de/projects/kosmos/) and by the EU through grant 619166 (FP7-STREP nanoCOPS, see http://www.fp7-nanoCOPS.eu/).

References

1. Günther, M., Sandu, A.: Multirate generalized additive Runge Kutta methods. Numer. Math. doi:10.1007/s00211-015-0756-z
2. Hachtel, C., Bartel, A., Günther, M.: Efficient simulation for electrical-thermal systems via multirate-MOR. In: Proceedings of SCEE 2014, Wuppertal (2014)
3. Sandu, A., Günther, M.: A generalized-structure approach to additive Runge-Kutta methods. SIAM J. Numer. Anal. **53**(1), 17–42 (2015)

Iterative Software Agent Based Solution of Multiphysics Problems

Matthias Jüttner, André Buchau, Desirée Vögeli, Wolfgang M. Rucker, and Peter Göhner

Abstract A novel approach is presented using software agents for an iterative and distributed solution of multiphysics problems. Overall convergence is achieved by using the individual capabilities of interworking agents. Every agent solves a partial single physics problem based on specialized, commercial or in-house code. The autonomy of each agent allows a physics adapted solution process without the need of a predefined solver sequence. The applied software agents are described in detail. Here, we focus on weak uni- and bidirectional field coupled multiphysics problems. This framework can also be used for node or boundary coupling as well as for optimising partial physics simulation. A coupled 3D electromagnetic wave propagation and heat transfer problem inside a waveguide is examined as numerical example.

1 Introduction

Methods for simulating single physics problems on high-performance computers were state of the art for many years. During the last years, tools were extended to cluster, cloud and graphical processing unit (GPU) computing to achieve further parallelism [1]. Recent developments combine different single physics implementations to a multiphysics framework by considering them as black boxes [2]. Improvements on software maintenance and functionality were achieved on costs of performance and memory usage [3]. For a practical usage, expert knowledge is needed in the fields of physics, their coupling and the numerical solution. However, engineers as users are experts within one or maybe a few physics. Therefore,

M. Jüttner (✉) • A. Buchau • W.M. Rucker
University of Stuttgart, Institute for Theory of Electrical Engineering, Pfaffenwaldring 47, 70569 Stuttgart, Germany
e-mail: matthias.juettner@ite.uni-stuttgart.de; andre.buchau@ite.uni-stuttgart.de; rucker@ite.uni-stuttgart.de

D. Vögeli • P. Göhner
University of Stuttgart, Institute of Industrial Automation and Software Engineering, Pfaffenwaldring 47, 70569 Stuttgart, Germany
e-mail: desiree.voegeli@ias.uni-stuttgart.de; peter.goehner@ias.uni-stuttgart.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_13

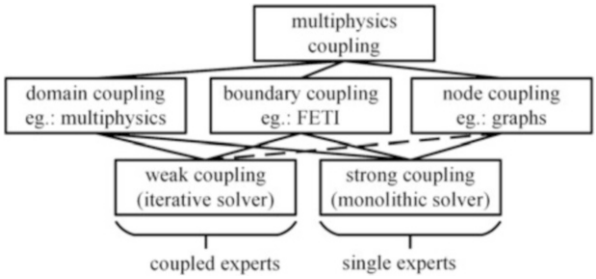


Fig. 1 Different methods of multiphysics coupling shown for two physics. The segregation of a monolithic multiphysics problem also represents a common way for parallelization. A central unit combines the partial results for an overall solution

an initial partitioning of a multiphysics problem (as in Fig. 1) seems odd at the beginning.

In practice, models are step-wise extended to consider multiphysics effects. This step-wise development starts from multiple independent physical models and uses shared variables to couple independent models to a multiphysics system. For solving several multiphysics problems, a monolithic as well as a segregated approach lead in practice to a valid solution [4]. For parallelizing the monolithic approach, the problem must be partitioned, while the segregated approach is natively parallel. Only connections of former independent problems lead to sequential dependencies.

Here, the work flow of distributed interacting single physics experts is projected into a multiphysics simulation environment. This system handles different physics with new encapsulated software agents and automatically coupling the physics. The agents autonomously interact with each other and share collective values. With this, a 3D coupled electromagnetic wave propagation and heat transfer problem inside a waveguide is solved exemplary. The hereinafter presented framework also promotes a physics based parallel calculation. In Sect. 2 an overview about software agents and their design is given. An explanation how that system is used for solving multiphysics problems is given in Sect. 3. In Sect. 4, the solver systems capabilities are demonstrated by a numerical example. A conclusion is given in Sect. 5.

2 Software Agent System

Software agents are encapsulated (software) entities with individual goals [5]. They are well tested in automation technologies for solving complex and distributed problems. A software agent tries to reach its goal by acting autonomously. It interacts with other agents of the system and its environment, while keeping a persistent state. The following list presents the main concepts of agents.

- Encapsulation: An agent encapsulates information. It has a certain knowledge of its environment and of its own capabilities.
- Persistence: An agent has its own control flow and keeps its internal state during lifetime. It is independent of an external activation.
- Autonomy: An agent is able to act autonomously and make decisions by itself.

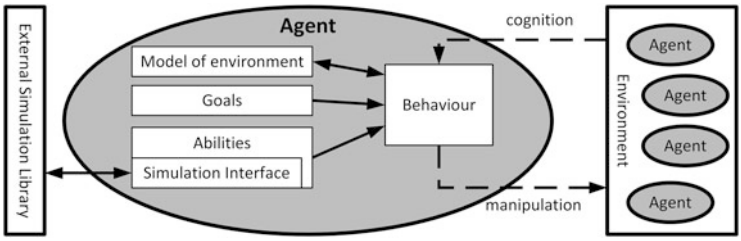


Fig. 2 Internal structure of a software agent. This allows the agent to act autonomously based on its abilities, and interact with other agents and its environment to solve complex tasks in a very flexible way

- Interaction: An agent can interact with other agents of the system. By doing this, agents are able to combine their knowledge and collaborate.
- Activity: An agent reacts to changes in its environment and can evoke changes.
- Goal-oriented: An agent has own goals that may change during lifetime. It is able to plan and execute activities by itself and react to situations by changing its plan.

If several agents work together, the system is called multi-agent system (MAS). Its setup can change during runtime. The internal structure [6] is shown in Fig. 2.

In the following, software agents are used as physics experts. They couple single physics simulations to a multiphysics problem. An interface to an external simulation library enables the agent to manipulate the model, couple it with other physics and control and supervise the attached solver within the simulation library [7]. An early attempt for 2D boundary coupled systems is given in [8]. Here, the presented work handles weakly coupled systems with different experts. Problems solvable with monolithic solvers only, are handled by a single expert (see Fig. 1). For establishing a coupling between the agents, the agents share information about coupling and calculation capabilities. This description provides information about calculation resources, numerical methods, solvable equations, possible boundary conditions, provided results, and derived values as a list. Implementing the agents was done using corresponding design rules [5]. The programming language must handle the complexity of agents' communication, provide the agents itself, manage the attached simulation interface, and handle exchanged numerical data in a powerful and parallel way. To use state of the art software development techniques, Java was chosen [9] together with the Java Agent DEvelopment framework (JADE) [10].

3 Solver System

For practical reasons, two types of software agents are required. A coordination agent (CO) splits the XML-file based multiphysics problem, created with nowadays computer aided design (CAD) tools into multiple single physics problems. Multiple

calculation agents (CA) cooperatively solve the coupled sub-problems. For the finite element method (FEM), the problem is given as

$$\mathbf{K}\mathbf{u} = \mathbf{b}, \tag{1}$$

$\mathbf{K}$ represents the stiffness matrix, $\mathbf{u}$ the solution and $\mathbf{b}$ the load. For a multiphysics problem $\mathbf{K}$ is usually not symmetric due to different influences between the physics. For a problem with two physics, u can be grouped and the problem reformulated as

$$(C \circ K)\, u = \begin{bmatrix} \mathbf{C}_{11}\mathbf{K}_1 & \mathbf{C}_{12}\mathbf{K}_{12} \\ \mathbf{C}_{21}\mathbf{K}_{21} & \mathbf{C}_{22}\mathbf{K}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}. \tag{2}$$

$\circ$ representing the Hadamard product and $\mathbf{C}$ an activation matrix for the coupling. An uncoupled problem has a $\mathbf{C}$ equal to the identity matrix $\mathbf{I}$. For a fully coupled system all non-diagonal matrices (e.g. $\mathbf{C}_{12}, \mathbf{C}_{21}$) become $\mathbf{I}$. In a loop wise coupled system, the main and upper diagonal matrices become $\mathbf{I}$, including the element of the first column and last row. For more than two physics, this fits best for an iterative sequential solution. If $\mathbf{K}$ includes further couplings (eg. $\mathbf{K}_{24}$), a parallelization is possible and automatically applied with this approach. Initially, no coupling is considered $\mathbf{C}_{12} = \mathbf{C}_{21} = 0$ and two CAs are used for this problem.

$$\begin{matrix} \textit{Agent 1} \\ \textit{Agent 2} \end{matrix} \quad \text{solves} \quad \begin{bmatrix} \mathbf{u}_1^0 \\ \mathbf{u}_2^0 \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1^{-1}\mathbf{b}_1^0 \\ \mathbf{K}_2^{-1}\mathbf{b}_2^0 \end{bmatrix} \tag{3}$$

in parallel. Each agent uses its own backbox simulation environment for its partial problem. Tests with different environments or solvers can be performed simultaneously by additional agents. The fastest agent for a partial problem survives. The fastest agent for a partial problem currently survives. As soon as any agent finished its calculation (e.g. *agent 1*), all agents get informed about an available result and derived values. Conditions are a first time calculated result or changes in the result $\mathbf{u}_1$ compared to a previous calculation cycle $\mathbf{u}_1^*$. Based on its own features list, each agent decides whether to couple or to ignore and continue calculating. In case of coupling material dependent parameters, $\mathbf{K}_2$ is reassembled. If new sources gets available, the coupling matrix $\mathbf{C}_{21}$ changes to $\mathbf{I}$. The new problem

$$\begin{bmatrix} \mathbf{u}_1^1 \\ \mathbf{u}_2^1 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^0 \\ \mathbf{K}_2^{-1} \underbrace{\left(\mathbf{b}_2 - \mathbf{K}_{21}\mathbf{u}_1^0\right)}_{\mathbf{b}_2^1} \end{bmatrix} \tag{4}$$

is solved, while calculated intermediate results are used as initial values for further calculations. Equation (4) can be seen as a first iterative step solving Eq. (2) using Jacobi method. The new $\mathbf{b}_2^1$ handles non-linear coupling between the physics. The strength of coupling changes during an iterative process [11]. Stabilising the system should be possible with relaxation methods like Aitken Δ^2 or gradient

based ones [12]. Obviously, at least one partial problem must converge during the iterative solution process. The iterative method ends, if the relative changes for $\mathbf{u}_i$ or derived values are below a limit ε_i, i representing the agent number. In Fig. 3 the unidirectional result propagation implementation for two CAs is shown. If more than one expert with the same knowledge works on a problem, methods like the Finite Element Tearing and Interconnection (FETI) domain decomposition approaches allows to engage the agents [13]. As more agents dealing with a problem, as further the parallelisation will be, limited by the communication overhead that is not considered here. Solver selection algorithms [14] as well as learning algorithms are imaginable. Adapted meshes for the different physics have been already tested [7]. Another application of this approach comes together with co-simulation and different time-steps [15] of the agents. In all cases, the individuality of the agents allows to optimize the process.

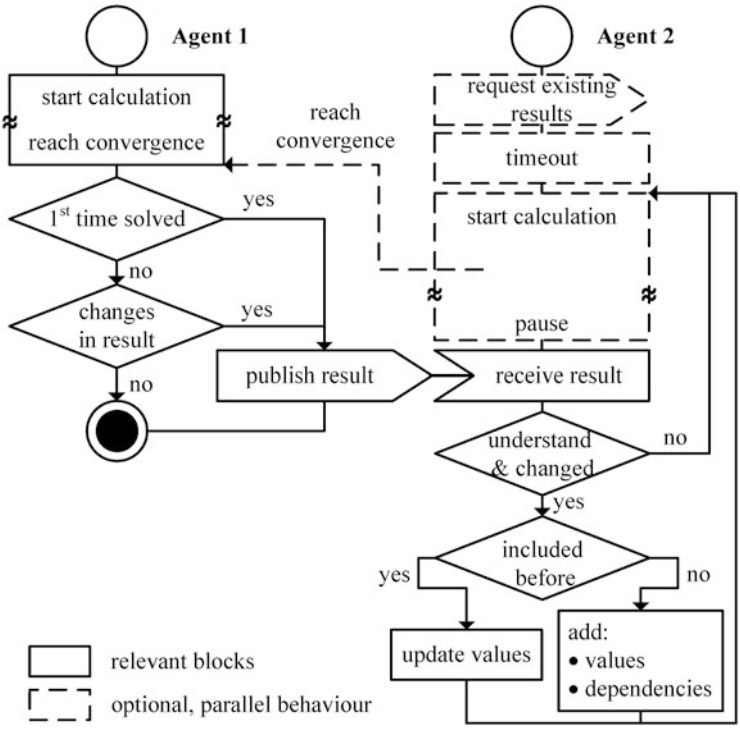


Fig. 3 Unidirectional result propagation process for two agents. *Agent 1* starts computing Eq. (3). *Agent 2* notices another agent working on the same problem and asks for existing results. If no results are provided, *agent 2* starts computing Eq. (3) in parallel. *Agent 1* finishes its calculation first and publishes the results to *agent 2*. This pauses its iterative solver to integrate the offered results, if it's possible. Afterwards, the calculation is continued until *agent 2* is ready to publish its results

4 Numerical Example

The solution process of a coupled electromagnetic wave propagation problem and a heat transfer problem is shown for a lossy dielectric within a waveguide surrounded by air. It demonstrates the principle of the iterative agent based solution of multiphysics problems. Here, three agents are needed. *Agent 0* represents a CO, *agent 1* and *agent 2* CAs. In Fig. 4 the MAS setup is shown.
The agents run on an Intel(R) Core(TM) i7-2600 CPU with 4 cores, max. 3.4 GHz, 16 GB (1333 MHz) RAM and Microsoft Windows 8.1 Enterprise 64-bit. *Agent 1* handles the electromagnetic wave problem according to

$$\Delta\mathbf{E} + \mu_r k_0^2(\varepsilon_r - \frac{\mathrm{j}\sigma}{\omega\varepsilon_0})\mathbf{E} = 0. \tag{5}$$

Here μ_r is the relative permeability, k_0 the wave number of free space, ε_r the relative permittivity, σ the electrical conductivity, ω the angular frequency, and ε_0 the free space permittivity. Eq. (5) is solved in the frequency domain within the waveguide. All over the model, the thermal problem is considered. It is defined by

$$\kappa\Delta T + Q = 0. \tag{6}$$

and solved by *agent 2* for a stationary case. κ represents the thermal conductivity and Q is a heat source. According to the FEM approach, the electric field strength $\mathbf{E}$ and the temperature T are the dependent variables $\mathbf{u}_1$ and $\mathbf{u}_2$ in Eq. 2. A convective

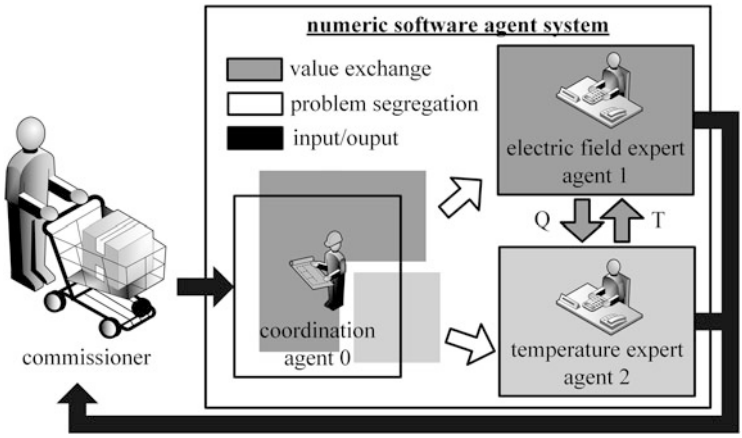


Fig. 4 Setup of the MAS for a coupled two physics problem. The commissioner hands over the multiphysics problem and receives the simulation results. The coordination agent distributes the problem and the calculation agents solve parts of the problem, they are versed to do. Exchanging value allows a coupled iterative solution

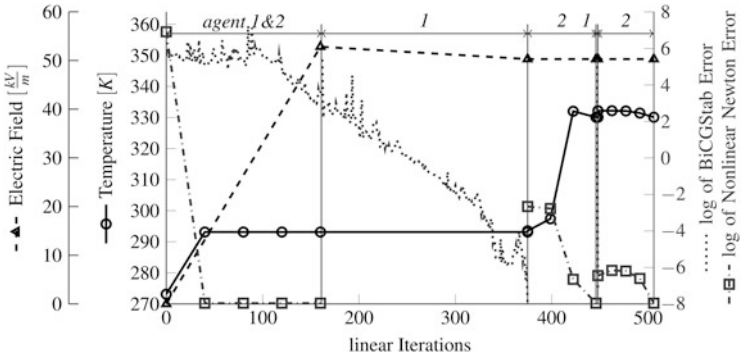


Fig. 5 Solver sequence for the coupled problem including the dependent variables (*left*), the problem dependent errors (*right*), the global iteration counter (*bottom*) and the active agents (*top*)

heat flux with the heat flux coefficient h at the boundaries given as

$$\mathbf{n} \cdot \kappa \nabla T = h(T_{ext} - T) \tag{7}$$

makes a stationary simulation possible. Coupling is dynamically established by a heat source Q representing the total power dissipation density in *agent 2* and the temperature dependent electric conductivity $\sigma(T)$ in *agent 1*. The slow heating process (within seconds) compared to the high frequency wave propagation (10 GHz) allows to consider the heat source Q as constant over time. The numerical solver is chosen from *agent 1* to be BICGStab and from *agent 2* to be a non-linear Newton method combined with a FGMRES. *Agent 0* segregates the multiphysics problem into two single physics problems and distributes them to *agent 1* and *agent 2*. After receiving the problems, *agent 1* and *agent 2* start computing in parallel (Fig. 5).

Values between the marked points for temperature T and the electric field $\mathbf{E}$ are linear interpolated. Here, Eq. (6) is successfully solved first. Due to the temperature dependent electric conductivity $\sigma(T)$ at *agent 1*, results of *agent 2* have to be considered in *agent 1*. Once a solution for *agent 1* is found, *agent 2* is informed about the results. Now, the total power dissipation density of the electromagnetic wave is available and can used as heat source Q in Eq. (6). The bidirectional coupling leads to a loop. Table 1 shows the maximum node wise difference of the exchanged values compared to the previous values. Due to the small changes ε_2 for the temperature, the loop ends. Additionally, a comparison between the agent based solver system and a segregated solver for a given iterative sequence is given. Identical meshes and a BiCGStab solver for both agents are used. The error is computed as maximum node wise difference of the solution vectors.

306 linear iterations were necessary to solve the electric field problem in a purely sequential process. A computation time advantage of the agent based solver is gained by solving the initially uncoupled problems in parallel. The computation of *agent 1* is interrupted when the results of *agent 2* get available (see Fig. 5). Here

Table 1 Solver sequence for the waveguide

Agent	Variable	Integrate	Max. difference	Lin. iterations	Max. error	Rel. error %
2	T	None	First	36	5×10^{-14}	2×10^{-14}
1	**E**	New Source(T)	First	306	5×10^{-4}	6×10^{-4}
2	T	New Source(Q)	25 K	77	0.71	0.21
1	**E**	Update(T)	7.64 $\frac{\mathrm{W}}{\mathrm{m}^3}$	1	5×10^{-4}	7×10^{-4}
2	T	Update(Q)	6×10^{-6} K	43	0.72	0.22

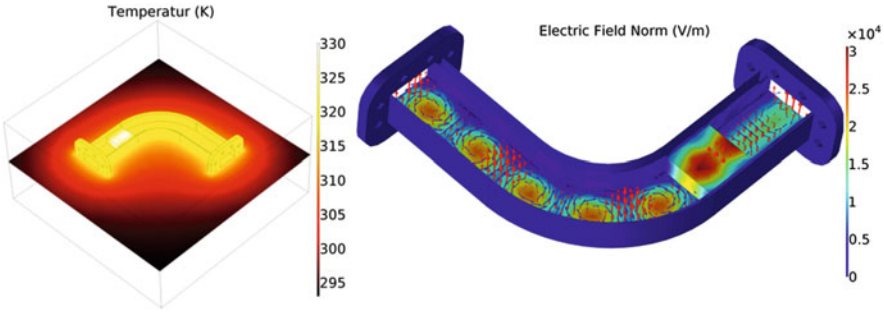


Fig. 6 Visualisation of results from *agent 2* and *agent 1*

agent 1 was interrupted after 160 linear iterations and only 212 additional iterations were needed to solve Eq. (4) after integrating results of *agent 2*. This shows, that iterations are spared, if partial results with final values are integrated before finishing the calculation, and more than two agents are working at a problem. The results of the solved waveguide problem for a mode 10 transverse electromagnetic wave (TE10) at 10 GHz and a convective heat flux at the boundaries of 1 $\frac{\mathrm{W}}{\mathrm{m}^2 \cdot \mathrm{K}}$ are shown in Fig. 6.

5 Conclusion

The step-wise development of multiphysics problems enables a parallelized way of solving coupled multiphysics problems. Based on the idea of interworking experts, several requirements were discussed for implementing this software system. Motivated by the affinity of multi-agent systems to the expert system, an algorithm for uni- and bidirectional coupling was presented. Details about their implementations as well as advantages of the system were given. The solution of a practical example finally demonstrates the performance of the presented expert system. Engaging more agents to further parallelize and optimize the solution process is a future task. Same holds for the selection mechanism of the numerical solver used in each agent. Using the system to solve strongly coupled problems with attached weakly coupled physics is now possible.

References

1. Griebel, M., Zumbusch, G.: Parallel multigrid in an adaptive PDE solver based on hashing and space-filling curves. Parallel Comput. **25**(7), 827–843 (1999)
2. Uekermann, B., Bungartz, H., Gatzhammer, B., Mehl, M.: A parallel, black-box coupling algorithm for fluid–structure interaction, pp. 1–12. CIMNE (2013)
3. Drashansky, T.T., Joshi, A., Rice, J.R., Houstis E.N., Weerawarana, S.: A multiagent environment for MPSEs. In: Conference on Parallel Processing for Scientific Computing (1997)
4. Heil, M., Hazel, A.L., Boyle, J.: Solvers for large-displacement fluid–structure interaction problems: segregated versus monolithic approaches. Comput. Mech. **43**, 91–101 (2008)
5. Jennings, N.R.: Agent-oriented software engineering. in Multiple Approaches to Intelligent Systems, pp. 4–10. Springer, Berlin/Heidelberg (1999).
6. Maurmaier, M., Wagner, T., Development of Embedded Software Systems with Structured Components and Active Composition Support. In: OMER 3 (2005)
7. Jüttner, M., Buchau, A., Rauscher, M. , Rucker, W.M., Göhner, P.: Iterative solution of multiphysics problems using software agents designed as physics experts. In: ISTET'13 (2013)
8. Drashansky, T.T., Joshi, A., Rice, J. R.: SciAgents-an agent based environment for distributed, cooperative scientific computing. In: ICTAI'95, pp. 452–459 (1995)
9. Villacis, J.: A note on the use of Java in scientific computing. ACM SIGAPP Appl. Comput. Rev. **7**(1), 14–17 (1999)
10. Bellifemine, F.L., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE. Wiley , Chichester (2007)
11. Markert, B.: Weak or strong: on coupled problems in continuum mechanics. Habilitation, Universität Stuttgart (2010)
12. Küttler, U., Wall, W.A.: Fixed-point fluid–structure interaction solvers with dynamic relaxation. Comput. Mech. **43**(1), 61–72 (2008)
13. Mandel, J., Dohrmann, C.R., Tezaur, R.: An algebraic theory for primal and dual substructuring methods by constraints. Appl. Numer. Math. **54**(2), 167–193 (2005)
14. Ewald, R., Himmelspach, J., Uhrmacher, A.M.: An algorithm selection approach for simulation systems. 22nd Workshop on Principles of Advanced and Distributed Simulation (2008)
15. Schöps, S.: Multiscale modeling and multirate time-integration of field/circuit coupled problems. PhD thesis, Bergische UniversitätWuppertal & Katholieke Universiteit Leuven (2011)

Simulation of Thermomechanical Behavior Subjected to Induction Hardening

Qingzhe Liu, Thomas Petzold, Dawid Nadolski, and Roland Pulch

Abstract Induction hardening is one of the most important heat treatments of steel components. This paper presents a mathematical and numerical model developed for a coupled problem of Maxwell's equations describing the electromagnetic fields, the balance of momentum which determines internal stresses and deformations resulting from thermoelasticity and phase transformation induced plasticity, a rate law to determine the distribution of different phases and the heat equation to determine the temperature distribution in the workpiece. The equations are solved using a finite element method. A good agreement between the simulation results and experiment performed to determine the deformation is observed. In addition, the distribution of residual stresses after the heat treatment is well predicted.

1 Introduction

For many applications in mechanical engineering, especially in automotive industry, there is a growing demand in components made of steel sheets. Therefore, to improve the quality of boundary layers of these sheets is a significant task since one must carefully control the process precisely in order not to harden the complete sheet which may lead to undesirable fatigue effects.

Surface hardening is a well known method for enhancing mechanical properties of steel components. The aim of this heat treatment is to increase the hardness of the boundary layers of components made from steel by rapid heating and subsequent quenching. The reason why the hardness increases relies on a change

Q. Liu (✉) • R. Pulch
Institut für Mathematik und Informatik, Ernst Moritz Arndt Universität Greifswald, 17487 Greifswald, Germany
e-mail: liuq@uni-greifswald.de; pulchr@uni-greifswald.de

T. Petzold
Weierstraß-Institut, Mohrenstr. 39, 10117 Berlin, Germany
e-mail: Thomas.Petzold@wias-berlin.de

D. Nadolski
Stiftung Institut für Werkstofftechnik IWT, Badgasteiner Str. 3, 28359 Bremen, Germany
e-mail: nadolski@iwt-bremen.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_14

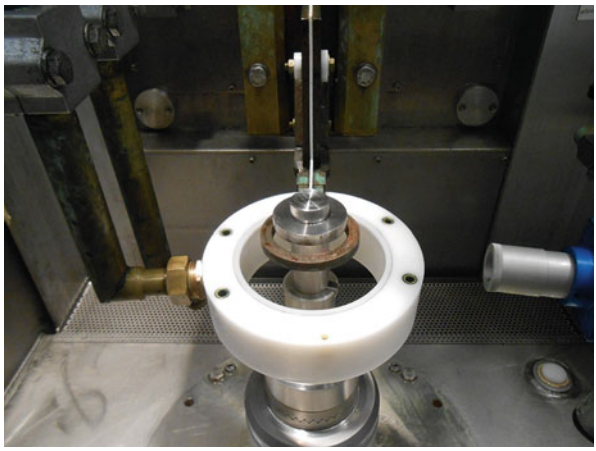


Fig. 1 Induction hardening of a disc (by Stiftung Institut für Werkstofftechnik IWT, Bremen)

in the microstructure of the workpiece during the surface hardening which produces the desired hardening effect.

Depending on heat sources there are different surface hardening procedures. The most import ones are flame hardening, laser hardening and induction hardening. In comparison to flame and laser, induction hardening is advantageous with regard to energy consumption because of the Joule effect resulting from eddy currents.

Figure 1 shows such an experimental set up which consists of an induction coil (inductor), an alternating current power supply, a cooling apparatus and the workpiece (in this experimental set up a disc made from steel) itself as basic components. During the heating stage of this process the inductor is connected to the power supply, the flow of the alternating current through the induction coil induces eddy currents inside the workpiece that lead to increase in temperature due to the Joule effect. Then the current is switched off and the workpiece is quenched by cooling liquid which leads to the desired hardening effect on the boundary layer of the workpiece.

A mathematical model for induction surface hardening accounts for the electromagnetic effects as well as for the thermomechanical behavior and the phase transitions that are caused by enormous changes in temperature during the heat treatment.

The paper is organized as follows. In Sect. 2 we give a brief survey of the complete mathematical model of induction hardening which has been investigated intensively in [2]. Section 3 is devoted to the numerical discretization of the problem. Here a finite element method is applied to solve the system of partial differential equations. The aspect arising from different time scales which needs to be considered in the simulation is addressed. Section 4 focuses on the simulation results of the coupled problem of electromagnetics and thermomechanics in the process of inductive heating for discoid samples made of steel 42CrMo4 (AISI 4140). In comparison to [7] in which only the quenching process has been considered we implement the full procedure, i.e., the induction heating and following

quenching process. Especially we present numerical results of mechanical effects like the residual stress distribution as well as thermally induced distortions where the TRIP (transformation induced plasticity) was involved since this effect is significant in the induction hardening (cf. [6]). From the simulation results a good agreement with experiments according to the deformation is observed and the distribution of residual stresses after the heat treatment is well predicted. Further numerical results of gears samples made of the material AISI 4140 including the determination of the most important properties of the parts for industrial practice, e.g. hardness pattern, residual stresses and distortion have been presented in [3]. The paper ends with a short conclusion and an outlook.

2 The Mathematical Model

For the complete process of heating and cooling we consider the model components corresponding to the electromagnetic field, the temperature evolution, the phase transformations as well as the mechanical deformations and stresses. It accounts for a coupled problem of Maxwell's equations describing the electromagnetic fields, the balance of momentum which determines internal stresses and deformations caused by thermoelasticity and TRIP and the heat equation describing the evolution of temperature distribution in the workpiece. Figure 2 depicts the interrelations among these physical model components. To model the coupled problem of electromagnetics and thermomechanics we first define spatial computational domains. Let $G \subset \mathbb{R}^3$ be a domain which surrounds the inductor Ω and the workpiece Σ.

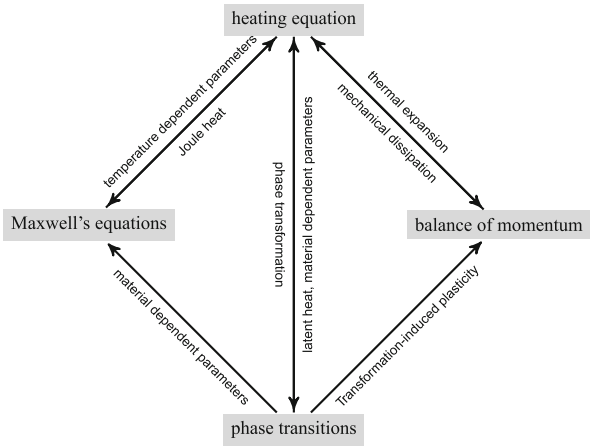


Fig. 2 Mathematical subproblems for induction hardening and their interplays

The electromagnetic effects in G are described by Maxwell's equations that consist of a system of partial differential equations with respect to the electric field $\mathbf{E}$, the magnetic induction $\mathbf{B}$, the magnetic field $\mathbf{H}$ and electric displacement field $\mathbf{D}$, i.e.:

$$\begin{aligned} \operatorname{curl} \mathbf{E} &= -\partial_t \mathbf{B} \\ \operatorname{div} \mathbf{B} &= 0 \\ \operatorname{curl} \mathbf{H} &= \mathbf{J} + \partial_t \mathbf{D} \\ \operatorname{div} \mathbf{D} &= \varsigma \end{aligned} \tag{1}$$

where $\mathbf{J}$ denotes the current density and ς the charge density. In addition Ohm's law yields

$$\mathbf{J} = \gamma \mathbf{E}$$

where γ is the electric conductivity and by constitutive laws we obtain

$$\mathbf{D} = \varepsilon \mathbf{E}, \quad \mathbf{B} = \mu \mathbf{H}$$

with material dependent electrical permittivity ε and magnetic permeability μ. We introduce the magnetic vector potential $\mathbf{A}$ such that

$$\mathbf{B} = \operatorname{curl} \mathbf{A},$$

and impose the Coulomb gauge

$$\operatorname{div} \mathbf{A} = 0.$$

Then following [2] we employ the vector potential formulation of Maxwell's equations which has been derived based on Helmholtz decomposition. More details can be found in [3].

With regard to phases, at the beginning of the heating process the workpiece consists of a mixture of ferrite, pearlite, and bainite. At the end of the heating process the outer layers of the workpiece have been transformed to austenite. The phase evolution of austenite is described along the ideas given in [9]. Then upon rapid quenching the austenite fraction is transformed to martensite. The rate laws describing phase evolutions during cooling have been presented in [6, 7].

The thermomechanical behavior in the complete process can be modeled by laws of energy balance and balance of momentum (cf. [3]). The coupling interface between temperature and deformation is thermal expansion and backward mechanical dissipation.

In summary, the governing equations of the electromagnetic field, the temperature evolution, the mechanical deformations and stresses as well as the steel phase transformations read as follows:

$$\begin{aligned} \gamma \partial_t \mathbf{A} + \operatorname{curl} \mu^{-1} \operatorname{curl} \mathbf{A} - \mathbf{J}_{\text{src}} &= \mathbf{0}, \quad \text{in } G \\ \rho c_\varepsilon \partial_t \theta - \operatorname{div} k \nabla \theta &= F, \quad \text{in } \Sigma \\ -\operatorname{div} \boldsymbol{\sigma} &= \mathbf{0}, \quad \text{in } \Sigma \\ \dot{\mathbf{z}} - \mathbf{f}(\mathbf{z}, \theta, t) &= \mathbf{0}, \quad \text{in } \Sigma \\ \dot{\boldsymbol{\varepsilon}}^{\text{trip}} - g(\boldsymbol{\sigma}, \theta, \mathbf{z}, \dot{\mathbf{z}}) &= 0, \quad \text{in } \Sigma \end{aligned} \tag{2}$$

where the variables $(\mathbf{A}, \theta, \boldsymbol{\sigma}, \mathbf{z}, \boldsymbol{\varepsilon}^{\text{trip}})$ denote the magnetic vector potential, the temperature, the stress tensor, phase fraction and phase transformation induced plasticity (TRIP) strain, respectively. For isotropic materials the stress tensor $\boldsymbol{\sigma}$ which is a matrix valued function admits a expression in terms of the displacement $\mathbf{u}$ (cf. [3, 7]). Here the divergence of the tensor $\boldsymbol{\sigma}$ is a vector field defined by the divergence for each row of the tensor matrix. The material dependent parameters $(\gamma, \mu, \rho, c_\varepsilon, k)$ denote electrical conductivity, magnetic permeability, density of the workpiece, specific heat and heat conductivity. $\mathbf{J}_{\text{src}}$ denotes the source current density satisfying $-\operatorname{div} \mathbf{J}_{\text{src}} = 0$, F summarizes the source term caused by Joule heat, mechanical dissipation and latent heat due to phase transitions. Here the vector potential formulation of Maxwell's equations is taken into account, the heating equation has been derived from energy balance, the deformation equation is based on balance of momentum, the rate of change of phase fractions $\mathbf{f}$ results from the Johnson-Mehl-Avrami equation (cf. [8]) and Schröder's approach, see e.g. [5]. The equation describing the evolution of TRIP $\boldsymbol{\varepsilon}^{\text{trip}}$ is derived from the Franitza-Mitter-Leblond proposal (cf. [1]).

All material parameters depend strongly upon the temperature θ and phase distribution $\mathbf{z}$. The intermediate coupling interface of heating equation [the second equation of (2)] and mechanical equation [the third equation of (2)] comprises the thermal strain, denoted by $\boldsymbol{\varepsilon}^{\text{th}}$ which corresponds to the mechanical strain resulting from temperature change and can be expressed by thermal expansion coefficient and the mechanical dissipation, i.e.,

$$\sigma : (\dot{\boldsymbol{\varepsilon}}^{\text{th}} + \dot{\boldsymbol{\varepsilon}}^{\text{trip}})$$

which reactively influences temperature and is involved in the source term of heating equation. Moreover, the term of Joule heat $\gamma |\partial_t \mathbf{A}|^2$ couples electromagnetics and thermomechanics.

3 Numerical Discretization

The workpiece boundary is dissected into a part τ_s which is free from any acting force and a part τ_u where the workpiece is fixed. The method of lines (MOL) is applied for discretization of Eq. (2). The first step is to discretize the partial differential equations with respect to space while keeping the time variable continuous. Here the spatial discretization is achieved by tetrahedral mesh generation.

Using curl-conforming finite elements we introduce the solution space of the vector potential $\mathbf{A}$:

$$\mathbb{H}(\text{curl}, G) = \left\{\mathbf{v} : G \to \mathbb{R}^3 \middle| \mathbf{v}, \text{curl } \mathbf{v} \in [L^2(G)]^3 \text{ and div } \mathbf{v} = 0, \mathbf{v} \times \mathbf{n} = \mathbf{0} \text{ on } \partial G\right\}$$

where $\mathbf{n}$ denotes the normal to the boundary. For the temperature θ and the phase fraction $\mathbf{z}$ we use classical H^1-conforming elements

$$H^1(\Sigma) = \left\{v : \Sigma \to \mathbb{R} \middle| v \in L^2(\Sigma), \nabla v \in [L^2(\Sigma)]^3\right\}$$

while the displacement $\mathbf{u}$ is approximated by vector-valued H^1 elements

$$X^u(\Sigma) = \left\{\mathbf{v} : \Sigma \to \mathbb{R}^3 \middle| \mathbf{v} \in [H^1(\Sigma)]^3, \mathbf{v} \cdot \mathbf{n}|_{\tau_s} = 0, \mathbf{v}|_{\tau_u} = \mathbf{0}\right\}.$$

With these definitions in mind we use a finite element method (FEM) to calculate the unknowns by computing their projections on corresponding finite dimensional subspaces. More precisely, the FE-discretized system (2) is already a system of DAEs for the variables: vector potential, temperature, displacement, phase fractions and TRIP. Concerning discretization in time we solve the heat equation together with the balance of momentum and the ODEs describing the phase transition and TRIP using a 'large' time step Δt resulting from the typical time scale of the heat conduction. To solve the electromagnetic problem in the time interval Δt we use a time step $\delta t \ll \Delta t$ that is related to the source term of the vector potential equation. Here we use a time stepping scheme of order two with time step δt. For more details we refer the reader to [3].

4 Simulation and Experimental Verification

The numerical simulations are carried out on a disc with diameter 47.7 mm made of steel 42CrMo4 (Fig. 3). From symmetry reasons we restrict ourselves to compute only a segment with an angle of $\frac{\pi}{20}$ (cf. Fig. 4). The cross sections are subject to the symmetric boundary conditions, that means the displacement on the symmetric cuts equals zero in normal direction and the normal directional space derivatives of the displacement along the cross sections are zero (cf. [7]). All material parameters

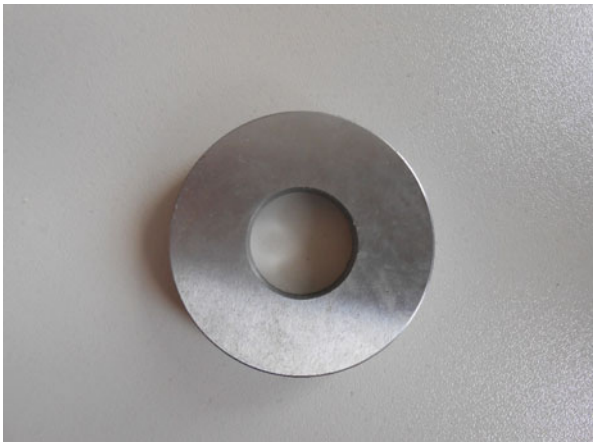


Fig. 3 Disc geometry provided by Stiftung Institut für Werkstofftechnik IWT, Bremen

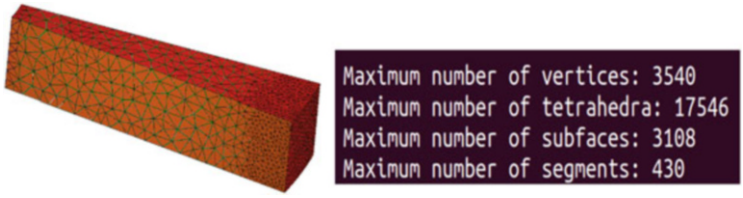


Fig. 4 The reduced computational domain with FE-mesh

associated with 42CrMo4 for the simulations are provided by IWT (Stiftung Institut für Werkstofftechnik, Bremen), and parameters for phase transitions are taken from [4]. All numerical results presented here accompanied with the thermally induced deformation are scaled by 40 to improve their visualizations. According to experimental setting we use a medium frequency 12 kHz with power 100 kW, relative power 63 % and current 575 A, and assume that the surrounding room temperature is 20 °C.

The simulation results of such an induction heating process are visualized. Figure 5 shows progressive temperature values at different heating stages. Owing to the enormous increase in temperature the workpiece suffers from gross distortion caused by thermal expansion during heating. The subsequent cooling process leads to thermal contraction as well as TRIP. Figure 6 shows the corresponding Euclidean norm of the displacement at the beginning of cooling stage and the end of cooling, respectively.

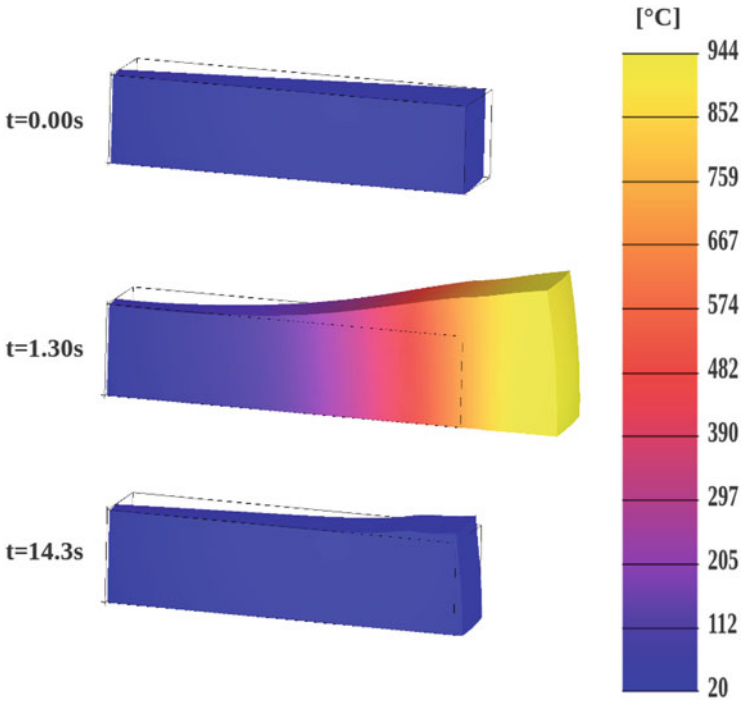


Fig. 5 Temperature evolution at the beginning of heating ($t = 0.00$ s), the beginning of cooling ($t = 1.3$ s), and the end of cooling ($t = 14.3$ s). The deformation is scaled by 40

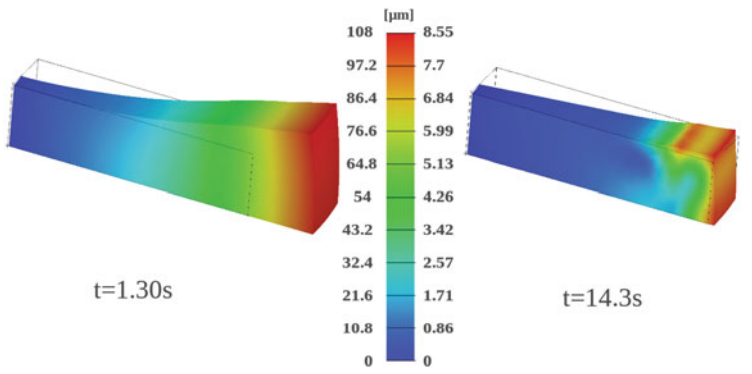


Fig. 6 Euclidean norm of the displacement at the beginning of cooling ($t = 1.3$ s), and the end of cooling ($t = 14.3$ s). The deformation is scaled by 40

Besides, the size change of disc diameter at $t = 14.3$ s (the end of cooling) has been calculated. In Fig. 7 it is obvious that at the boundary layer a maximal stretch of size 9.9 µm is observed. Compared with the original size of the workpiece the dimensional change is relatively slight. A comparison with

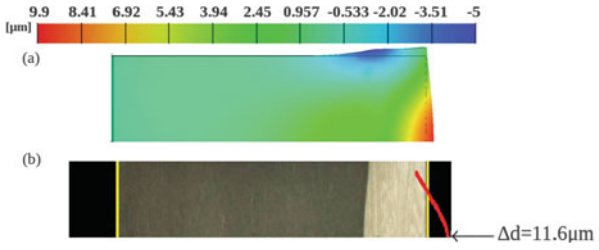


Fig. 7 Size changes of disc diameter; (**a**) simulated results scaled by 40, (**b**) experimental measurements performed by Stiftung Institut für Werkstofftechnik IWT, Bremen; deformations are depicted by the contour which is magnified 200 times

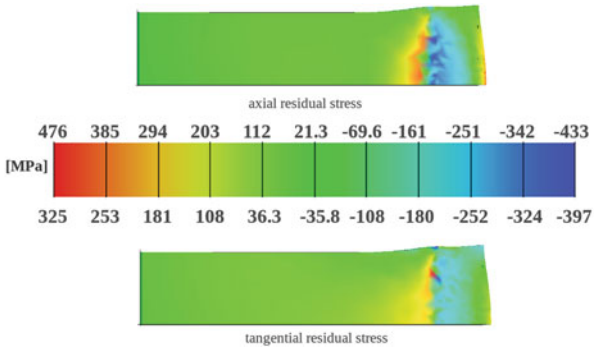


Fig. 8 Axial and tangential residual stresses at the sectional symmetry plane after induction heat treatment

experimentally measured values shows a very good conformance since the absolute error $|11.6 - 9.9|$ µm (cf. Fig. 7) is negligible in industrial practice.

The compressive residual stress distribution in the workpiece which is generated during the phase transformations is considered as one of the most important effects for the enhancement of strength. A numerical result of axial and tangential residual stresses after induction hardening is presented in Fig. 8.

5 Conclusions

In the mathematical treatment of the complete process of induction hardening including heating and quenching stages, a coupled problem of electromagnetics, thermomechanics and phase transitions is taken into account. The numerical simulations based on an FEM are carried out to predict the temperature evolution as well as mechanical behaviors.

Since only simple symmetric workpieces (disc) have been considered in the simulation, the solution with symmetric boundary conditions is not dependent on the angular coordinate and the segment angular openness is not relevant. Consideration of workpieces with complex geometries, helical gears for instance, should be a further application.

In addition, investigations of the effect of uncertain data for the simulation results, optimal control of the inductive heating under consideration of the growth of the high temperature phase austenite remain open problems.

Acknowledgements This research is a part of the project MeFreSim (Modeling, Simulation and Optimization of Multi-Frequency Induction Hardening) funded by Bundesministerium für Bildung und Forschung (BMBF). Furthermore we thank our industrial cooperation partners ZF Friedrichshafen AG and Dr. H. Stiele at EFD Induction GmbH for the technical support.

References

1. Fischer, F.D., Reisner, G., Werner, E., Tanaka, K., Cailletaud, G., Antretter, T.: A new view on transformation induced plasticity (TRIP). Int. J. Plast. **16**, 723–748 (2000)
2. Hömberg, D.: A mathematical model for induction hardening including mechanical effects. Real World Appl. **5**, 55–90 (2004)
3. Hömberg, D., Liu, Q., Urquizo, J.M., Nadolski, D., Petzold, T., Schmidt, A., Schulz, A.: Simulation of multi-frequency-induction-hardening including phase transitions and mechanical effects. Weierstraß-Institut für Angewandte Analysis und Stochastik, Leibniz-Institut im Forschungsverbund Berlin e.V. (2014, preprint). ISSN 2198-5855
4. Mioković, T.: Analyse des Umwandlungsverhaltens bei ein- und mehrfacher Kurzzeithärtung bzw. Laserstrahlhärtung des Stahls 42CrMo4. Dissertation, Universität Karlsruhe, Shaker Verlag Aachen, Band 2005, 25 (2005) ISBN: 3-8322-4689-4, Erschienen: Dezember 2005
5. Schröder, R.: Untersuchung zur Spannungs- und Eigenspannungsausbildung beim Abschrecken von Stahlzylindern. Dissertation, University of Karlsruhe (1985)
6. Urquizo, J.M., Liu, Q., Schmidt, A.: Quenching simulation for the induction hardening process—Thermal and mechanical effects. Berichte aus der Technomathematik, University of Bremen (2013)
7. Urquizo, J.M., Liu, Q., Schmidt, A.: Simulation of quenching involved in induction hardening including mechanical effects. Comput. Mater. Sci. **79**, 639–649 (2013)
8. Visintin, A.: Mathematical models of solid-solid phase transitions in steel. IMA J. Appl. Math. **39**, 143–157 (1987)
9. Wolff, M., Böhm, M., Böttcher, S.: Phase transformations in steel in the multi-phase case - general modelling and parameter identification. Technical Report 07-02, Universität Bremen, Berichte aus der Technomathematik (2007)

Tools for Aiding the Design of Photovoltaic Systems

Timo Rahkonen and Christian Schuss

Abstract This paper presents a collection of tools for aiding the design and system simulations of fixed and mobile photovoltaic energy harvesting systems. The presented tools help to estimate the available power, and to study the requirements of the maximum power point tracking. The effects caused by panel self-heating and active bypassing in series connected panel systems are studied.

1 Introduction

Photovoltaic (PV) modules are commonly used for harvesting renewable electrical energy, and numerous tools exist to estimate the performance of fixed installations [1, 2]. However, for simulating the performance of fast maximum power point tracking (MPPT) algorithms and partial shading on a moving PV installation, for example, a dedicated simulation setup needs to be built. This paper collects together all the core functionalities needed for such simulations, and discusses especially non-static effects. Section 2 reviews how the achievable insolation is calculated. Section 3 concentrates on the modelling of PV, emphasizing the double-diode behavior, and the effects of self-heating and bypassing. Finally, Sect. 4 concludes the paper.

2 Estimating the Insolation

A good estimate of Sun's trajectory is needed to maximize the energy collection of fixed solar panels. The most precise Equation of Time (EoT) models take into account the eccentricity of Earth's orbit and the axis tilt that cause a peak +/−5 min analemma error between the solar and local mean time. Yet, for estimating the daily insolation a very high time precision is not really needed, because the Sun's

T. Rahkonen (✉) • C. Schuss
University of Oulu, Oulu, Finland
e-mail: timo.rahkonen@oulu.fi; cschuss@ee.oulu.fi

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_15

azimuth and elevation trajectory during the day matters, but not the exact time of the solar noon, for example. Hence, a relatively simple algorithm [3, 4] can be used to estimate solar coordinates instead of the more precise equation-of-time like [5].

Much more important is the effect of systematic shadowing and average atmospheric conditions. Average insolation data is available from many measurement stations [6, 7], and also the aeronautical METAR weather reports [8] from the nearest airport can be used to estimate the average cloudiness and percentage of clear sky. Aviation typically reports cloud cover on a scale of (0–8)/8.

Our implementation for predicting daily and yearly insolation consists of the following. An equation-of-time function returns the Sun's azimuth (AZ) and elevation (EL) angle. Using the elevation angle, the effective atmospheric mass (AM) is calculated according to [9] to include the atmospheric losses, as illustrated in Fig. 1. The amount of diffuse (non-direct) illumination depends on the albedo of the surrounding terrain and the elevation of the viewing angle, but here a fixed 10 % is used for simplicity.

Next the Sun's position data is compared against a shadow mask storing the nearby obstacle profile: If the Sun is below the shadow elevation in a given azimuth angle, the direct sun ray is blocked and only the diffuse illumination is received. The obstacle profile can be collected by a mobile camera for example, if the field-of-view of the camera is known.

Altogether, the received radiation power P can be calculated by

$$P = I_o \times T_{\mathrm{AM}} \times ((1 - D) \times ShM \times A_{\mathrm{PV}} \times cos(\theta) + D \times A_{\mathrm{PV}}) \tag{1}$$

where I_o is the direct irradiance (about 1 kW/m^2), T_{AM} is the transmittance due to AM losses, D is the relative amount of diffuse illumination (here 0.1), ShM is a 0/1 shadow mask blocking or enabling the direct illumination, A_{PV} is the area of the PV panel and θ is the angle between PV's normal vector and Sun's direction vector. Due to the diffuse illumination term, even when the line-of-sight is blocked, a clear sky

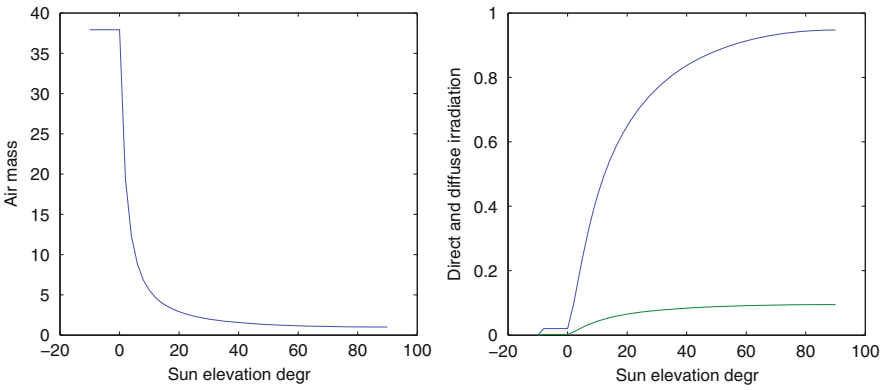


Fig. 1 Model of atmospheric mass (AM) and the direct and diffuse irradiation vs. Sun's altitude

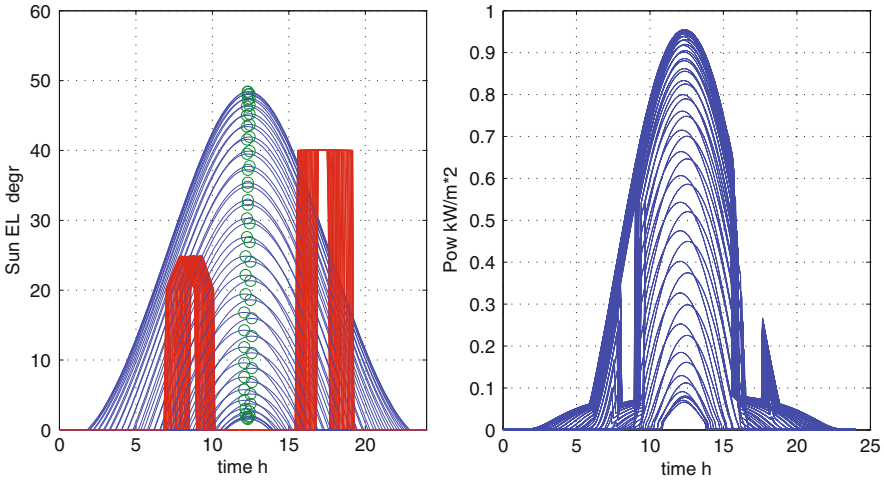


Fig. 2 Sun's trajectory and obstacles (*left*) and incident power and diffuse power (*right*) vs. time throughout the year in Oulu, Finland (65N)

gives diffuse irradiation equal to $D \times I_o$. Now it is easy to sweep the PV orientation to find the best orientation for achieving the maximum insolation. After the $P(time)$ response is known, it is easy to integrate the daily and yearly energy collection.

An example of using Eq. (1) is shown in Fig. 2 where the Sun's trajectory is plotted during the year in Oulu, Finland at lat. 65N. Data is calculated every 5 min, and once a week. Due to location near the arctic circle the Sun hardly sets in June, but its elevation is always lower than 50°, and around Christmas the Sun rises only a couple of degrees above the horizon. The red curve shows an example obstacle profile, and it is seen that in summer time the Sun shines above the neighboring building in the morning, but otherwise the building blocks the direct irradiation. The effect of this is shown in the right plot where the hourly irradiation is shown for a PV facing south at an elevation angle of 40°. The shallow tails model the diffuse illumination that comes from the sky and is not shadowed when the direct line-of-sight is broken or when the Sun is already on the back side of the PV.

The importance of the diffuse illumination is further illustrated in Fig. 5 where a small PV has been directly illuminated only for 3 h a day. In this case the diffuse illumination contributes 25 % of the total daily insolation.

The above calculations were verified by comparing them with measurements in Fig. 3, where the elevation of a south-facing PV is varied. It can be seen that in the beginning of April a vertical tracking does not affect much to the energy harvesting, as the maximum throughout the day is achieved at an elevation angle of ca. 30°. Instead, horizontal tracking would help in collecting the morning and evening sunlight.

The above is mostly aimed for fixed PV installations, but equally well in a mobile arrangement (e.g. on car's rooftop or in a bike) you need to know Sun's altitude.

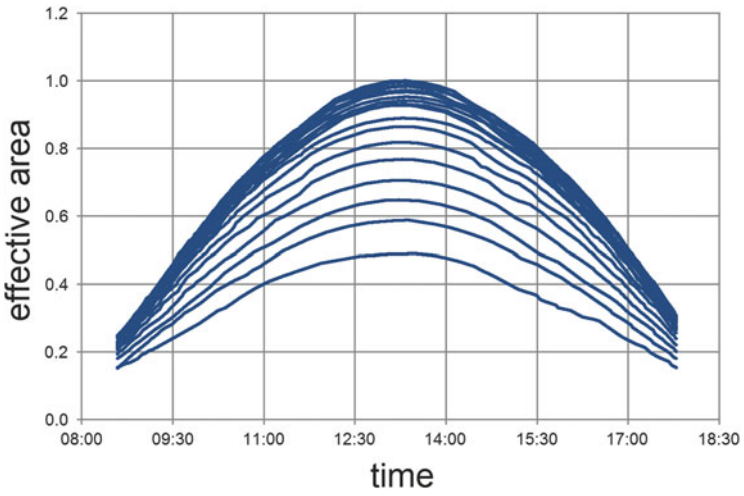


Fig. 3 Measured effective area of a south-facing PV with different elevation angles at 3 April. The *lowest curve* is flat horizontal alignment, then in 5° steps

Instead, the shadow profile is more complex and frequent, and this emphazises the need for a very quickly reacting maximum power point tracking.

3 Modeling the PV

3.1 The PV Model

A PV panel is usually modeled at circuit level by a parallel combination of the photocurrent source and a pn diode, and one series and one parallel resistance. The nonlinear nature of the diode current equation makes the solution of voltage iterative. To avoid convergence problems caused by the steep exponential response, the iteration steps taken by Newton-Raphson iteration algorithm need to be limited.

The I-V curve of the diode easily shows a double-diode behaviour due to recombination effects. This is illustrated in Fig. 4 where the short-circuit current I_{sc} and open-circuit voltage V_{oc} of a small garden lamp PV has been collected during the day. On a $V_{oc} - log(I_{sc})$ axis the points follow two lines with different slopes. The slope at low current levels is more shallow, and is caused by carrier recombination inside the PV. This can be easily modeled by presenting the diode current as a sum of two exponential functions with different parameters. This is inherently built into the Spice diode model, where it is controlled by parameters NR and ISR (emission coefficient and gain terms of the second diode, respectively).

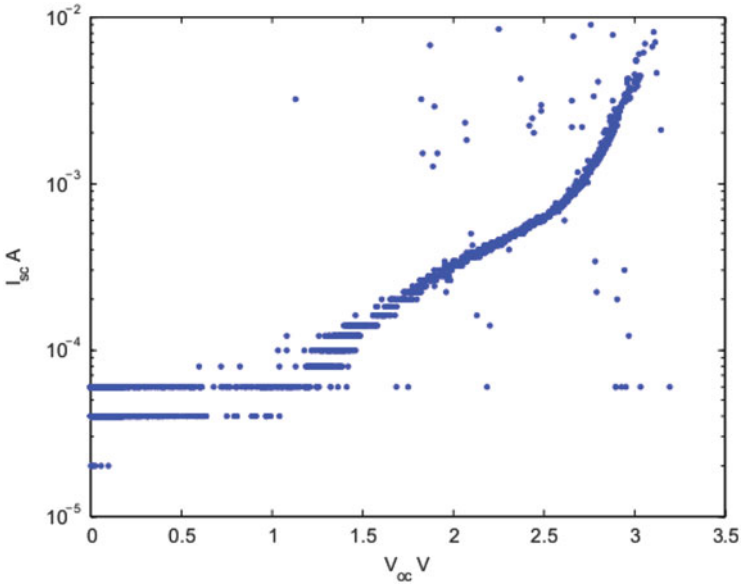


Fig. 4 Measured photovoltaic $log(I_{sc}) - V_{oc}$ pairs during a day

3.2 Ambient Effects in the PV

The power available from photovoltaics is very dependent on two ambient conditions: The irradiance and the temperature. The calculation of the insolation was shown above. Yet, also the effects of air/PV interface are important—for example, a cheap plastic cover of PV garden lamps can easily attenuate 15 % of the incoming light. Moreover, surface reflection depends on the angle of arrival, and this causes additional losses when the sunlight is almost parallel to the PV. At input angles higher than 70° the actual illumination can be some tens of percent lower than suggested by Eq. (1). This effect can easily be included into (1).

The characteristics of the diode depends heavily on the ambient temperature: The diode voltage typically drops by −1...−2 mV/degrees with increasing temperature, and a 30° change in the temperature can vary the available power by 10 %. Hence it is important to model temperature effects. Here not only the ambient temperature matters: the dark PV suffers from serious self-heating in bright daylight and warms up heavily. It is also worth noting that a PV transfers some power away from the module in electrical form: A PV biased in the maximum peak operating point heats less than a PV biased at V=V_{oc}, for example.

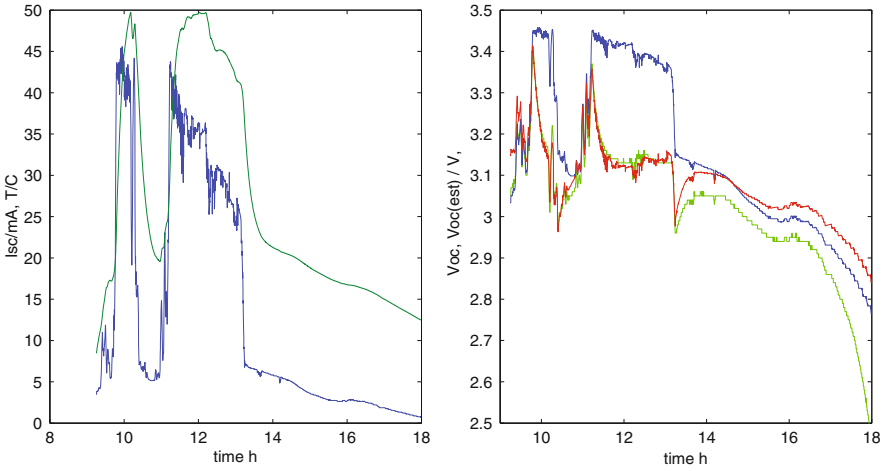


Fig. 5 *Left*: Measured I_{sc} (*blue*) and estimated junction temperature (*green*). *Right*: V_{oc} measured (*green*), estimated by $log(I_{sc})$ (*red*) and with the estimated junction temperature (*red*)

3.3 Estimating the Self-Heating

When measuring the response of PV panels, I_{sc} is very linearly proportional to the illumination, and V_{oc} is a good indication of the junction temperature of the PV. Figure 5 shows an example where a small PV is placed in sunlight in a cold day in March. When the Sun comes out of shadow, both the optical current I_{sc} and the open circuit voltage V_{oc} of the PV increase, but within 20 min the PV has heated up so much that the V_{oc} has reduced almost back to its initial value.

As the self-heating has 5–10 % effect to the operating point and achievable power, we built a lowpass thermal model similar to one thermal resistance and capacitance [10], where the junction temperature lags the incoming power (estimated by the measured I_{sc}). In the right plot of Fig. 5 the blue curve is the measured V_{oc}, green is the $V_{oc,est1}$ estimated on current and constant temperature, and the red one utilizes a self-heating model with a single 10 min time constant. The estimated junction temperature is shown in green in left plot. This modelling is important especially if the behavior of an energy harvester is analyzed in a vehicle, where shadow and illumination change quickly and often. During the first minutes in sunlight the efficiency is highest, and the MPP tracker needs to utilize that.

3.4 Modelling and Modifying the Connections

We typically have several PV modules in series or parallel connection. In a parallel connection a partial shadowing causes a proportional decrease in I_{sc} but only a slight decrease in V_{oc} value. In a (more typical) series connection one shadowed module

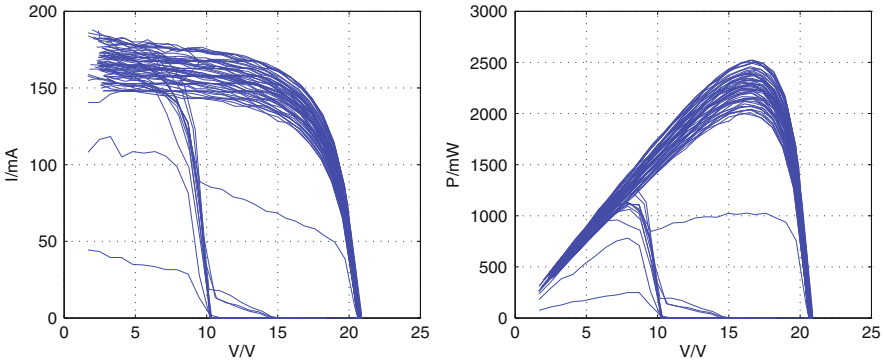


Fig. 6 Measured V-I curve and V-P curves of a partially shadowed PV with two series connected blocks and bypass diodes [11]

limits the current passing in all PVs. In this case it is very quickly beneficial to bypass the shadowed PV entirely, either by using bypass diodes or active switches. Bypassing causes a clear drop in the V_{oc} but allows the current flow, as shown in the measurements in Fig. 6.

Analyzing the connections and bypass circuits is easy in a circuit simulator that automatically solves the resulting voltages, but experimenting different controllers is difficult. To aid the system simulations, we built in Matlab a Newton-Raphson solver, where an arbitrary number of PVs and their bypass diodes can be simulated, together with changing illumination profiles and MPPT. The importance of this is that a partial shadowing causes a big jump in the maximum power operating point, and typical MPPT algorithms can react slowly or even get stuck to the wrong local maximum. Hence, modelling of this effect is needed for developing of fast and robust MPPT units.

3.5 Modelling the MPPT Requirements

The power of the PV depends on its operating voltage, and the voltage V_{mpp} for maximum power depends rather linearly on the ambient temperature and logarithmically on the strength of the illumination. The maximum is rather broad (made even more constant because of self-heating), and hence it makes sense to check how much the operating point may vary to maintain e.g. 90 % efficiency. This is illustrated on the right plot of Fig. 7 where the peak operating voltages and $+/-10\,\%$ limits are shown for -20 (upper) and $+80°$ (lower) temperatures for 1:1000 illumination range. It can be seen that a 1:10 illumination changes do not cause more than 10 % loss in efficiency even without any MPPT (provided that the operating power at maximum illumination is slightly below the maximum). A 1:100 illumination change causes already a similar operating point change as

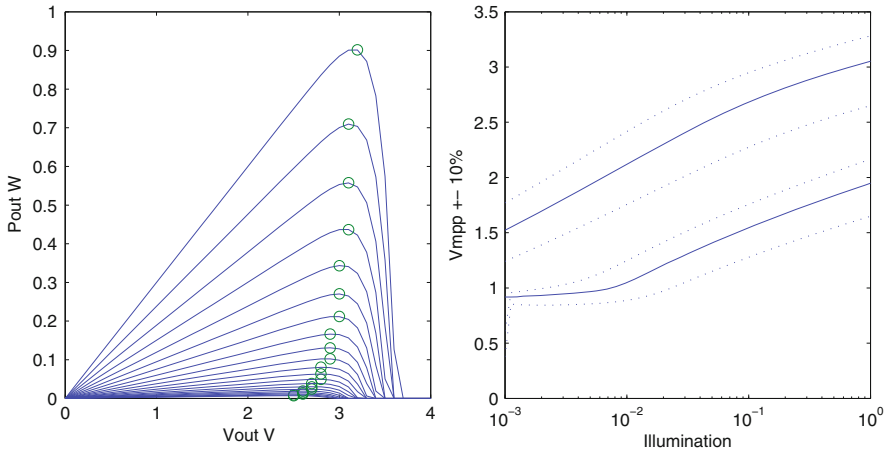


Fig. 7 *Left*: Simulated P-V curve and maximum power operating point of a PV, *right*: Maximum power operating point Vmpp and the limits of 90 % efficiency vs. illumination, at −20 and +80° temperatures

a 50° temperature change, and definitely needs to be compensated for. These kind of variations can be observed when driving a car on a road surrounded by trees or buildings, and in this case the MPPT controller needs to react to these very quickly.

4 Summary

We have built models for PV system simulation both into circuit simulator and for system simulations in Matlab environment, with an emphasis in studying moving PV systems with quickly varying operating conditions. In fixed installations the emphasis is on calculating Sun's position, obstacles, and directing the panels so that they collect maximum amount of energy. In movable installations the direction is arbitrary, but any way Sun's altitude is important to know. In either case, the amount of diffuse illumination is important.

Regarding PVs we showed two effects that affect the requirements set for the maximum power point tracking. First, the panels show self-heating with ca. 10 min time constant, and this both affects the achievable power and emphasizes the speed of the MPPT: If the PV is dominantly in shadow, it is very beneficial to react quickly to the first glimpse of direct sunlight before the PV heats up and efficiency drops. Second, the active bypassing of shadowed series connected panels causes very large changes in the optimum operating point, and the MPPT algorithm needs to be able to react quickly to these.

Acknowledgements Financial support of Infotech-DP doctoral programme is acknowledged.

References

1. Comparison of available solar photovoltaic software. Online document (2011). http://www.appropedia.org/Solar_photovoltaic_software/.Cited10Oct2014
2. Klise, G.T., Stein, J.S.: Models used to assess the performance of photovoltaic systems. SANDIA REPORT SAND2009-8258. Sandia National Laboratories Albuquerque, New Mexico 87185 and Livermore, CA 94550 (2009)
3. Khavrus, V., Shelevytsky, I.: Introduction to solar motion geometry on the basis of a simple model. Phys. Educ. **45**(6), 641 (2010). doi:10.1088/0031-9120/45/6/010
4. Whitman, A.M.: A simple expression for the equation of time. J. North American Sundial Society **14**, 29–33 (2007)
5. Grena, R.: An algorithm for the computation of the solar position. Solar Energy **82**(5), 462–470 (2008). ISSN 0038-092X. http://dx.doi.org/10.1016/j.solener.2007.10.001
6. National Solar Radiation Data Base (2015). http://rredc.nrel.gov/solar/old_data/nsrdb/
7. GAISMA website (2015). http://www.gaisma.com/en/location/oulu.html
8. ADDS - Aviation Digital Data Service METAR reports (2015). http://www.aviationweather.gov/adds/metars/
9. Kasten, F., Young, A.T.: Revised optical air mass tables and approximation formula. Appl. Opt. **28**, 4735–4738 (1989)
10. Ye, F.: Simulation and modeling-a CAD model for the BJT with self heating. IEEE Circuits Devices Mag. **7**(3), 7–9 (1991). doi:10.1109/101.79790
11. Rahkonen, T., Schuss, C., Hietanen, M., Kotikumpu, T., Mustajarvi, J., Myllymaki, A.: Electronics for characterizing and using photovoltaics. In: Proceedings of NORCHIP, Tampere, Finland, October 2014, pp. 1,4, 27–28 (2014). doi:10.1109/NORCHIP.2014.7004719

Part IV
Model Order Reduction

In Model Order Reduction focus is on error estimates and in balancing between moments (and thus size of the reduced model in the neighbourhood of an expansion point) and in using more expansion points. This applies both to frequencies as well as to parameter values. Here two papers address these points. Another point is how a reduced model is used. Is a bigger system with feed-back between output to input of the reduced model still stable: can the reduced model be made passive? And can application of a reduced model be beneficial in a system with multirate dynamics? These are considered in two other papers. For the use of parametric Model Order Reduction in Uncertainty Quantification we refer to the next part of of the book.

The paper by L. Feng, P. Meuris, W. Schoenmaker and P. Benner: *Parametric and Reduced-Order Modeling for the Thermal Analysis of Nanoelectronic Structures*, tackles parametric or parameterized Model Order Reduction. The method, developed by the Max Planck Institute, Magdeburg, Germany, was succesfully applied to an industrial heat problem of a protective package example, provided by the company MAGWEL, Leuven, Belgium. One assumes linear state-space formulations in which the occuring matrices depend on a (suitably chosen, abstract) parameter vector $\mathbf{p} = (p_1, \ldots, p_m)^T$, where the matrices depend linearly on $p_1, \ldots, p_m$. By this one can always evaluate the matrices when they have been determined in advance for a set $p_1^k, \ldots, p_m^k$, $k = 1, \ldots m$. How to select the p_i^k is part of an MOR process where adaptively expansion points for the frequency and for the parameters are chosen according to minimizing an a posteriori error estimate. This error estimate applies to any MOR approach. Here it is demonstrated in a Krylov-method framework for which a very efficient algorithm was developed for dealing with the expansion in the parameters. Inherently there is a choice between using more moments of expansions or using more expansion points - the choice affects the sparsity of the result of the reduced matrices.

The paper by S. Grivet-Talocia, A. Ubolli, A. Chinea and M. Bandinu: *On tuning passive black-box macromodels of LTI systems via adaptive weighting* considers linear, time-invariant systems for reduction and enforces passivity in a postprocessing step. The approach is demonstrated to a power distribution network. For given least-square weights and given scattering matrices $\hat{\mathbf{S}}_k$ between incident

($\mathbf{a}(j\omega)$) and reflected ($\mathbf{b}(j\omega)$) power waves with $\mathbf{b}(j\omega_k) = \hat{\mathbf{S}}_k\mathbf{a}(j\omega_k)$ one determines the transfer function $\mathbf{S}$ via Vector Fitting. Next for embedding the system into a bigger one it is assumed that $\mathbf{a} = \mathbf{Mb} + \mathbf{Nu}$ where $\mathbf{M}$ introduces feed-back effects and $\mathbf{u}$ is the independent source. The output of the bigger system is $\mathbf{y} = \mathbf{Pb} + \mathbf{Qu}$. The found $\mathbf{S}$ and $\mathbf{S}_k$ lead to corresponding $\mathbf{H}$ and $\mathbf{H}_k$ as transfers between $\mathbf{u}$ and $\mathbf{y}$. The least-square weights are adapted such that frequencies for which the errors between $\mathbf{H}$ and $\mathbf{H}_k$ are important are emphasized. Here the passivity enforcement comes into play for which some approaches are discussed.

The paper by E. Rita Samuel, L. Knockaert and T. Dhaene: *Multipoint Model Order Reduction using Reflective Exploration* exploits an adaptive technique for determining frequency points as expansion points, which is available in a toolbox developed at the University of Ghent during the last years.[1] It is demonstrated for the Krylov method PRIMA. For a given expansion point the order of the moment expansion can be increased by considering the difference between the last two approximation models. For the error a root mean square relative error at a discrete grid is considered. As long as the difference is too large the order is increased with the number of the ports. When the correction is small enough a new expansion point is selected from frequencies on the discrete grid where the error between the transfers of the last model and the original model is largest. The approach is applied to a transmission line model with six ports.

The paper by C. Hachtel, A. Bartel and M. Günther: *Interface Reduction for Multirate ODE-Solver* aims to an interesting goal: combining model order reduction to facilitate purposes to efficiently time integrate systems with different dynamics. These can be systems where the dynamics is different at different geometrical locations. It can also be a partitioning in coordinates of the unknown solution vector that involves different physical quantities. In both cases it is interesting to reduce a large part with low dynamics (slow part) and combine that with the remaining active part. Actually the slow part is linearized. Next the interface between active and slow part is considered. Such an interface offers more general coupling between models (not just between boundary values). Model Order Reduction automatically reveals which interface unknowns are more important than others. The approach is demonstarted on an example consisting of a fast circuit that is coupled to heat by a detailed heat model (slow part) of a resistor.

[1] http://www.sumo.intec.ugent.be/.

Parametric and Reduced-Order Modeling for the Thermal Analysis of Nanoelectronic Structures

Lihong Feng, Peter Meuris, Wim Schoenmaker, and Peter Benner

Abstract In this work, we discuss the parametric modeling for the thermodynamic analysis of components of nanoelectronic structures and automatic model order reduction of the consequent parametric models. Given the system matrices at different values of the parameters, we introduce a simple method of extracting system matrices which are independent of the parameters, so that parametric models of a class of linear parametric problems can be constructed. Then the reduced-order models of the large-scale parametric models are automatically obtained using a posteriori output error bounds for the reduced-order models. A thermal problem with conductance variations is studied as an example to illustrate the proposed parametric modeling and model order reduction techniques.

1 Introduction

Parameter variations have become essential and have to be taken into account in today's design of micro- and nano-electronic (-mechanical) problems as well as coupled electro-thermal problems. In design processes, modeling and simulation at many values of the parameters are necessary. For many simulation tools, modeling and simulation have to be done at several instances of the parameters. That means, given fixed values of the parameters, a certain numerical discretization method (e.g., the finite element method) is set up for that value, and numerical integration is then performed to get the output response corresponding to that value. The only available data from the software often are the (mass, damping) matrices corresponding to certain samples of the parameters.

It is desired that a single discretized system is valid for all possible values of the parameters, so that discretization does not have to be implemented anew for

L. Feng (✉) • P. Benner
Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany
e-mail: feng@mpi-magdeburg.mpg.de; benner@mpi-magdeburg.mpg.de

P. Meuris • W. Schoenmaker
Magwel NV, 3000 Leuven, Belgium
e-mail: peter.meuris@magwel.com; wim.schoenmaker@magwel.com

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_16

each fixed value, which can save much simulation time. In this paper, we propose a simple method of extracting matrices which are independent of any fixed value of the parameters. These matrices constitute a realization of a single parametric system which can be repeatedly used in simulation for any variations of the parameters. The approach is in particular suitable for the thermal analysis of nanoelectronic structures, as the parameters here often enter in a linear affine form which is needed for this extraction of a parametric model.

Simulating the consequent parametric system is, however, still very time consuming, because of the large scale of the system. We propose to use parametric model order reduction (PMOR) to compute a small reduced-order model (ROM), so that a single ROM is accurate for all possible values of the parameters. It is therefore cheap and sufficient to simulate only the ROM.

Different PMOR methods have been proposed so far; a survey of PMOR methods can be found in [2]. In this paper, we use a multi-moment-matching PMOR method [1] to construct the reduced-order model. These methods are popular in practical applications since they are easy to implement and need less computations than most of the other methods. Furthermore, we propose to use an a posteriori output error bound [4] to automatically construct the ROM. This makes model order reduction automatic.

The paper is organized as follows. In Sect. 2, we propose a simple method of extracting the state-space representation of a class of parametric problems. Section 3 reviews the basic idea of PMOR method based on multi-moment-matching. Section 4 and shows an algorithm that adaptively implements the multi-moment-matching PMOR method based on an a posteriori output error bound for the ROM. Section 5 addresses a thermal problem of a package, where the thickness of the top layer of the package varies and is taken as a parameter. We show that using any three samples of the parameters, one can easily extract a linear parametric system for the problem. A parametric ROM is automatically obtained using the proposed algorithm in Sect. 4, and meets the requirements of accuracy and compactness. Section 6 concludes the paper.

2 Parametric Modeling

In this section we introduce a method for extracting system matrices of a class of parametric problems, so that the parametric representation of the models in state-space form can be derived. Assume that the parametric problem can be generally described by the following differential equation,

$$\partial_t u(t,z;p) + \mathscr{L}[u(t,z;p)] = 0,\ t \in [0,T],\ z \in \Omega,\ p \in \mathscr{P},$$

where $\mathscr{L}[\cdot]$ is a linear spatial differential operator, $p = (p_1, \ldots, p_m)$ is a vector of parameters, $\Omega \subset \mathbb{R}^d (d = 1,2,3)$ is the spatial domain and $\mathscr{P} \in \mathbb{R}^m$ is the parameter domain.

Using finite-element simulation software, discretization in space can be done only at a fixed value p^* of p, and one may get the discretized system

$$\begin{aligned} E(p^*)\tfrac{dx(t,p)}{dt} &= A(p^*)x(t,p) + B(p^*)u(t), \\ y(t,p) &= C(p^*)x(t,p) + D(p^*)u(t), \end{aligned}$$

where only $E(p^*), A(p^*) \in \mathbb{R}^{n\times n}$, $B(p^*) \in \mathbb{R}^{n\times l_I}$, $C(p^*) \in \mathbb{R}^{l_O\times n}$, and $D(p^*) \in \mathbb{R}^{l_O\times l_I}$ at a fixed value p^* of p are available. Here, $x \in \mathbb{R}^n$ is the state vector, and $y \in \mathbb{R}^{l_O}$ is the output response. For design purposes, the simulation results at many fixed values of p should be derived and analyzed. If simply using the software, then the discretization must be repeated at many values of p. To avoid repeated discretization in space, and hence to save design time, it is desired that a parametric representation of the model is available.

We will show that if $E(p), A(p), B(p), C(p), D(p)$ are in the form of

$$M(p) = M_1 p_1 + \ldots + M_m p_m, \tag{1}$$

then one can easily compute $M_1, \ldots, M_m$ based on the data of $M(p)$ at m fixed values of p. Here and below, $M(p)$ stands for any of the matrices $E(p), A(p), B(p), C(p), D(p)$. Hence, the parametric representation of (2) is available, i.e.,

$$\begin{aligned} E(p)\tfrac{dx(t,p)}{dt} &= A(p)x(t,p) + B(p)u(t), \\ y(t,p) &= C(p)x(t,p) + D(p)u(t). \end{aligned} \tag{2}$$

The discretized parametric model in (2) prevents repeated discretization at all values of p. Notice that the parameters p_i may be abstract parameters, such as functions of the geometrical and/or physical parameters. Please also see the numerical example in Sect. 5, where $p_2 = 1/p$, and p is the layer thickness of the package.

Suppose that m groups of matrices $E(p^{a_i}), A(p^{a_i}), B(p^{a_i}), C(p^{a_i}), D(p^{a_i})$ have been obtained (e.g., by simulation software) at m different samples p^{a_i}, $i = 1, \ldots, m$. Using the formulation in (1), one can get a group of equations as below,

$$\begin{aligned} M_1 p_1^{a_1} + \ldots + M_m p_m^{a_1} &= M(p^{a_1}), \\ &\vdots, \\ M_1 p_1^{a_m} + \ldots + M_m p_m^{a_m} &= M(p^{a_m}). \end{aligned}$$

The above equations can be re-written as

$$(P_m \otimes I_n) \begin{pmatrix} M_1 \\ \vdots \\ M_m \end{pmatrix} = \begin{pmatrix} M(p^{a_1}) \\ \vdots \\ M(p^{a_m}) \end{pmatrix}, \tag{3}$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and

$$P_m = \begin{pmatrix} p_1^{a_1} & \dots & p_m^{a_1} \\ \vdots & & \vdots \\ p_1^{a_m} & \dots & p_m^{a_m} \end{pmatrix} \in \mathbb{R}^{m\times m}.$$

It is possible to select the samples p^{a_i} whose corresponding $m \times m$ matrix P_m in (3) is nonsingular, such that

$$\begin{pmatrix} M_1 \\ \vdots \\ M_m \end{pmatrix} = (P_m^{-1} \otimes I_n) \begin{pmatrix} M(p^{a_1}) \\ \vdots \\ M(p^{a_m}) \end{pmatrix},$$

where the property of tensor product (Kronecker product): $(U\otimes Q)^{-1} = U^{-1}\otimes Q^{-1}$, $\forall U \in \mathbb{R}^{n_U\times n_U}$, and $Q \in \mathbb{R}^{n_Q\times n_Q}$ invertible, is used. Finally, the matrices M_i, $i = 1, \dots, m$, can be easily computed from the following equations,

$$\begin{aligned} M_1 &= M(p^{a_1})\tilde{p}_{11} + \dots + M(p^{a_m})\tilde{p}_{1m}, \\ &\vdots, \\ M_m &= M(p^{a_1})\tilde{p}_{m1} + \dots + M(p^{a_m})\tilde{p}_{mm}. \end{aligned} \tag{4}$$

Here,

$$P_m^{-1} = \begin{pmatrix} \tilde{p}_{11} & \dots & \tilde{p}_{1m} \\ \vdots & & \vdots \\ \tilde{p}_{m1} & \dots & \tilde{p}_{mm} \end{pmatrix}.$$

One important property of the above computation is that it is independent of the large dimension n. For any large-scale matrices in (2), only the inverse of a small $m \times m$ matrix P_m is needed to compute all M_i, for $M_i = E_i, A_i, B_i, C_i, D_i$, $i = 1, \dots, m$.

Simulating the system in (2) may still take a lot of time when the dimension n is large and when it has to be simulated at many samples of p. In the next section we propose to use PMOR to construct a parametric reduced-order model, so as to replace the original large system in (2) in simulations. Since the size of the reduced-order model is usually much smaller than n, simulation can be accomplished in a much shorter and reasonable time period.

3 PMOR Based on Multi-Moment-Matching

There are various PMOR methods in the literature, among which the method based on multi-moment-matching is probably the most easy one to implement, and has a low computational complexity [1]. The multi-moment-matching PMOR method computes a projection matrix V based on the series expansion of the state vector x in the frequency domain. The system in (2) in the frequency domain is

$$\begin{aligned}(sE(p) - A(p))x(s,p) &= B(p)u(s),\\ y(s,p) &= C(p)x(s,p) + D(p)u(s).\end{aligned} \tag{5}$$

Given expansion points $p^0 = [p_1^0, \ldots, p_m^0]$, and s_0, $x(s,p)$ in (5) can be expanded as

$$\begin{aligned}x(s,p) &= [I - (\sigma_1 G_1 + \ldots + \sigma_m G_m + \sigma_{m+1} G_{m+1} + \ldots + \sigma_{2m} G_{2m})]^{-1} B_M u(s)\\ &= \sum_{i=0}^{\infty} (\sigma_1 G_1 + \ldots + \sigma_{2m} G_{2m})^i B_M u(s),\end{aligned}$$

where $\sigma_i = sp_i - s_0 p_i^0$, $\sigma_{m+i} = p_i - p_i^0$, $G_i = -[s_0 E(p^0) - A(p^0)]^{-1} E_i$, $G_{m+i} = [s_0 E(p^0) - A(p^0)]^{-1} A_i$, $i = 1, 2, \ldots, m$, and $B_M = [s_0 E(p^0) - A(p^0)]^{-1} B(p)$.

Defining $R_0 = [s_0 E(p^0) - A(p^0)]^{-1}[B_1, \ldots, B_m]$ and $R_j = [G_1, \ldots, G_p] R_{j-1}$, $j = 1, \ldots, q$, a matrix V_{s_0,p^0}, whose columns form an orthonormal basis of the subspace spanned by the R_i's, is computed as

$$\text{range}\{V_{s_0,p^0}\} = \text{span}\{R_0, R_1, \ldots, R_q\}_{s_0,p^0}. \tag{6}$$

Using $V := V_{s_0,p^0}$, we obtain the parametric reduced-order model via Galerkin projection,

$$\begin{aligned}V^T E(p) V \frac{dx_r(t,p)}{dt} &= V^T A(p) V x_r(t,p) + V^T B(p) u(t),\\ y_r(t,p) &= V^T C(p) V x_r(t,p) + D(p) u(t).\end{aligned}$$

Notice that the number of columns in R_j increases exponentially with j. When the number of the parameters in p is larger than 2, or when there are many inputs, multiple point expansion should be used to keep the size of the reduced-order model as small as possible. The idea is straight forward. Given a group of expansion points $s_i, p^i, i = 0, \ldots, k$, (the superscript i for p is not a power, it only indicates the ith expansion point), a matrix V_{s_i,p^i} can be computed for each pair s_i, p^i as

$$\text{range}\{V_{s_i,p^i}\} = \text{span}\{R_0, R_1, \ldots, R_{q_r}\}_{s_i,p^i}. \tag{7}$$

The final projection matrix V is obtained from the orthogonalization of all matrices V_{s_i,p^i},

$$V = \text{orth}\{V_{s_0,p^0}, \ldots, V_{s_k,p^k}\}. \tag{8}$$

For similar accuracy, the number q_r in (7) can usually be taken much smaller than q in (6). Consequently, the size of the reduced-order model can be kept small.

Different expansion points s_i, p^i may lead to reduced-order models with different accuracy. In the next section, we introduce a technique for adaptively selecting the expansion points according to an a posteriori error bound $\Delta(s,p)$ for the ROM. The error bound guarantees the reliability of the reduced-order model, while providing a way of automatically constructing the ROM.

4 Adaptively Selecting the Expansion Points

For a multiple-input and multiple-output (MIMO) system, the error bound $\Delta(s,p)$ is defined as

$$\Delta(s,p) = \max_{ij} \Delta_{ij}(s,p).$$

Here $\Delta_{ij}(s,p)$ is the error bound for the (i,j)th entry of the transfer function matrix of the ROM, i.e.,

$$|H_{ij}(s,p) - \hat{H}_{ij}(s,p)| \le \Delta_{ij}(s,p).$$

For a single-input and single-output system, there is no need to take the maximum. $\Delta_{ij}(s,p)$ can be computed as

$$\Delta_{ij}(s,p) = \frac{||r_i^{du}(s,p)||_2 ||r_j^{pr}(s,p)||_2}{\beta(s,p)} + |(\hat{x}^{du})^* r_j^{pr}(s,p)|,$$

where $r_j^{pr}(s,p) = b_j - [sE - A]\hat{x}^{pr}$, $\hat{x}^{pr} = V(sV^TEV - V^TAV)^{-1}V^Tb_j$, $r_i^{du}(s,p) = -c_j^T - [\bar{s}E^T - A^T]\hat{x}^{du}$, $\bar{s}$ is the conjugate of s, and $\hat{x}^{du} = -V^{du}[\bar{s}(V^{du})^TE^TV^{du} - (V^{du})^TA^TV^{du}]^{-1}(V^{du})^Tc_j^T$. Here for ease of notation, p is dropped from the matrices $E(p), A(p), B(p), C(p)$. b_j is the jth column of $B(p)$. c_i is the ith row of C(p). The variable $\beta(s,p)$ is the smallest singular value of the matrix $sE(p) - A(p)$. The matrix V^{du} can be computed, for example, using (7) and (8), by replacing $R_0, \ldots, R_{qr}$ with $R_0^{du}, R_1^{du}, \ldots, R_{qr}^{du}$, where the matrices $s_iE(p^i) - A(p^i)$ in $R_0, \ldots, R_{qr}$ are substituted by $\bar{s}_iE^T(p^i) - A^T(p^i)$, and E_j by E_j^T, A_j by A_j^T, B_j by B_j^T, $j = 1, \ldots, m$. The derivation of $\Delta(s,p)$ is detailed in [4].

Given the error bound $\Delta(s,p)$ for the ROM, the expansion points p^i, s_i can be adaptively selected, and the projection matrix V can be automatically computed as shown in Algorithm 1. It is worth pointing out that although the error bound is parameter-dependent, many p-independent terms constituting the error bound can be precomputed only once, and repeatedly used in the algorithm for the many samples of p in Ξ_{train}, e.g., the terms $V^TM_1V, \ldots, V^TM_mV$, etc..

Algorithm 1 Adaptively selecting expansion points $\hat{s}$, $\hat{p}$, and automatically computing V

$V = []$; $V^{du} = []$;
Choose some $\varepsilon_{tol} < 1$; set $\varepsilon = 1$;
Choose Ξ_{train}: a set of samples of s and p, taken over the interesting domain;
Choose initial expansion points: $\hat{s}$, $\hat{p}$;
while $\varepsilon > \varepsilon_{tol}$ **do**
  range$(V_{\hat{s}}, \hat{p}) = \text{span}\{R_0, R_1, \ldots, R_{q_r}\}_{\hat{s},\hat{p}}$;
  range$(V^{du}_{\hat{s},\hat{p}}) = \text{span}\{R^{du}_0, R^{du}_1, \ldots, R^{du}_{q_r}\}_{\hat{s},\hat{p}}$;
  $V = \text{orth}\{V, V_{\hat{s},\hat{p}}\}$;
  $V^{du} = \text{orth}\{V^{du}, V^{du}_{\hat{s},\hat{p}}\}$;
  $(\hat{p}, \hat{s}) = \arg \max_{s,p \in \Xi_{train}} \Delta(s, p)$;
  $\varepsilon = \Delta(\hat{s}, \hat{p})$;
end while.

5 Numerical Experiments

We take a thermal model of a package [3] to study the proposed techniques. Integrated circuits are put into protective packages to allow easy handling and assembly onto printed circuit boards and to protect the devices from damage. The heat flowing in the package is produced in the integrated circuit and in the electrical leads.

A finite-integration technique (FIT) for the modeling of the package leads to thermal fluxes that are proportional to the dual areas of the mesh cells and inversely proportional to the lengths of the edges in the mesh cells. Therefore, when considering meshes that are topologically equivalent for different package thicknesses, the parametric dependence of the matrices will take the form as

$$M(p) = M_0 + pM_1 + 1/pM_2,$$

where p is the package thickness. The second term originates from the linear dependence of dual areas corresponding to the cell edges perpendicular to the thickness, whereas the third term originates from dual areas associated to cell edges tangential to the thickness orientation.

It is clear that the above formulation is a special case of (1), where p_1 is fixed as $p_1 = 1$, and $p_2 = p$, $p_3 = 1/p$. Only the inverse of a 3×3 matrix needs to be computed. After the inverse of the 3×3 matrix is obtained, the equations in (4) can be used to compute M_0, M_1, M_2. The final parametric system is in the form of (2). It is a MIMO system, with 34 inputs and 68 outputs.

Furthermore, Algorithm 1 is employed to automatically compute the parametric reduced-order model. We used 6 samples of $p \in (0, 100]$, and one sample of $s = 2\pi fj$, $f \in [0, 10^8]$: $s_0 = 200\pi j$, $j = \sqrt{-1}$, to constitute the training set Ξ_{train} in Algorithm 1. The algorithm essentially selects the expansion points for p, since we force a single expansion point for s. There are two iterations, and two expansion

points are selected for p. The reduced model is of size $r = 58$. For each selected expansion point, we construct V_{s_i,p^i} with only two terms R_0 and R_1 ($q_r = 1$ in Steps 6–7, Algorithm 1), in order to avoid the exponential increase in R_j, $j > 1$. Table 1 lists the iterations and the error bounds at each iteration step. Figures 1 and 2 plot the temperature and the current at two different parts of the package. The temperature is of big magnitude, while the current is of very small magnitude, showing that there is no current at that part of the circuit. The reduced model catches the accuracy of both at 120 samples of p, and 100 time steps for each sample.

Table 1
$V_{s_i,p^i} = \text{span}\{R_0, R_1\}_{s_i,p^i}$,
$i = 1, 2, \varepsilon_{tol} = 10^{-3}$,
$n = 8549$, $r = 58$

Iteration i	(s_0, p^i)	$\Delta(s_0, p^i)$
1	$(0.3834, 200\pi j$	0.0153
2	$(0.0677, 200\pi j$	5×10^{-4}

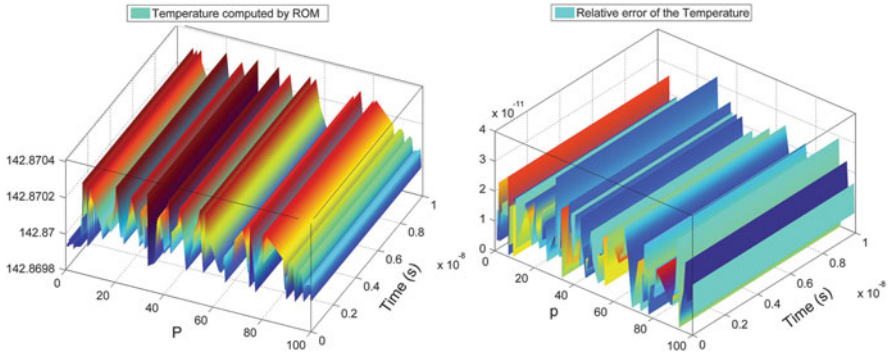


Fig. 1 FOM $n = 8549$, ROM $r = 58$, inputs 34, outputs 68. *Left*: temperature computed by the ROM. *Right*: relative error of the temperature computed by the ROM

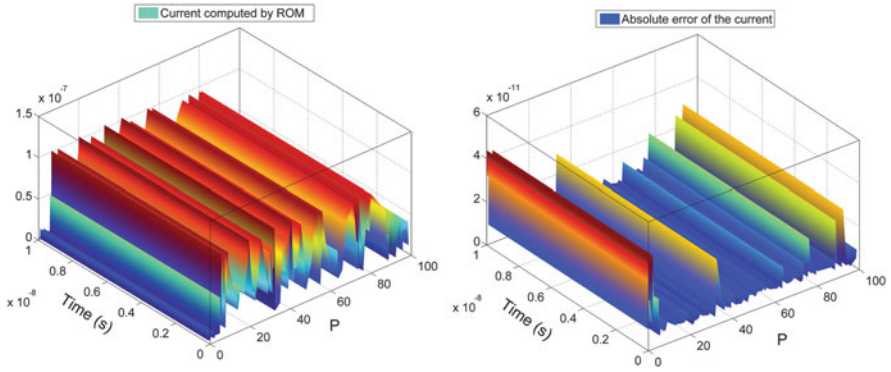


Fig. 2 FOM $n = 8549$, ROM $r = 58$, inputs 34, outputs 68. *Left*: current computed by the ROM. *Right*: Absolute error of the current computed by the ROM

6 Conclusions

We have proposed a simple automatic parametric modeling technique for a class of linear parametric problems, and shown that automatic parametric model order reduction can be realized with the guidance of an a posteriori error bound. The above techniques have been successfully applied to a thermal problem of a package. A compact and reliable reduced-order model has been automatically obtained, which offers the possibility of being integrated into dedicated electro-thermal simulation software to accelerate design automation.

Acknowledgements This work is supported by the collaborative project nanoCOPS, Nanoelectronic COupled Problems Solutions, supported by the European Union in the FP7-ICT-2013-11 Program under Grant Agreement Number 619166.

References

1. Benner, P., Feng, L.: A robust algorithm for parametric model order reduction based on implicit moment matching. In: Quarteroni, A., Rozza, G. (eds.) Springer MS &A Series. Reduced Order Methods for Modeling and Computational Reduction, vol. 8, pp. 159–185. Springer International Publishing, Switzerland (2014)
2. Benner, P., Gugercin, S., Willcox, L.: A survey of model reduction methods for parametric systems. SIAM Review **57**(4), 483–531 (2015)
3. Benner, P., Feng, L., Schoenmaker, W., Meuris, P.: Parametric modeling and model order reduction of coupled problems. ECMI Newsl. **56**, 68–69 (2014). Available from http://www.mafy.lut.fi/EcmiNL/issues.php
4. Feng, L., Antoulas, A.C., Benner, P.: Some a posteriori error bounds for reduced order modelling of (non-)parametrized linear systems MPI Magdeburg Preprint MPIMD/15-17, October 2015. Available from: http://www.mpi-magdeburg.mpg.de/preprints/.

On Tuning Passive Black-Box Macromodels of LTI Systems via Adaptive Weighting

Stefano Grivet-Talocia, Andrea Ubolli, Alessandro Chinea, and Michelangelo Bandinu

Abstract This paper discusses various approaches for tuning the accuracy of rational macromodels obtained via black-box identification or approximation of sampled frequency responses of some unknown Linear and Time-Invariant system. Main emphasis is on embedding into the model extraction process some information on the nominal terminations that will be connected to the model during normal operation, so that the corresponding accuracy is optimized. This goal is achieved through an optimization based on a suitably defined cost function, which embeds frequency-dependent weights that are adaptively refined during the model construction. A similar procedure is applied in a postprocessing step for enforcing model passivity. The advantages of proposed algorithm are illustrated on a few application examples related to power distribution networks in electronic systems.

1 Introduction

Several engineering design flows are often based on a partial knowledge of the dynamic behavior of individual devices, components, or subsystems. This situation arises when such components are measured with finite resolution, when the corresponding responses are obtained from finite-precision numerical simulation of first-principle field equations, or even when these responses are available from a component vendor. In order to use such components, suitable simulation models are required, in order to verify full system performance since early design stages.

In this work, we concentrate on electronics applications, for which reliable models of the Power Distribution Network (PDN) at chip, package, board and system level are required [1–3]. The PDN can be regarded as a large-scale Linear and Time-Invariant (LTI) dynamic system [4, 5]. A first-principle formulation would

S. Grivet-Talocia (✉) • A. Ubolli
Department of Electronics and Telecommunications, Politecnico di Torino, C. Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: stefano.grivet@polito.it; andrea.ubolli@polito.it

A. Chinea • M. Bandinu
IdemWorks s.r.l., C. Trapani 16, 10139 Torino, Italy
e-mail: a.chinea@idemworks.com; m.bandinu@idemworks.com

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_17

lead to a state-space or descriptor formulation with billions of states and hundreds of inputs/outputs. However, such detailed first-principle descriptions are usually not available to the power integrity engineer, who is responsible for compliance verification at the system level. Moreover, even if such descriptions were available, the resulting complexity of system-level verification would be overwhelming. Hence, there is a strong need for accurate and broadband reduced-order models.

We concentrate here on the construction of state-space PDN macromodels in a black-box setting, via identification from a finite set of frequency response samples. The main tool that we employ is frequency-domain rational approximation, for which several good algorithms exist, such as Vector Fitting [6–10], followed by a postprocessing step aimed at enforcing passivity [11–15]. Passivity is in fact a fundamental requirement for ensuring model robustness and global stability of successive system-level transient simulations.

The main problem that we address is the sensitivity of the state-space macromodel to the termination networks to which the model will be connected during normal operation. This sensitivity may be the root cause for major accuracy degradation, so that a model that is very accurate in the input-output representation that is adopted for its construction may result quite inaccurate during normal operation. This degradation results from the feedback mechanisms that the terminations induce on the model dynamics [16, 17].

We propose a simple algorithm to alleviate this accuracy degradation, based on the definition of suitably and adaptively defined frequency-dependent weights, which are used to construct an optimized cost function embedding information on the nominal termination scheme for the model. Minimization of this cost function during model identification and passivity enforcement leads to an effective compensation of the model sensitivity, with resulting improved accuracy. Various examples from real applications demonstrate the effectiveness of this approach.

2 Problem Statement

Let us consider a P-port PDN system, known through a set of K frequency samples of its $P \times P$ scattering matrix

$$\hat{\mathbf{S}}_k \approx \hat{\mathbf{S}}(\mathrm{j}\omega_k), \quad k = 1, \ldots, K. \tag{1}$$

The scattering representation is such that $\mathbf{b}(\mathrm{j}\omega) = \hat{\mathbf{S}}(\mathrm{j}\omega)\mathbf{a}(\mathrm{j}\omega)$ where $\mathbf{a}, \mathbf{b}$ are the power waves that are incident into and reflected from the structure, respectively. This representation is preferred here since it is guaranteed to exist for any LTI system. We want to construct a regular state-space model

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{a}(t) \\ \mathbf{b}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{a}(t) \end{aligned} \tag{2}$$

with transfer (scattering) matrix

$$\mathbf{S}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}\,, \tag{3}$$

so that

- a cumulative least squares fitting error with respect to the original data (1)

$$E_w^2 = \sum_{k=1}^{K} E_{w,k}^2 = \sum_{k=1}^{K} w_k^2 \|\mathbf{S}(\mathrm{j}\omega_k) - \hat{\mathbf{S}}_k\|_F^2 \tag{4}$$

  is minimized, where w_k are appropriate frequency-dependent weights and $_F$ denotes the Frobenius norm;
- the model is passive, so that

$$\sigma_{\max}(\mathbf{S}(\mathrm{j}\omega)) = \|\mathbf{S}(\mathrm{j}\omega)\|_2 \le 1\,, \quad \forall \omega \in \mathbb{R}\,, \tag{5}$$

  where $\sigma_{\max}$ denotes the maximum singular value of its matrix argument.

In standard applications, the weights in (4) are uniformly set to $w_k = 1$, or at best to $w_k = 1/\varsigma_k$ when the variance ς_k^2 of noise affecting raw data is known. Here, we want to construct these weights such that a second objective is met. We assume that the nominal termination scheme is fully known and characterized in the frequency-domain as

$$\begin{aligned} \mathbf{a}(s) &= \mathbf{M}(s)\mathbf{b}(s) + \mathbf{N}(s)\mathbf{u}(s), \\ \mathbf{y}(s) &= \mathbf{P}(s)\mathbf{b}(s) + \mathbf{Q}(s)\mathbf{u}(s), \end{aligned} \tag{6}$$

where $\mathbf{u}$ is a vector collecting independent sources embedded in the termination network, $\mathbf{y}$ collects the output variables of interest, and $\mathbf{M}, \mathbf{N}, \mathbf{P}, \mathbf{Q}$ are suitable transfer matrices. Note that the port inputs $\mathbf{b}$ of the termination network (6) are the outputs of the macromodel (2), and viceversa. Our objective is minimization of the error

$$\Delta^2 = \sum_{k=1}^{K} \Delta_k^2 = \sum_{k=1}^{K} \|\mathbf{H}(\mathrm{j}\omega_k) - \hat{\mathbf{H}}_k\|_F^2\,, \tag{7}$$

where $\mathbf{H}(\mathrm{j}\omega_k)$ and $\hat{\mathbf{H}}_k$ are the transfer functions between input $\mathbf{u}$ and output $\mathbf{y}$, based on the model $\mathbf{S}(\mathrm{j}\omega_k)$ of (3) and on the raw data $\hat{\mathbf{S}}_k$ of (1), respectively.

3 Iterative Rational Approximation via Adaptive Weighting

A simple first-order approximation of the relationship between the frequency-dependent model error E_k and transfer function error Δ_k leads to

$$\Delta_k \approx \mathscr{S}_k E_k, \tag{8}$$

where $\mathscr{S}_k$ can be interpreted as a sensitivity of $\mathbf{H}(\mathrm{j}\omega_k)$ with respect to perturbations in the model responses $\mathbf{S}(\mathrm{j}\omega_k)$ under nominal termination conditions (6). Therefore, if we set $w_k = \mathscr{S}_k$ and we minimize (4) during model construction, we expect that the resulting model will achieve an equivalent minimization of (7). As documented in [16], this approach still leaves margins for improvement, in addition to requiring the explicit computation of the sensitivity. We resort to a simpler and more effective iterative approach, based on the following steps.

1. At the first iteration $\mu = 0$, we initialize the weights as $w_k^{(0)} = 1$ for all k.
2. For each iteration $\mu = 0, 1, \ldots$, we compute a state-space macromodel (3) by minimizing (4). This is obtained by a standard application of the Vector Fitting (VF) algorithm [6–10].
3. Once the model is available, the corresponding frequency-dependent transfer function error $\Delta_k^{(\mu)}$ is computed. If $\Delta_k^{(\mu)} < \delta$ at all frequencies, where δ is the desired target accuracy, the iteration is stopped.
4. Otherwise, a new frequency-dependent weight for next iteration is defined as

$$w_k^{(\mu+1)} = w_k^{(\mu)} \cdot \mathscr{F}(\Delta_k^{(\mu)}), \tag{9}$$

   where $\mathscr{F} : \mathbb{R}^+ \mapsto \mathbb{R}^+$ denotes a non-decreasing function such that $\mathscr{F}(\xi) = 1$ for $\xi \leq \delta$. Then, the iteration index is increased $\mu \leftarrow \mu + 1$, and the scheme is restarted from step 2.

The redefinition of the weights in (9) further emphasizes those frequencies for which the transfer function error is significant, without affecting the other frequencies. The result of this process is both the termination-tuned model $\mathbf{S}(s)$ and the corresponding set of optimal weights w_k. The convergence properties of this iteration are related to the specific choice of $\mathscr{F}$. A detailed convergence analysis is in progress and will be documented in a future report.

4 Passivity Enforcement

Once a state-space macromodel is available, its passivity should be verified. We perform this check by computing the set $\mathscr{I}$ including all purely imaginary eigenvalues $\lambda_i = \mathrm{j}\omega_i$ of the associated Hamiltonian matrix [12] (we assume

$\|\mathbf{D}\|_2 \leq 1$)

$$\mathscr{M} = \begin{pmatrix} \mathbf{A} + \mathbf{B}(\mathbf{I} - \mathbf{D}^\mathsf{T}\mathbf{D})^{-1}\mathbf{D}^\mathsf{T}\mathbf{C} & \mathbf{B}(\mathbf{I} - \mathbf{D}^\mathsf{T}\mathbf{D})^{-1}\mathbf{B}^\mathsf{T} \\ -\mathbf{C}^\mathsf{T}(\mathbf{I} - \mathbf{D}\mathbf{D}^\mathsf{T})^{-1}\mathbf{C} & -\mathbf{A}^\mathsf{T} - \mathbf{C}^\mathsf{T}\mathbf{D}(\mathbf{I} - \mathbf{D}^\mathsf{T}\mathbf{D})^{-1}\mathbf{B}^\mathsf{T} \end{pmatrix}. \tag{10}$$

If $\mathscr{I}$ is empty, the model is already passive and no other action is required. Otherwise, the model needs to be corrected to eliminate local passivity violations, intended as violations of condition (5) within localized frequency bands $\Omega_i = (\omega_i, \omega_{i+1})$. The boundary points of each violation band Ω_i correspond to the imaginary part of some Hamiltonian eigenvalue in set $\mathscr{I}$.

The passive model to be determined is parameterized by perturbing the state-output map $\tilde{\mathbf{C}} = \mathbf{C} + \Delta\mathbf{C}$, corresponding to a model perturbation

$$\tilde{\mathbf{S}}(s) = \mathbf{S}(s) + \Delta\mathbf{S}(s), \quad \Delta\mathbf{S}(s) = \Delta\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\,. \tag{11}$$

A set of local passivity constraints is obtained by considering each individual singular value trajectory $\sigma_r(\mathrm{j}\omega)$ that exceeds one within a given violation band Ω_i, finding its local maximum $\bar{\sigma}_{i,r} = \sigma_r(\mathrm{j}\bar{\omega}_{i,r})$ with $\bar{\omega}_{i,r} \in \Omega_i$, and linearizing the relationship between this singular value and the decision variables $\Delta\mathbf{C}$. Imposing that this linearized singular value falls below one gives the linear inequality constraints

$$\mathbf{z}_{i,r}^\mathsf{T}\mathrm{vec}(\Delta\mathbf{C}) \leq 1 - \bar{\sigma}_{i,r}, \quad \forall i, r\,, \tag{12}$$

to be enforced concurrently while minimizing the model perturbation (11).

Most existing passivity enforcement schemes [11–15] aim at minimizing the $\mathscr{L}_2$ norm of the model perturbation, which can be characterized as

$$\|\Delta\mathbf{S}\|_2^2 = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \|\Delta\mathbf{S}(\mathrm{j}\omega)\|_2^2\,\mathrm{d}\omega = \mathrm{tr}\left(\Delta\mathbf{C}\,\mathbf{G}_c\,\Delta\mathbf{C}^\mathsf{T}\right), \tag{13}$$

where $\mathbf{G}_c$ is the controllability Gramian of the original model. Minimization of (13) subject to (12) optimizes the model accuracy, but may degrade the accuracy of the target transfer function $\mathbf{H}(s)$, since no weighting is considered. We propose two different approaches to overcome this limitation.

The first approach is to consider a frequency-weighted controllability Gramian $\mathbf{G}_w$ instead of $\mathbf{G}_c$ in (13). This Gramian is constructed based on an augmented state-space system providing a realization of

$$\Delta\mathbf{S}_w(s) = \Delta\mathbf{S}(s)F(s)\,, \tag{14}$$

where $F(s)$ is a minimum-phase transfer function such that $|F(\mathrm{j}\omega_k)|^2 \approx w_k^2$, where w_k are the optimal weights from the fitting. More details can be found in [17].

A second and more straightforward approach is to construct a data-based cost function. We consider the model deviation at frequency $\mathrm{j}\omega_k$, which we write as

$$\mathscr{E}_k^2 = \left\| \tilde{\mathbf{S}}(\mathrm{j}\omega_k) - \hat{\mathbf{S}}_k \right\|_F^2 = \left\| \Delta\mathbf{C}\,\mathbf{K}_k + \mathbf{S}(\mathrm{j}\omega_k) - \hat{\mathbf{S}}_k \right\|_F^2 , \tag{15}$$

where $\mathbf{K}_k = (\mathrm{j}\omega_k\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$, and where $\hat{\mathbf{S}}_k$ are the original frequency samples. Based on this expression, we define a weighted cost function as

$$\mathscr{E}^2 = \sum_{k=1}^{K} w_k^2 \mathscr{E}_k^2 , \tag{16}$$

to be minimized subject to the passivity constraints (12).

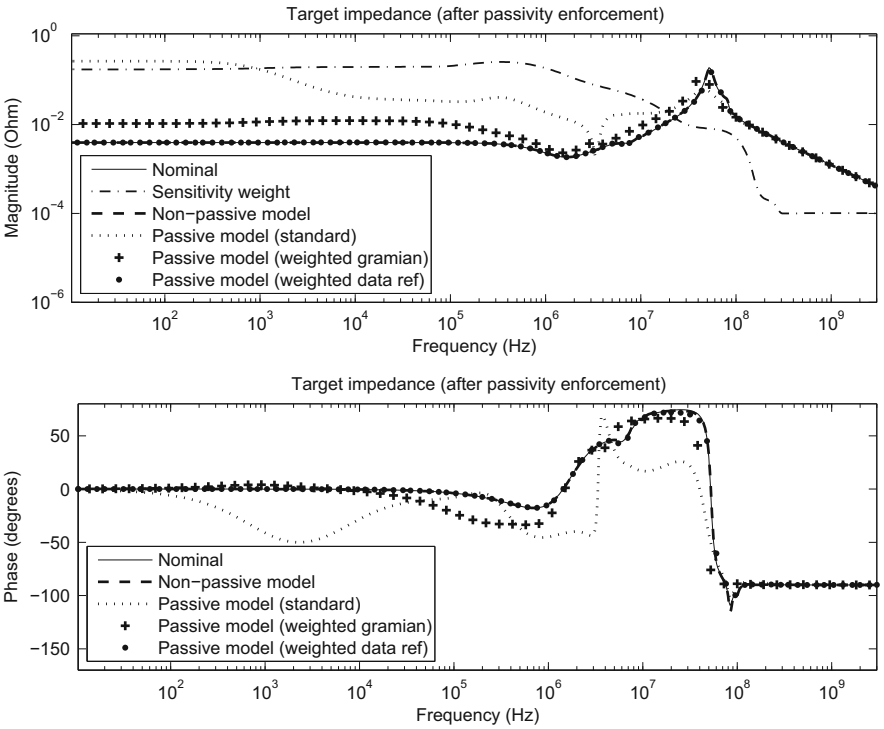


Fig. 1 Magnitude (*top*) and phase (*bottom*) of the input impedance for different models of PDN example 1, compared to the nominal impedance. See text for a detailed description

5 Numerical Examples

We apply the proposed passive model identification process to two different PDN structures, whose scattering responses are available through a broadband electromagnetic simulation. In both cases, the nominal termination conditions are also available in terms of current sources with an RC internal impedance to represent on-chip loading, various decoupling capacitors of different sizes to be placed at the package and board ports, and one Voltage Regulator Module (VRM). The transfer function of interest is the input impedance observed from one of the on-chip ports, subject to the above loading conditions at all other ports.

Figure 1 reports magnitude and phase of the reference (exact) PDN impedance for the first structure (thin solid line), based on nominal terminations, and computed using the raw scattering data. This response is compared to the non-passive model obtained from the proposed iteratively reweighted rational approximation (dashed line). We see that the accuracy of this initial model is excellent. The passive model obtained by perturbation based on a standard cost function (13) is seriously degraded (dotted line), as can be justified by the (rescaled) sensitivity function, also depicted in the top panel (dash-dotted line). The model obtained using a frequency-weighted

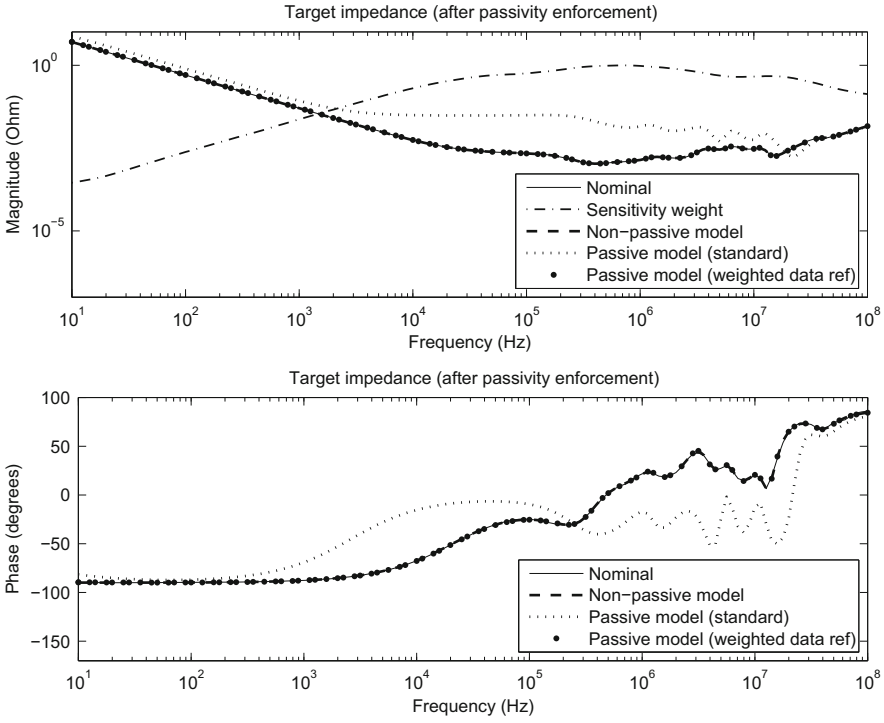


Fig. 2 As in Fig. 1, but for PDN example 2

Gramian (plus markers) shows some improvement, but only using the proposed data-based cost function we are able to match almost perfectly the reference (black dot markers).

Similar conclusions can be drawn from a second application example, which refers to a different PDN structure, with similar overall characteristics and nominal termination scheme. The corresponding curves are depicted in Fig. 2.

6 Conclusions

We have presented a simple approach for the identification of broadband black-box macromodels of LTI systems subject to passivity constraints, and with an input-output accuracy tuned to particular loading conditions. The proposed algorithm is based on a set of adaptively defined frequency-dependent weights, which are used in both rational approximation and passivity enforcement stages of model identification. Numerical results obtained for two chip-package power distribution networks demonstrate the excellent performance of proposed technique.

References

1. Swaminathan, M., Engin, A.E.: Power Integrity Modeling and Design for Semiconductors and Systems. Prentice Hall, Englewood Cliffs, NJ (2007)
2. Swaminathan, M., Kim, J., Novak, I., Libous, J.P.: Power distribution networks for system-on-package: status and challenges. IEEE Trans. Adv. Packag. **27**(2), 286–300 (2004)
3. Su, H., Sapatnekar, S.S.; Nassif, S.R.: Optimal decoupling capacitor sizing and placement for standard-cell layout designs. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **22**(4), 428–436 (2003)
4. Swaminathan, M., Chung, D., Grivet-Talocia, S., Bharath, K., Laddha, V., Xie, J.: Designing and modeling for power integrity. IEEE Trans. Electromagn. Compat. **52**(2), 288–310 (2010)
5. Kose, S., Friedman, E.G.: Distributed on-chip power delivery. IEEE J. Emerging Sel. Top. Circuits Syst. **2**(4), 704–713 (2012)
6. Gustavsen, B., Semlyen, A.: Rational approximation of frequency responses by vector fitting. IEEE Trans. Power Delivery **14**(3), 1052–1061 (1999)
7. Deschrijver, D., Haegeman, B., Dhaene, T.: Orthonormal vector fitting: a robust macromodeling tool for rational approximation of frequency domain responses. IEEE Trans. Adv. Packag. **30**(2), 216–225 (2007)
8. Deschrijver, D., Mrozowski, M., Dhaene, T., De Zutter, D.: Macromodeling of multiport systems using a fast implementation of the vector fitting method. IEEE Microwave Wireless Compon. Lett. **18**(6), 383–385 (2008)
9. Chinea, A., Grivet-Talocia, S.: On the parallelization of vector fitting algorithms. IEEE Trans. Compon. Packag. Manuf. Technol. **1**(11), 1761–1773 (2011)
10. Grivet-Talocia, S., Olivadese, S.B., Triverio, P.: A compression strategy for rational macromodeling of large interconnect structures, EPEPS 2011, San Jose, CA, pp. 53–56, October 23–26, 2011
11. Coelho, C.P., Phillips, J., Silveira, L.M.: A Convex programming approach for generating guaranteed passive approximations to tabulated frequency-data. IEEE Trans. CAD **23**(2), 293–301 (2004)

12. Grivet-Talocia, S.: Passivity enforcement via perturbation of Hamiltonian matrices. IEEE Trans. CAS-I **51**(9), 1755–1769 (2004)
13. Saraswat, D., Achar, R., Nakhla, M.: Global passivity enforcement algorithm for macromodels of interconnect subnetworks characterized by tabulated data. IEEE Trans. VLSI Syst. **13**(7), 819–832 (2005)
14. Grivet-Talocia, S., Ubolli, A.: On the generation of large passive macromodels for complex interconnect structures. IEEE Trans. Adv. Packag. **29**(1), 39–54 (2006)
15. Gustavsen, B., Semlyen, A.: Enforcing passivity for admittance matrices approximated by rational functions. IEEE Trans. Power Syst. **16**(1), 97–104 (2001)
16. Grivet-Talocia, S., Ubolli, A., Bandinu, M., Chinea, A.: An iterative reweighting process for macromodel extraction of power distribution networks. In: IEEE 22nd Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS 2013), San Jose, CA, pp. 125–128, October 27–30, 2013
17. Ubolli, A., Grivet-Talocia, S., Bandinu, M., Chinea, A.: Sensitivity-based weighting for passivity enforcement of linear macromodels in power integrity applications. In: DATE 2014 - Design, Automation and Test in Europe, Dresden, Germany, pp. 1–6, March 24–28, 2014

Multipoint Model Order Reduction Using Reflective Exploration

Elizabeth Rita Samuel, Luc Knockaert, and Tom Dhaene

Abstract Reduced order models obtained by model order reduction methods must be accurate over the whole frequency range of interest. Multipoint reduction algorithms allow to generate accurate reduced models. In this paper, we propose the use of a reflective exploration technique for obtaining the expansion points adaptively for the reduction algorithm. At each expansion point the corresponding projection matrix is computed. Then, the projection matrices are merged and truncated based on their singular values to obtain a compact reduced order model. Three conductor transmission line example is used to illustrate the technique.

1 Introduction

For the accurate modeling of modern integrated circuits and high-speed systems, electromagnetic (EM) methods [1–3] have become an indispensable analysis and design tool. However, a major drawback of EM method is that it usually generate very large systems of equations. The optimization and simulation of these large scale models is computationally expensive, not to say unfeasible. Therefore, model order reduction (MOR) techniques are crucial to reduce the size of large scale models and the computational cost of the simulations, while retaining the important physical features of the original system.

The basic idea of MOR techniques is to reduce the size of a system described by ordinary differential equations, but preserve the dominant behavior of the original system. MOR techniques, for instance, the asymptotic waveform evaluation (AWE) [4], Krylov subspace projection based algorithms [5–7], and truncated balanced realization (TBR) methods [8] have been topics of intense research in the EM modeling field in recent years.

Multipoint MOR methods have been developed over the years [5, 9–11], which allows to generate accurate reduced models over the whole frequency range of interest. This paper focuses on the adaptive selection of the expansion points using

E.R. Samuel (✉) • L. Knockaert • T. Dhaene
Ghent University, iMinds, Gaston Crommenlaan 8 Bus 201, 9050 Gent, Belgium
e-mail: elizabeth.ritasamuel@ugent.be; lizita3@gmail.com; luc.knockaert@ugent.be; tom.dhaene@ugent.be

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_18

a reflective exploration (RE) technique. It is a selective sampling algorithm, where the model is improved incrementally using the best possible data at each iteration, allowing it to propose candidate exploration points [12]. An error-based exploration is performed to find the expansion points. After obtaining the expansion points, the corresponding projection matrices are computed using any of the Krylov based MOR techniques. The projection matrices are then merged and truncated based on their singular values to obtain a more compact projection matrix. Then the reduced order models are obtained by congruence transformation using the compact projection matrix.

This paper is organized as follows. Section 2 gives a brief overview of the PRIMA algorithm [7] and of multipoint model order reduction. In Sect. 3 the proposed technique using RE with model compacting is described. Finally in Sect. 4 the proposed method is illustrated using three conductor transmission line example.

2 Brief Overview of Multipoint MOR

The PRIMA algorithm [7] is used for obtaining the projection matrices.

2.1 PRIMA

Consider a MIMO descriptor system of the form

$$\begin{aligned} \mathbf{E}\dot{x}(t) &= \mathbf{A}x(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t). \end{aligned} \tag{1}$$

The transfer function is

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}. \tag{2}$$

Let s_0 be a suitably chosen expansion point such that the matrix $s_0\mathbf{E} - \mathbf{A}$ is nonsingular. Then the transfer function can be rewritten as:

$$\begin{aligned} \mathbf{H}(s) &= \mathbf{C}(s_0\mathbf{E} - \mathbf{A} + (s - s_0)\mathbf{E})^{-1}\mathbf{B} + \mathbf{D} \\ &= \mathbf{C}(\mathbf{I} + (s - s_0)\mathbf{M})^{-1}\mathbf{R} + \mathbf{D} \end{aligned} \tag{3}$$

where $\mathbf{M} = (s_0\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}$, $\mathbf{R} = (s_0\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$. The qth block Krylov-subspace is given by

$$\mathscr{K}_q(M, R) = colspan[\mathbf{R}\ \mathbf{MR}\ \mathbf{M^2R}\ \ldots\ \mathbf{M^{(q-1)}R}]. \tag{4}$$

This yields the projection matrix V_q, which is the column orthogonal matrix computed from the Krylov subspace $\mathscr{K}_q(M, R)$, from which, using congruence transformation (5), the reduced state-space matrices ($\mathbf{A}_q, \mathbf{E}_q, \mathbf{B}_q, \mathbf{C}_q, \mathbf{D}_q$) are obtained:

$$\mathbf{A}_q = V_q^T \mathbf{A} V_q, \quad \mathbf{E}_q = V_q^T \mathbf{E} V_q, \quad \mathbf{B}_q = V_q^T \mathbf{B}, \quad \mathbf{C}_q = \mathbf{C} V_q, \quad \mathbf{D}_q = \mathbf{D}. \tag{5}$$

2.2 Multipoint Projection Matrix

After MOR, the resulting model must not only be accurate at one frequency point but over the whole frequency range of interest. For this reason, the multipoint reduction algorithm is used [9]. At each expansion point, the projection matrix is computed as described in Sect. 2.1. Then, for N expansion points, the corresponding projection matrices V_{q_i} $(i = 1, 2, \ldots, N)$ are merged to give;

$$V_{comm} = colspan[V_{q_1}\ V_{q_2}\ \ldots\ V_{q_N}]. \tag{6}$$

The merged projection matrix is not truncated using its singular values during the iterative procedure of the reflective exploration. But the matrix is truncated after all the expansion points are adaptively chosen which is described in the following section.

3 Proposed Technique

3.1 Reflective Exploration

The process of selecting samples and building the model in an adaptive way is referred to as reflective exploration (RE) [12]. RE is an effective technique when it is very expensive to obtain the frequency response of the model from EM simulators. For the exploration a reflective function is required to select a new sample. The proposed algorithm uses the root mean square (RMS) error (7) between the obtained best models as the reflective function.

$$Err_{est}^{(I)} = \sqrt{\frac{\sum_{k=1}^{K_s} \sum_{i=1}^{P_{in}} \sum_{j=1}^{P_{out}} \frac{\left|H_{I,(ij)}(s_k) - H_{I-1,(ij)}(s_k)\right|^2}{|H_{I,(ij)}(s_k)|^2}}{P_{in} P_{out} K_s}} \tag{7}$$

where, K_s, P_{in} and P_{out} are the number of frequency samples considered on a dense grid, input and output ports of the system, respectively. The exploration consists of an adaptive modeling loop and an adaptive sampling loop.

1. Adaptive Modeling Loop: The algorithm starts with two expansion points at ω_{min} and ω_{max} of the frequency range of interest. The reduced order q at these points

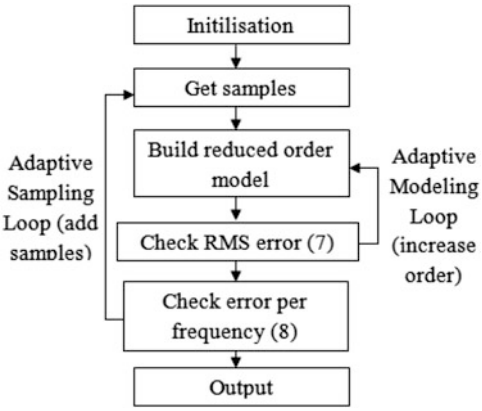


Fig. 1 Flowchart for reflective exploration

is equal to the number of ports of the system for the first iteration, I. Then with a common projection matrix as explained in Sect. 2, the reduced model is obtained. Then in the next iteration again the projection matrix is computed for a reduced order equal to two times the number of ports of the system. If the RMS error between the two best models (i.e., the model obtained in the Ith and the $(I-1)$th iteration) exceeds an estimated error threshold δ_{est}, then the reduced order q is again increased by the number of ports for the respective expansion points.

2. Adaptive Sampling Loop: When the difference in RMS error between the Ith and $(I-1)$th, is less than a threshold δ_{comp}, a new expansion point is selected. For selecting the new expansion point a dense grid is considered for the frequency range and the error per frequency is computed by taking the norm $l2$, of the frequency response of the best model (H_I) and the original model (H_{act}):

$$Err_{s_k} = norm(H_{act}(s_k) - H_I(s_k), 2); \quad k = 1, \ldots, K_s. \tag{8}$$

The frequency point on the grid at which Err_{s_k} is maximum is considered as the new expansion point. It is important to consider a constant dense grid for the frequency throughout the algorithm. This process is iteratively repeated until the RMS error between the original frequency response and the reduced model is 10^{-3}. Figure 1 shows the reflective exploration algorithm.

3.2 Model Compacting

After obtaining the best reduced order model from the iterative procedure, it might be possible to further compact the model with the information obtained from the singular values σ of V_{comm} (6). The economy-size svd is computed for the common projection matrix V_{comm} (6), to obtain the singular values σ of the merged projection matrix. In matlab the economy-sized svd is computed as shown:

$$\mathbf{U}\Sigma\mathbf{V}^T = svd(\mathbf{V}_{comm}, 0) \tag{9}$$

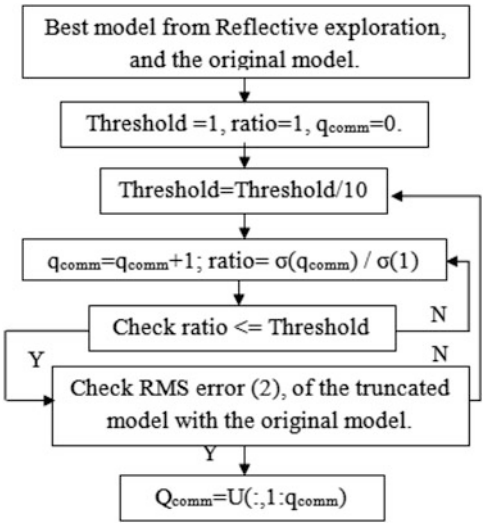


Fig. 2 Flowchart for the truncation of the projection matrix

Here, **U** and **V** are orthogonal matrices, which are known as the left and right singular values. The diagonal of Σ gives the singular values σ of the system. The reduced order for the system is defined based on the first q_{comm} significant singular values of $\mathbf{V}_{comm}$, which is computed by adaptively setting a threshold to the ratio of the singular values to the largest singular value as shown in Fig. 2. The ROM obtained by the truncation of the merged projection matrix with respect to the singular value, is compared with the best model obtained from reflective exploration. If the RMS error is less than 10^{-4}, then we shall truncate the singular values, else we keep the model with the reduced order obtained using the reflective exploration. The compact projection matrix $\mathbf{Q_{comm}}$ is equal to the left singular value **U** where the column is truncated to a size q_{comm} based on the significance of the singular values.

$$\mathbf{Q_{comm}} = \mathbf{U}(:, 1 : q_{comm}). \tag{10}$$

Figure 2 shows the flowchart for the truncation of the singular values. After computing the compact projection matrix $\mathbf{Q_{comm}}$, through congruence transformation (5) on the original system (1) a reduced order model is obtained.

4 Numerical Results

Three conductor transmission lines of six ports described by an original state-space of order 1203 is considered as shown in Fig. 3. The frequency range considered is 1 kHz–1 GHz.

Fig. 3 Three conductor transmission line

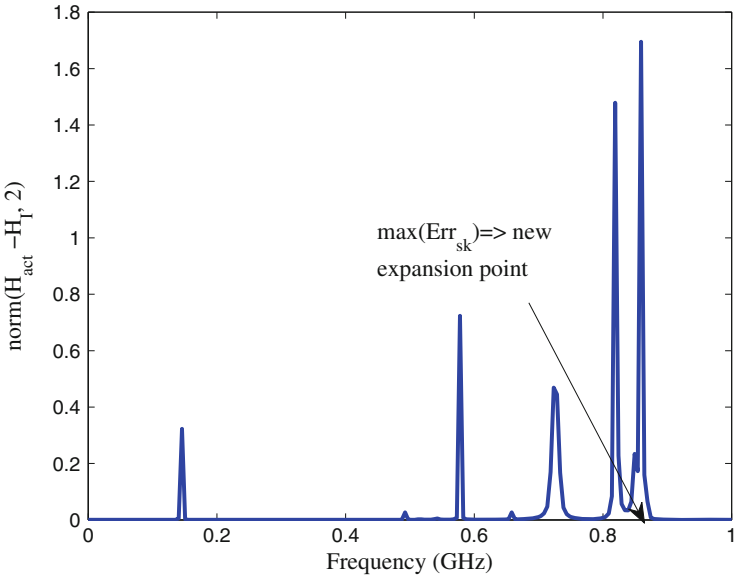


Fig. 4 Error per frequency used to select the new expansion point for the adaptive sampling loop

The sampling starts by considering two samples at the minimum and maximum frequency. The reduced order for the first iteration is equal to 6. Then as discussed in Sect. 2, the frequency response is computed using a common projection matrix. In the next step the frequency response is computed for the same samples with an increased order i.e; it is increased by the number of ports. Then we compute the difference in response between the two models using (7). The error obtained is 4.1979, which is greater than $\delta_{est} = 10^{-3}$, the threshold set for the estimated error. Therefore, the algorithm continues to increase the order till the difference between successive estimated error is less than $\delta_{comp} = 10^{-1}$. Then the adaptive sampling loop starts to find the new sample, by computing the norm of the frequency responses of the two best models (8) over 200 samples of the frequency range. The new sample point is considered at the frequency at which the error (8) is maximum as shown in Fig. 4.

Table 1 RMS error (7) after each modeling loop in RE

Sample	2	3	4
RMS error (7)	1.7	9.8×10^{-2}	1.2×10^{-3}

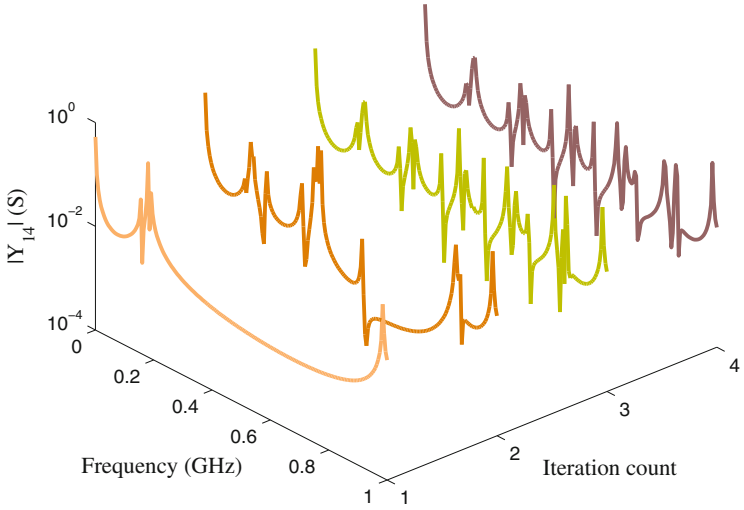


Fig. 5 Magnitude of Y_{14} for the different iteration steps of RE

Then again the frequency response is computed with all the samples and also with increment of the reduced order by the number of ports. Similarly in this manner the sampling process is iterated till the estimated error (7) is less than threshold value $\delta_{est} = 10^{-3}$. Table 1 shows the number of samples used during each iteration of RE to achieve an estimated error less than δ_{est}. Figure 5, shows the admittance Y_{14} obtained during the RE for different iterations. A best model of dimension 96 is obtained with sour samples within a CPU time of 11.2 s on an Intel$^{(R)}$ Core$^{(TM)}$ 2 Duo P8700 2.53 GHz machine with 2 GB RAM and has been implemented in Matlab R2012b on the Windows 7 platform.

Then the model is compacted as described in Sect. 3.2 w.r.t. the singular values to a ROM of dimension 61 with a RMS error (7) of 2.3×10^{-3}. Thus the reflective exploration technique with model compacting was able to automate the generation of expansion points to obtain an accurate and compact reduced order model (Fig. 6).

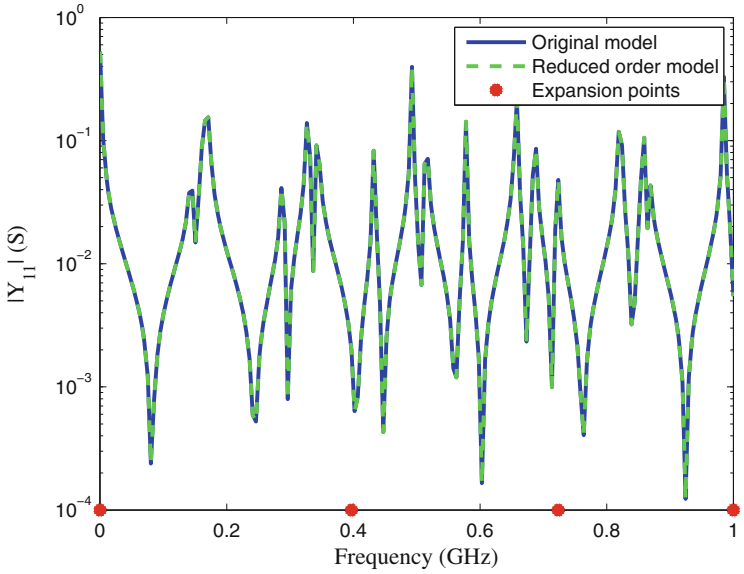


Fig. 6 Magnitude of Y_{11}

5 Conclusions

Multipoint model order reduction algorithms generates reduced order models that are accurate over the whole frequency range of interest. Reflective exploration technique is proposed in this paper for obtaining the expansion points adaptively and also for choosing the reduced order per expansion point for the multipoint reduction algorithm. For each expansion point the corresponding projection matrix is computed and then the projection matrices are merged and truncated based on their singular values to obtain a compact reduced order model. The technique has been illustrated with a coupled transmission line example.

Acknowledgements This work was supported by the Research Foundation Flanders (FWO-Vlaanderen) and by the Interuniversity Attraction Poles Programme BESTCOM initiated by the Belgian Science Policy Office.

References

1. Harrington, R.F.: Field Computation by Moment Methods. Macmillan, New York (1968)
2. Ruehli, A.: Equivalent circuit models for three-dimensional multiconductor systems. IEEE Trans. Microwave Theory Tech. **22**(3), 216–221 (1974)
3. Jin, J.M.: The Finite Element Method in Electromagnetics, 2nd edn. Wiley, New York (2001)

4. Pillage, L., Rohrer, R.: Asymptotic waveform evaluation for timing analysis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **9**(4), 352–366 (1990)
5. Gallivan, K., Grimme, E., Dooren, P.V.: A rational Lanczos algorithm for model reduction. Numer. Algorithms **12**(1), 33–63 (1996)
6. Knockaert, L., De Zutter, D.: Laguerre-SVD reduced-order modeling. IEEE Trans. Microwave Theory Tech. **48**(9), 1469–1475 (2000)
7. Odabasioglu, A., Celik, M., Pileggi, L.: PRIMA: passive reduced order interconnect macromodeling algorithm. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **17**(8), 645–654 (1998)
8. J.R. Phillips, L.M. Silveira, Poor man's TBR: a simple model reduction scheme. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **24**(1), 43–55 (2005)
9. F. Ferranti, M. Nakhla, G. Antonini, T. Dhaene, L. Knockaert, A.E. Ruehli, Multipoint full-wave model order reduction for delayed PEEC models with large delays. IEEE Trans. Electromagn. Compat. **53**(4), 959–967 (2011)
10. Burke, G.J., Miller, E.K., Chakrabarti, S., Demarest, K.: Using model-based parameter estimation to increase the efficiency of computing electromagnetic transfer functions. IEEE Trans. Magn. **25**(4), 2807–2809 (1989)
11. Zhao, W.H., Pang, G.K.H., Wong, N.: Automatic adaptive multi-point moment matching for descriptor system model order reduction. In: International Symposium on VLSI Design, Automation, and Test, 22–24 April 2013
12. Beyer, U., Frank, U.B.: Data exploration with reflective adaptive models. Comput. Stat. Data Anal. **22**(2), 193–211 (1996)

Interface Reduction for Multirate ODE-Solver

Christoph Hachtel, Andreas Bartel, and Michael Günther

Abstract For systems of ordinary differential equations, where the components exhibit a largely differing dynamic behaviour, multirate methods exploit this structure to gain computational efficiency. A model order reduction applied to a single subsystem keeps the dimension in the coupling unchanged. In the case of stiff subsystems, where Jacobians are needed, the computational effort remains high. The here presented *interface reduction* approach is a promising way to turn the reduced dimension into an improved efficiency for multirate time domain simulation.

1 Introduction

The starting point is the following system of ordinary differential equations (ODEs)

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \tag{1}$$

with components of highly different dynamic behaviour. A multirate method exploits this special structure to compute the numerical approximation in a more efficient way. Thus, the system is split according to the dynamical behaviour:

$$\begin{aligned} \dot{\mathbf{y}}_A &= \mathbf{f}_A(\mathbf{y}_A, \mathbf{y}_L, t), \\ \dot{\mathbf{y}}_L &= \mathbf{f}_L(\mathbf{y}_A, \mathbf{y}_L, t), \end{aligned} \tag{2}$$

where $\mathbf{y}_A \in \mathbb{R}^{n_A}$ denote the fast changing, active components and $\mathbf{y}_L \in \mathbb{R}^{n_L}$ the slow changing, latent components (of $\mathbf{y}$). The partitioning is either given by the underlying physical properties of the modeled system or has to be detected e.g. by

C. Hachtel (✉)
Lehrstuhl für Angewandte Mathematik und Numerische Analysis, Fachbereich C Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany
e-mail: hachtel@math.uni-wuppertal.de

A. Bartel • M. Günther
Applied Math. & Numerical Analysis, Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: bartel@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_19

usage of error estimators [1] or based on step size control strategies. A multirate method integrates the slow part with a large macro step H and the fast changing subsystem with a small micro step h with $h \ll H$. Macro and micro step size are computed using the same strategies as for the underlying (singlerate) integration scheme(s). A partitioning into more than two subsystems is possible. For simplicity of notation, we restrict ourselves to two subsystems. It is obvious that multirate methods can be even more efficient if the dimension of the slow subsystem is very large compared to the dimension of the active one.

For a system with a fixed and given partitioning, a model order reduction (MOR) of the slow, high dimensional subsystem promises another gain of efficiency for the time domain simulation. Here, the challenging part is to combine the reduced dimension and the coupling interface with the other non-reduced subsystems.

The work is organised as follows. First we give an introduction to multirate compound step methods and we repeat briefly the well-known concepts of MOR. The combination of both techniques forms the heart of our work: multirate interface reduction. For this setting, a multiphysics application fits to illustrate the capabilities of this new approach. We are considering a regularised, academic test circuit with thermal active and dependent elements and give first numerical results.

2 Multirate Compound-Step Methods

Given the partitioned ODE (2), the crucial part of multirate schemes is the realisation of the coupling between the subsystems. In fact, this is one of the distinguishing features of these techniques. In (2) and in all later equations, the coupling terms are printed in colour. Multirate integration schemes for implicit methods were first presented by Gear and Wells [2], where the coupling is simply achieved by inter- and extrapolating the unknown values. Though this approach seems to be a natural choice, several problems concerning the coupling terms appear. In the last years methods that achieve the multirate integration by using a dynamic refinement strategy became popular. These schemes integrate the whole system with a large step size H. By using error estimators, the step size is only refined for those components for which a given accuracy is not reached. Savcenco [1] uses embedded Runge-Kutta schemes for error estimation, Constantinescu and Sandu [3] are using Richardson extrapolation. These methods can handle systems for which a partitioning according to the dynamic behaviour is not known a priori.

Here we follow the idea of compound step methods, which were first developed using Runge-Kutta schemes by Kværnø and Rentrop [4] and then expanded to W-methods [5]. The main idea is to compute the macro-step $\mathbf{y}_L(t_0 + H)$ and the first micro step $\mathbf{y}_A(t_0 + h)$ coupled together in one compound step. The remaining micro steps $\mathbf{y}_A(t_0 + ih),\ i = 2, \ldots, m$, can either be computed by interpolating the slow components or by using a dense output formula for the slow part. Compound step methods can be used for systems with a stronger coupling than the inter-/extrapolation methods of [2]. Mixed-multirate compound step methods

[6] allow the usage of different integration schemes for compound and remaining micro steps. So the single methods can be chosen according to the properties of the subsystems.

Linear (simply diagonal) implicit compound step methods like in [5] can be used for (at least moderately) stiff systems only by the computational cost of solving one system of linear equations per time step. The simplest version is the multirate linear implicit Euler method, [3]: the method reads for the compound step

$$\begin{pmatrix} h\dfrac{\partial \mathbf{f}_A}{\partial \mathbf{y}_A} - \mathbf{I}_A & \dfrac{h}{m}\dfrac{\partial \mathbf{f}_A}{\partial \mathbf{y}_L} \\ mH\dfrac{\partial \mathbf{f}_L}{\partial \mathbf{y}_A} & H\dfrac{\partial \mathbf{f}_L}{\partial \mathbf{y}_L} - \mathbf{I}_L \end{pmatrix} \begin{pmatrix} \mathbf{y}_A(t_0+h) - \mathbf{y}_A(t_0) \\ \mathbf{y}_L(t_0+H) - \mathbf{y}_L(t_0) \end{pmatrix} = \begin{pmatrix} h\mathbf{f}_A(\mathbf{y}_A(t_0), \mathbf{y}_L(t_0)) \\ H\mathbf{f}_L(\mathbf{y}_A(t_0), \mathbf{y}_L(t_0)) \end{pmatrix} \tag{3}$$

and for the remaining micro steps holds

$$\left(h\,\frac{\partial \mathbf{f}_A}{\partial \mathbf{y}_A} - \mathbf{I}_A\right)\mathbf{k}_{A,i} = -h\,\mathbf{f}_A(\mathbf{y}_A(t_0+ih), \tilde{\mathbf{y}}_L(t_0+ih)), \qquad i = 1,\ldots,m-1 \tag{4}$$

with $\mathbf{k}_{A,i} = \mathbf{y}_A(t_0 + (i+1)h) - \mathbf{y}_A(t_0 + ih)$ and $\tilde{\mathbf{y}}_L$ the interpolated values of the slow components. The coupling between the slow and the active subsystem is realised by the off-diagonal elements in the coefficient matrix of the system of linear equations in (3). The multirate method (3)–(4) is of order one, but also in higher order compound step methods with underlying linear implicit integration schemes, e.g. [5, 6], the coupling is partly realised by the off-diagonal blocks of the coefficient matrix. We can only expect high improvements in the computational effort by applying multirate schemes if the number of active components is much smaller than the number of slow components ($n_A \ll n_L$). So the question arises whether one can exploit this structure for a more efficient computation not only by using larger step sizes for the slow component but also reducing the dimension of the slow part by MOR.

3 Model Order Reduction (MOR)

Since we only expect small variations in the slow components, we assume here that the slow part is linear or at least linearised over a given macro step. This might be not true for all nonlinear cases but covers relevant application. Nevertheless one has to take care of the choice of step size while linearising a nonlinear model.

Now the partitioned system (2) can be rewritten with system matrix $\mathbf{A} \in \mathbb{R}^{n_L,n_L}$, input matrix $\mathbf{B} \in \mathbb{R}^{n_L,n_A}$ and output matrix $\mathbf{C} \in \mathbb{R}^{n_L,n_L}$:

$$\dot{\mathbf{y}}_A = \mathbf{f}_A(\mathbf{y}_A, \mathbf{y}_L, t) \tag{5}$$

$$\dot{\mathbf{y}}_L = \mathbf{A} \cdot \mathbf{y}_L + \mathbf{B} \cdot \mathbf{y}_A \tag{6}$$

$$\mathbf{y}_L = \mathbf{C} \cdot \mathbf{y}_L. \tag{7}$$

Next we apply MOR to the internal variable $\mathbf{y}_L$. To this end, the system matrices are projected on a low dimensional subspace by biorthogonal projection matrices $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n_L,r}$ with $r \ll n_L$. One ends up with a reduced slow subsystem

$$\dot{\mathbf{y}}_{L,r} = \mathbf{W}^T\mathbf{A}\mathbf{V} \cdot \mathbf{y}_{L,r} + \mathbf{W}^T\mathbf{B} \cdot \mathbf{y}_A \tag{8}$$

$$\tilde{\mathbf{y}}_L = \mathbf{C}\mathbf{V} \cdot \mathbf{y}_{L,r}. \tag{9}$$

The motivation of applying MOR is to obtain a small dimensional variable $\mathbf{y}_{L,r}$ while the output $\tilde{\mathbf{y}}_L$ shall be approximated sufficiently accurate. The way how the projection matrices $\mathbf{V}, \mathbf{W}$ are computed are defined by the MOR method, for further details see [7]. For the multirate-MOR setting, the usage of a certain MOR method is not mandatory so the user can choose his favorite method.

Notice that the dimension of the output variable in (9), i.e., the dimension of the coupling interface slow to active, will be not reduced in this setting. In fact, with a non-reduced interface we cannot expect large improvements of the computational efficiency solving the system of linear equations in the compound step (3) by using a reduced slow subsystem. So we have to find a way to transfer the reduced dimension to the coupling interface to gain efficiency in the compound step.

4 Interface Reduction

Often the active components do not depend on the detailed information of every single slow component. So we may replace the coupling interface $\mathbf{y}_L$ in (5) by a low dimensional input $\mathbf{u}_A = \mathbf{g}(\ldots, \mathbf{z}_L, \ldots)$ while $\mathbf{z}_L$ denotes the output of the slow subsystem. The same can be made for the slow part (7). Adopting the notation for coupled linear systems from [8], we get

$$\dot{\mathbf{y}}_A = \mathbf{f}_A(\mathbf{y}_A, \mathbf{u}_A, t) \qquad \dot{\mathbf{y}}_L = \mathbf{f}_L(\mathbf{y}_L, \mathbf{u}_L, t) := \mathbf{A} \cdot \mathbf{y}_L + \mathbf{B} \cdot \mathbf{u}_L \tag{10}$$

$$\mathbf{u}_A = \mathbf{g}(\mathbf{z}_A, \mathbf{z}_L, \mathbf{u}, t) \qquad \mathbf{u}_L = \mathbf{K}_{LA} \cdot \mathbf{z}_A + \mathbf{K}_{LL} \cdot \mathbf{z}_L + \mathbf{H} \cdot \mathbf{u} \tag{11}$$

$$\mathbf{z}_A = \mathbf{h}(\mathbf{y}_A, t) \qquad \mathbf{z}_L = \mathbf{C} \cdot \mathbf{y}_L \tag{12}$$

with input $\mathbf{u}_X \in \mathbb{R}^{q_X}$, global input $\mathbf{u}$, output $\mathbf{z}_X \in \mathbb{R}^{p_X}$ and coupling matrices $\mathbf{K}_{LX}$, $X \in \{A, L\}$. The coupling functions $\mathbf{g}, \mathbf{h}$ and matrices $\mathbf{K}_{LA}, \mathbf{C}$ are not given by the system itself. Thus for the multirate setting they must be defined by the

user exploiting some underlying properties, e.g. physical laws. These modifications in the multirate setting will not change the diagonal blocks in the compound step coefficient matrix (3), but for the off-diagonal blocks the mixed derivatives change into

$$\frac{\partial \mathbf{f}_A}{\partial \mathbf{y}_L} = \frac{\partial \mathbf{f}_A}{\partial \mathbf{u}_A} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{y}_L} = \frac{\partial \mathbf{f}_A}{\partial \mathbf{u}_A} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{z}_L} \cdot \frac{\partial \mathbf{z}_L}{\partial \mathbf{y}_L} = \frac{\partial \mathbf{f}_A}{\partial \mathbf{u}_A} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{z}_L} \cdot \mathbf{C}. \tag{13}$$

$$\frac{\partial \mathbf{f}_L}{\partial \mathbf{y}_A} = \frac{\partial \mathbf{f}_L}{\partial \mathbf{u}_L} \cdot \frac{\partial \mathbf{u}_L}{\partial \mathbf{y}_A} = \frac{\partial \mathbf{f}_L}{\partial \mathbf{u}_L} \cdot \frac{\partial \mathbf{u}_L}{\partial \mathbf{z}_A} \cdot \frac{\partial \mathbf{z}_A}{\partial \mathbf{y}_A} = \mathbf{B} \cdot \mathbf{K}_{LA} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{y}_A}. \tag{14}$$

On the right hand sides of (13) and (14) we find matrix products of the dimensions:

$$(n_A \times q_A) \cdot (q_A \times p_L) \cdot (p_L \times n_L) \tag{15}$$

$$(n_L \times q_L) \cdot (q_L \times p_A) \cdot (p_A \times n_A). \tag{16}$$

In a multirate context the dimension n_A is supposed to be small. If the interface functions $\mathbf{g}, \mathbf{h}, \mathbf{K}_{LA}, \mathbf{C}$ are chosen such that the dimension of their codomains are small, then only one large dimension remains, namely the number of the slow components n_L. However, as we saw in Sect. 3, we can compute a reduced model of dimension r for the slow part and use matrices $\mathbf{B}_r$ and $\mathbf{C}_r$ in the mixed derivatives of (13)–(14).

Using this framework we expect higher efficiency in a time domain simulation. If we apply a MOR technique for which any error bounds are known also the error due to MOR can be handled. Nevertheless the replacement of $\mathbf{y}_X$ to $\mathbf{u}_X$ can influence the numerical properties of the integration method in particular the stability, which is not yet investigated.

5 Simulation

To apply the theoretical considerations of the above sections, we use as benchmark example the electric-thermal test circuit of [9] with the modifications given in [10]. It is a small electric circuit, in which some elements are modeled temperature dependent. The circuit diagram is given in Fig. 1 (left). Due to electric current, the resistor $R(\mathbf{T})$ is heated and so the resistance of this device changes. The characteristic curve of the diode is also temperature dependent. The voltages are modeled by a nodal analysis using Kirchhoff's laws. For the temperature of the resistor (wire), the 1-D heat equation is semi-discretised using a finite volume approach, see Fig. 1 (right). Finally we get a system of ordinary differential equations like in (1) in terms of the unknowns $\mathbf{y} = [u_3, u_4, e, \mathbf{T}]$, where u_3, u_4 denote the voltages at node 3 and 4, e is the dissipated energy in the thermal dependent resistor and $\mathbf{T}$ the vector of temperatures in the semi-discretised resistor. The multirate behaviour of this system is given by the physical properties: the voltages and the dissipated energy change

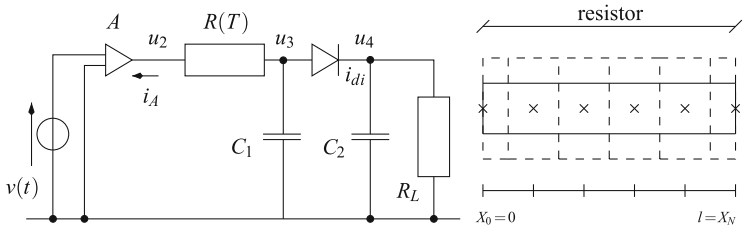


Fig. 1 Circuit diagram (*left*), and finite volume discretised resistor (*right*)

very fast (with source of the network), and the temperature in the resistor changes much slower. Hence the partitioning according to the dynamical behaviour is quite natural:

$$[\dot{u}_3, \dot{u}_4, \dot{e}] = \mathbf{f}_A([u_3, u_4, e], \mathbf{T}, t) \tag{17a}$$

$$\dot{\mathbf{T}} = \mathbf{f}_L([u_3, u_4, e], \mathbf{T}, t). \tag{17b}$$

As mentioned in [10], the subsystem of the semi-discretised heat equation (17b) is not a-priori linear, but it can be easily linearised without loss of much accuracy. Hence a linear model order reduction of the thermal subsystem is possible.

For (17), the computational cost of the compound step (3) depends on the number of discretisation points of the spatial variable of the thermal subsystem. If a high accuracy is demanded this dimension can be large and the computational cost increases. So the question is how the coupling interface can be modified such that the dimension of the input of the active part and the output of the slow part is small.

The heating of the resistor, caused by the electric current, is computed by the dissipated power p. The electric subsystem is computing the total dissipated energy e in on macro step H. The ratio e/H defines the averaged power, which we use for coupling [9]. Hence we add an output function to the active subsystem: $\mathbf{h}([u_3, u_4, e], t) = e/H$. To compute e, we have either to calculate differences of e or we have to assign zero as the initial value for each macro step. If H is adjusted by a step size control, it has to be handled as an independent parameter.

For the coupling interface slow to active, one has to consider the thermal dependent, physical parameters, which are necessary in the circuit model and which can be computed by a linear model. In our case, these are the total resistance $R(\mathbf{T})$ and the diode's temperature T_{di}. Additional input functions for the slow and the active part are not necessary with this choice of coupling interfaces. As global input variable $\mathbf{u}$ we have the source voltage $v(t)$ which is used in the active, electric subsystem only. These modifications in the interface of the coupled system (17)

lead to

$$[\dot{u}_3, \dot{u}_4, \dot{e}] = \mathbf{f}_A([u_3, u_4, e], \mathbf{u}_A, t) \qquad \dot{\mathbf{T}} = \mathbf{A} \cdot \mathbf{T} + \mathbf{B} \cdot \mathbf{u}_L \tag{18a}$$

$$\mathbf{u}_A = [R(\mathbf{T}), T_{di}, v(t)]^T \qquad \mathbf{u}_L = p \tag{18b}$$

$$p = \mathbf{h}([u_3, u_4, e], t) = e/H \qquad [R(\mathbf{T}), T_{di}] = \mathbf{C} \cdot \mathbf{T}. \tag{18c}$$

For this system the off-diagonal blocks of the Jacobian matrix in the compound step (3) become much smaller. Inspecting the dimensions like in (15) gives for $\frac{\partial \mathbf{f}_A}{\partial \mathbf{y}_L}$ the matrix sizes $(3 \times 2) \cdot (2 \times n_L)$ and for $\frac{\partial \mathbf{f}_L}{\partial \mathbf{y}_A}$ the dimension $(n_L \times 1) \cdot (1 \times 3)$. Now, a model order reduction can decrease the number of thermal variables from n_L to a significant smaller number r. No large dimensional terms occur in this setting so we expect a large gain concerning the computational effort using compound step multirate methods for this multiphysics application.

For the simulation of the system we use the mixed multirate compound step method of [6] which consists of a third order for the compound and a fourth order linear implicit method for the remaining micro steps. For the model order reduction we chose balanced truncation. We implemented the system and the integration methods in Matlab 2013a. All relevant simulation parameters are listed in Table 1 and also the computation time can be seen there. The table shows the necessity of an interface reduction when combining a multirate scheme with a model order reduction: Only applying a model order reduction increases the computation time due to the loss of special matrix structures (sparsity, band structure). Interface reduction and MOR can decrease the computation time to 25 %. Here, we are interested in two physical sizes: One is the temperature of the diode and the other is the highest temperature in the resistor which is found at its middle. Figure 2 shows the relative error of the multirate solution to the reference solution of these two physical sizes. Figure 3 shows the voltage curve at node 3. The error is very small and we can say that our method decreases the computation time significantly with only a very small loss of accuracy.

Table 1 Simulation parameters and computation time for full order model (FOM) and reduced order model (ROM) for simulation time [0 s, 0.12 s]

Model			Monolithic		Interface reduced	
Parameters	H	m	FOM $n_L = 50$	ROM $r = 5$	FOM $n_L = 50$	ROM $r = 5$
Singlerate	$5 \cdot 10^{-5}$	1	4.81 s	5.65 s	2.65 s	1.91 s
Multirate	$2.5 \cdot 10^{-4}$	5	3.44 s	5.00 s	1.36 s	1.18 s

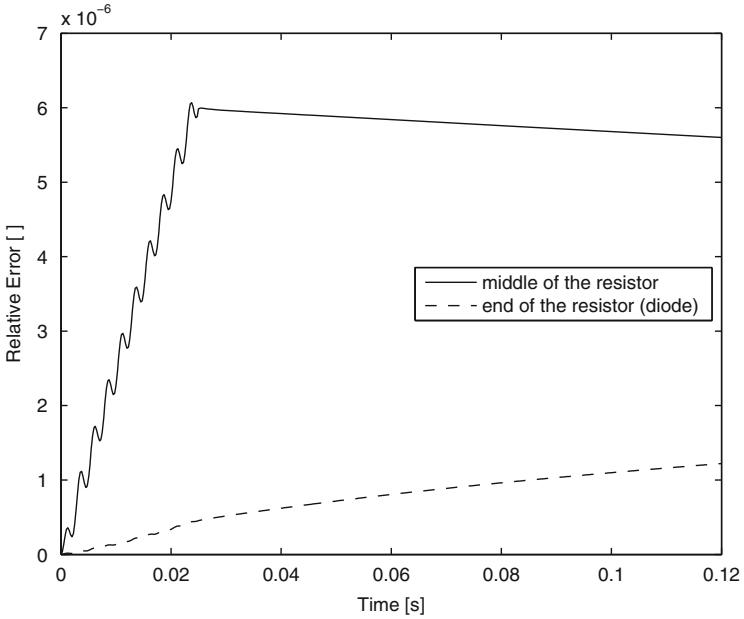


Fig. 2 Relative errors

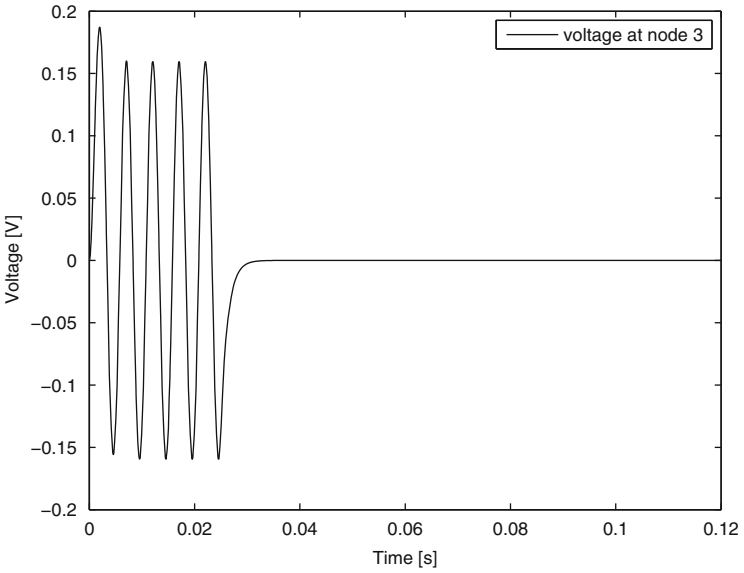


Fig. 3 Voltage at node 3

6 Conclusion

Applying a model order reduction to a subsystem of a monolithic model the computation time does not decrease as expected in a multirate time domain simulation. In this paper we pointed out why a MOR in a unmodified multirate framework does not lead to computational improvements. Furthermore, by introducing interface reduction, we presented a way how the reduced dimension in a model order reduced subsystem can be exploited also for multirate compound step methods. We put up interface reduction approach to an academic multiphysics test system. The observed error is not yet understood, its sources will be analysed in the future. Hence interface reduction modifies the multirate ODE framework stability of the multirate compound step method (cf. [11]) cannot be guaranteed any more so further work about this open point is necessary.

Acknowledgements This work was supported by the Research Network KoSMos: *Model Reduction Based Simulation of coupled PDAE Systems* funded by the German Federal Ministry of Education and Science (BMBF), grant no. 05M13PXA. Responsibility for the contents of this publication rests with the authors.

References

1. Savcenco, V.: Multirate numerical integration for ordinary differential equations. Ph.D. thesis, Universiteit van Amsterdam (2008)
2. Gear, C.W., Wells, D.R.: Multirate linear multistep methods. BIT **24**(4), 484–502 (1984)
3. Constantinescu, E., Sandu, A.: On extrapolated multirate methods. In: Fitt, A.D., Norbury, J., Ockendon, H., Wilson, E. (eds.) Progress in Industrial Mathematics at ECMI 2008. Mathematics in Industry, pp. 341–347. Springer, Berlin/Heidelberg (2010)
4. Kværnø, A., Rentrop, P.: Low order multirate runge-kutta methods in electric circuit simulation. URL www.math.ntnu.no/preprint/numerics/1999/N2-1999.ps (1999). Preprint No. 2/99
5. Bartel, A., Günther, M.: A multirate w-method for electrical networks in state-space formulation. J. Comput. Appl. Math. **147**(2), 411–425 (2002)
6. Bartel, A.: Multirate row methods of mixed type for circuit simulation. In: van Rienen, U., Günther, M. (eds.) Scientific Computing in Electrical Engineering. Lecture Notes in Computational Science and Engineering, pp. 241–249. Springer, Berlin (2001)
7. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia (2005)
8. Reis, T., Stykel, T.: Stability analysis and model order reduction of coupled systems. Math. Comput. Model. Dyn. Syst. **13**(5), 413–436 (2007)
9. Bartel, A., Günther, M., Schulz, M.: Modeling and discretization of a thermal-electric test circuit. In: Antreich, K., Bulirsch, R., Gilg, A., Rentrop, P. (eds.) Modeling, Simulation and Optimization of Integrated Circuits. International Series of Numerical Mathematics, vol. 146, pp. 187–201. Birkhäuser, Basel (2003)
10. Hachtel, C., Günther, M., Bartel, A.: Model order reduction for multirate ode-solvers in a multiphysics application (2014). BUW Preprint 14/22, accepted for ECMI 2014 proceedings
11. Kværnø, A.: Stability of multirate runge-kutta schemes. Int. J. Differ. Equ. Appl. **1A**, 97–105 (2000)

Part V
Uncertainty Quantification

In all papers in this part Uncertainty Quantification exploits Polynomial Chaos Expansions. In two papers (parametric) Model Order Reduction techniques are successfully applied to reduce systems, that have to be solved, in size. The expansions are also of use in expanding special coefficient functions needed to evaluate systems for optimization and for which sensitivity has to be efficiently determined. And, finally, it is observed that uncertainty can also affects convergence in combining numerical procedures from different sources.

The paper by S. Clénet: *Approximation Methods to solve Stochastic Problems in Computational Electromagnetics* gives an introduction to the use of Polynomial Chaos Expansions in which the solution of a parametric problem is expressed in a basis of orthonormal polynomials that are evaluated at the parameter value. It is illustrated for the vector potential in the case of static Maxwell equations, both for a non-intrusive method, like Stochastic Collocation, as well as for an intrusive approach, like Stochastic Galerkin. The success of the expansion depends first on how fast the series converges, so that a finite expansion can be used as approximation. However also the dimension of the parameter space counts as well as methods for integrating several integrals. Different quadrature formulas along the different coordinate axes in the parameter space can be used. Coefficient functions in the equations can be approximated by finite expansions as well, or can be expressed as sum of separable functions, like $G(\mathbf{p}) = \sum_{j=1}^{T} \prod_{k=1}^{K} u_k^j(p_k)$, in which $\mathbf{p} = (p_1, \ldots, p_K)^T$ and where the $u_k^j(p_k)$ are unknowns to the problem and are looked for in a low dimensional space. Also Proper Generalized Decomposition is described. All these methods aim to reduce the memory size needed to determine the coefficients in any expansion. Actually these are approaches that have nice links to Model Order Reduction.

The paper by P. Benner and M.W. Hess: *Reduced Basis Modeling for Uncertainty Quantification of Electromagnetic Problems in Stochastically Varying Domains* is a paper where a parametric Model Order Reduction technique successfully is applied, in this case by Reduced Basis Modeling. The reduced model reduces the costs of evaluations by Monte Carlo Simulations or by Stochastic Collocation to analyze

the uncertainty in the model with respect to small variations in geometry. A crucial point is that all geometry is viewed as being mapped from a reference geometry on which one can assemble the system matrices and the use of an affine transformation to map to a particular realization. This approach assumes that each degree of freedom in the mesh has the same meaning under affine transformations and that the field solutions for different geometries are in the same functional space. Thus the dimension of the mesh remains the same for each actual geometry. The Stochastic Collocation is performed with Hermite Genz-Keister sparse grids, generated by the Smolyak algorithm. These types of methods can exhibit a mean convergence rate of $\mathcal{O}((\log n)^p/n)$, using n points for p-dimensional integrands, while for Monte Carlo the mean convergence rate is of the order $\mathcal{O}(1/\sqrt{n})$. The approach is demonstrated for a stripline model of a coplanar waveguide involving ten geometrical parameters.

The paper by R. Pulch: *Model Order Reduction for Stochastic Expansions of Electric Circuits* considers linear time-invariant dynamical systems in which all matrices depend on some parameter $\mathbf{p}$, with solution $\mathbf{x}(t, \mathbf{p})$ and output vector $\mathbf{y}(t, \mathbf{p})$. Stochastic Collocation (SC) involving quadrature leads to separate systems for the different nodes (parameter values) of the quadrature rule. One can write this in one big system for $\mathbf{x}(t) = (\mathbf{x}(t, \mathbf{p}_1), \ldots, \mathbf{x}(t, \mathbf{p}_q))$ and with output the time-dependent coefficients for the generalized polynomial expansion of $\mathbf{y}(t, \mathbf{p})$. This system has a block-tridiagonal structure and all sub-systems inherit stability from the original dynamical system. Stochastic Galerkin (SG) provides a system of a similar size, which is fully coupled. The author has proved that the SG system sometimes looses the stability. This loss of stability does not affect the convergence of the SG method on compact time intervals. Yet, then the asymptotic behaviour becomes incorrect in time. This does not occur for the SC approach. For both cases, to the big systems a Model Order Reduction technique is applied, in this case Balanced Truncation exploiting an ADI implementation. Demonstration is made to a bandpass filter with single input-single output involving 11 parameters. A Stroud quadrature rule of order 5 was used for SC and a sparse grid of level 3 based on Legendre quadrature for SG.

The paper by P. Putek, K. Gausling, A. Bartel, K.M. Gawrylczyk, J. ter Maten, R. Pulch, and M. Günther: *Robust topology optimization of a permanent magnet synchronous machine using multi-level set and stochastic collocation methods* considers topology optimization for a permanent magnet (PM) synchronous machine with material uncertainties, in this paper the reluctivities in the iron, the air-gaps and in the PM. The variations of the (non)linear material characteristics are modeled by the Polynomial Chaos Expansion method. During the iterative optimization process, the shapes of the rotor poles, represented by zero-level sets, are simultaneously optimized by redistributing the iron and the magnet material over the design domain. The gradient directions of the multi-objective function are evaluated by utilizing the Continuous Design Sensitivity Analysis (CDSA). The constraints are composed of the mean and the standard deviation, which are provided by Stochastic Collocation (SC). Incorporating the SC into the level set method allows to use already existing deterministic solvers. Demonstration is made to a two-dimensional problem of a

low cogging torque design of an Electrically Controlled Permanent Magnet Excited Synchronous Machine.

The paper by K. Gausling and A. Bartel: *First Results for Uncertainty Quantification in Co-Simulation of Coupled Electrical Circuits* considers how uncertainty may affect convergence of numerical procedures. Co-simulation operates on time windows $[T_n, T_n + H]$ and tries to compute the overall solution iteratively by decoupling. After integration over the time window, one obtains new time profiles for the unknowns of all parts of the partition. With these new time profiles one can re-start the co-simulation process over the same time window to further update the profiles. Thus one solves the subsystems multiple times. The benefit is that one can use larger time windows than by using one iteration and performing stepsize control. Co-simulation applied to coupled ordinary differential equations always converges. The situation is different for coupled differential-algebraic equations. In such cases convergence can only be guaranteed if a contraction condition is fulfilled. The theory of co-simulation shows that its stability and its rate of convergence is directly influenced by a) the sequence in which the subsystems are computed and b) by the coupling interface. Creating a successful splitting already can be a piece of art. Now, in case of uncertainties in parameters, the contraction factor will become stochastic. Demonstration is made for a 2-level RLC network.

We finally remark that over the last years a popular library for Uncertainty Quantification has been provided by DAKOTA, developed at Sandia National Laboratories.[1]

[1] https://dakota.sandia.gov/.

Approximation Methods to Solve Stochastic Problems in Computational Electromagnetics

Stéphane Clénet

Abstract To account for uncertainties on model parameters, the stochastic approach can be used. The model parameters as well as the outputs are then random fields or variables. Several methods are available in the literature to solve stochastic models like sampling methods, perturbation methods or approximation methods. In this paper, we propose an overview on the solution of stochastic problems in computational electromagnetics using approximation methods. Some applications will be presented in order to illustrate the possibilities offered by the approximation methods but also their current limitations due to the *curse of dimensionality*. Finally, recent numerical techniques proposed in the literature to face the *curse of dimensionality* are presented for non-intrusive and intrusive approaches.

1 Introduction

Applying a discretisation scheme (Finite Element Method-FEM, Finite Integration Technique-FIT, ...) to solve the Maxwell equations leads to valuable tools for understanding and predicting the features of electromagnetic devices. With the progress in the fields of numerical analysis, CAD and postprocessor tools, it is now possible to represent and to mesh very complex geometries and also to take into account more realistic material behaviour laws with non-linearities, hysteresis .... Besides, computers have nowadays such capabilities that it is common to solve problems with millions of unknowns. The modelling error due to the assumptions made to build the mathematical model (the set of equations) and the numerical errors due especially to the discretisation (by a FEM for example) can be negligible. Consequently, in some applications represented by very accurate models (the modelling and the numerical errors are negligible), if a gap exists between the measurements, assuming perfect, and the results given by the numerical model, it comes from deviations on input parameters which are not in the "real world" equal to their prescribed values. The origins of these deviations are numerous and

S. Clénet (✉)
L2EP, Arts et Métiers ParisTech, 8, Bd Louis XIV, 59046 Lille Cedex, France
e-mail: stephane.clenet@ensam.eu

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_20

are related to either a lack of knowledge (epistemic uncertainties) or uncontrolled variations (aleatoric uncertainties). For example, mechanical parts are manufactured with dimensional tolerances whereas some dimensions, such as air gaps in electric machines, are critical as they strongly influence performance. Besides uncertainties in material composition, the material characteristics which change with uncontrolled environmental factors (humidity, pressure, etc.) are also often unknown [41]. Even if the environmental factors are perfectly known, in some situations, the behaviour law parameters cannot be identified because measurements are not possible under the right experimental conditions. Consequently, to be more realistic, numerical models must now be able to take into account uncertainties.

The stochastic approach which consists in representing the uncertain parameters as random variables, (the output variables are then also random variables) is one possible way to model and to evaluate the influence of the uncertainties on the parameters. Monte Carlo Simulation methods or perturbation methods are available to solve stochastic problems since early 1950s [22, 34]. In the 1990s, researches on quantification of uncertainties in numerical models using approximation methods first began in the field of mechanical and civil engineering [19]. In the 2000s, this approach has met a growing interest with the development of approximation methods based especially on truncated polynomial chaos expansions that offer a higher convergence rate than the Monte Carlo Simulation Method if the model outputs present a sufficient regularity versus the input parameters.

In this paper, we propose a survey on the solution of stochastic problems in computational electromagnetics using approximation methods. First, we present the deterministic model based on FEM then the stochastic model is derived when the input parameters are considered as random variables. The approximation method is introduced which consist in finding a solution in a finite dimensional functional space. Different numerical techniques, available in the literature, are described to solve the stochastic problem. Then, a description of applications of the stochastic approach in the field of computational electromagnetics is proposed in order to illustrate the capabilities of such approach but also its current limitations particularly due to the *curse of dimensionality*. Finally, recent numerical techniques proposed in the literature to face the *curse of dimensionality* are presented.

2 Presentation of the Problem

2.1 Deterministic Problem

In the following, we will address the magnetostatic problem but the results can be easily extended to other static and quasi static field problems. In the following, the aim is to introduce notations when the magnetostatic problem is solved numerically using the vector potential formulation and FEM. The partial differential equations

to be solved on a domain D are:

$$\mathbf{curl}\ \mathbf{H}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) \tag{1}$$

$$div\mathbf{B}(\mathbf{x}) = 0 \tag{2}$$

with **H** the magnetic field, **B** the magnetic flux density and **J** the current density that is assumed to be known. In addition, boundary conditions on **H** and **B** are added and also the behaviour law of the material which will be assumed to be written in the form:

$$\mathbf{H}(\mathbf{x}) = \nu(\mathbf{x})\mathbf{B}(\mathbf{x}) \tag{3}$$

with ν the reluctivity. To solve the problem, the vector potential formulation can be used:

$$\mathbf{curl}[\nu(\mathbf{x})\mathbf{curl}\mathbf{A}(\mathbf{x})] = \mathbf{J}(\mathbf{x}) \tag{4}$$

with **A** the vector potential. To find an approximate solution of this equation, FEM is often applied. We seek for an approximation **A** of the vector potential in the edge element space such that:

$$\mathbf{A}(\mathbf{x}) = \sum_{i=0}^{\mathrm{N}} a_i \mathbf{w}_i(\mathbf{x}) \tag{5}$$

with N the number of Degrees of Freedom (DoF's), $\mathbf{w}_i$ the edge shape functions and a_i unknown real coefficients. By applying the Galerkin method to a weak form of (4):

$$\int_D \nu(\mathbf{x})\mathbf{curl}\ \mathbf{A}(\mathbf{x}).\mathbf{curl}\ \mathbf{w}_i(\mathbf{x})dx = \int_D \mathbf{J}(\mathbf{x})\mathbf{w}_i(\mathbf{x})dx \qquad \forall i \in [1;\mathrm{N}] \tag{6}$$

Replacing **A** by its expression (5) in (6), a system of N linear equations with N unknown coefficients a_i is obtained which can be written in the form:

$$\mathbf{S}\,\mathbf{A} = \mathbf{F} \tag{7}$$

with **S** the stiffness matrix (NxN), **F** the source vector (Nx1) and **A** the vector of the coefficients a_i. We should mention that non-homogeneous boundary conditions can be taken into account, additional entries are then added to the source vector **F**. The coefficients s_{ij} of **S** and f_i of **F** satisfy:

$$s_{ij} = \int_D \nu(\mathbf{x})\mathbf{curl}\mathbf{w}_j(\mathbf{x}).\mathbf{curl}\mathbf{w}_i(\mathbf{x})dx \quad f_i = \int_D \mathbf{J}(\mathbf{x}).\mathbf{w}_i(\mathbf{x})dx \tag{8}$$

Once the equation system (6) is solved, local quantities like the magnetic flux density distribution or global quantities of interest like the flux, the torque can be calculated in a post processing step.

2.2 Stochastic Problem

In the deterministic case, the input parameters like the dimensions related to the geometry of the device, the material characteristics and the electromagnetic sources are supposed to be perfectly known. If the input parameters of the model are subject to variability, the solution of (7) will be also subject to variability. The stochastic approach enables to quantify this variability. When accounting for the uncertainties using the stochastic approach, the input parameters are then modelled by random variables $\mathbf{p}(\theta)$ with θ an elementary event. The joint probability density function is supposed to be known (or each marginal probability density functions if the random variables are independent). The outputs of the electromagnetic model become then random and should be characterized. A stochastic partial differential equation system is generally numerically solved by applying, like in the deterministic case (see Sect. 2.1), a semi-discretisation in space [see (6)]. The DoF's a_i of the vector potential [see (5)], which were real numbers in the deterministic case, becomes random variables $a_i[\mathbf{p}(\theta)]$. The matrix $\mathbf{S}$ and the vector $\mathbf{F}$ have random entries $s_{ij}[\mathbf{p}(\theta)]$ and $f_i[\mathbf{p}(\theta)]$ and the unknown vector $\mathbf{A}$ is random and satisfies:

$$\mathbf{S}[\mathbf{p}(\theta)]\mathbf{A}[\mathbf{p}(\theta)] = \mathbf{F}[\mathbf{p}(\theta)] \tag{9}$$

As already mentioned above, the input parameters $\mathbf{p}(\theta)$ of the model are related either to the geometry or to the behaviour laws of the material or to the sources (including non-homogeneous boundary conditions). Taking into account the randomness on the source is quite straightforward especially when the deterministic problem is linear [33]. In the following, we will assume that the sources are deterministic. For the other kinds of randomness, the problem is more complicated. The processing of uncertain geometries is slightly different than the processing of uncertain behaviour laws and requires additional treatments. The most natural way to account for randomness on the geometry consist in remeshing according to the deformation but the remeshing leads to a discontinuous solution in the space of the input parameters and can create additional numerical noise which can disturb the random solution. Alternatives have been proposed in the literature [28–30, 37, 38, 52] to avoid remeshing. In the following, we will focus mainly on uncertainties on the behaviour laws. However, the quantification methods presented in the following can be applied to solve problems with random geometries as mentioned previously.

To solve (9), sampling techniques, like the Monte Carlo Simulation Methods (MCSM) [22, 34], or perturbation methods [23, 42] can be applied. In this paper, we will focus only on the approximation methods which are well fitted to solve (9) when the entries of the vector $\mathbf{A}(\mathbf{p})$ are smooth functions of the input parameters $\mathbf{p}$.

3 Approximation Methods

We denote G the quantity of interest which can be an unknown of the problem (a value of the circulation $a_i[\mathbf{p}(\theta)]$ of the vector potential along an edge i), a local quantity like the value of the magnetic field or the Joule losses at one point of the domain or a global quantity derived from the magnetic fields like the magnetic flux flowing through a stranded inductor or the magnetic energy. An approximation of the quantity G which is a function of the random input parameters $\mathbf{p}(\theta)$, is sought in a finite dimensional function space of $\mathbf{p}(\theta)$ that is to say:

$$G[\mathbf{p}(\theta)] = \sum_{i=0}^{\mathrm{P}} g_i \Psi_i[\mathbf{p}(\theta)] \tag{10}$$

with g_i coefficients to determine. The approximation functions $\Psi_i[\mathbf{p}(\theta)]$ can be chosen in different finite dimensional spaces [1, 33]. If the output G has a finite variance and is sufficiently "smooth", polynomial expansions are well suited. If it exists some singularities (for example in the case random geometry), other approximation spaces should be introduced [26]. Approximations based on the Polynomial Chaos Expansion (PCE) are currently the most used in engineering. PCE was first introduced by Wiener [50] to represent Gaussian processes. In [51], Xiu et al. proposed a more general approach by referring to the Wiener-Askey scheme. A PCE requires the random components $p_i(\theta)$ of the vector $\mathbf{p}(\theta)$ to be independent. If it is not the case, alternatives are proposed in the literature either to modify the approximation space or to express the vector $\mathbf{p}(\theta)$ as a function of a vector $\mathbf{p'}(\theta)$ of independent random variables (using isoprobabilistic transformation for example). In the following, we will assume the random variables $p_i(\theta)$ independent with a probability density function (pdf) $f_i(y)$. The size of the random vector $\mathbf{p}(\theta)$ will be equal to K. We denote E[X(θ)] the expectation of the random variable X(θ) [the expectation of X(θ) is equal to the mean of X(θ)]. We introduce now the monovariate orthogonal polynomial $\psi_i^l(y)$ of order l associated to the parameter $p_i(\theta)$. The polynomials $\psi_i^l(y)$ are orthogonal with respect to the pdf $f_i(y)$ that is to say:

$$E[\psi_i^l(p_i(\theta))\psi_i^m(p_i(\theta))] = \int_{-\infty}^{+\infty} \psi_i^l(y)\psi_i^m(y) f_i(y)dy = \delta_{lm} \tag{11}$$

with δ_{lm} the kronecker symbol. The determination of the monovariate polynomials $\psi_i^l(y)$ is not an issue whatever the pdf of $f_i(y)$ (see [51]). We define now the set of multivariate orthogonal polynomials $\Psi_{\boldsymbol{\alpha}}[\mathbf{p}(\theta)]$ with $\boldsymbol{\alpha}$ a K-tuple such that:

$$\Psi_{\boldsymbol{\alpha}}(\mathbf{p}(\theta)) = \prod_{i=1}^{K} \psi_i^{\alpha_i}(p_i(\theta)) \text{ with } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)\ \alpha_i \in \mathrm{N} \tag{12}$$

Since $\mathbf{p}(\theta)$ is a vector of independent random variables, the multivariate polynomials $\Psi_{\boldsymbol{\alpha}}(\mathbf{p}(\theta))$ are orthogonal with respect to the joint probability density function $\prod_{i=1}^{K} f_i$ and we have $\mathrm{E}[\Psi_{\boldsymbol{\alpha}}(\mathbf{p}(\theta))\Psi_{\boldsymbol{\beta}}(\mathbf{p}(\theta))]=0$ if $\alpha \neq \beta$. If the random variable $G[\mathbf{p}(\theta)]$ has a finite variance, the PCE refers to the representation of $G[\mathbf{p}(\theta)]$ as a linear combination of multivariate polynomials $\Psi_{\boldsymbol{\alpha}}(\mathbf{p}(\theta))$:

$$G[\mathbf{p}(\theta)] = \sum_{\alpha_1=0}^{+\infty} \cdots \sum_{\alpha_K=0}^{+\infty} g_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}[\mathbf{p}(\theta)] \tag{13}$$

In practice, the expansion (13) is truncated up to the multivariate polynomials of order p (the sum $\alpha_1 + \ldots + \alpha_K$ is lower or equal than p). The total number of multivariate polynomials P to be considered is:

$$\mathrm{P} = \frac{(\mathrm{K}+\mathrm{p})!}{\mathrm{K!p!}} \tag{14}$$

In Table 1, we have reported the number of multivariate polynomials P of the space of approximation as a function of the maximum polynomial order p and the number K of input parameters. We can see that P increases exponentially with K which is usually so-called the *curse of dimensionality*.

In the following, to simplify the notation, the multivariate polynomials $\Psi_{\boldsymbol{\alpha}}(\mathbf{p}(\theta))$ will be indexed by an integer i ($1 \leq i \leq P$) instead of the K-tuple $\boldsymbol{\alpha}$. The function $G[\mathbf{p}(\theta)]$ is approximated by a truncated expansion given by (10) of orthogonal multivariate polynomials defined by (12). As already mentioned previously, after applying the semi-discretisation in space, the terms $a_i[\mathbf{p}(\theta)]$ of the decomposition of the vector potential $\mathbf{A}[\mathbf{x}, \mathbf{p}(\theta)]$ are random [see (5)]. Each term $a_i[\mathbf{p}(\theta)]$ is approximated using a truncated PCE (13). Finally, the vector potential $\mathbf{A}[\mathbf{x}, \mathbf{p}(\theta)]$ is approximated by the expression:

$$\mathbf{A}[\mathbf{x}, \mathbf{p}(\theta)] = \sum_{i=0}^{\mathrm{N}} \sum_{j=0}^{\mathrm{P}} a_{ij} \Psi_j(\mathbf{p}(\theta)) \mathbf{w}_i(\mathbf{x}) \tag{15}$$

The number of coefficients a_{ij} is equal to NxP. It is not seldom to meet in practise deterministic models with a number of unknowns N of order 10^5. According to (14) and Table 1, the unknown number NxP can be quickly very huge (of order 10^8) if the number K of random input parameters is higher than a dozen.

Table 1 Example of the multivariate polynomial number as a function of the maximum multivariate polynomial order p and the number of random inputs K

	$p=1$	$p=2$	$p=3$	$p=4$
$K=2$	3	6	10	15
$K=5$	6	21	56	126
$K=10$	11	66	286	1001
$K=20$	21	231	1771	10,626

In a postprocessing step, quantities of interest (energy, flux,...) can be also expressed using (10). Among the method proposed in the literature to determine these coefficients, some are called non-intrusive since they encapsulate a deterministic model in an environment of stochastic procedures. A preprocessor generates a sample of parameter values according to their probability density function. A deterministic model is then run for each set of parameter values of the sample and a new sample of output values is then obtained. From this sample, a postprocessor determines the approximation of the output. Collocation [14], regression [5] and projection methods belongs to this group of non-intrusive methods. Some stochastic methods, so-called intrusive methods, require to access to the heart of the deterministic model to be implemented like the Spectral Stochastic Finite Element Method. In the following, to illustrate the main principles of intrusive and non-intrusive methods, we will present the projection method and the Spectral Stochastic Finite Element Method.

3.1 A Non-intrusive Method: Projection Method

Since the polynomials $\Psi_i(\mathbf{p}(\theta))$ are orthogonal, the coefficients a_{ij} satisfy:

$$a_{ij} = \frac{E[a_i(\mathbf{p}(\theta))\Psi_j(\mathbf{p}(\theta))]]}{E[\Psi_j^2(\mathbf{p}(\theta))]]} \tag{16}$$

$$= \frac{\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} a_i(p_1, \dots, p_K)\Psi_j(p_1, \dots, p_K)f_1(p_1)\dots f_K(p_K)dp_1 \dots dp_K}{\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \Psi_j^2(p_1, \dots, p_K)f_1(p_1)\dots f_K(p_K)dp_1 \dots dp_K}$$

The determination of a_{ij} yields the calculation of multidimensional integrals. The denominator of (16) can be calculated generally analytically ($d_j = E[\Psi_j^2(\mathbf{p}(\theta))]$) but not the numerator. Different methods can be used to approximate the multidimensional integral: MCSM, Gauss quadrature methods, sparse grid methods, adaptive integration schemes... [3, 27]. All of them yield the following expression for the approximation:

$$a_{ij} = \frac{\sum_{l=1}^{Q} a_i(\mathbf{p}^l)\Psi_j(\mathbf{p}^l)w^l}{d_j} \tag{17}$$

where w^l are the weights and $\mathbf{p}^l = (p_1^l, \dots, p_K^l)$ the Q evaluation points. The model (9) is solved for Q sets of the input parameters $\mathbf{p}^l$ to determine $a_i(\mathbf{p}^l)$ that is to say that the deterministic model (7) has to be solved Q times with $\mathbf{p}^l$ as input parameters. One should notice that Q can increase dramatically with K. Let consider for example a Gauss quadrature of order q_i along the random direction i associated to each parameter p_i ($1 \leq i \leq K$). We denote by p_i^l $1 \leq l \leq q_i$ the evaluation points and w_i^l $1 \leq l \leq q_i$ the associated weights. The points are the roots of the

polynomial $\psi_i^q(y)$ of order q introduced in (11). A multidimensional quadrature can be obtained by tensorizing the monodimensional gauss quadratures along each random dimension. In that case, the number of evaluation points Q is equal to $q_1 q_2 \dots q_K$ and so increases exponentially with the number K of parameters. The number of evaluation points can be reduced by using sparse grids like Smolyak cubature [46] but the exponential increasing with the dimension remains.

The coefficients g_i of the approximation of any quantity of interest $G[\mathbf{p}(\theta)]$, like the flux or the force, can be determined using the same approach. If G is the only quantity of interest for the user, there is no need to access to the $a_i[\mathbf{p}(\theta)]$'s. The deterministic model is run Q times, as a black box, with the different parameter values $\mathbf{p}^l$ to calculate Q evaluations $G[\mathbf{p}^l]$. From the $G[\mathbf{p}^l]$'s, the coefficients g_i are approximated using a quadrature formula (see 17). The non-intrusive approach is very convenient because the coupling with existing deterministic models, especially commercial software, is straightforward.

One should note that the non-linearities on the behaviour laws are naturally taken into account within the deterministic model that is to say that the non-intrusive method is the same when dealing with either a linear model or a non-linear model.

3.2 Galerkin Method: Stochastic Finite Element Method

To solve stochastic partial differential equations, the Galerkin approach was first introduced in the early 1990s by Ghanem et al. in mechanics [19]. It consists in searching the solution in a tensorial space $W(D) \otimes \mathbb{P}_P^K$ with $W(D)$ the standard finite element space used in the deterministic case and $\mathbb{P}_P^K$ the space of approximation of random variables spanned by the basis functions $(\Psi_i[\mathbf{p}(\theta)])_{1 \le i \le P}$ introduced previously [see (10)]. In magnetostatics, the vector potential is sought in a space generated by the basis function $\Psi_j(\mathbf{p}(\theta))\mathbf{w}_i(\mathbf{x})$. The solution should satisfy a weak form of the initial problem. Let consider again our magnetostatic problem, the weak form (6) is extended in the stochastic case and can be written [9]:

$$E\left[\int_D \nu(\mathbf{x}, \mathbf{p}(\theta))\mathbf{curl}\,\mathbf{A}(\mathbf{x}, \mathbf{p}(\theta)).\mathbf{curl}\mathbf{w}_i(\mathbf{x})dx\,\Psi_j(\mathbf{p}(\theta))\right] \tag{18}$$

$$= E\left[\int_D \mathbf{J}(\mathbf{x})\mathbf{w}_i(\mathbf{x})dx\,\Psi_j(\mathbf{p}(\theta))\right] \qquad \forall i \in [1;\mathrm{N}] \quad and \quad \forall j \in [1;\mathrm{P}]$$

Replacing $\mathbf{A}[(x), \mathbf{p}(\theta)]$ in (18) by its expression (15) and applying the weak formulation for the NxP test functions $\Psi_j(\mathbf{p}(\theta))\mathbf{w}_i(\mathbf{x})$, a NxP equation system is obtained:

$$\mathbf{S}_s\mathbf{A}_s = \mathbf{F}_s \tag{19}$$

with $\mathbf{S}_s$ a (NxP)x(NxP) matrix, $\mathbf{A}_s$ the (NxP) vector of the unknowns a_{ij} and $\mathbf{F}_s$ a (NxP) vector. The "intrusivity" of the method is related to the fact that the entries

of $\mathbf{S}_s$ and $\mathbf{F}_s$ are integral functions of $\Psi_j(\mathbf{p}(\theta))$ and $\mathbf{w}_i(\mathbf{x})$. Their calculation requires to have access to the procedures of the calculation of the terms s_{ij} and f_i [see (8)] of the deterministic model. The size of the system (NxP) can be extremely large preventing the storage of the matrix $\mathbf{A}_s$ and so its solution. If the reluctivity can be written as a sum of separable functions like,

$$\nu(\mathbf{x}, \mathbf{p}(\theta)) = \sum_{i=1}^{M} \nu_i[\mathbf{p}(\theta)] g_i(\mathbf{x}) \tag{20}$$

the system (19) can be rewritten taking advantage of the Kronecker product [43]. This representation of the reluctivity as a sum of separable functions can be obtained either during the process of probabilistic modelling of the input data or by applying a model reduction technique (Karuhnen-Loeve expansion for example). According to this new expression, the matrix $\mathbf{S}_s$ can be written in the form [16]:

$$\mathbf{S}_s = \sum_{i=1}^{M} \mathbf{C}_i \otimes \mathbf{D}_i \tag{21}$$

The memory space required can be significantly reduces by storing only the 2M matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ with $\mathbf{C}_i$ depending only on the functions $\mathbf{w}_i(\mathbf{x})$ and $\mathbf{D}_i$ on the functions $\Psi_j(\mathbf{p}(\theta))$. It should be noticed that the matrices $\mathbf{C}_i$ can be easily extracted from a deterministic standard finite element code. The determination of the matrix $\mathbf{S}_s$ does not require a high modification of the deterministic code and so the "intrusivity" of the Galerkin approach can be highly alleviated using expression based on separable functions. This approach can be extended to quasistatics. Besides, dedicated solvers can be employed to solve the Eq. (19) by taking advantage the expression (21) based on Kronecker products. Accounting for non-linearities in the Galerkin approach is more tricky than in the non-intrusive case but remains possible [44]. The Galerkin method, for given approximation spaces $W(D)$ and $\mathbb{P}_P^K$, minimizes the error of approximation in the "L2" sense which is not the case with other approximation methods based on the evaluations of the deterministic model (non intrusive methods like projection method, collocation method, regression method). However, when a multivariate double orthogonal polynomial expansion is used to approximate the stochastic dimension then the collocation and the Galerkin methods are equivalent [9].

4 Applications

Approximation methods have been already applied in computational electromagnetics to study EEG Source Analysis [17], Eddy Current in human body [18], Eddy Current Non Destructive Testing [3, 4], Accelerator Cavities and Magnets [2, 12, 45], Dosimetry [49], electrical machines [31, 39]... The development and

the application of such models have started in the early 2000s and know a growing interest in the community. The methods have been evaluated on academic examples [9, 44] but one can notice a trend towards more and more realistic applications which shows that the stochastic approach is getting more and more mature in the community of computational electromagnetics. In [17], the projection method has been applied uncertainties in the EEG source analysis. In [14], a 2D dosimetry problem has been tested by comparing several non-intrusive approximation methods and the Monte Carlo Simulation Method. It has been shown that the approximation methods enable to reduce dramatically the number of evaluation points compared to sampling technics. An Eddy Current-Non Destructive Testing problem where some material characteristics are assumed to be random has been addressed. Samples were not accessible for measurement (nuclear application) [35] to determine the conductivity and the permeability of material like magnetite deposit. The lack of knowledge was modelled by a stochastic approach considering the material characteristic as random variables. The aim was to determine the influence of this lack of knowledge on the model output, here the output sensor. In this application, a sensitivity analysis showed that only one material characteristic among the 6 considered has an influence on the variability of the sensor output. In other words, only the lack of knowledge of one material characteristic (p) has an influence on the accuracy of the model. Consequently, to improve the accuracy, investigation shall focus on the parameter p and not on the others. This study shows that the stochastic approach is a powerful tool for improving the accuracy of models by determining the input parameters whose uncertainties (due to a lack of knowledge) strongly influence the quantity of interest. It can also be very helpful to develop indicators based on measurements that are robust, that is to say that these indicators are few influenced by the variability introduced by the imperfections on the device studied. To solve this problem, the Galerkin method and a Projection method are compared [3, 4]. It shows that the Galerkin approach can be competitive compared to a non-intrusive approach. The influence of the lack of knowledge on the B(H) curve of the ferromagnetic material has been also addressed in the case of a turbo alternator [32]. The global sensitivity analysis based on the Sobol approach [11, 47, 48] allows to determine the most influential parameters of the B(H) curve. It appears that the magnetic flux density is the most influential but not the magnetic field H in the saturation area. The proposed approach provides the quantity of interest domain where the parameter uncertainties are the most influential and then allows to act in order to reduce their variability by increasing the accuracy of the measurement in the corresponding area.

The influence of the dimension and material characteristics variability on the performances of an electrical machines produced in mass is also studied when the number of random parameters is about a dozen [15, 31, 39]. The aim is to propose a methodology based on a stochastic approach to assess the influence of the variability of the manufacturing process on the performances of the electrical machines which can be applied in robust design. The tolerancing using the stochastic approach has been also studied for a permanent magnet machine [24].

5 Facing the Curse of Dimensionality

If we want to go further with the stochastic approach which can be very useful to solve numerous problems in engineering, the *curse of dimensionality* should be overcome in order to be able to deal with real world problems where the number of parameters is often greater than the dozen. In the following, we will present briefly methods that have been proposed recently to overcome this challenge. We will keep the distinction between non-intrusive and intrusive methods.

5.1 Non-intrusive Methods

First, to limit the number of calls of the deterministic model which grows exponentially with the number of random parameters, the number of quadrature points q_i (see Sect. 3.1) can not be the same along each random direction $p_i(\theta)$. In fact, if a parameter p_i has almost no influence on the variability of the quantity of interest G then G needs to be evaluated only on one quadrature point p_i^1 along the dimension i, which limits the number of evaluations Q (Q=$q_1 \times .. \times q_{i-1} \times 1 \times q_{i+1} \times \ldots \times q_K$). The number of quadrature points is optimized automatically based an error indicator which can be for example the value of the variance of the quantity of interest. Adaptive methods coupled with sparse grids and nested quadrature scheme have shown their efficiency on practical application [3]. However, with a high parameter number, the expansion based on truncated PCE becomes too large [see (14)]. To limit the number of terms, a sparse basis should be constructed which can be determined from the adaptive scheme or directly from a random sampling of the quantity of interest. In [6, 13], the most significant terms of the PCE are extracted using iterative algorithm aiming at reducing not only the error of approximation but also the number of terms of the expansion. These methods are efficient if a small fraction of coefficients g_i in the exact expression (10) of the quantity of interest are dominant.

Another alternative to reduce the number of terms of the expansion is to decompose the quantity of interest under a sum of separable functions $G[\mathbf{p}(\theta)] = \sum_{j=1}^{T} u_1^j[p_1(\theta)] \ldots .u_K^j[p_K(\theta)]$ with T the tensor rank. The functions $u_i^j[p_j(\theta)]$ are the unknowns of the problem and are sought in a one dimensional space for example the space generated by the polynomials $\psi_i^l[p_i(\theta)]$ [see (11)]. The calculation of the optimal low rank approximation (the value of T as smaller as possible) is a difficult task. Methods have been recently proposed in the literature to tackle this issue [40] for stochastic problems.

Finally, an adaptive interpolation technique is proposed in [8] to determine a sparse polynomial approximation using an iterative procedure. The evaluation points $\mathbf{p}_l$ are determined iteratively by comparing the error between the approximation and the full model. These evaluation points must satisfy an admissible

condition in order to obtain interpolant polynomials. For a class of parametric elliptic problems, a fast convergence of the method has been proved.

5.2 *Intrusive Methods*

We have seen that the application of the Galerkin Method requires the solution of a huge equation system (9) of size NxP. Under separability condition on the behaviour law, this system of equations can be written in the form of (21) which alleviates the storage space requirement. Dedicated solvers can be applied [4, 25] but it does not decrease the size of the equation system. Model Order Reduction Methods like Proper Orthogonal Decomposition (POD), Reduced Basis Method enables to reduce the stochastic problem (9) to solve to an order $R \leq\leq N$ (N is the number of DoF's of the spatial mesh)[21]. The unknown vector $\mathbf{A}[\mathbf{p}(\theta)]$ is approximated by $\sum_{i=1}^{R} a_i^r[\mathbf{p}(\theta)]\mathbf{A}_i^r$ with $\mathbf{A}_i^r$ solutions of (9) for a given set of parameters $(\mathbf{p}^1, \ldots, \mathbf{p}^R)$.[1] Replacing $\mathbf{A}[\mathbf{p}(\theta)]$ in (9) leads to an overdetermined system of N equations with R unknowns. Then, by applying the Galerkin method for example, a reduced equation system of R equations with R unknowns is obtained under the form $\mathbf{S}_r[\mathbf{p}(\theta)]\mathbf{A}^r[\mathbf{p}(\theta)] = \mathbf{F}[\mathbf{p}(\theta)]$. The R functions $a_i^r[\mathbf{p}(\theta)]$ becomes the unknowns which are then approximated by the expression (10) that is to say $a_i^r[\mathbf{p}(\theta)] = \sum_{j=0}^{\mathrm{P}} a_{ij}^r \Psi_i[\mathbf{p}(\theta)]$. The terms a_{ij}^r can be determined by applying the methods presented in (3.2) or (3.1). This approach has been applied to solve a dosimetry problem where the reduced basis method and a non-intrusive collocation have been combined [14]. The efficiency of the model order reduction method relies on the choice of the reduced basis spanned by the $\mathbf{A}_i^r$. Error indicators, available in the literature, can help for the determination of the reduced basis.

Another approach has been proposed in [36] and applied recently in electromagnetism in [10] called the Proper Generalized Decomposition (PGD) [7, 10]. The idea is to search a solution under the form:

$$\mathbf{A}[\mathbf{x}, \mathbf{p}(\theta)] = \sum_{i=1}^{T} a_i^{PGD}[\mathbf{p}(\theta)]\mathbf{A}_i^{PGD}(\mathbf{x}) \tag{22}$$

with $\mathbf{A}_i^{PGD}(\mathbf{x})$ in $W(D)$ and $a_i^{PGD}[\mathbf{p}(\theta)]$ in $\mathtt{P}_P^K$ [see (3.2)]. The couple of functions $(a_i^{PGD}[\mathbf{p}(\theta)],\mathbf{A}_i^{PGD}(\mathbf{x}))$ is determined iteratively from the previous couples $(a_j^{PGD}[\mathbf{p}(\theta)],\mathbf{A}_j^{PGD}(\mathbf{x}))$ $1 \leq j \leq i-1$. The process is stopped when the contribution of the couple $(a_i^{PGD}[\mathbf{p}(\theta)],\mathbf{A}_i^{PGD}(\mathbf{x}))$ is "sufficiently" small. The term $a_i^{PGD}[\mathbf{p}(\theta)]$ satisfies a system of P equations which depends on the terms $\mathbf{A}_i^{PGD}(\mathbf{x})$ and the

[1] The vectors $\mathbf{A}_i^r$ must be linearly independent to enforce the uniqueness of the solution of the reduced problem. If it is not the case, a Singular Value Decomposition (SVD) or a Gram-Schmidt process can be applied to obtain linearly independent vectors.

term $\mathbf{A}_i^{PGD}(\mathbf{x})$ a system of N equations which depends on the functions $a_i^{PGD}[\,\mathbf{p}(\theta)]$. The determination of $(a_i^{PGD}[\,\mathbf{p}(\theta)],\mathbf{A}_i^{PGD}(\mathbf{x}))$ requires the solution of two coupled equation systems of size N and P which are usually solved iteratively using a fixed point method. If T couples are required we can see that we have only T×(N+P) unknowns instead of NxP in the Galerkin approach [see (3.2)]. If the number T of couples to approximate correctly the solution is small, this method is very interesting in terms of memory storage and computation time. Moreover, under "separabilty" conditions on the behaviour law [see (20)], the term $a_i^{PGD}[\,\mathbf{p}(\theta)]$ can be sought under the following separable form:

$$a_i[\,\mathbf{p}(\theta)] = \prod_{j=1}^{K} a_{ij}^{PGD}[p_j(\theta)] \tag{23}$$

Then $a_i[\,\mathbf{p}(\theta)]$ is obtained by solving K one dimensional problems which avoid the curse of dimensionality when the number K of parameters is too large. The PGD remains intrusive in the sense that, to be implemented, numerous additional developments in a deterministic software are required. However, recently, a method has been proposed to compute an approximation of the solution based on simple evaluations of the residual of the deterministic problem [20].

6 Conclusion

In this paper, we have presented approximation methods to solve stochastic problems based on partial differential equations. Examples of application in computational electromagnetism have been presented showing that the stochastic approach based on approximation methods provide very useful tools for the study and the design of electromagnetic devices. It has been shown that when the number of random parameters is high, the approximation can leads to an unsolvable problem (*curse of dimensionality*). To face this issue, recent methods proposed in the literature have been listed.

Acknowledgements This work has been supported by the pole MEDEE funded by the Nord Pas de Calais Region and the European Union and also by the Arts et Métiers Fundation.

References

1. Babuska, I., Tempone, R., Zouraris, E.: Galerkin finite element approximation of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. **42**(2), 800–825 (2004)
2. Bartel, A., De Gersem, H., Hülsmann, T., Romer, U., Schops, S., Weiland, T.: Quantification of uncertainty in the field quality of magnets originating from material measurements. IEEE Trans. Magn. **49**(5), 2367–2370 (2013)

3. Beddek, K., Clénet, S., Moreau, O., Costan, V., Le Menach, Y., Benabou, A.: Adaptive method for non-intrusive spectral projection application on a stochastic eddy current NDT problem. IEEE Trans. Magn. **48**(2), 759–762 (2012)
4. Beddek, K., Clénet, S., Moreau, O., Le Menach, Y.: Solution of large stochastic finite element problems – application to ECT-NDT. IEEE Trans. Magn. **49**(5), 1605–1608 (2013)
5. Berveiller, M., Sudret, B., Lemaire, M.: Stochastic finite elements: a non-intrusive approach by regression. Eur. J. Comput. Mech. **15**(1–3), 81–92 (2006)
6. Blatman, G., Sudret, B.: Adaptive sparse polynomial chaos expansion based on least angle regression. J. Comput. Phys. **230**(6), 2345–2367 (2011)
7. Chinesta, F., Ladeveze, P., Cueto, E.: A short review on model order reduction based on proper generalized decomposition. Arch. Comput. Methods Eng. **18**(4), 395–404 (2011)
8. Chkifa, A., Cohen, A., Schwab, C.: High-dimensional adaptive sparse polynomial interpolation and application to parametric PDEs. Found. Comput. Math. **14**, 601–633 (2014)
9. Clénet, S., Ida, N., Gaignaire, R., Moreau, O.: Solution of dual stochastic static formulations using double orthogonal polynomials of static field. IEEE Trans. Magn. **46**(8), 3543–3546 (2010)
10. Codecasa, L., Di Rienzo, L.: Generalised spectral decomposition approach to a stochastic finite integration technique electrokinetic formulation. In: CEM 2014, London (2014)
11. Crestaux, T., Le Maitre, O., Martinez, J.M.: Polynomial chaos expansion for sensitivity analysis. Reliab. Eng. Syst. Saf. **94**(7), 1161–1172 (2009)
12. Deryckere, J., Masschaele, B., De Gersem, H., Steyaert, D.: Stochastic response surface method for dimensioning accelerator cavities. In: OIPE 2012, Gent (2012)
13. Doostan, A., Owhadi, H.: A non-adapted sparse approximation of PDEs with stochastic inputs. J. Comput. Phys. **230**, 3015–3034 (2011)
14. Drissaoui, A., Lanteri, S., Lévêque, P., Musy, P., Nicolas, L., Perrussel, R., Voyer, D.: A stochastic collocation method combined with a reduced basis method to compute uncertainties in numerical dosimetry. IEEE Trans. Magn. **48**(2), 563–566 (2012)
15. Fratila, M., Ramarotafika, R., Benabou, A., Clénet, S., Tounzi, A.: Stochastic post-processing calculation of iron losses – application to a PMSM. COMPEL **32**(4), 1383–1392 (2013)
16. Gaignaire, R., Clénet, S., Moreau, O., Guyomarch, F., Sudret, B.: Speeding up in SSFEM computation using Kronecker tensor products. IEEE Trans. Magn. **45**(3), 1432–1435 (2009)
17. Gaignaire, R., Crevecoeur, G., Dupreé, L., Sabariego, R.V., Dular, P., Geuzaine, C.: Stochastic uncertainty quantification of the conductivity in EEG source analysis by using polynomial chaos decomposition. IEEE Trans. Magn. **46**(8), 3457–3460 (2010)
18. Gaignaire, R., Scorretti, R., Sabariego, R.V., Geuzaine, C.: Stochastic uncertainty quantification of Eddy currents in the human body by polynomial chaos decomposition. IEEE Trans. Magn. **48**(2), 451–454 (2012)
19. Ghanem, R., Spanos, P.D.: Stochastic Finite Elements: A Spectral Approach. Dover, New York (2003)
20. Giraldi, L., Liu, D., Matthies, H.G., Nouy, A.: To be or not to be intrusive? The solution of parametric and stochastic equations – proper generalized decomposition. arXiv:1405.0875v1 [math.NA] (2014)
21. Haasdonk, B., Urban, K., Wieland, B.: Reduced basis methods for parameterized partial differential equations with stochastic influences using the Karhunen-Loeve expansion. J. Uncertain. Quantif. **1**(1), 79–105 (2013)
22. Hammersley, J.M., Handscomb, D.C.: Monte Carlo Methods. Chapman and Hall, London (1964)
23. Harbrecht, H., Schneider, R., Schwab, C.: Sparse second moment analysis for elliptic problems in stochastic domains. Numer. Math. **109**, 385–414 (2008)
24. Kim, Y., Hong, J., Hur, J.: Torque characteristic analysis considering the manufacturing tolerance for electric machine by stochastic response surface method. IEEE Trans. Ind. Appl. **39**(3), 713–719 (2003)

25. Le Maitre, O., Knio, O.M.: Spectral Methods for Uncertainty Quantification with Applications to Computational Fluid Dynamics. Springer Series Scientific Computation. Springer, Dordrecht (2010)
26. Le Maitre, O.P., Knio, O.M., Najm, H.N., Ghanem, R.G.: Uncertainty propagation using Wiener-Haar expansions. J. Comput. Phys. **197**, 28–57 (2004)
27. Liu, M., Gao, Z., Hesthaven, J.S.: Adaptive sparse grid algorithms with applications to electromagnetic scattering under uncertainty. Appl. Numer. Math. **61**(1), 24–37 (2011)
28. Mac, H., Clénet, S., Mipo, J.C., Moreau, O.: Solution of static field problems with random domains. IEEE Trans. Magn. **46**(8), 3385–3388 (2010)
29. Mac, H., Clénet, S., Mipo, J.C.: Transformation method for static field problem with random domains. IEEE Trans. Magn. **47**(5), 1446–1449 (2011)
30. Mac, H., Clénet, S., Mipo, J.C.: Comparison of two approaches to compute magnetic field in problems with random domains. IET Sci. Meas. Technol. **6**(5), 714–721 (2012)
31. Mac, H., Clénet, S., Zheng, S., Coorevits, T., Mipo, J.C.: On the geometric uncertainties of an electrical machine: stochastic modeling and impact on the performances. In: COMPUMAG 13, Budapest (2013)
32. Mac, H., et al.: Influence of uncertainties on the B(H) curves on the flux linkage of a turboalternator. Int. J. Numer. Modell. **27**, 385–399 (2014)
33. Matthies, H.G., Keese, A.: Galerkin method for linear and non-linear elliptic stochastic partial differential equations. Comput. Methods Appl. Mech. Eng. **194**, 1295–1331 (2005)
34. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, H.A., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
35. Moreau, O., Beddek, K., Clénet, S., Le Menach, Y.: Stochastic non destructive testing simulation: sensitivity analysis applied to material properties in clogging of nuclear power plant steam generator. IEEE Trans. Magn. **49**(5), 1873–1876 (2013)
36. Nouy, A.: A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations. Comput. Methods Appl. Mech. Eng. **196**(37–40), 4521–4537 (2007)
37. Nouy, A., Clément, A.: eXtended stochastic finite element method for the numerical simulation of heterogeneous materials with random material interfaces. Int. J. Numer. Methods Eng. **83**(10), 1312–1344 (2010)
38. Nouy, A., Clément, A., Schoefs, F., et al.: An extended stochastic finite element method for solving stochastic partial differential equations on random domains. Comput. Methods Appl. Mech. Eng. **197**(51–52), 4663–4682 (2008)
39. Offermann, P., Mac, H., Nguyen, T.T., Clénet, S., De Gersem, H., Hameyer, K.: Uncertainty quantification and sensitivity analysis in electrical machines with stochastically varying machine parameters. In: CEFC 14, Grenoble (2014)
40. Rai, P., Chevreuil, M., Nouy, A., Lebrun, R.: A regression based method using sparse low rank approximations for uncertainty propagation. In: 7th International Conference on Sensitivity Analysis of Model Output-SAMO (2013)
41. Ramarotafika, R., Benabou, A., Clénet, S., Mipo, J.C.: Experimental characterization of the iron losses variability in stators of electrical machines. IEEE Trans. Magn. **48**(4), 629–1632 (2012)
42. Romer, U., Schops, S., Weiland, T.: Approximation of moments for the nonlinear manetoquasistatics problem with material uncertainties. IEEE Trans. Magn. **50**(2), 417–420 (2014)
43. Rosseel, E., Vandewalle, S.: Iterative solvers for the stochastic finite element method. SIAM J. Sci. Comput. **32**(1), 372–397 (2010)
44. Rosseel, E., De Gersem, H., Vandewalle, H.: Spectral stochastic simulation of a ferromagnetic cylinder rotating at high speed. IEEE Trans. Magn. **47**(5), 1182–1185 (2011)
45. Schmidt, C., Flisgen, T., Heller, T., Van Rienen, U.: Comparison of techniques for uncertainty quantification of superconducting radio frequency cavities. In: International Conference on Electromagnetics in Advanced Applications 2014 (ICEAA 2014), pp. 117–120 (2014)
46. Smolyak, S.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dokl. Akad. Nauk SSSR **4**, 240–243 (1963)

47. Sobol, I.M.: Sensitivity estimates for non linear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55**, 271–280 (2001)
48. Sudret, B.: Global sensitivity analysis using polynomial chaos expansions. Reliab. Eng. Syst. Saf. **93**, 964–979 (2008)
49. Voyer, D., Musy, F., Nicolas, L., Perrussel, R.: Probabilistic methods applied to 2D electromagnetic numerical dosimetry. Int. J. Comput. Math. Electr. Electron. Eng. **27**(3), 651–667 (2008)
50. Weiner, N.: The homogeneous chaos. Am. J. Math. **60**(4), 897–936 (1938)
51. Xiu, D., Karniadakis, G.: The Wiener Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)
52. Xiu, D., Tartakovsky, D.M.: Numerical methods for differential equations in random domains. SIAM J. Sci. Comput. **3**, 1167–1185 (2006)

Reduced Basis Modeling for Uncertainty Quantification of Electromagnetic Problems in Stochastically Varying Domains

Peter Benner and Martin W. Hess

Abstract The reduced basis method (RBM) is a model order reduction technique for parametrized partial differential equations (PDEs) which enables fast and reliable evaluation of the transfer behavior in many-query and real-time settings. We use the RBM to generate a low order model of an electromagnetic system governed by time-harmonic Maxwell's equations. The reduced order model then makes it feasible to analyze the uncertainty in the model by a Monte-Carlo simulation. Stochastic collocation is employed as a second technique to estimate the statistics. In particular the combination of model order reduction and stochastic collocation allows low computation times compared to Monte-Carlo simulations. We compare the accuracy of Monte-Carlo simulation Hermite Genz-Keister stochastic collocation and the RBM to compute the transfer function under uncertain geometric parameters.

1 Introduction

As the simulation of integrated circuits requires a significant amount of computational power, the simulation of large-scale models benefits from using model order reduction (MOR) techniques. The original system size of order 10^4 and higher is typically reduced to a dimension of less than 100, which allows to examine the frequency response of parametrized systems using the reduced order model. Of particular interest are small random variations in geometry, due to inaccuracies in the production process. A possible extension of this work is a coupling to a heat problem, which introduces a temperature-dependent conductivity, and would add a natural nonlinearity. The influence of the geometric variations is measured in the expectation and variance of the transfer function.

As a sample application we consider a coplanar waveguide, which is governed by time-harmonic Maxwell's equations. The parametric model reduction technique

P. Benner • M.W. Hess (✉)
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany
e-mail: benner@mpi-magdeburg.mpg.de; hessm@mpi-magdeburg.mpg.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_21

we use is the reduced basis method (RBM), which generates low order models to parametrized partial differential equations (PDEs). In recent years, the RBM has been developed to apply to a wide range of problems, see [9] for an overview.

Section 2 introduces the model problem, Sect. 3 reviews the uncertainty quantification problem, while Sect. 4 briefly covers the reduced basis model reduction. The main contribution of the paper is in Sect. 5, where numerical comparisons are performed between Monte-Carlo simulations, Hermite Genz-Keister stochastic collocation and the use of RBM in the computation of the expected transfer behavior and the standard deviation. Section 6 concludes our findings.

2 Model Problem

We treat the second order time-harmonic formulation of Maxwell's equations in the electric field E, in the frequency domain,

$$\nabla \times (\mu^{-1} \nabla \times E) + j\omega\sigma E - \omega^2 \varepsilon E = j\omega J \quad \text{in} \quad \Omega, \tag{1}$$

subject to essential boundary conditions $E \times n = 0$ on Γ_{PEC} and Neumann boundary conditions $(\nabla \times E) \times n = 0$ on Γ_{PMC}, where $\partial\Omega = \Gamma_{\text{PEC}} \cup \Gamma_{\text{PMC}}$ in general. The boundary Γ_{PEC} applies to metal boundaries. A perfect electric conductance (PEC, i.e. $\sigma \to \infty$) is assumed on the boundary Γ_{PEC}. As a consequence, the tangential parts of the electric field vanish, see [10]. In our model of a coplanar waveguide, all boundaries are PEC.

The source current density is denoted by J, the imaginary number j, the frequency ω and the material coefficients are the permeability μ, conductivity σ, and permittivity ε. The field solution is sought in the space $H(\text{curl})$, see [10].

The parameter dependent weak form, with a test function w applied to (1), is established with the sesquilinear form

$$a(E, w; \nu) = \left(\mu^{-1} \nabla \times E, \nabla \times w\right) + j\omega \left(\sigma E, w\right) - \omega^2 \left(\varepsilon E, w\right) \tag{2}$$

using the complex L_2-inner product $(\cdot, \cdot)$ over Ω and linear form $f(w; \nu) = j\omega (J, w)$ as

$$a(E(\nu), w; \nu) = f(w; \nu) \quad \forall w \in \mathscr{X}, \tag{3}$$

using the function space

$$\mathscr{X} = \{u \in H(\text{curl}) | u \times n = 0 \quad \text{on} \quad \Gamma_{\text{PEC}}\}. \tag{4}$$

The parameter vector ν is introduced to denote parametric dependence in frequency ω and geometry. In particular, the sesquilinear form $a(E, w; \nu)$ depends on frequency and geometric variations and the linear form $f(w; \nu)$ on the frequency.

After discretization with $H(\text{curl})$-conforming Nédélec finite elements [10], solving (3) reduces to solving a parameter-dependent sparse linear system $A(\nu)x(\nu) = j\omega b(\nu)$ for the state vector $x(\nu)$, which represents the electric field solution $E(\nu)$ in the discrete space X.

Splitting the state vector x into real and complex parts $x = x_{real} + jx_{imag}$, the complex linear system can be rewritten as an equivalent system of twice the dimension over the real numbers. This leads to a real and symmetric system matrix, which is the form we will use in our computations [5]. The parametric dependence on ν carries over through this transformation. The geometric variations lead to an affine parameter dependence in the bilinear form (2), see [6] for a single, deterministic geometric parameter. The affine decomposition for a single geometry parameter is then extended to multiple parameters by splitting the computational domain into distinct parts and applying the transformation given in [6] to each subdomain.

2.1 Coplanar Waveguide

As an example model, we consider the coplanar waveguide, shown in Fig. 1. The model setup is contained in a shielded box Ω with perfect electric conducting (PEC) boundary. We consider three perfectly conducting strip lines as shown in the geometry. The system is excited at a discrete port (located at $x = 70\,\text{mm}$, $y = 5\,\text{mm}$ and extending from $z = 0\,\text{mm}$ to $z = 10\,\text{mm}$) and the output is taken at a discrete

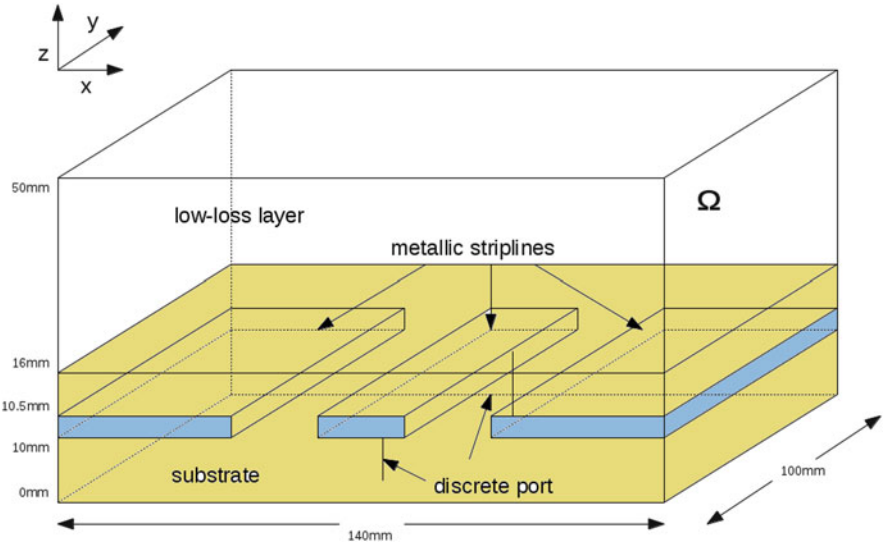


Fig. 1 Geometry of coplanar waveguide. The forcing term J is defined at the discrete port

port on the opposite end of the middle strip line (located at $x = 70\,\text{mm}$, $y = 95\,\text{mm}$ and extending from $z = 0\,\text{mm}$ to $z = 10\,\text{mm}$). These discrete ports are used to model input and output currents/voltages. The metal parts, i.e., the shielded box and the boundaries of the three metal sheets constitute Γ_{PEC}. As a consequence, the interiors of the metallic striplines are not part of the computational domain Ω. The metallic strip lines are immersed in a substrate (at coordinates $z <= 16\,\text{mm}$) of conductivity $\sigma = 0.02\,\text{S/m}$ and relative permittivity $\varepsilon_r = 4.4$. In the low-loss upper layer (at coordinates $z > 16\,\text{mm}$), the conductivity is $\sigma = 0.01\,\text{S/m}$ and relative permittivity is $\varepsilon_r = 1.07$. The relative permeability is $\mu_r = 1$ in the entire domain.

As deterministic parametric variation we look at the frequency $\omega \in [1.3, 1.6]\,\text{GHz}$. The full order simulation has been performed with the finite element package FEniCS, [7]. For our numerical experiments, we used a discretization size of 52,134 degrees of freedom leading to the linear systems $A(\nu)x(\nu) = b(\nu)$.

3 Uncertainty Quantification

Let $(\Omega_p, \mathscr{F}, \mathscr{P})$ denote a probability space. Given is a square integrable random variable $Y : \Omega_p \to \mathbb{R}^p$ with probability density function f and a function $g : \mathbb{R}^p \to \mathbb{R}^d$, where p is the number of geometric parameters. The function g corresponds to a mapping of realizations of a random variable to the output of the electromagnetic system such that $g(Y)$ also is square-integrable, cf. [1].

We consider only geometric stochastic variations. However, the computational approach is also applicable to more general parametric variations. We use Monte-Carlo simulation, Hermite Genz-Keister stochastic collocation and show how reduced basis model reduction can enhance the computational speed. Stochastic collocation computes the expectation by a quadrature rule

$$\mathbb{E}(g(Y)) = \int_{\Gamma} g(x)f(x)dx \approx \sum_{i=1}^{n} g(\xi_i)w_i, \tag{5}$$

where the realizations ξ_i are the sample points, n denotes the sample size and the weights w_i are determined using the probability density function f. The number of sample points depends on the quadrature rule and the number of parameters in an irregular fashion, see [1] and the references therein. Monte-Carlo simulations use equally weighted samples, which are generated using the underlying distribution. A drawback of the Monte-Carlo simulation is its mean convergence rate of $O(1/\sqrt{n})$, while the stochastic collocation is performed with Hermite Genz-Keister sparse grids, generated by the Smolyak algorithm. These types of methods can exhibit a mean convergence rate of $O((\log n)^p/n)$, see [8].

A reduced basis can be computed a priori to generate a reduced order model for the parametric domain of interest. The reduced order model can then be used to perform the stochastic collocation and Monte-Carlo simulation at the sample points.

This significantly reduces computational costs, as each sample point evaluation uses the reduced model. The dominating computational effort then lies in the number of system solves used to generate the reduced model.

4 Reduced Basis Method for Time-Harmonic EM-Problems

The aim of the RBM is to determine a low order space $X_N \subset X$ of dimension N, which approximates the parametric solution manifold $M^\nu = \{E(\nu) \in X | \nu \in \mathscr{D}\}$ well. The reduced basis method composes the space X_N of snapshot solutions $E(\nu_i)$. Given such a space X_N, it is possible to obtain accurate approximations $E_N(\nu) \in X_N$ to $E(\nu)$ by projecting (3) onto X_N, i.e., solve

$$a(E_N(\nu), w_N; \nu) = f(w_N; \nu) \quad \forall w_N \in X_N. \tag{6}$$

The affine decomposition of the bilinear form is given as

$$a(E(\nu), w; \nu) = \sum_{q=1}^{Q_a} \Theta_a^q(\nu) a^q(E(\nu), w), \tag{7}$$

see [9] on how this can be established analytically.

Using the affine decomposition allows evaluating the error estimator with an algorithmic complexity that is independent of the full order discretization size, cf. [9].

The reduced space X_N is built iteratively by a greedy sampling. Starting from an initial reduced space (in our example this is spanned by the snapshots at the expected values of the stochastic parameters), an error indicator is evaluated over the parametric domain. The next snapshot will then be chosen where the maximum of the error indicator is attained. The maximum dimension N is found, once the error indicator reaches a desired tolerance. The error indicator considers the norm of the residual $r(w, \nu) = f(w; \nu) - a(E_N(\nu), w, \nu)$ weighted with a $(\mu_{\omega_p}, \sigma_{\omega_p})$ Gaussian probability density function. We use the subscript ω_p to distinguish between expectation and permeability as well as standard deviation and conductivity. The error indicator is computed at each parameter sample location in each iteration step of the model reduction. We refer to [9] for the efficient computation of the norm of the residual making use of the affine decomposition (7), while the weighted RB has been introduced in [3]. For all $\omega_p \in \Omega_p$, let $\Omega(\omega_p)$ denote the random domain with boundary $\partial\Omega(\omega_p)$. We employ a mapping to a deterministic domain $\overline{\Omega}$ for a reference parameter configuration $\overline{\nu}$ such that we can assemble the system matrices for the domain $\overline{\Omega}$ and use an affine transformation to map to a particular realization $\Omega(\omega_p)$. This approach assumes that each degree of freedom in the mesh has the same meaning under affine transformations and that the field solutions for different geometries are in the same functional space. Thus the dimension of the mesh remains the same for each actual geometry. The geometric variations for the coplanar waveguide are shown in Fig. 2.

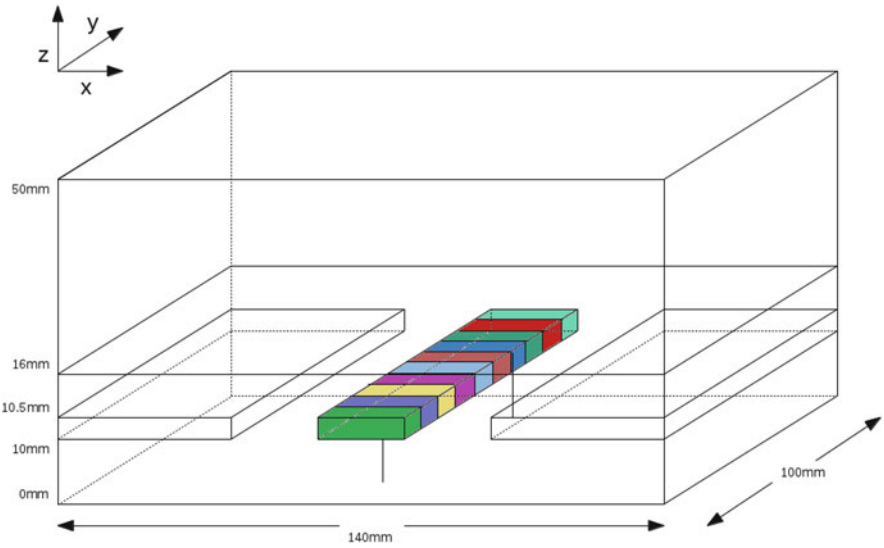


Fig. 2 In the reference configuration, the width of the middle stripline is uniformly 6 mm. The width of the middle stripline varies for each colored part independently in the model with ten geometric parameters. In [6] the affine transformations are shown for a deterministic geometric parameter in the coplanar waveguide. In the model with ten geometric parameters, this transformation is applied to each subsection

5 Numerical Experiments

Each stochastic parameter is modeled such that a corresponding part of the middle stripline width varies as a ($\mu_{\omega_p} = 6$, $\sigma_{\omega_p} = 0.1$) normally distributed random variable. Each stochastic parameter is assumed to be stochastically independent from the others. In Fig. 3 the transfer function is shown for different discretizations of the geometric variation. These results were obtained by Monte-Carlo simulation. To show the capabilities of the RBM in Uncertainty Quantification, we focus on the example with two stochastic parameters (Fig. 4). We employ the SGMGA package [2] for computation of Hermite Genz-Keister integration points and weights.

In Table 1 the Hermite Genz-Keister (HGK) stochastic collocation of order 4 serves as a reference solution for the expectation. As the Monte-Carlo (MC) simulation coincides with other methods only to a degree of 10^{-3}, the HGK of 4th order is a more viable reference choice. The Monte-Carlo simulation of the full system took 14,000 samples. As comparison, the weighted RB has been used in a Monte-Carlo simulation with 14,000 and 20,000 solves, where the computation time is negligible, as the resulting systems are only of size 73×73. This RB model size originates from the greedy sampling, which was set to terminate once the change in the error indicator is $O(1)$. For the stochastic collocation RB the reduced system has been evaluated at the Hermite Genz-Keister (HGK) points. The

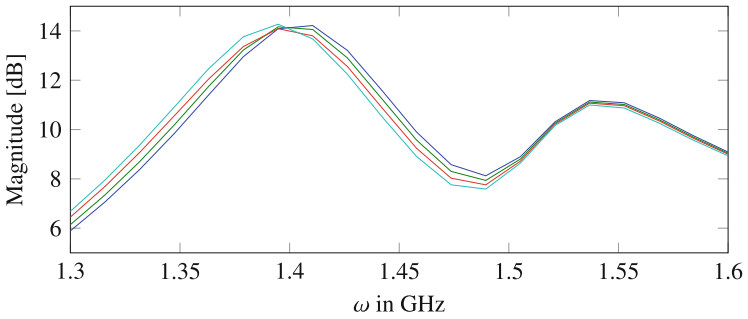


Fig. 3 Magnitude plot over [1.3, 1.6] GHz. The mean of the norm of the transfer function and the phase is shown for four cases: without geometric variation (*blue*), two geometric parameters (*green*), three geometric parameters (*red*) and ten geometric parameters (*light blue*). Computed by Monte-Carlo simulation with a standard deviation of $\sigma_{\omega_p} = 0.1$. Each parameter varies independently as a ($\mu_{\omega_p} = 6$, $\sigma_{\omega_p} = 0.1$) normally distributed random variable

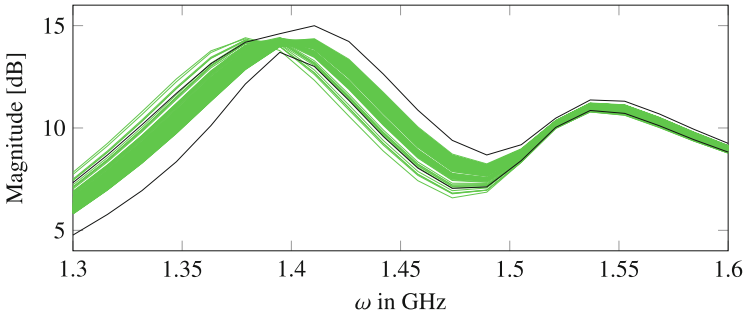


Fig. 4 Relizations of the Monte-Carlo sampling with two geometric parameters (*green*). In *black* the $\pm 3\sigma_{\omega_p}$ deviations from the mean. Of the realizations, 97 % are within the $\pm 3\sigma_{\omega_p}$ deviation

reduced basis results confirm with an accuracy of 10^{-4}. However, the computational complexity does not only scale with the solved linear systems. The RB method additionally requires the evaluation of the residual over the parametric domain, which increases computational complexity by a factor between two and three.

6 Conclusion

The RB approach to Uncertainty Quantification shows the potential to significantly reduce the computational costs. While a weighted sampling gave accurate results in the computed example, this result can be certified by using error estimators in the statistical quantities, see [4] for more details. However, as the Monte-Carlo simulation with a Reduced Basis is quickly performed, an heuristic stopping criterion can be used, in the sense that when the oscillations in the expectation become small, the RB enrichment can stop.

Table 1 Comparison of methods using the stochastic collocation with Hermite Genz-Keister (order 4) rule as reference

Method	Linear solves	Mean expectation	Mean rel. error	Mean std. deviation
HGK (4th)	2020	3.373268	Reference	0.1015
HGK (3rd)	900	3.373268	$1.8 \cdot 10^{-6}$	0.1015
HGK (2nd)	420	3.373266	$3.3 \cdot 10^{-5}$	0.1004
HGK (4th) RB	73	3.373436	$5.4 \cdot 10^{-4}$	0.1009
MC RB (20k)	73	3.373168	$8.1 \cdot 10^{-4}$	0.0963
MC	14,000	3.372528	$1.1 \cdot 10^{-3}$	0.1160
MC RB (14k)	73	3.383087	$2.1 \cdot 10^{-2}$	0.0959

The 'mean expectation' shows the arithmetic mean of the computed expected transfer function as an indicator to compare the results. The 'mean rel. error' shows the mean of the relative error in the transfer function with respect to the chosen reference solution. The 'mean std. deviation' shows the arithmetic mean of the computed sample standard deviation over the frequency range. Each parameter configuration requires 20 solves to resolve the transfer function. The table is ordered with respect to accuracy of the methods

Acknowledgements This work is supported by the collaborative project nanoCOPS, Nanoelectronic COupled Problems Solutions, supported by the European Union in the FP7-ICT-2013-11 Program under Grant Agreement Number 619166.

References

1. Benner, P., Schneider, J.: Uncertainty quantification for Maxwell's equations using stochastic collocation and model order reduction. Int. J. Uncertain. Quantif. **5**(3), 195–208 (2015)
2. Burkardt, J.: SGMGA - Sparse Grid Mixed Growth Anisotropic Rules. http://people.sc.fsu.edu/jburkardt/m_src/m_src.html (2009)
3. Chen, P., Quarteroni, A., Rozza, G.: A weighted reduced basis method for elliptical partial differential equations with random input data. SIAM J. Numer. Anal. **51**(6), 3163–3185 (2014)
4. Haasdonk, B., Urban, K., Wieland, B.: Reduced basis methods for parametrized partial differential equations with stochastic influences using the Karhunen-Loève expansion. SIAM J. Uncertain. Quantif. **1**, 79–105 (2013)
5. Hess, M.W., Benner, P.: Fast evaluation of time harmonic Maxwell's equations using the reduced basis method. IEEE Trans. Microw. Theory Tech. **61**, 2265–2274 (2013)
6. Hess, M.W., Benner, P.: A reduced basis method for microwave semiconductor devices with geometric variations. COMPEL Int. J. Comput. Math. Electr. Electron. Eng. **33**(4), 1071–1081 (2014)
7. Logg, A., Mardal, K.-A., Wells, G.N.: Automated Solution of Differential Equations by the Finite Element Method. Springer, Heidelberg (2012)
8. Papageorgiou, A., Traub, J.F.: Faster evaluation of multidimensional integrals. Comput. Phys. **11**(6), 574–578 (1997)
9. Rozza, G., Huynh, D.B.P., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Arch. Comput. Methods Eng. **15**, 229–275 (2008)
10. Zaglmayr, S.: High order finite element methods for electromagnetic field computation. Ph.D. thesis, Johannes Kepler University of Linz (2006)

Model Order Reduction for Stochastic Expansions of Electric Circuits

Roland Pulch

Abstract We consider dynamical systems modelling linear electric circuits. Physical parameters are replaced by random variables for an uncertainty quantification. The random process satisfying the dynamical system exhibits an expansion into a series with orthonormal basis polynomials. We apply quadrature formulas to determine an approximation of the unknown coefficient functions. The separate systems for the different nodes of a quadrature rule are reinterpreted as a single large system to enable a potential for model order reduction. For comparison, the stochastic Galerkin method is also investigated for the same problem. We focus on balanced truncation techniques for a reduction of the state space in the large systems. Numerical results are presented using a band pass filter.

1 Introduction

The mathematical modelling of electric circuits yields dynamical systems of ordinary differential equations (ODEs) or differential algebraic equations (DAEs), see [5]. In case of huge dimensions of the state space, model order reduction (MOR) is used to decrease the complexity of the problems, see [1, 3]. Moreover, random variables can be included to describe uncertainties of physical parameters, i.e., the resulting stochastic problem becomes more extensive than the original deterministic formulation. We investigate an approach for an MOR of the stochastic problem, which can be used for both small and large dynamical systems.

We consider linear ODE models with random parameters. Hence the solution of the dynamical system represents a random process. We use an expansion of this random process into an orthonormal basis following the technique of the polynomial chaos, see [13].

On the one hand, the stochastic Galerkin (SG) method yields a larger coupled system, which has to be solved just once to obtain an approximation of the expansion of the random process, see [8, 9]. These large systems have been

R. Pulch (✉)
Institute for Mathematics and Computer Science, Ernst-Moritz-Arndt-Universität Greifswald, Walther-Rathenau-Str. 47, D-17489 Greifswald, Germany
e-mail: pulchr@uni-greifswald.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_22

reduced successfully by moment matching techniques in [10]. On the other hand, the stochastic collocation (SC) approach, see [9, 14], is based on a sampling scheme or a quadrature formula (QF), cf. [4, 12]. Therein, a large number of separate dynamical systems have to be solved, whereas also an approximation of the unknown coefficient functions in the expansion is produced. We take the separate systems as a single large system with a specific input-output behaviour to achieve a high potential for MOR. A reduction of the dimension of the system is performed by balanced truncation, see [1, Chap. 7].

We examine a test example, where the electric circuit represents a band pass filter. Since error bounds are available for an MOR by balanced truncation, it is sufficient to investigate the decay of the Hankel singular values to confirm the potential for an efficient reduction.

2 Stochastic Modelling

We start from the linear time-invariant dynamical system

$$\begin{aligned} \dot{\mathbf{x}}(t,\mathbf{p}) &= A(\mathbf{p})\mathbf{x}(t,\mathbf{p}) + B(\mathbf{p})\mathbf{u}(t) \\ \mathbf{y}(t,\mathbf{p}) &= C\mathbf{x}(t,\mathbf{p}) \end{aligned} \tag{1}$$

with the state variables $\mathbf{x} \in \mathbb{R}^n$, the output variables $\mathbf{y} \in \mathbb{R}^q$, the input signals $\mathbf{u} \in \mathbb{R}^k$ and matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{q\times n}$. The matrices A, B include the physical parameters $\mathbf{p} \in \Pi \subseteq \mathbb{R}^p$. Thus both the states and the outputs depend on time as well as the parameters. The dimension n of the ODEs (1) can be small or large in the following. Generalisations to descriptor formulations, which typically represent DAEs, cf. [5], are straightforward.

We assume that the parameters exhibit uncertainties. For example, imperfections of an industrial production cause variations of the parameters due to miniaturisation. To quantify the uncertainties, the parameters are replaced by independent random variables. Let $\rho : \Pi \to \mathbb{R}$ be their joint density function. We obtain an associated Hilbert space $L^2(\Pi, \rho)$ for functions depending on the random parameters. Now the aim is to determine statistics of the outputs like the expected value and the variance, for example, or more sophisticated quantities.

To compute probabilistic integrals approximately, a sampling technique or a QF yields nodes $\{\mathbf{p}_1, \ldots, \mathbf{p}_s\} \subset \Pi$ and weights $w_1, \ldots, w_s \in \mathbb{R}$. Consequently, the dynamical systems

$$\dot{\mathbf{x}}(t,\mathbf{p}_j) = A(\mathbf{p}_j)\mathbf{x}(t,\mathbf{p}_j) + B(\mathbf{p}_j)\mathbf{u}(t) \tag{2}$$

have to be resolved for $j = 1, \ldots, s$.

Let an orthonormal basis $\{\Phi_1(\mathbf{p}), \ldots, \Phi_m(\mathbf{p})\}$ be given for some subspace in $L^2(\Pi, \rho)$. Typically, orthogonal polynomials are chosen following the concept of

the polynomial chaos, see [13]. An expansion of the outputs with respect to the basis reads as

$$\tilde{\mathbf{y}}(t,\mathbf{p}) := \sum_{i=1}^{m} \mathbf{v}_i(t)\Phi_i(\mathbf{p}) \quad \text{with} \quad \mathbf{v}_i(t) := \int_{\Pi} \mathbf{y}(t,\mathbf{p})\Phi_i(\mathbf{p})\rho(\mathbf{p})\,\mathrm{d}\mathbf{p} \tag{3}$$

provided that the exact output $\mathbf{y}$ is in $L^2(\Pi,\rho)$. The above probabilistic integration is done component-wise. Our aim is to compute the coefficients $\mathbf{v}_i$ in (3), since they generate an approximation $\tilde{\mathbf{y}}$ of the exact outputs $\mathbf{y}$. Moreover, the expected value and the variance can be reproduced.

We approximate the coefficient functions of the stochastic expansion (3) by the QF, i.e.,

$$\mathbf{v}_i(t) \doteq \sum_{j=1}^{s} w_j \Phi_i(\mathbf{p}_j) C\mathbf{x}(t,\mathbf{p}_j) \tag{4}$$

for $i = 1,\ldots,m$. Now a single system is constructed, which describes this approach completely. Let $\hat{\mathbf{x}}(t) := (\mathbf{x}(t,\mathbf{p}_1),\ldots,\mathbf{x}(t,\mathbf{p}_s)) \in \mathbb{R}^{ns}$. Furthermore, we consider the outputs $\hat{\mathbf{v}}(t) := (\mathbf{v}_1(t),\ldots,\mathbf{v}_m(t)) \in \mathbb{R}^{qm}$. Using (2) for $j = 1,\ldots,s$ as well as (4) for $i = 1,\ldots,m$, we define the larger system

$$\begin{aligned} \dot{\hat{\mathbf{x}}}(t) &= \hat{A}\hat{\mathbf{x}}(t) + \hat{B}\mathbf{u}(t) \\ \hat{\mathbf{v}}(t) &= \hat{C}\hat{\mathbf{x}}(t). \end{aligned} \tag{5}$$

The system (5) consists of s separate subsystems (2), which are coupled only by the definition of the outputs (4). Hence the matrix $\hat{A} \in \mathbb{R}^{ns\times ns}$ has a block diagonal structure. It holds that $\hat{B} \in \mathbb{R}^{ns\times k}$. The formulas (4) yield the matrix $\hat{C} \in \mathbb{R}^{qm\times ns}$.

Alternatively, the SG approach generates a large system of the form (5) with dimension nm, which is fully coupled, cf. [8, 9]. Therein, the unknowns $\hat{\mathbf{x}}$ represent an approximation of the coefficient functions in the expansion of the random process $\mathbf{x}(t,\mathbf{p})$ with respect to the basis functions, whereas the outputs $\hat{\mathbf{v}}(t)$ yield an approximation of the coefficient functions in (3) again. The SG system sometimes looses the stability, although all systems (1) are stable, see [11]. This loss of stability does not affect the convergence of the SG method on compact time intervals. Yet the asymptotic behaviour becomes incorrect in time. In contrast, the system (5) inherits obviously the stability of the systems (1).

3 Model Order Reduction

In both the QF approach and the SG method, the dynamical system (1) changes into a stochastic model with a much larger dimension of the state space. It follows that both types of systems exhibit a high potential for MOR. The inputs for the stochastic

models (5) coincide with the inputs of the original system (1). Hence the number of inputs is much smaller than the dimension of the state space in (5), namely $k \ll ns$. However, the number of outputs increases by the factor m of the number of basis functions. Nevertheless, the ratio of outputs to state variables is $\frac{qm}{nm} = \frac{q}{n}$ for the SG scheme and $\frac{qm}{ns}$ for the QF, which is the same as in the original system (1) and less or equal provided that the number s of nodes is sufficiently large, respectively.

For linear time-invariant dynamical systems, the state space can be reduced by moment matching techniques or balanced truncation, for example, see [1, 3, 6]. Once a reduced order model (ROM) of our system (5) is constructed, a transient simulation directly yields an approximation of the coefficient functions and thus we obtain the desired approximation (3) of the random process $\mathbf{y}(t, \mathbf{p})$. This strategy can be seen as an alternative to parametric MOR. In a parametric MOR, the system (1) is reduced in a first step, while preserving the dependence on all $\mathbf{p} \in \Pi$. In a second step, the probabilistic integrations are applied to the ROM. Yet this parametric MOR can be applied only if the original system (1) is already large.

We focus on an MOR by balanced truncation, where the algorithm coincides for both the QF and the SG method. Given a system (5), the controllability Gramian W_C and the observability Gramian W_O are defined by the Lyapunov equations

$$\hat{A} W_\mathrm{C} + W_\mathrm{C} \hat{A}^\top = -\hat{B}\hat{B}^\top \quad \text{and} \quad \hat{A}^\top W_\mathrm{O} + W_\mathrm{O} \hat{A} = -\hat{C}^\top \hat{C}. \tag{6}$$

We require symmetric decompositions

$$W_\mathrm{C} = Z_\mathrm{C} Z_\mathrm{C}^\top \quad \text{and} \quad W_\mathrm{O} = Z_\mathrm{O} Z_\mathrm{O}^\top \tag{7}$$

of the solutions from (6). The singular value decomposition $USV^\top = Z_\mathrm{C}^\top Z_\mathrm{O}$ yields the Hankel singular values $S = \mathrm{diag}(\sigma_1, \sigma_2, \ldots)$ with $\sigma_\ell \geq \sigma_{\ell+1}$. A truncation is done via $S_\mathrm{red} := \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, where just the r largest singular values are kept. We obtain the transformation matrices

$$P := S_\mathrm{red}^{-\frac{1}{2}} V_\mathrm{red}^\top Z_\mathrm{O}^\top \quad \text{and} \quad Q := Z_\mathrm{C} U_\mathrm{red} S_\mathrm{red}^{-\frac{1}{2}}$$

with $U_\mathrm{red}, V_\mathrm{red}$ containing just the first r columns of U, V. The matrices $\hat{A}_\mathrm{red} := P\hat{A}Q$, $\hat{B}_\mathrm{red} := P\hat{B}$, $\hat{C}_\mathrm{red} := \hat{C}Q$ define a lower-dimensional dynamical system of size r.

On the one hand, the symmetric decompositions (7) can be calculated directly by the Cholesky factorisation, for example. However, the Gramians are dense matrices independent of the structure of the matrix $\hat{A}$ in our two types of methods. Thus the Cholesky factors are also dense and a high computational effort occurs for large systems. On the other hand, iterative methods yield low-rank approximations for the factors $Z_\mathrm{C}, Z_\mathrm{O}$ in (7), which is often more efficient. A popular algorithm is the alternating directions implicit (ADI) iteration, see [1, Chap. 12.4]. In this scheme, matrix-vector-multiplications as well as linear system solves have to be done for matrices $\hat{A} - \mu I$ with the identity I and shift parameters μ. The QF approach implies

a much lower computational effort within these linear algebra operations than the SG method, since the matrices are block diagonal for the QF. Yet both QF and SG suffer from the large number of outputs within a straightforward ADI method.

4 Numerical Results for a Test Example

We apply the band pass filter depicted in Fig. 1 and its mathematical model from [7]. The dynamical system (1) exhibits the dimension $n = 6$, single input and single output ($q = k = 1$). All capacitances, inductances and resistances are chosen as random variables, where independent uniform distributions with 20 % variation around the mean values are used. Thus $p = 11$ random parameters appear. Figure 2 shows the Bode plot of the transfer function for the selected mean values.

For the output of the stochastic model, we arrange an orthonormal basis using all multivariate Legendre polynomials up to degree three. Hence $m = 364$ basis functions are involved. We apply the Stroud quadrature scheme of order 5 with $s = 243$ nodes, see [12]. Now the system (5) exhibits the dimension $ns = 1458$ of the state space and the number of outputs becomes $qm = 364$. The SG method

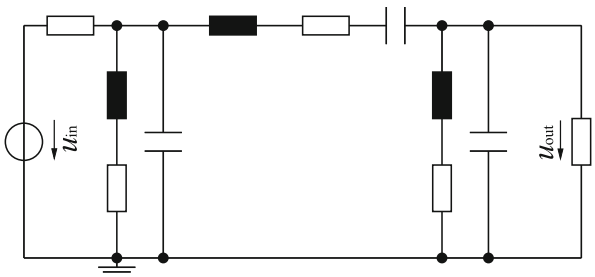


Fig. 1 Circuit of a band pass filter with *L-C-Π* element

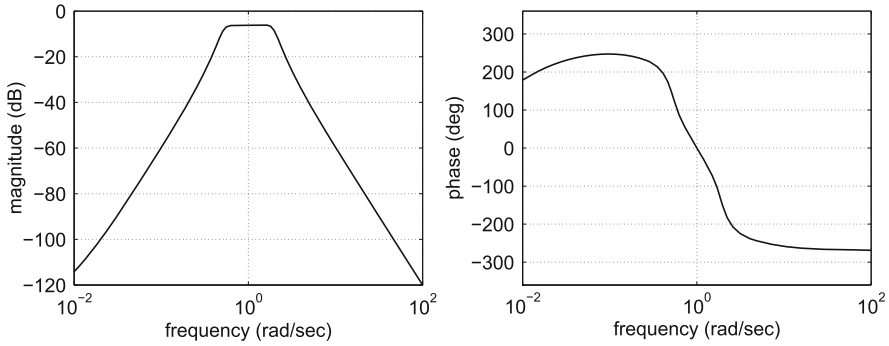


Fig. 2 Bode diagramme for the transfer function of the band pass filter

results in a system (5) of dimension $nm = 2184$ with the same number of outputs. A sparse grid of level 3 with 3103 nodes from the Smolyak construction based on the one-dimensional Gauss-Legendre quadrature, see [4], yields the matrices in the system (5) of the SG method. We use this highly accurate quadrature, since the number of nodes does not affect the dimension of the coupled system. Thus the quadrature error becomes negligible in comparison to the other error sources in the SG method. Figure 3 depicts the sparsity pattern of the matrices $\hat{A}$ in both techniques. To illustrate the dynamics, Fig. 4 shows the eigenvalues of the matrices $\hat{A}$. In particular, the two systems are stable. The CPU-times for the computation of the matrices in the systems (5) is 1 s for QF and 73 s for SG.

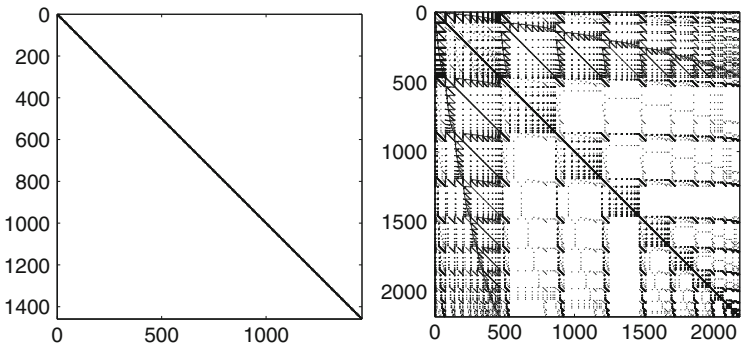


Fig. 3 Matrix $\hat{A}$ in the system (5) for the QF (*left*) and the SG method (*right*). The percentage of non-zero elements is 0.2 % and 1.1 %, respectively

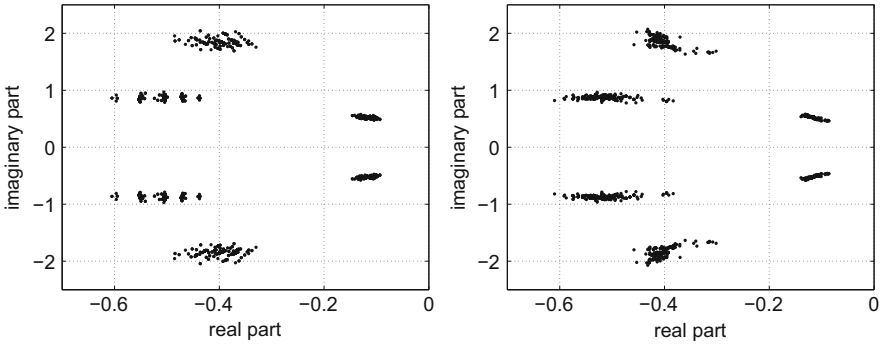


Fig. 4 Eigenvalues of the matrix $\hat{A}$ in the system (5) for the QF (*left*) and the SG method (*right*)

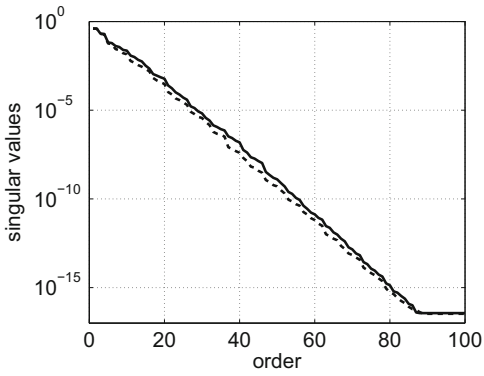


Fig. 5 Dominant Hankel singular values of the system (5) in QF (*solid line*) and in SG (*dashes line*)

Based on the controllability Gramian and the observability Gramian, we determine the Hankel singular values of the system (5) using a direct algorithm with Cholesky factorisations of the Gramians. Figure 5 illustrates the 100 largest singular values. We recognise that less than 90 singular values are above the machine precision in both approaches. Since the Hankel singular values decrease rapidly, the dimension of the systems (5) can be reduced efficiently by balanced truncation. To include 99.9 % of the amount of all singular values in a ROM, it is sufficient to reduce to the dimension $r = 19$ for the QF system and to $r = 18$ for the SG system (i.e. $(\sigma_1 + \cdots + \sigma_r)/(\sigma_1 + \sigma_2 + \cdots) \geq 0.999$). The CPU-times for the balanced truncation technique were 36 s for QF and 160 s for SG due to the different size of the systems.

To get an impression of the input-output behaviour, the absolute values of the transfer function of the ROM from the QF approach are shown in Fig. 6. The function for degree zero represents an approximation of the expected value, whereas all other functions yield an approximation of the variance.

We compare briefly the accuracy of the MORs. The transfer function $H(\mathrm{i}\omega, \mathbf{p})$ of the dynamical system (1) represents a scalar complex-valued random process due to $k = q = 1$. The expected value as well as the standard deviation of this transfer function are computed approximately by the ROMs. A reference solution is determined using a sparse grid of level 4 with 25,653 nodes applied to the original system (1). Table 1 demonstrates the maximum differences in the frequency interval $\omega \in [0.1, 10]$ on the imaginary axis, where the real part and the imaginary part of the transfer function are analysed separately. We observe a higher accuracy in the SG method.

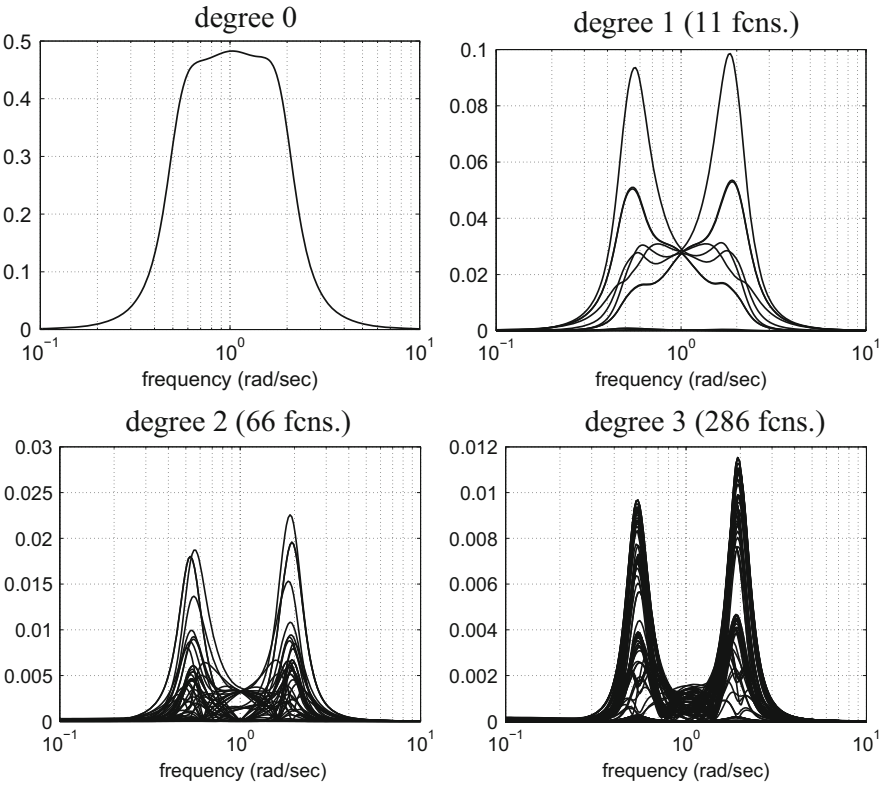


Fig. 6 Absolute values of the components of the transfer function for different polynomial degrees reconstructed from the ROM by the QF

Table 1 Maximum differences between ROM and a reference solution for approximations of expected value and of standard deviation belonging to transfer function at frequencies $\omega \in [0.1, 10]$

		Galerkin	Quadrature
Expected value	Real part	5.03e-05	4.12e-04
	Imaginary part	4.85e-05	4.20e-04
Standard deviation	Real part	3.14e-04	2.33e-02
	Imaginary part	2.77e-04	1.33e-02

5 Conclusions

A stochastic model of linear dynamical systems can be solved by an expansion with orthonormal basis functions. The stochastic Galerkin method yields a larger fully coupled system, where MOR is applicable. We reinterpreted a quadrature rule as a large weakly coupled system such that MOR is also feasible. We examined both approaches in the reduction of a test example. Using balanced truncation via

a direct method, the efficiency of both techniques agrees roughly when comparing the accuracy and the computational work. An open question is if the quadrature approach may cause significant savings in the computational effort of an iterative method, because the required system matrix exhibits just a block diagonal form. Furthermore, a strategy for systems with many outputs but few inputs proposed in [2] should be investigated for both the Galerkin method and the quadrature schemes, since these assumptions are satisfied in our stochastic problem.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia (2005)
2. Benner, P., Schneider, A.: Balanced truncation model order reduction for LTI systems with many inputs or outputs. In: Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems, MTNS 2010, pp. 1971–1974, Budapest, 5–9 July 2010
3. Benner, P., Hinze, M., ter Maten, E.J.W. (eds.): Model Reduction for Circuit Simulation. Springer, Dordrecht (2011)
4. Gerstner, T., Griebel, M.: Numerical integration using sparse grids. Numer. Algorithms **18**, 209–232 (1998)
5. Günther, M., Feldmann, U.: CAD based electric circuit modeling in industry I: mathematical structure and index of network equations. Surv. Math. Ind. **8**, 97–129 (1999)
6. Ionutiu, R., Lefteriu, S., Antoulas, A.C.: Comparison of model reduction methods with applications to circuit simulation. In: Ciuprina, G., Ioan, D. (eds.) Scientific Computing in Electrical Engineering SCEE 2006. Mathematics in Industry, vol. 11, pp. 3–24. Springer, Berlin/Heidelberg (2007)
7. Kessler, R.: Aufstellen und numerisches Lösen von Differential-Gleichungen zur Berechnung des Zeitverhaltens elektrischer Schaltungen bei beliebigen Eingangs-Signalen. Online document. http://www.home.hs-karlsruhe.de/~kero0001/aufst6/AufstDGL6hs.html (2014). Cited 9 Sep 2014
8. Pulch, R.: Polynomial chaos for linear differential algebraic equations with random parameters. Int. J. Uncertain. Quantif. **1**, 223–240 (2011)
9. Pulch, R.: Stochastic collocation and stochastic Galerkin methods for linear differential algebraic equations. J. Comput. Appl. Math. **262**, 281–291 (2014)
10. Pulch, R., ter Maten, E.J.W.: Stochastic Galerkin methods and model order reduction for linear dynamical systems. Int. J. Uncertain. Quantif. **5**, 255– 273 (2015)
11. Sonday, B.E., Berry, R.D., Najm, H.N., Debusschere, B.J.: Eigenvalues of the Jacobian of a Galerkin-projected uncertain ODE system. SIAM J. Sci. Comput. **33**, 1212–1233 (2011)
12. Stroud, A.: Approximate Calculation of Multiple Integrals. Prentice Hall, Englewood Cliffs (1971)
13. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)
14. Xiu, D., Hesthaven, J.S.: High order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**, 1118–1139 (2005)

Robust Topology Optimization of a Permanent Magnet Synchronous Machine Using Multi-Level Set and Stochastic Collocation Methods

Piotr Putek, Kai Gausling, Andreas Bartel, Konstanty M. Gawrylczyk, E. Jan W. ter Maten, Roland Pulch, and Michael Günther

Abstract The aim of this paper is to incorporate the stochastic collocation method (SCM) into a topology optimization for a permanent magnet (PM) synchronous machine with material uncertainties. The variations of the non-/linear material characteristics are modeled by the Polynomial Chaos Expansion (PCE) method. During the iterative optimization process, the shapes of rotor poles, represented by zero-level sets, are simultaneously optimized by redistributing the iron and magnet material over the design domain. The gradient directions of the multi-objective function with constraints, composed of the mean and the standard deviation, is evaluated by utilizing the continuous design sensitivity analysis (CDSA) with the SCM. Incorporating the SCM into the level set method yields designs by using already existing deterministic solvers. Finally, a two-dimensional numerical result demonstrates that the proposed method is robust and effective.

P. Putek (✉) • K. Gausling
Chair of Applied Mathematics & Numerical Analysis, University of Wuppertal, Wuppertal, Germany
e-mail: putek@math.uni-wuppertal.de; gausling@math.uni-wuppertal.de

A. Bartel • E. Jan W. ter Maten • M. Günther
Applied Math. & Numerical Analysis, Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: bartel@math.uni-wuppertal.de; termaten@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

K.M. Gawrylczyk
Department of Electrotechnology and Diagnostic, West Pomeranian University of Technology, Szczecin, Poland
e-mail: Konstanty.Gawrylczyk@zut.edu.pl

P. Putek • R. Pulch
Ernst-Moritz-Arndt-Universität Greifswald, Greifswald, Germany
e-mail: pulchr@uni-greifswald.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_23

1 Introduction

Nowadays, permanent-magnet (PM) machines have become more popular due to their attractive features such as a high performance, efficiency, and power density [1]. Therefore, they have found a broad use in industrial applications such as robotics, computer peripherals, industrial drivers or automotive industry, for example, in commercialized hybrid vehicles with different hybridization level, e.g. [5, 8]. However, this type of motor construction suffers inherently from a relatively high level of acoustic noise and mechanical vibration. In the case of a PM machine, the interaction between the stator slot driven air-gap performance harmonics and the magnet driven magnetomotive force (MMF) harmonics is mainly responsible for producing a high cogging torque (CT). On the other hand, the torque ripple developed in electromagnetic torque is caused by the cogging torque and harmonic contents in the back-electromotive force (EMF). In addition, magnetic saturations in the stator and rotor cores with the converted related issue may further disturb the electromagnetic torque of the machine [2]. Therefore, the designers aim above all at blackucing the torque fluctuations. In turn, this may significantly affect the machine performance.

In this paper, we focus on optimizing topology a PM machine, as the machine topology itself is a major contributor to the electromagnetic torque fluctuation. Because the result of the design procedure is strongly affected by the unknown material characteristics [12], the uncertainties in modeling the soft ferromagnetic material are taken into account. In some applications [11], especially the relative permeability of the magnetic material itself should be accounted to model more accurately the magnetic flux density of permanent magnets. This parameter is also in our model assumed as uncertain. The novel aspect of the proposed method is the incorporation of stochastic modeling into the topology optimization method for the low cogging torque (CT) design of an Electric Controlled Permanent Magnet Excited Synchronous Machine (ECPSM).

2 Model Description

In the design of a PM machine, the shape/fabrication and the placement of magnets, iron poles and air-gaps primarily determine the torque characteristic. A part of an assembly drawing of such a device, considered as a case study in our paper, is given in Fig. 1. The structure of the rotor comprises two almost identical parts, which differ only in the magnetization direction of the constructed PM poles of the rotor. The key feature of the machine is the installation of an additional DC control coil that is fixed in the axial center of the machine, between two laminated stators. The proper supply of this coil by the DC-chopper enables to control the effective excitation of the machine. In the end, this results in a field weakening of 1:4, which is of great importance in electric vehicles applications (Fig. 1).

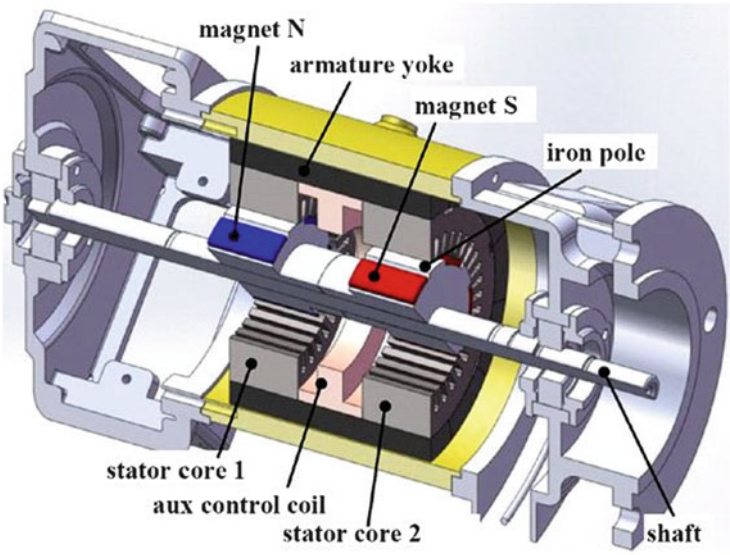


Fig. 1 Cross-section of an ECPSM and its main parameters (surface-mounted PM rotor, three-phase windings, fixed excitation control auxiliary coil) [8]

The magnetic behavior can be described in terms of the unknown magnetic vector potential (MVP) $\mathbf{A} = (0,\ 0,\ A)$ for the quasi-linear curl-curl equation. In fact, in order to reduce the computational burden, we consider a two dimensional (2-D) model that is additionally simplified by neglecting the eddy current phenomena, i.e., $(\sigma \frac{\partial A}{\partial t} = 0)$. Then, the curl-curl equation becomes a Poisson equation

$$\nabla \cdot \left(\upsilon(\mathbf{x},\ |\nabla A(\mathbf{x})|^2)\nabla A(\mathbf{x}) - \upsilon_{\mathrm{PM}}\mathbf{M}(\mathbf{x})\right) = J(\mathbf{x}), \quad \mathbf{x} \in \mathrm{D} \subset \mathbb{R}^2, \tag{1}$$

equipped with periodic boundary conditions Γ_{PBC} on ∂D in order to further decrease the computational burden. This establishes the computational domain D, cf. Fig. 3. Here, the current density is denoted by $J \in L^2(\mathrm{D})$ and the remanent flux density of the PM is denoted by $\mathbf{M}$. Furthermore, the reluctivity υ is as a real parameter, which describes the isotropic material relation $\mathbf{H} = \upsilon(|\mathbf{B}|^2)\mathbf{B}$ between the flux density $\mathbf{B} = \nabla \times \mathbf{A}$ and the field strength $\mathbf{H}$. The parameter υ depends on $|\mathbf{B}| = |\nabla \mathbf{A}|$. In the air-gap, the vacuum reluctivity $\upsilon(|\mathbf{B}|^2) = \upsilon_0$ is taken into account. The quality of the design of a PM motor, on the one hand, is assessed by the cogging torque fluctuation T. This quantity is calculated by using the Maxwell stress tensor method [1] as the function of the rotor position θ

$$T(\theta) = \upsilon_0 \oint_S \mathbf{r} \times \left((\mathbf{n} \cdot \mathbf{B}(\mathbf{x}))\mathbf{B}(\mathbf{x}) - \frac{|\mathbf{B}(\mathbf{x})|^2}{2}\mathbf{n}\right) \mathrm{d}S, \tag{2}$$

where $\mathbf{n}$ is the unit outward normal vector and S denotes any closed integration surface in the air-gap surrounding the rotor and $\mathbf{r}$ denotes the position vector. Its main contributor is the machine topology. Additionally in the bi-objective optimization problem, the root mean square (*rms*) value of the magnetic field density calculated in the air-gap along the path l is treated as the second criterion but only in an approximate way as in [9]

$$B^2_{\mathrm{r-rms}} = \frac{1}{L}\int_{\theta_1}^{\theta_2} |B_r|^2 \mathrm{d}l \geq \alpha, \tag{3}$$

where the coefficient α denotes an assumed level of the magnetic flux density in the air-gap (the fraction of $B_{\mathrm{r-rms}}$ calculated for the initial configuration), L refers to the length of the path l (from θ_1 to θ_2).

A further difficulty regards the reluctivity υ: it is discontinuous across material borders and it is nonlinear in ferromagnetic materials. Moreover, the ferromagnetic material characteristics (deduced from measurements) suffers from uncertainties [12]. In certain applications, especially the relative permeability of the magnetic material should be modeled to obtain a more accurate magnetic flux density of permanent magnets [11]. Since the uncertainties affect the results of the design procedure, we have to include these uncertainties to enable a robust design. That is, the reluctivity becomes a random field. To this end, we consider the following parameters as uncertain $\boldsymbol{\upsilon} := \left(\upsilon_{\mathrm{PM}}, \upsilon_{\mathrm{Fe}}, \upsilon_{\mathrm{air-gap}}\right) := \mathbf{p}$ in the stochastic reluctivity model. The uncertainty of $\upsilon_{\mathrm{air-gap}}$ is significant from the mathematical viewpoint; it could account for inaccuracies of the gap or material inside the gap.

3 Stochastic Forward Problem

For uncertainty quantification, we modify the parameters $\boldsymbol{\upsilon} : \Omega \to \Pi \subset \mathbb{R}^3$ using independent random variables $\boldsymbol{\upsilon}(\boldsymbol{\xi})$, defined on some probabilistic space $(\Omega, \mathscr{F}, \mathbb{P})$ with a joint density $\rho : \Pi \to \mathbb{R}$. In our case, it will be a uniform distribution (ranging $\pm 10\,\%$ around the respective nominal values).[1] Consequently, the direct problem is governed by the random-dependent PDEs system, considered here for

[1] For the UQ, the stochastic reluctivity model for the iron pole with the higher perturbation than analyzed in the paper [10] was applied. Due to the used Stroud quadrature formulas, the same distribution had to be assumed with a relatively high variance based on [11] for the reluctivity of a PM. The last parameter was rather of "the mathematical relevance" and simulates the high impact of the air-gap parameters into the electromagnetic torque [12].

no load state (with the excitation density current $J = 0$)

$$\begin{cases} \nabla \cdot \left(\upsilon_{\mathrm{Fe}}\left(\mathbf{x}, |\nabla A(\mathbf{x})|^2, \xi_1\right) \nabla A(\mathbf{x})\right) = 0, & \text{in } \mathrm{D}_{\mathrm{Fe}}, \\ \nabla \cdot \left(\upsilon_{\mathrm{air-gap}}\left(\mathbf{x}, \xi_2\right) \nabla A(\mathbf{x})\right) = 0, & \text{in } \mathrm{D}_{\mathrm{air}}, \\ \nabla \cdot \left(\upsilon_{\mathrm{PM}}(\xi_3) \nabla A(\mathbf{x})\right) = \nabla \cdot \upsilon_{\mathrm{PM}}(\xi_3) \mathbf{M}(\mathbf{x}), & \text{in } \mathrm{D}_{\mathrm{PM}}, \end{cases} \tag{4}$$

where $A : \mathrm{D} \times \Omega \to \mathbb{R}$ with $\mathrm{D} = \mathrm{D}_{\mathrm{air}} \cup \mathrm{D}_{\mathrm{Fe}} \cup \mathrm{D}_{\mathrm{PM}}$, becomes a random field. The statistical information like the expected value for a function $f : \Pi \to \mathbb{R}$ reads as

$$\langle f(\boldsymbol{\upsilon}) \rangle := \mathbb{E}\left[f(\boldsymbol{\upsilon})\right] = \int_{\Pi} f(\boldsymbol{\upsilon}(\boldsymbol{\xi})) \rho(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}, \tag{5}$$

provided that the integral is finite. Furthermore, for two functions $f, g : \Pi \to \mathbb{R}$ this operator yields an inner product $\langle f, g \rangle := \mathbb{E}\left(f(\boldsymbol{\upsilon}) g(\boldsymbol{\upsilon})\right)$ on $L^2(\Omega)$, see e.g. [7, 17]. If each component υ_i exhibits a finite second moment, then the random field A can be expanded in the truncated polynomial chaos (PC) series [17]

$$A\left(\mathbf{x}, \boldsymbol{\upsilon}\right) = \sum_{i=0}^{N} v_i\left(\mathbf{x}\right) \Phi_i\left(\boldsymbol{\upsilon}\right) \tag{6}$$

with a priori unknown coefficient functions v_i. Here, the basis functions $(\Phi_i)_{i \in \mathbb{N}}$ with $\Phi_i : \Pi \to \mathbb{R}$ are orthonormal polynomials, i.e., $\langle \Phi_i(\boldsymbol{\upsilon}), \Phi_j(\boldsymbol{\upsilon}) \rangle = \delta_{ij}$ with the Kronecker delta δ_{ij}. To calculate v_i the SCM with Stroud quadrature formula [13, 18] is used. The basic concept is to provide the solution of the deterministic problem at each quadrature grid point $\boldsymbol{\upsilon}^{(k)}$, $k = 0, \ldots, K$. The Stroud rules yield a relatively small number of grid points for a quadrature of a fixed order. Thus, finally we approximate statistical quantities like the mean and the standard deviation

$$\mathbb{E}\left[\mathrm{A}\left(\mathbf{x}, \boldsymbol{\upsilon}\right)\right] \doteq \mathrm{v}_0(\mathbf{x}), \quad \mathrm{std}\left[A\left(\mathbf{x}, \boldsymbol{\upsilon}\right)\right] \doteq \sqrt{\sum_{i=1}^{N} |\mathrm{v}_i(\mathbf{x})|^2} \tag{7}$$

by using a multi-dimensional quadrature rule with corresponding weights w_k

$$\mathrm{v}_i(\mathbf{x}) := \left\langle \mathrm{A}\left(\mathbf{x}, \boldsymbol{\upsilon}\right), \Phi_i(\boldsymbol{\upsilon}) \right\rangle \approx \sum_{k=0}^{K} w_k \mathrm{A}\left(\mathbf{x}, \boldsymbol{\upsilon}^{(k)}\right) \Phi_i(\boldsymbol{\upsilon}^{(k)}). \tag{8}$$

4 Multi-Level Set Representation

The level set method, first proposed in [6], has recently found a wide application in electrical engineering to address the design, shape and topology optimization problems, see e.g. [4, 8]. To trace the two interfaces between different materials

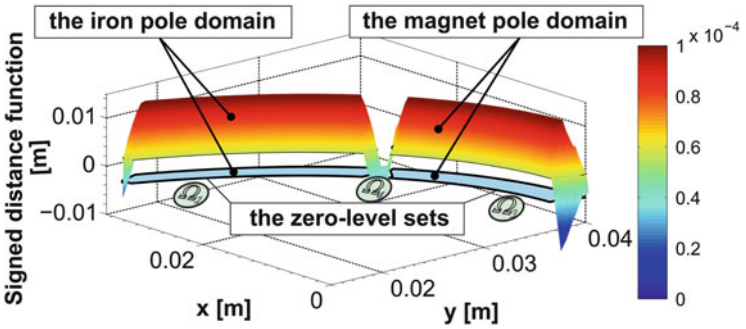


Fig. 2 Distribution of the signed distance function

with some assumed variations such as air, iron and PM poles of rotor, the modified multilevel set method (MLSM) has been used [9, 16]. Thus, we extend this framework into the situation, where the material parameter exhibit some uncertainty. These domains are described using two signed distance functions

$$\begin{aligned} \mathrm{D}_1 &= \{\mathbf{x} \in \mathrm{D} | \phi_1 > 0 \text{ and } \phi_2 > 0\}, \ \mathrm{D}_2 = \{\mathbf{x} \in \mathrm{D} | \phi_1 > 0 \text{ and } \phi_2 < 0\}, \\ \mathrm{D}_3 &= \{\mathbf{x} \in \mathrm{D} | \phi_1 < 0 \text{ and } \phi_2 > 0\}, \ \mathrm{D}_4 = \{\mathbf{x} \in \mathrm{D} | \phi_1 < 0 \text{ and } \phi_2 < 0\}, \end{aligned} \tag{9}$$

with $\phi(x)$ a signed distance function that is shown on Fig. 2.[2] In this situation, the reluctivity υ and the remanent flux density coefficient b_r (of the PM-material) read as

$$\begin{aligned} \upsilon(\boldsymbol{\phi}, \boldsymbol{\xi}) = \upsilon_1(\xi_1) H(\phi_1) H(\phi_2) + \upsilon_2(\xi_2) H(\phi_1) (1 - H(\phi_2)) + \\ + \upsilon_3(\xi_3) (1 - H(\phi_1)) H(\phi_2) + \upsilon_4(\xi_4) (1 - H(\phi_1)) (1 - H(\phi_2)), \end{aligned} \tag{10}$$

$$\begin{aligned} b_r(\boldsymbol{\phi}) = b_{r1}(H(\phi_1) H(\phi_2) + b_{r2}(H(\phi_1) (1 - H(\phi_2)) + \\ + b_{r3}((1 - H(\phi_1)) H(\phi_2) + b_{r4} (1 - H(\phi_1)) (1 - H(\phi_2)) \end{aligned} \tag{11}$$

with the Heaviside function $H(\cdot)$. The evolution of ϕ_i is described by the Hamilton-Jacobi-type equation [6] (during optimization with pseudo-time t)

$$\frac{\partial \phi_i}{\partial t} = -\nabla \phi_i (\mathbf{x}, t) \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = V_{n,i} |\nabla \phi_i|, \tag{12}$$

where $V_{n,i}$ is the normal component of the zero-level set velocity corresponding to the objective function (14) and the forward problem (4). Figure 2 shows the exemplary the distance function in fifth iteration of the optimized process, where

[2] Notice, D_4 is an auxiliary set. We need in our application $\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3$, only.

shapes of rotors poles (the blue shape with black lines) are described by the zero-level set.

5 Robust Topology Optimization Problem

The cogging torque minimization in the 2-D magnetostatic setting can be equivalently formulated as minimization of the magnetic energy W_r variation [3, 9]. The advantage of the latter formulation is the calculation of the sensitivity in an efficient way as follows [3]:

$$\frac{\partial W_\mathrm{r}}{\partial \mathbf{p}} = \int_\gamma (\upsilon_1 - \upsilon_2)\mathbf{B}_1 \cdot \mathbf{B}_2 - (\mathbf{M}_1 - \mathbf{M}_2) \cdot \mathbf{B}_2 \mathrm{d}\gamma, \qquad \text{in D} \tag{13}$$

with υ_1 and υ_2 the reluctivities for different domains. Since the energy operator is self-adjoint, the dual and primary problem are the same. However, for the shape optimization problem constrained by the elliptic PDEs (4) with random material variations, the magnetic energy is defined as

$$W_\mathrm{r}(\upsilon(\phi_1, \phi_2, \boldsymbol{\xi})) = \int_\mathrm{D} \mathbf{B}(\phi_1, \phi_2)\mathbf{H}(\phi_1, \phi_2)\mathrm{d}\mathbf{x} + \sum_{i=0}^{I=2} \beta_i \mathrm{TV}(\phi_\mathrm{i}), \tag{14}$$

which is subjected to the constraint (3) with $\mathbf{B}_r$ replaced by $\mathbf{B}_r(\phi_1, \phi_2)$, while the $\mathrm{TV}(\cdot)$ denotes the Total Variation regularization with the coefficients β_i that account for controlling the geometrical complexity of obtained shapes [8, 16]. Finally, this constraint has been introduced approximately to the optimization problem as two area constraints (for each rotor pole separately), which are involved in the level set method scheme, see, e.g., [4, 9].[3] Furthermore, we formulate the optimal shape optimization in the framework of the robust optimization [19] using the statistical moments such as the expectation and the standard deviation

$$\begin{aligned} \min_{\boldsymbol{\phi}} &: \mathbb{E}\left[W_\mathrm{r}(\boldsymbol{\upsilon})\right] + \kappa_1 \sqrt{\mathrm{Var}\left[W_\mathrm{r}(\boldsymbol{\upsilon})\right]} \\ \text{s.t.} &: \mathbf{K}\left(\boldsymbol{\upsilon}^k\right)\mathbf{A}^k = \mathbf{f}^k,\ k = 0, \ldots, K, \\ &\quad \upsilon_{\max j} \le \upsilon_j \le \upsilon_{\min j},\ j = 1, 2, \end{aligned} \tag{15}$$

where κ_1 is a prescribed parameter, $\mathbf{K}$ denotes the stiffness matrix (at $K + 1$ quadrature grid points). In this case, it is possible to calculate the total derivative of the function Eq. (14) based on only the analysis of the forward model in the collocation points and taking Eqs. (10), (11) and (13) and then (8), (7) into account.

[3] In our paper, they were defined in an analogical way as in [9]: $G_1(\boldsymbol{\phi}) = |\mathrm{D_{FE}}|/|\mathrm{D_{FE_0}}| - S_\mathrm{FE} = 0$ and $G_2(\boldsymbol{\phi}) = |\mathrm{D_{PM}}|/|\mathrm{D_{PM_0}}| - S_\mathrm{PM} = 0$ with the prescribed coefficients S_FE and S_PM.

The similar approach, but for different type of the functional was used in [14, 15] for the solution of stochastic identification/control problems for constrained PDEs with random input data.

6 Numerical Results

The procedure described in the previous section has been applied to design the rotor poles of the ECPSM for no-load state. The main parameters of the machine are given in Table 1. The initial configuration of the ECPSM machine is depicted in Fig. 3 (left). The quantities that are taken subject to variations are the reluctivity of the iron pole and the PM pole. Also the reluctivity of the air-gap is assumed to be uncertain. To model the uncertainty, we choose a uniform distribution of the reluctivity with a maximum deviation from a nominal value $\upsilon(\mathbf{x}, |\nabla \mathbf{A}|^2)$ of 10 %. The application of Stroud-5 points for a system of the ECPSM machine with three parameters yields $K + 1 = 19$ sample points $\{\boldsymbol{\xi}_i\}_{i=0}^{18}$ in the three-dimensional parameter space. The

Table 1 Main parameters of the ECPSM topology

$2p$: number of poles	12
r_{ostat}: outer radius of the stator	67.5 mm
r_{istat}: inner radius of the stator	41.25 mm
l_{as}: axial length one part of the stator	35 mm
w_{oslot}: width of the slot opening	4.0 mm
ns: number of slots	36
m: number of phases	3
t_m: thickness of magnets ($NdFeB, B_r = 1, 2T$) 3.0 mm	

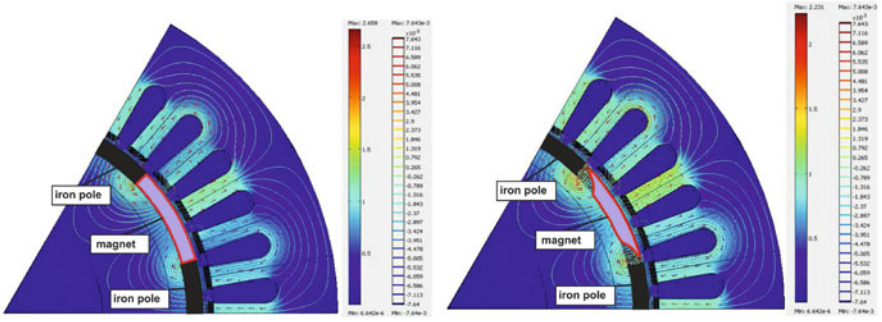


Fig. 3 Topology of the ECPSM: initial (*left*), optimal (*right*)

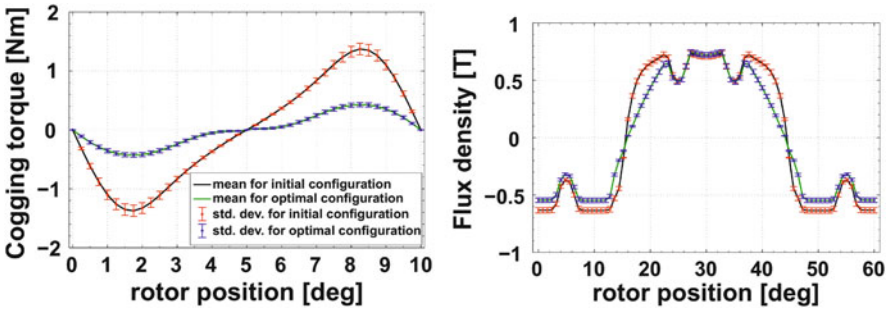


Fig. 4 Mean and standard deviation for initial and optimized topology of the ECPSM: cogging torque (*left*), flux density under magnet and iron poles calculated in the air-gap (*right*)

optimized rotor poles are shown in Fig. 3 (right).[4] For the optimal configuration the CT is calculated over a half of the period to assess the stator teeth interaction with the rotor poles, shown in Fig. 4 (left). The pick value of the CT expected value is reduced about 69 %.

7 Conclusion

This paper demonstrated the incorporation of the SCM into the MLSM for the robust topology optimization of a PM synchronous machine. For this purpose, the shape of rotor poles was investigated. This methodology resulted in the minimization of the level of noise and vibrations by the significant reduction of both the rectified mean of the CT (70 %) and the standard deviation, while taking variations with respect to manufacturing tolerances/imperfections into account. Unfortunately, the rectified value of the flux density calculated in the air-gap decreased around 17 %, which is a drawback of the used approach. This work also highlights the effectiveness of the proposed methodology.

Acknowledgements The project nanoCOPS (Nanoelectronic COupled Problems Solutions) is supported by the European Union in the FP7-ICT-2013-11 Program under the grant agreement number 619166 and the SIMUROM project is supported by the German Federal Ministry of Education and Research (05M13PXB).

[4]The construction of the PM machine under consideration was a topic of the scientific project called *"The Electrically Controlled Permanent Magnet Excited Synchronous Machine (ECPSM) with application to electro-mobiles"* under the Grant No. N510 508040, where among others a small prototype of the deterministically optimized machine with the similar topology as obtained in our paper was investigated.

References

1. Gieras, F., Wing, M.: Permanent Magnet Motor Technology, Wiley, New York (2008)
2. Islam, M.S., Islam, R., Sebastian, T.: Experimental verification of design techniques of permanent-magnet synchronous motors for low-torque-ripple applications. IEEE Trans. Ind. Appl. **47**(1), 88–95 (2011)
3. Kim, D., Sykulski, J., Lowther, D.: The implications of the use of composite materials in electromagnetic device topology and shape optimization. IEEE Trans. Magn. **45**, 1154–1156 (2009)
4. Lim, S., Min, S., Hong, J.P.: Low torque ripple rotor design of the interior permanent magnet motor using the multi-phase level-set and phase-field concept. IEEE Trans. Magn. **48**(2), 907–909 (2012)
5. Makni, Z., Besbes, M., Marchand, C.: Multiphysics design methodology of permanent-magnet synchronous motors. IEEE Trans. Veh. Technol. **56**(4), 1524–1530 (2007)
6. Osher, S.J., Sethian, J.A.: Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
7. Pulch, R.: Stochastic collocation and stochastic Galerkin methods for linear differential algebraic equations. J. Comput. Appl. Math. **262**, 281–291 (2014)
8. Putek, P., Paplicki, P., Palka, R.: Low cogging torque design of permanent magnet machine using modified multi-level set method with total variation regularization. IEEE Trans. Magn. **50**(2), 657–660 (2014)
9. Putek, P., Paplicki, P., Palka, R.: Topology optimization of rotor poles in a permanent-magnet machine using level set method and continuum design sensitivity analysis. COMPEL **66**, 1–21 (2014)
10. Römer, U., Schöps, S., Weiland, T.: Approximation of moments for the nonlinear magnetoquasistatic problem with material uncertainties. IEEE Trans. Magn. **50**(2), 7010204 (2014)
11. Rovers, J.M.M., Jansen, J.W., Lomonova, E.A.: Modeling of relative permeability of permanent magnet material using magnetic surface charges. IEEE Trans. Magn. **49**, 2913–2919 (2013)
12. Sergeant, P., Crevecoeur, G., Dupré, L., van den Bossche, A.: Characterization and optimization of a permanent magnet synchronous machine. COMPEL **28**, 272–284 (2008)
13. Stroud, A.H.: Some fifth degree integration formulas for symmetric regions. Math. Comput. **20**(93), 90–97 (1966)
14. Teckentrup, A.L., Jantsch, P., Webster, C.G., Gunzburger, M.: A multilevel stochastic collocation method for partial differential equations with random input data. pp. 1404–2647 (2014, Preprint). arXiv:1404.2647
15. Tiesler, H., Kirby, R.M., Xiu, D., Preusser, T.: Stochastic collocation for optimal control problems with stochastic PDE constraints. SIAM J. Numer. Anal. **50**, 2659–2682 (2012)
16. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford and Shah model. Int. J. Comput. Vis. **50**(3), 271–293 (2002)
17. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)
18. Xiu, D., Hesthaven, J.S.: High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**(3), 1118–1139 (2005)
19. Yao, W., Chen, X., Luo, W., van Tooren, M., Guo, J.: Review of uncertainty-based multidisciplinary design optimization methods for aerospace vehicles. Prog. Aerosp. Sci. **47**, 450–479 (2011)

First Results for Uncertainty Quantification in Co-Simulation of Coupled Electrical Circuits

Kai Gausling and Andreas Bartel

Abstract This paper combines uncertainty quantification with co-simulation numerically. Our focus is mainly on the behavior of the stochastic quantities during the iterations in the co-simulation for a test circuit with uncertain parameters. For this purpose we first classify the coupling structure of co-simulation model the test circuit by using standard theory. Using the gPC expansion for the stochastic process, we analyze the contraction and the rate of convergence of the co-simulation process.

1 Introduction

Co-simulation is an important method for coupled systems in time domain. Especially, when dedicated simulation tools for the subsystems are available, it is a relevant option. Co-simulation is performed on certain time periods or windows. Thereby, each time integration for a part of the unknowns assumes that values for the another unknowns are available. If the time window is just one time step one can tune accuracy by applying stepsize control. Our approach of co-simulation allows for larger time windows. Then, after integration over the time window, we have new time profiles for the unknowns of all parts of the partition. With these new time profiles we can re-start the co-simulation process over the same time window to further update the profiles. Thus we solve multiple times the subsystem.

Co-simulation applied to coupled ordinary differential equations (ODEs) always converges, see, e.g., [4]. The situation is different for coupled differential-algebraic equations (DAEs). In such cases convergence can only be guaranteed if a contraction condition is fulfilled, see, e.g., [1]. The theory of co-simulation shows that its

K. Gausling (✉)
Chair of Applied Mathematics/Numerical Analysis, Bergische Universität Wuppertal, 42119 Wuppertal, Germany
e-mail: gausling@math.uni-wuppertal.de

A. Bartel
Applied Math. & Numerical Analysis, Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: bartel@math.uni-wuppertal.de

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4_24

stability and its rate of convergence is directly influenced by a) the sequence in which the subsystems are computed and b) by the coupling interface, see, e.g., [3].

Co-simulation operates on time windows $[T_n, T_n + H]$ and tries to compute the overall solution iteratively by decoupling. Let (k) denote the current iteration and $(k-1)$ the old iterates, a co-simulation scheme can be encoded as:

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}, \mathbf{z}) & &\leftrightarrow \dot{\tilde{\mathbf{y}}} = \mathbf{F}\left(\tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{z}}^{(k)}, \tilde{\mathbf{y}}^{(k-1)}, \tilde{\mathbf{z}}^{(k-1)}\right) \\ 0 &= \mathbf{g}(\mathbf{y}, \mathbf{z}) & & \quad\; 0 = \mathbf{G}\left(\tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{z}}^{(k)}, \tilde{\mathbf{y}}^{(k-1)}, \tilde{\mathbf{z}}^{(k-1)}\right) \end{aligned} \tag{1}$$

employing splitting functions $\mathbf{F}, \mathbf{G}$ with compatibility $\mathbf{F}(\mathbf{y}, \mathbf{z}, \mathbf{y}, \mathbf{y}) = \mathbf{f}(\mathbf{y}, \mathbf{z})$ and similarly for $\mathbf{G}$. The actual successful splitting is usually part of the game and a piece of art. Then the contraction condition to guarantee convergence reads, e.g., [1, 2]:

$$\alpha := \|\mathbf{G}_{z^{(k)}}^{-1} \mathbf{G}_{z^{(k-1)}}\|_2 < 1 \tag{2}$$

with partial Jacobians $\mathbf{G}_u$. It is still open, how uncertainties in coupled systems change the contraction properties. In general, α (2) may depend on components from the model. In such cases contraction depends on the balance between several parameters. Consequently, dealing with uncertain components in the co-simulation may change the contraction condition (2), that is, α will become stochastic.

Our paper is arranged as follows: In Sect. 2 we consider a linear test circuit with uncertainties, where no algebraic constraint depend on old algebraic iterates (see Sect. 3). Thus (2) holds for all further considerations. Section 4 provides an introduction to the gPC as one suitable technique. Section 5 gives insight into our simulation settings. Finally in Sect. 6, we discuss our simulation results, especially the rate of convergence when co-simulation is applied in stochastic approaches.

2 Circuit Modeling and Uncertain Test Circuit

Usually, a mathematical model for electric circuits is obtained by modified nodal analysis (MNA), e.g., [5]. This leads to a DAE

$$\mathbf{E}(\mathbf{p})\,\dot{\mathbf{x}} + \mathbf{A}(\mathbf{p})\,\mathbf{x} = \mathbf{f}(t)$$

with dynamic part $\mathbf{E}$, static part $\mathbf{A}$, sources $\mathbf{f}$ and unknown $\mathbf{x}$ containing the node potentials and some branch currents. Here, the matrices $\mathbf{E}$, $\mathbf{A}$ depend on physical parameters $\mathbf{p} = (p_1, \ldots, p_q)^T$, which we assume to be uncertain. These parameters are assumed to be independent random variables. Our test example is the 2-level RLC network, Fig. 1, with uncertain components $\mathbf{p} = (R_1, R_2, C_1, C_2)^T$. We

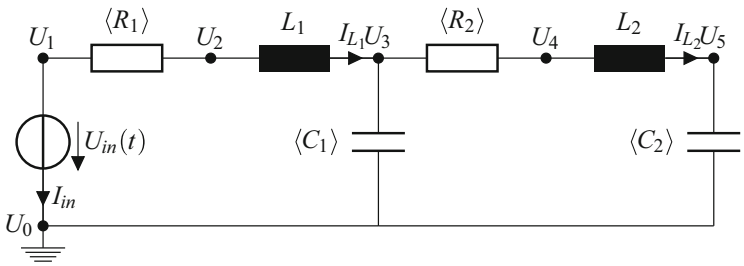


Fig. 1 Uncertain 2-level RLC circuit (reference model). $\langle \cdot \rangle$ indicate uncertain parameters

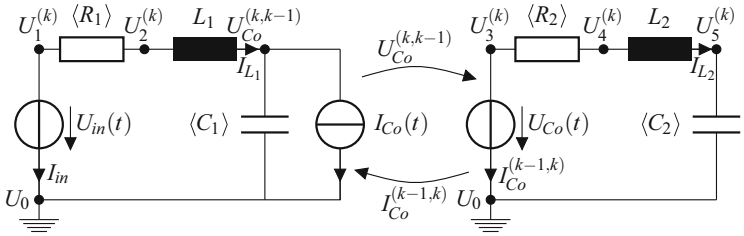


Fig. 2 Split 2-level RLC network with source-coupling for a Gauss-Seidel type of co-simulation (with uncertain R_1, R_2, C_1, C_2). Super-indices refer to the iteration count. The coupling variables have two super-indices (e.g. $U_{Co}^{(k,k-1)}$), which covers the two cases: (1) subsystem 1 first, and (2) subsystem 2 first

consider the two stochastic models for $\mathbf{p}$, i.e., Gaussian and uniform distribution: $(i = 1, 2)$

$$\begin{aligned} & R_i \sim \mathcal{N}(10\text{k}\Omega, \sigma^2 R_i), && C_i \sim \mathcal{N}(1\text{pF}, \sigma^2 C_i), \\ \text{or} \quad & R_i \sim \mathcal{U}(10\text{k}\Omega - \delta R_i, 10\text{k}\Omega + \delta R_i), && C_i \sim \mathcal{U}(1\text{pF} - \delta C_i, 1\text{pF} + \delta C_i). \end{aligned} \tag{3}$$

Furthermore we assume inductance $L = 1\text{mH}$ and supply voltage $U_{in}(t) = 1\text{V} \cdot \cos(\omega t)$ with $\omega = 2\pi \cdot 5 \cdot 10^3$ Hz. Now MNA yields a DAE of index-1. To apply co-simulation, we use source coupling [2] to split the system into two coupled networks at node U_3, see Fig. 2. Notice, both subsystems can be described by the same (index-1) DAE. The subsystem at the left receives information from the subsystem at the right by a current source and provides input to the system at the right by a voltage source. The exchange of information between both subsystems is organized by the additional variables U_{Co} and I_{Co}. Using a Gauss-Seidel type of co-simulation, we have to define, which system is computed first.

3 Abstract Coupling Analysis

To analyze co-simulation, one can use 'standard theory', e.g., [6]. To this end, we express the circuit model of Fig. 2 in semi-explicit form with variables

$$\begin{aligned}\dot{\mathbf{y}}_1 &= \mathbf{f}_1(\mathbf{y}_1, \mathbf{z}_1, \mathbf{z}_2),\ 0 = \mathbf{g}_1(\mathbf{y}_1, \mathbf{z}_1), \quad \mathbf{y}_1 := \left[U_{Co},\ I_{L_1}\right]^T, \mathbf{z}_1 := \left[U_1,\ U_2,\ I_{in}\right]^T.\\ \dot{\mathbf{y}}_2 &= \mathbf{f}_2(\mathbf{y}_2, \mathbf{z}_2),\ 0 = \mathbf{g}_2(\mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_2), \quad \mathbf{y}_2 := \left[U_5,\ I_{L_2}\right]^T, \mathbf{z}_2 := \left[U_3,\ U_4,\ I_{Co}\right]^T.\end{aligned} \tag{4}$$

Applying technique of MNA one obtains equations for $\mathbf{f}_1,\ \mathbf{g}_1,\ \mathbf{f}_2$ and $\mathbf{g}_2$:

$$\textbf{Subs. 1: } 0 = \begin{pmatrix} 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&C_1&0\\ 0&0&0&0&L_1\\ 0&0&0&0&0 \end{pmatrix}\begin{pmatrix}\dot{U}_1\\ \dot{U}_2\\ \dot{I}_{in}\\ \dot{U}_{Co}\\ \dot{I}_{L_1}\end{pmatrix} + \begin{pmatrix} -G_1 & G_1 & 1 & 0 & 0\\ G_1 & -G_1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & -1\\ 0 & 1 & 0 & -1 & 0\\ -1 & 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix}U_1\\ U_2\\ I_{in}\\ U_{Co}\\ I_{L_1}\end{pmatrix} - \begin{pmatrix}0\\ 0\\ -I_{Co}(t)\\ 0\\ -U_{in}(t)\end{pmatrix}, \tag{5}$$

$$\textbf{Subs. 2: } 0 = \begin{pmatrix} 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&C_2&0\\ 0&0&0&0&L_2\\ 0&0&0&0&0 \end{pmatrix}\begin{pmatrix}\dot{U}_3\\ \dot{U}_4\\ \dot{I}_{co}\\ \dot{U}_5\\ \dot{I}_{L_2}\end{pmatrix} + \begin{pmatrix} -G_2 & G_2 & 1 & 0 & 0\\ G_2 & -G_2 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & -1\\ 0 & 1 & 0 & -1 & 0\\ -1 & 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix}U_3\\ U_4\\ I_{co}\\ U_5\\ I_{L_2}\end{pmatrix} - \begin{pmatrix}0\\ 0\\ 0\\ 0\\ -U_{co}(t)\end{pmatrix}, \tag{6}$$

where $\partial\mathbf{g}_1/\partial\mathbf{z}_1$ and $\partial\mathbf{g}_2/\partial\mathbf{z}_2$ are regular. $\mathbf{y_1},\ \mathbf{y_2}$ define the differential and $\mathbf{z_1},\ \mathbf{z_2}$ the algebraic variables. Depending on what subsystem is computed first, we obtain different splitting schemes. For **subsystem 1 first** it reads:

$$\mathbf{F}(\cdot,\cdot,\cdot,\cdot) := \begin{bmatrix}\mathbf{f}_1(\mathbf{y}_1^{(k)}, \mathbf{z}_1^{(k)}, 0, \mathbf{z}_2^{(k-1)})\\ \mathbf{f}_2(0, 0, \mathbf{y}_2^{(k)}, \mathbf{z}_2^{(k)})\end{bmatrix},\ \mathbf{G}(\cdot,\cdot,\cdot,\cdot) := \begin{bmatrix}\mathbf{g}_1(\mathbf{y}_1^{(k)}, \mathbf{z}_1^{(k)}, 0, 0)\\ \mathbf{g}_2(\mathbf{y}_1^{(k)}, 0, \mathbf{y}_2^{(k)}, \mathbf{z}_2^{(k)})\end{bmatrix}. \tag{7}$$

Notice that no algebraic constraint depends on old algebraic iterates. Thus the contraction factor α vanishes for the splitting scheme (7). Furthermore it becomes obvious, that introducing uncertainty in our co-simulation model does not manipulate the properties of contraction. Consequently, convergence is guaranteed for the splitting schemes (7) involving uncertainties by using a time step size H small enough.

4 Generalized Polynomial Chaos (gPC)

To compute the stochastic quantities of our uncertain model, the gPC expansion method is used. The gPC expansion involving a finite number of P summands reads:

$$f(t,\mathbf{p}) \approx f_{\mathrm{gPC}}(t,\mathbf{p}) := \sum_{j=0}^{P-1} f_j(t)\,\Phi_j(\mathbf{p}), \tag{8}$$

with unknown time-dependent coefficient functions $f_j(t)$ and basis polynomials $\Phi_j(\mathbf{p})$, see, e.g., [7]. The polynomial basis represents an orthogonal system, which depends of the random parameters. Due to the orthogonality of the basis, it is easy to show that the mean and variance of the response respectively read:

$$\mathbb{E}[f(t,\mathbf{p})] = f_0(t), \quad \mathrm{Var}\left[f_{\mathrm{gPC}}(t,\mathbf{p})\right] = \sum_{j=1}^{P-1} f_j^2(t)\,\mathbb{E}\left[\Phi_j^2(\mathbf{p})\right]. \tag{9}$$

The costly part of the gPC expansion is to determine the unknown coefficient functions. For this purpose we employ stochastic collocation, e.g., [7]. The total sensitivity coefficients, which denote the interactions between several parameters, can be derived by regrouping the coefficient functions and subsequent normalization.

5 Numerical Simulation

For all our investigations, a co-simulation is studied in one time window $[t_0, t_0 + H]$ with $t_0 = 0.4$ ms. To obtain an adequate quality of approximation on H, a gPC expansion with maximum polynomial order three is used, thus momenta up to order three can be detected. We apply the stochastic collocation method which belongs to the family of non-intrusive methods. We use the Legendre-quadrature rule, see, e.g., [7], of order five based on a tensor-product grid in probability space, which requires to solve the model 81 times. Notice that these are sample points of Ω.

Our algorithm works in the following manner: For each sample-point out of Ω, the reference model is solved in time domain up to t_0 to obtain initial values which are close to the solution. Now we start co-simulation with k iteration steps for each sample on $[t_0, t_0 + H]$ using the corresponding initial values. Furthermore, constant extrapolation of the initial value is used for the initial guess $x^{(0)}(t)$ on the time window. Finally, we compute the stochastic momenta (depending on step k).

6 Numerical Results

Using MATLAB®we first investigate the error behavior related to the stochastic process in the output voltage U_5 using different levels of uncertainties for the splitting scheme (7). For this purpose, we consider a range of deviations for the resistances and capacitances, which are typical in electrical engineering. To this end we are focusing our attention on the error in the total sensitivity coefficients. The error of the solutions on the n-th time window after k iterations $x_c^{(k)}(t)$ is measured by comparing with a reference solution $x_{\text{ref}}(t)$ computed by a monolithic simulation: $\Delta_n^{(k)}(t) = x_{\text{ref}}(t) - x_c^{(k)}(t)$, $\delta_n^{(k)} := \|\Delta_n^{(k)}\|_{2,\infty}$. Thus, the $l^2 -$ norm is applied to all unknowns for each point of time followed by the infinity norm which refers the largest total error in H. To this end, we assume that the biggest discrepancy is located at the end of the time window. Please note that our co-simulation observations yet only apply to the time window H, which means that there is no error transport between several time windows. For a quantitative evaluation we compute the average error over all total sensitivity coefficients. As uncertainty we suppose uniformly distributed parameters with the same variation between $\delta R_1 = \delta R_2 = 0.1\ (10\,\%) - 0.5\ (50\,\%)$ for the resistances and $\delta C_1 = \delta C_2 = 0.1\ (10\,\%) - 0.5\ (50\,\%)$ for the capacitances around the nominal respective values.

Figure 3 shows the average error for $k = 1, 3, 5, 10$ iterations in the co-simulation. It becomes obvious, that for a high level of uncertainty the error becomes larger. Furthermore, a continuous improvement in the error up to $k = 10$ can be observed, especially in cases of high uncertainties for C_i and R_i. Accordingly, small uncertainties in our co-simulation model requires a smaller number of iterations, where the level of uncertainty in the capacitances mainly controls the rate of convergence.

Next we investigate the contraction and the rate of convergence regarding all node potentials $U_1, \ldots, U_5$ by calculating the expectation and standard deviation for each quantity. Afterwards, a deterministic solution will be computed by using the reference circuit, given in Fig. 1. In the case of uniformly distributed parameters, we suppose an variation about $\delta R_i = 0.1\ (10\,\%)$ for the resistances and $\delta C_i = 0.5\ (50\,\%)$ for the capacitances. Figure 4 shows convergence for splitting scheme (7). Thus all quantities have nearly the same rate of convergence for window sizes $10^{-8}\ s < H < 10^{-4}\ s$. It becomes obvious that a further reduction of the window size does not reduce the error in the expectation and standard deviation. This behavior differs to its deterministic solution, where an improvement up to the machine precision is achieved. The reasons for this are diverse: the usage of Gauss-Legendre quadrature formulas of order five produces a numerical quadrature error in each coefficient function $f_j(t)$ of (8). Furthermore, the accuracy of the stochastic process is limited by using a finite number of summands in the gPC expansion.

In order to investigate the performance of contraction, we decrease the window size by 10 % down to $[0.4,\ 0.49]$ ms. For our tests the minimum error is bounded by the time integrator precision of 10^{-3} with which we solve the subsystems. Figure 4 shows the performance of contraction for different quantities measured

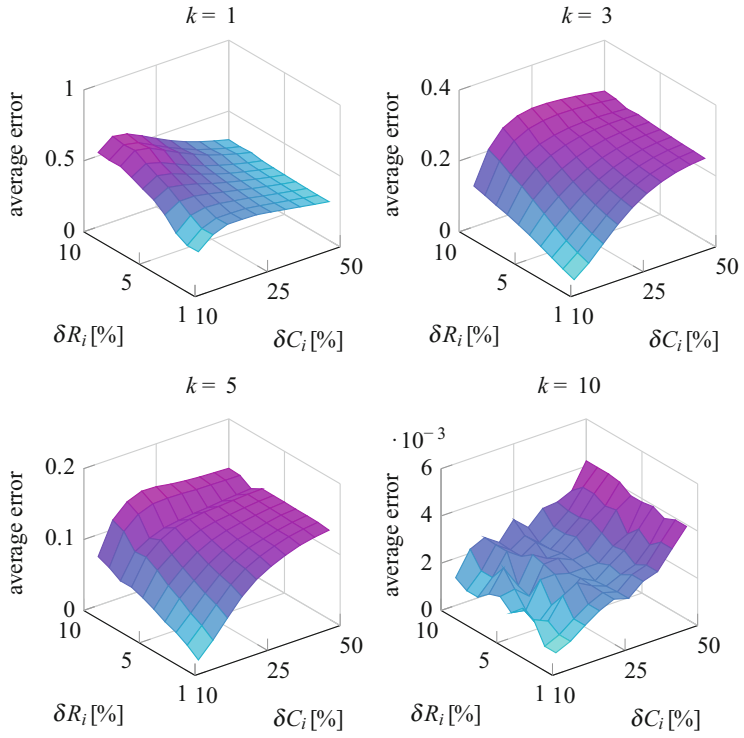


Fig. 3 Average error over all total sensitivity coefficients obtained by comparing with a reference solution for $k = 1, 3, 5, 10$ iterations over $H = 0.1$ ms. Uniform distribution (Legendre polynomials), $R_i \sim \mathcal{U}(10\text{k}\Omega - \delta R_i, 10\text{k}\Omega + \delta R_i)$, $C_i \sim \mathcal{U}(1\text{pF} - \delta C_i, 1\text{pF} + \delta C_i)$

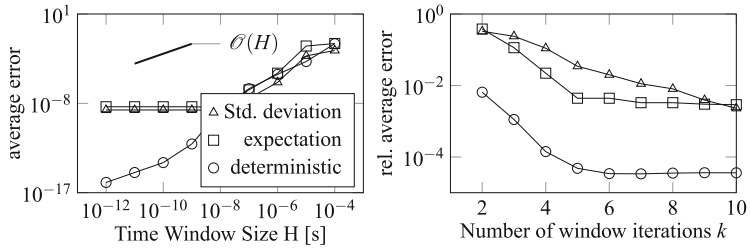


Fig. 4 (*left*) Convergence in expectation, standard deviation and in the deterministic solution concerning the node potentials U_1, U_2, U_3, U_4, U_5 after 0.4 ms for different time step sizes H with four iterations per time window. (*right*) Contraction measured by the relative average error in dependence of the iterations k on the time window [0.4, 0.49] ms

by the relative average error. It becomes apparent that the performance for the expectation is much better than for the standard deviation. Here, expectation is already reproduced after five iteration steps, whereas the standard deviation requires

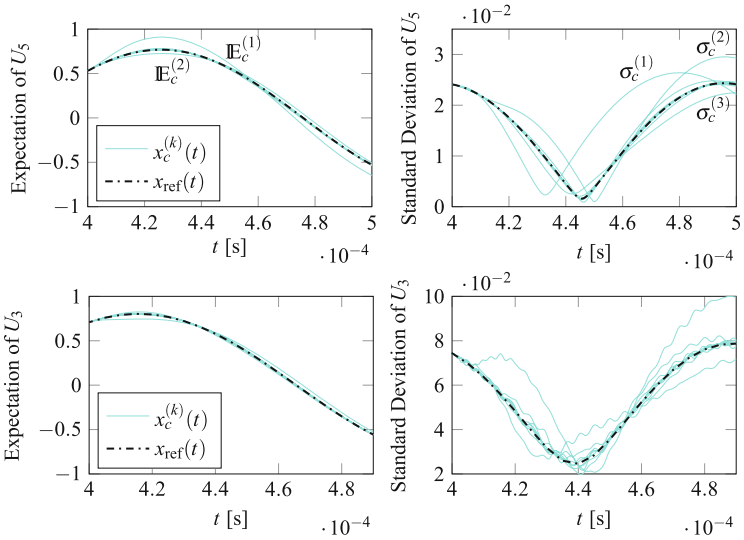


Fig. 5 Expectation and standard deviation for U_5, U_3 using different numbers of iteration steps for subsystem 1 first where uniformly distributed components are introduced. $\mathbb{E}_c^{(k)}$ and $\sigma_c^{(k)}$ denotes the solution of expectation and standard deviation over H for increasing number of iterations k

about ten iterations to achieve the maximum precision of $\delta_n^{(k)} = 10^{-3}$. Due to the definition of the expectation, which is exactly represented by the first coefficient function $f_0\,(t)$, see (9), it is resolved with a higher quadrature order than the standard deviation. Hence, in contrast to the standard deviation no approximation error is caused by using a finite number of coefficient functions in (8). This can explain Fig. 4 right.

An example is given in Fig. 5, where the expectation and standard deviation is presented only for the output voltage of each subsystem, i.e. node potential U_3 and U_5, over the time window $[0.4,\ 0.49]$ ms. As uncertainty we choose $R_i \sim \mathscr{U}(10\text{k}\Omega - 10\,\%, 10\text{k}\Omega + 10\,\%)$, $C_i \sim \mathscr{U}(1\text{pF} - 50\,\%, 1\text{pF} + 50\,\%)$. Our calculation aim is to provide a accurate band that must contain the solution of U_3 and U_5. Here, expectation is well approximated already after $k = 3$ iteration steps, whereas the standard deviation requires more than five iterations to achieve an error of approximately $\delta_n^{(k)} = 10^{-3}$. In addition there are oscillations in the standard deviation of U_3 over H, which are not be further analysed in this paper. However, tests have shown that the oscillation can be minimized by reducing the window size. All our investigations hold also by using Gaussian distribution settings given in (3).

7 Conclusions

We have shown for our test case, that the number of iterations which are needed to achieve a predefined accuracy is mainly controlled by the level of uncertainty. Co-simulation models with higher uncertainties naturally require a larger number of iterations. Furthermore, uncertainties in the capacitances have a greater impact than uncertainties in the resistances regarding the rate of convergence. Concerning our test example, the speed of contraction for expectation and standard deviation differs from each other. Thus, different stochastic quantities requires a different number of iterations to archive a suitable accuracy in co-simulation.

It is a future aim to combine co-simulation and UQ for electrical circuits, where the contraction factor α does not vanish.

Acknowledgements This work is supported by the German Federal Ministry of Education and Research (BMBF) in the research projects SIMUROM, http://www.simurom.de, (05M13PXB) and KoSMos, http://scwww.math.uni-augsburg.de/projects/kosmos, (05M13PXA).

References

1. Arnold, M., Günther, M.: Preconditioned dynamic iteration for coupled differential-algebraic systems. BIT (Ind.) **41**, 1–25 (2001)
2. Bartel, A., Brunk, M., Günther, M., Schöps, S.: Dynamic iteration for coupled problems of electric circuits and distributed devices. SIAM J. Sci. Comput. (Ind.) **35**(2), 315–335 (2013)
3. Bartel, A., Brunk, M., Schöps, S.: On the convergence rate of dynamic iteration for coupled problems with multiple subsystems. J. Comput. Appl. Math. (Ind.) **262**, 14–24 (2014)
4. Burrage, K.: Parallel Methods for Systems of Ordinary Differential Equations. Clarendon Press, Oxford (1995)
5. Feldmann, U., Günther, M.: CAD-based electric-circuit modeling in industry I: mathematical structure and index of network equations. Surv. Math. Ind. **8**(2), 97–129 (1999)
6. Gausling, K., Bartel, A.: Analysis of the contraction condition in the co-simulation of a specific electric circuit. BUW Preprint 14/32, accepted for ECMI 2014 proceedings
7. Xiu D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)

Index

Adaptive epansion point selection, 175
Adaptive weighting, 165
Additive RK-schemes, 115
Algebraic multigrid preconditioner, 83

Balanced truncation, 223
Bifurcation, 21
Boundary integral equations, 73

Circuit simulation, 13, 31, 243
Circuit synthesis, 31
Co-simulation, 243
Cogging torque, 233
Computational Fluid Dynamics (CFD), 103
Conjugate gradients, 83
Coupled problem, 133
Coupled problems, 115, 123
Curse of dimensionality, 199
CUSP library, 83

Device modelling, 91
Differential-algebraic equations (DAEs), 21
Dirac billard resonator, 43
Discontinuous Galerkin, 53
Domain decomposition method, 53
Dry Transformer, 103

Edge elements, 53
Eigenfields, 43
Electromagnetics, 43, 53, 63, 73, 83
Electronic transport, 91
Energy harvesting, 143

Finite integration technique (FIT), 43
Finite-element simulation, 63
Frequency domain, 43

GARK, 115
Generalized polynomial chaos (gPC) expansion, 243
GPU implementation, 83
Graphene, 3, 91
Green's function formalism, 3

High-frequency scattering, 73

Induction hardening, 133
Interface reduction, 185
Intrusive approach, 199

Latency, 13
Level set method, 233
Linear electric circuits, 223
Low-frequency stability, 63
LTI systems, 165

Maximum power point tracking, 143
Model order reduction, 31, 155, 165, 175, 223
Model order reduction (MOR), 185
Multi-objective optimization, 103

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4

Multiphysics problem, 123, 133
Multiphysics simulation, 31, 115
Multirate methods, 185
Multirate schemes, 115
Multirate simulation, 13
Multiscale, 3

Non equilibrium green function, 91
Non-intrusive approach, 199
Non-matching three-dimensional grids, 53
Nonlinear circuit, 21

Parallelisation, 123
Parametric modelling, 155
Passivity, 165
Phase transitions, 133
Photovoltaic, 143
Potential formulation, 63

Quadratic turning points, 21
Quantum transport, 3

Radio frequency (RF), 13, 31
Reduced basis method (RBM), 215
Reflective exploration, 175

Software agents, 123
Spectral methods, 73
Stochastic Collocation, 233
Stochastic collocation, 215

Thermal analysis, 155
Thermal/Pressure Networks, 103
Thermoelasticity, 133
Topology optimization, 233
Transfer function, 215

Uncertainty quantification, 199, 215, 223, 233, 243

Vector fitting, 165

Wavelets, 13

Authors Index

Ackermann Lee, W., 43

Bandinu, M., 165
Banova, T., 43
Bartel, A., 185, 233, 243
Benner, P., 155, 215
Bittner, K., 13
Blaszczyk, A., 103
Brachtendorf, H.-G., 13
Buchau, A., 123

Casagrande, R., 53
Chinea, A., 165
Ciuprina, G., 31
Clénet, S., 199
Clemens, M., 83
Cranganu-Cretu, B., 103
Cremasco,A., 103

Deretzis, I., 3
Dhaene, T., 175
Di Barba, 103
Diţă, B., 31
Dyczij-Edlinger, 63

Farle, O., 63
Feng, L., 155

Göhner, P., 123
Günther, M., 115, 185, 233

García de la Vega, I., 21
Gausling, K., 233, 243
Gawrylczy, K. M., 233
Grivet-Talocia, S., 165

Hachtel, Ch., 115, 185
Hess, M.W., 215
Hiptmair, R., 53

Isvoranu, D., 31

Jüttner, M., 123
Jerez-Hanckes, C., 73
Jochum, M., 63

Knockaert, L., 175
Kula, S., 31

La Magna, A., 3
Liu, Q., 133
Lup, A.-S., 31

Meuris, P., 155

Nadolski, D., 133

Ostrowski, J., 53

A. Bartel et al. (eds.), *Scientific Computing in Electrical Engineering*, Mathematics in Industry 23, DOI 10.1007/978-3-319-30399-4

Petzold, Th., 133
Pinto, J., 73
Pulch, R., 133, 223, 233
Putek, P., 233

Rahkonen, T., 143
Riaza, R., 21
Richter, Ch., 83
Romano, V., 3
Rucker, W. M., 123

Samuel, E. R., 175
Sandu, A., 115
Schöps, S., 83
Schoenmaker, W., 155
Schuss, Ch., 143
Sorohan, S., 31

Tournier, S., 73

Ubolli, A., 165

Vögeli, D., 123
van Rienen, U., 91

Weiland, Th., 43
Winkelmann, Ch., 53
Wu, W., 103

Zheng, D., 91