Reduced Order Multirate Schemes for Coupled Systems

Dissertation

zur Erlangung
des akademischen Grades
eines

Doktors der Naturwissenschaften
- Dr. rer. nat. -

der
Fakultät für Mathematik und Naturwissenschaften
der
Bergischen Universität Wuppertal (BUW)
vorgelegt von

Marcus W.F.M. Bannenberg

betreut von: Prof. Dr. rer. nat. Michael Günther (BUW)

Wuppertal 2022

Acknowlegdements

First and foremost, I would like to thank my supervisor prof. dr. Michael Günther for believing in me and guiding me in my research endeavours. Thank you for your time and your excellent supervision, for without it this thesis would not be here.

Next I would like to thank STMicroelectronics, and especially Angelo Ciccazzo and Giuliana Gangemi for hosting me and helping me understand the fascinating world of microelectronics.

From the faculty at the Bergische Universität Wuppertal I would like to thank all my coworkers. Especially dr. Jan ter Maten, our discussions over lunch have been pivotal in my research and helped me so much along the way.

To dr. Fotios Kasolis I would like to express my gratitude. It turns out that Laocoön was wrong after all when he stated "Timeo Danaos et dona ferentes", as your gift of knowledge was a most welcome one.

I am indebted to the funding given by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 765374, ROMSOC. All the ESRs from the ROMSOC project have been a massive support, sharing our experiences and elevating the conferences and workshops to a higher level.

Finally I would like to express my deepest gratitude to my loving parents and sister. This thesis is dedicated to you.

Summary

In the context of time-domain simulation of multiphysical integrated circuits, one often encounters large systems of coupled differential-algebraic equations. To keep the simulation times of these systems feasible, a multitude of techniques can be applied exploiting different characteristics of the underlying systems.

As there are different natural phenomena occurring at once inside these circuits, one of the exploited characteristics is the difference of time scales for each of these phenomena. This is done through multirate time integration. Another way of drastically improving the feasibility of these simulations is by incorporating model order reduction techniques. These model order reduction techniques aim to reduce the computational complexity of mathematical models in numerical simulations.

This thesis presents novel work on the combination of both model order reduction and multirate time-integration into reduced order multirate schemes. To construct these combined schemes different types of coupling structure and model order reduction approaches are considered. Furthermore, two different types of approaches for the numerical integration are applied and tested with numerical experiments.

Two nonlinear model order reduction techniques are discussed, proper orthogonal decomposition and maximum entropy snapshot sampling. They are applied both to systems of ordinary differential equations as well as differential-algebraic equations. For he maximum entropy snapshot sampling a method is presented for the estimation of the reduction parameter. For the multirate approach both Runge-Kutta as well as backward differentiation formula methods have been considered.

Results have been achieved in the numerical analysis of the reduced order multirate schemes and convergence has been proven. These results have been verified in both academic and industrial experiments. The numerical method described in this thesis shows clear computational advantages over regular integration methods.

Zusammenfassung

Bei der Simulation multiphysikalischer integrierter Schaltungen im Zeitbereich stößt man häufig auf große Systeme gekoppelter differential-algebraischer Gleichungen. Um die Simulationszeiten dieser Systeme praktikabel zu halten, kann eine Vielzahl von Techniken angewandt werden, die verschiedene Eigenschaften der zugrunde liegenden Systeme ausnutzen.

Da in diesen Schaltkreisen verschiedene natürliche Phänomene gleichzeitig auftreten, ist eine der genutzten Eigenschaften die Differenz der Zeitskalen für jedes dieser Phänomene. Dies geschieht durch eine mehrstufige Zeitintegration. Eine weitere Möglichkeit, die Durchführbarkeit dieser Simulationen drastisch zu verbessern, ist die Einbeziehung von Techniken zur Reduzierung der Modellordnung. Diese Techniken zur Reduzierung der Modellordnung zielen darauf ab, die Rechenkomplexität der mathematischen Modelle in numerischen Simulationen zu verringern.

In dieser Arbeit werden neue Arbeiten zur Kombination von Modellordnungsreduktion und Multirate-Zeitintegration in Multirate-Schemata reduzierter Ordnung vorgestellt. Um diese kombinierten Schemata zu konstruieren, werden verschiedene Arten von Kopplungsstrukturen und Ansätze zur Modellordnungsreduktion betrachtet. Darüber hinaus werden zwei verschiedene Arten von Ansätzen für die numerische Integration angewandt und mit numerischen Experimenten getestet.

Es werden zwei Verfahren zur Reduktion nichtlinearer Modellordnungen diskutiert, nämlich die orthogonale Zerlegung und das Maximum Entropy Snapshot Sampling. Sie werden sowohl auf Systeme gewöhnlicher Differentialgleichungen als auch auf differential-algebraische Gleichungen angewandt. Für das Maximum-Entropie-Snapshot-Sampling wird eine Methode zur Schätzung der Reduktionsparameter vorgestellt. Für den Multirate-Ansatz wurden sowohl Runge-Kutta- als auch Rückwärtsdifferenzierungsformelverfahren berücksichtigt.

Es wurden Ergebnisse in der numerischen Analyse der Multirate-Schemata reduzierter Ordnung erzielt und die Konvergenz wurde nachgewiesen. Diese Ergebnisse wurden sowohl in akademischen als auch in industriellen Experimenten verifiziert. Die in dieser Arbeit beschriebene numerische Methode zeigt deutliche rechnerische Vorteile gegenüber regulären Integrationsmethoden.

Samenvatting

In de context van tijddomein simulatie van multifysische geïntegreerde circuits, komt men vaak grote stelsels van gekoppelde differentiaal-algebraïsche vergelijkingen tegen. Om de simulatietijden van deze systemen haalbaar te houden, kan een veelvoud aan technieken worden toegepast die gebruik maken van de verschillende eigenschappen van de onderliggende systemen.

Aangezien er verschillende natuurverschijnselen tegelijk optreden in deze circuits, is een van de geëxploiteerde kenmerken het verschil in tijdschaal voor elk van deze verschijnselen. Deze exploitatie vindt plaats door middel van multirate tijdsintegratie. Een andere manier om de simulatietijd van deze simulaties drastisch te verbeteren is door gebruik te maken van model order reduction technieken. Deze model order reduction technieken hebben als doel de computationele complexiteit van de wiskundige modellen in numerieke simulaties te verminderen.

Dit proefschrift presenteert nieuw werk betreffende de combinatie van zowel model order reduction als multirate tijdintegratie in de multiratenschema's met verminderde orde. Om deze gecombineerde schema's te construeren worden verschillende soorten koppelstructuren en model order reduction benaderingen overwogen. Verder worden twee verschillende soorten benaderingen voor de numerieke integratie overwogen en zijn deze getest met numerieke experimenten.

Twee niet-lineaire model order reduction technieken worden besproken, namelijk proper orthogonal decomposition en maximum entropy snapshot sampling. Deze worden zowel toegepast op stelsels van gewone differentiaalvergelijkingen als op differentiaal-algebraïsche vergelijkingen. Voor de maximum entropy snapshot sampling techniek wordt een methode gepresenteerd voor de schatting van de reductieparameter. Voor de multirate integratie methoden komen zowel Runge-Kutta- als Backward Differential Formula in aanmerking.

Er zijn resultaten bereikt bij de numerieke analyse van de multiratenschema's met verminderde orde en de convergentie is bewezen. Deze resultaten zijn geverifieerd in zowel academische als industriële experimenten. De in dit proefschrift beschreven numerieke methode vertoont duidelijke rekenkundige voordelen ten opzichte van reguliere integratiemethoden.

Contents

List of Figures

List of Tables

Introduction

Motivation

Nowadays, a world without microchips seems inconceivable. This fundamental building block of our electronic devices plays one of the most, if not the utmost important role in our digital world. From our communication systems to navigation and even our financial infrastructure, integrated circuits are ingrained in every aspect of our lives. For the design and manufacturing of these devices, numerical simulation techniques are necessary, and thus the need for mathematical modelling and simulation arises.

Integrated circuits, or microchips, are made of silicon. Silicon is a natural semiconductor, which means that under certain conditions, it conducts electricity; whilst under others, it acts as an insulator. These different electrical properties of silicon can be achieved by the addition of impurities, which is done by a process called doping. These characteristics make it an ideal material for making transistors, which are devices that can be used to amplify of switch electronic signals or electrical power.

To use silicon, it must be processed. The silicon needs to reach a level of 99.999% purity, which is done through chemical processes. Heat is applied to create a purified silicon melt, and then it's grown into a mono-crystalline ingot, a salami-shaped bar of silicon. These ingots can be enormous, with measurements going up to 2 meter in length and weighing almost 500 kilograms. The ingots are sawed into super-thin wafers, less than 10 human hairs thick, by a diamond saw. The wafers are then polished in a number of steps until they're smooth and their surface has a mirror like surface. The wafer then undergoes a complex process, so the design of the microchip can be transferred onto the wafer. This design is transferred using photolithography, which projects each of the thousands of interconnected layers onto the wafer. Each one of the layers of the microchip is electrically connected to the next with billions of transistors, each with an unique circuit pattern. To make such etchings, a deep ultraviolet light source is used, hence the name photolithography.

Subsequently, each of the layers of the microchip are etched, polished, and integrated. The non-silicon elements in these layers, also known as the dopants, are used to further alter the electrical properties of each layer. When the silicon is exposed to the proper amounts of other elements, heat and pressure, it reacts by adjusting its conductivity. This essential step ensures that the chip works as intended by it's design and is not in danger of using too much or too little charge.

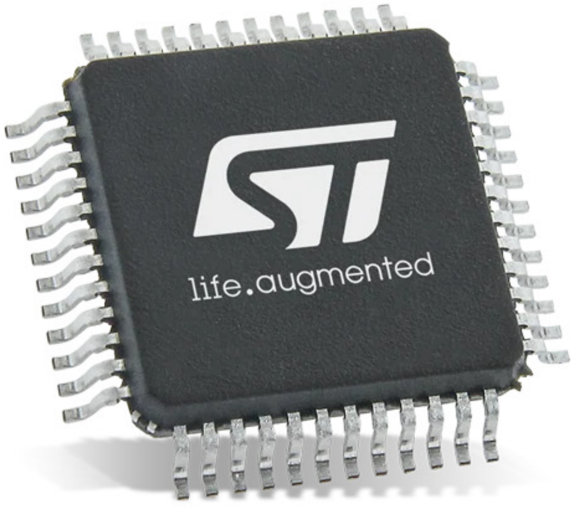

Figure 1: *The final version of a microchip. The L99ASC03G Multifunctional System IC is a brushless/sensorless 3-phase motor pre-driver for automotive applications, developed by STMicroelectronics.*

The conductive paths between layers are constructed by coating the whole microchip with metal, for which usually aluminium is used. Then the photolithographic process is used again to remove all of this coating but the conductive pathways. In case of larger microchips, this process may also involve multiple layers of conductors separated by glass.

Each individual node on the microchip is then tested for functionality. Should one of the nodes malfunction or behave abnormally, then the entire microchip has to be discarded. But if all parts are validated and function as expected, it's checked and marked to moved onto the final step, packaging. During the packaging the wafers are cut and the wires are attached. To protect the microchip from the elements and damage it is further encased. The final product is the integrated chip that we know and can be found in all of our electronic devices, see Figure 1.

As this process of developing a microchip is such a complex and time consuming endeavour, it is most important that the circuits are adequately designed. Designing integrated circuits has become very complex over the last decades. The number of transistors, one of the essential building blocks of both analogue and digital circuits, has grown nearly exponentially since their inception. However, the size of the microchips remained nearly constant, or even decreased. Thus each single transistor has to be described very detailed and take all kinds of environmental effects into account to be accurately simulated. This desire for accuracy results in very complex models to describe the behaviour of the natural phenomena occurring inside the microchip. Now however, due to the ever increasing complexity of the circuits, we even run into limitations using computer aided design as the mathematical models become prohibitively large.

In the context of time-domain simulation of multiphysical integrated circuits, one often encounters large systems of coupled differential-algebraic equations (DAEs). To keep the simulation times of these systems feasible, a multitude of techniques can be applied exploiting different characteristics of the underlying systems. As there are different natural phenomena occurring at once inside these circuits, one of the exploited characteristics is the difference of time scales for each of these phenomena. This is done through multirate (MR) time integration. Another way of drastically improving the feasibility of these simulations is by incorporating model order reduction (MOR) techniques. MOR techniques aim to reduce the computational complexity of mathematical models in numerical simulations.

In this thesis we present novel twofold approach to efficiently simulate coupled nonlinear DAEs by combining these two techniques. By applying multirate time integration and partitioning the circuit, large slow subsystems lend themselves for model order reduction. The gained accuracy at the cost of a slight increase in computational effort provided by multirate integration is kept whilst the computational complexity is reduced by applying model order reduction. These new reduced order multirate integration schemes are the central theme of this thesis.

Previous Work

Model order reduction for industrial circuit simulation has been a tried and tested approach. Most of the model order reduction algorithms apply strictly to linear time-invariant systems. The most popular classes of these type of algorithms are Krylov sub-space methods, [39], [35], [38] and truncated balanced realisation methods, [25], [49]. These methods are well understood and have a long standing record of being applied in industry, [40], [30], [29], [12].

In contrast, in the field of nonlinear model order reduction there are still many steps to be made. One of the most promising approaches to nonlinear model order reduction is through the use of the proper orthogonal decomposition method, [51], [13], [42], [54], [36]. This method is a type of reduced basis model order reduction method that constructs such a basis from a sample of time or frequency-domain snapshots. The temporal or parametric dependent information that is used for reduced basis construction is extracted from these snapshots by use of a singular value decomposition. These snapshots are obtained from a high-fidelity simulation of the full order problem. Based on predefined cut-off criteria, left singular vectors are selected to form the reduced basis. Although this method is widely used for nonlinear model order reduction there are some remarks to be made.

The singular value decomposition is an inherently linear method and optimal in the least-squares sense, it might remove high-frequency components in its basis construction, whilst these are especially present and of importance in the context of circuit simulation. To circumvent this, a discrete reduced basis framework founded on the asymptotic properties of measure-preserving transformations is used, the maximum entropy snapshot sampling method. As first presented by Kasolis in [31], this method is bases on the so-called Grassberger-Procaccia correlation sum, [24], on recurrence quantification analysis, and on an estimate of the invariant Kolmogorov-Sinai entropy, [15], [45]. The reduction is obtained by constraining the entropy estimate to be a strictly increasing function on the time index. By then applying any orthonormalization process the reduced basis is obtained.

The roots of multirate time-integration techniques can be traced back to a paper by Rice, [37], where the step-sizes for the integration were adapted to the level of activity of subsystems. Since then many works followed, [23], [27], [28], relating to many times of different integration techniques. This thesis marks the first time that multirate time-integration techniques have been combined with model order reduction methods to receive significant reductions in computational time, whilst maintaining accuracy. This thesis specifically combines both Runge-Kutta and Backward

Differentiation Formula methods, [20], [2], [52], with proper orthogonal decomposition and maximum entropy snapshot sampling methods. A compound-first-step multirate approach is used for stability reasons.

Results

The combination of model order reduction techniques and multirate time integration has been shown to be a very promising approach for the simulation of industrial circuits. The main goals of this research project were to show both analytically and numerically that these type of schemes were capable of reducing simulation times, whilst maintaining accuracy. As a result five papers have been published.

Paper I The first publication titled "A combination of model order reduction and multirate techniques for coupled dynamical systems" has been printed in the SCEE proceedings, [47]. The paper relates to the combination of RK multirate integration and nonlinear model by proper orthogonal decomposition. These techniques were applied to a problem with a DAE-ODE coupled structure. The slow system was the ODE subsystem and this was to be reduced. The techniques proposed in this paper were numerically verified.

Paper II This publication titled "Coupling of model order reduction and multirate techniques for coupled dynamical systems" has been printed in Applied Mathematics Letters 112, 2021, [4]. In this paper a numerical analysis is presented regarding the convergence of reduced order multirate schemes applied to DAE-ODE structured problems. This is further validated by numerical experiments and the result is that order 1 convergence can be achieved with a reduced computational effort.

Paper III The third paper titled "Maximum entropy snapshot sampling for reduced basis modelling" was presented at the IGTE 2020 conference and has been written as in a joint effort with Fotios Kasolis and Markus Clemens from the Chair of Electromagnetic Theory at the Bergische Universität Wuppertal, [8]. This paper has been selected from the conference proceedings to be published in The International Journal for Computation and Mathematics in Electrical and Electronic Engineering. In this paper the maximum entropy snapshot sampling is applied in circuit simulation and a novel approach for the estimation of the reduction parameter is presented.

Paper IV Published in Applied Numerical Mathematics, the paper titled "Reduced order multirate schemes for coupled differential-algebraic systems" presents the numerical analysis for convergence of reduced order multirate schemes applied to DAE-DAE structured problems, [5]. The main result of this paper is the proven order 1 convergence with a BDF style approach to integration. Numerically this has been verified by experiments.

Paper V This paper titled "Reduced Order Multirate Schemes in Industrial Circuit Simulation" has been written on invitation for the Journal of Mathematics in Industry, as an extension of the first paper for the SCEE proceedings, [7]. This paper illustrates how to apply the theoretically proven concepts of reduced order multirate scheme in an industrial circuit simulation setting. A test case from STMicroelectronics regarding a photo-voltaic solar panel is studied and positive simulation time reduction has been achieved. The paper serves as an accumulation of the previously obtained results.

Besides the published papers multiple deliverables have been written related to the ROMSOC project. These papers document the open source benchmark cases and provide the reader with plug-and-play source code to use reduced order multirate schemes. The main achievement regarding these documents is the presented academic circuit simulation package presented in the final deliverable, [6].

Document Outline

This thesis functions as a comprehensive compendium of the aforementioned publications, as to give a complete overview of the development of reduced order multirate methods. The document is divided into five distinct chapters that each cover a facet of the whole mathematical framework.

In Chapter 1 we introduce the mathematical foundations of differential-algebraic equations. This special type of differential equations needs careful consideration as they may represent ill-posed problems. The chapter presents established results that are necessary for the subsequent discussions in the thesis and to validate the model order reduction and multirate techniques that we apply to these equations.

Chapter 2 is dedicated to the subject of integrated circuit simulation. In this chapter the mathematical equations describing the phenomena occurring inside a microchip are obtained through charge-oriented modified nodal analysis. Individual types of components are presented and examples of their model and behaviour are given. Furthermore, a brief description of the transient analysis, a certain type of circuit analysis, is given.

Chapter 3 gives a background on the model order reduction techniques applied in the scope of this thesis. Furthermore, the application of the maximum entropy snapshot sampling method is presented. For this application, a novel technique for the estimation of the reduction parameter is outlined.

Then multirate time-integration is discussed in Chapter 4. After the definition of regular multirate time-integration, the notion of reduced order multirate is given. Important convergence results from numerical analysis on these schemes are presented in this chapter and it functions as the last theoretical chapter.

In Chapter 5 numerical experiments are performed with the previously presented techniques for illustrative and explanatory purposes. This chapter verifies the theoretically claimed results from the previous chapters and shows the advantages of reduced order multirate integration schemes. Both academic and industrial test cases are shown.

The final chapter rounds up all the previously presented results and highlights the main take-aways. It concludes with and outlook on future perspectives in the fields of multirate and model order reduction and their combinations.

1

Differential-Algebraic Equations

Abstract

Differential-algebraic equations arise in the mathematical modelling of a variety of problems such as in multibody and flexible body mechanics or electrical circuit design. The mathematical models to describe these type of systems have the property that they are governed by both differential equations, describing the dynamics of the underlying system, as well as algebraic equations, which put constraints on the solutions to the differential equations. This chapter introduces the basic definitions and principles of differential-algebraic equations.

Introduction

Before introducing the numerical integration and model order reduction techniques the foundation of the mathematical framework for differential-algebraic equations (DAEs) is presented. In this chapter established results are discussed regarding differential-algebraic equations and their characteristics. It is essential to consider these special cases of differential equations as systems of DAEs may represent ill-posed problems. Therefore, more considerations are required to solve these systems compared to systems of ordinary differential equations (ODEs).

1.1 Basic Definitions

Let us start by giving the definition of them most general form of a nonlinear differential-algebraic equation.

Definition *1.1.1* (Differential-algebraic equation, DAE)**.** A general nonlinear differential-algebraic equation is defined by

$$F(t, x, \dot{x}) = 0, \tag{1.1a}$$

$$x(t_0) = x_0. \tag{1.1b}$$

with $F : I \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ with $n, m \in \mathbb{N}$ and $I \subseteq \mathbb{R}$ a compact interval. Provided with initial conditions $x(t_0) = x_0 \in \mathbb{R}^n$ it is an initial value problem.

Definition *1.1.2* (Solution and solvability)**.** Let x be a function defined by $x : I \to \mathbb{R}^n$, where $I \subseteq \mathbb{R}$ is a compact interval.

- The function x is a solution of the DAEs (1.1a), if it satisfies (1.1a) pointwise.
- The function x is a solution of the initial value problem (1.1), if x is a solution of (1.1a) and in addition satisfies the initial condition (1.1b).
- If there exists at least one solution to the initial value problem (1.1) it is called solvable.

The difference between a system of implicit ordinary differential equations, ODEs, and DAEs is that for ODEs the Jacobian matrix $\frac{\partial F}{\partial \dot{x}}$ is nonsingular. However, in the scope of integrated circuit simulation this is generally not the case. If the Jacobian matrix $\frac{\partial F}{\partial \dot{x}}$ is singular, there are some key aspects to keep in mind such as the choice of consistent initial conditions x_0 and the solvability of the initial value problem, as it consists of a mix of differential and algebraic equations. Furthermore, the DAEs arising in integrated circuit simulation are often encountered in the compact form

$$\dot{q}(x) + j(x) = 0, \tag{1.2}$$

with $q, j : \mathbb{R}^n \to \mathbb{R}^n$ or even semi linear implicit form

$$E\dot{x} = j(x), \tag{1.3}$$

where $E \in \mathbb{R}^{n \times n}$ and $x \in \mathbb{R}^n$.

Due to the singularity of the Jacobian matrix, systems of DAEs are more difficult to solve. As previously seen, the solution of the initial value problem has to satisfy a number of algebraic equations in the DAE setting. These algebraic constraints also apply at the initial time t_0. Therefore, the initial conditions also have to satisfy these constraints. Such initial conditions are called consistent.

Definition *1.1.3* (Consistent initial conditions)**.** The vector $x_0 \in \mathbb{R}^m$ is called a consistent initial condition of (1.1), if there exists a solution that fulfils $x(t_0) = x_0$ and must satisfy the algebraic constraints and even the differentiated hidden constraints of the system.

There are several different special classes of implicit DAEs. It is easy to recognise these distinct subclasses and they often appear in applications. These subclasses have a specific structure which makes them relatively simple, compared to fully implicit DAEs which can be very complex. One of these subclasses are semi-explicit DAEs.

Definition *1.1.4* (Semi-explicit DAEs)**.** For a pair of vectors $x \in \mathbb{R}^{n_x}$ and $z \in \mathbb{R}^{n_z}$ the semi-explicit DAEs has the form

$$\dot{x} = f(t, z, x), \tag{1.4}$$

$$0 = g(t, z, x), \tag{1.5}$$

with $f : I \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$, $g : I \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_z}$, with $n_x, n_z \in \mathbb{N}$ and $I \subseteq \mathbb{R}$ a compact interval.

The definition illustrates that semi-explicit DAEs can be considered as a system of differential equations that is combined with some algebraic equations. These algebraic equations, or algebraic constraints, define a manifold to which the solution is constrained. Therefore, DAEs can be interpreted as differential equations on manifolds. The good thing about semi-explicit DAEs is that the variables can be neatly divided into differential variables and algebraic variables.

From the semi-explicit form it is clear that the initial conditions have to satisfy the algebraic equations of the DAEs, as not every solution of (1.4) satisfies (1.5). To obtain these initial conditions a steady state analysis can be performed by solving

$$F(0, x_0, t_0) = 0. \tag{1.6}$$

for x_0. Note however that in the nonlinear case the Jacobian matrix may become numerically singular due to vanishing partial derivatives or in the case of bifurcation. Then such a steady state analysis cannot be performed. Finding these consistent initial conditions often requires the consideration of the derivatives of the DAEs. To classify this dependency the highest order of such a derivative that is necessary is called the differentiation-index.

Definition *1.1.5* (Differentiation-index)**.** The differential index k of a nonlinear, sufficiently smooth DAEs in general form (1.1), is the smallest k for which the system:

$$\begin{aligned} F(t, x, \dot{x}) &= 0, \\ \frac{d}{dt} F(t, x, \dot{x}) &= 0, \\ &\vdots \\ \frac{d^k}{dt^k} F(t, x, \dot{x}) &= 0, \end{aligned}$$

uniquely determines the variable $\dot{x}$ as a continuous function of x and t.

In general, if the index of a system of DAEs is greater than 1, then there are additional algebraic constraints which are not explicitly given. These hidden algebraic constraints can be derived by differentiation and algebraic transformations of the DAEs. For DAEs with an index equal to one, the stationary solution is a consistent solution.

It is obvious that a system of ODEs has differentiation-index 0. The index can be seen as a measure of the degree of singularity in the system. In general, the higher the index the more complex the problem and the more difficulties we are likely to encounter in solving the DAEs by a numerical method, as we shall see. In general, the differentiation-ndex is the most important and most used definition of the index of a DAEs. Therefore, through out the rest of the thesis index stands for differentiation-index.

Example *1.1.1*. The simplest form of a nonlinear semi-explicit DAE s

$$\begin{aligned} \dot{x} &= f(t, z, x), \\ 0 &= g(t, z, x), \end{aligned}$$

is guaranteed to be of index-1 by the assumption that the Jacobian

$$\frac{\partial g(t, z, x)}{\partial z} \text{ is nonsingular.}$$

The solution of this semi-explicit nonlinear system lies on the manifold defined by the algebraic constraints coming from g. Differentiation of this algebraic equation gives

$$0 = \frac{\partial g(t,z,x)}{\partial x}\dot{x} + \frac{\partial g(t,z,x)}{\partial z}\dot{y} + \frac{\partial g(t,z,x)}{\partial t}.$$

Now by using that the Jacobian $\frac{\partial g(t,z,x)}{\partial z}$ is nonsingular, the system can be written as a regular ODEs that is solvable for $\dot{x}$ and $\dot{y}$.

1.2 Problems in Implicit Form

Very often, differential-algebraic problems arising in practice are not at once in the semi-explicit form. A system of DAEs in semi linear implicit form

$$E\dot{y} = j(y), \tag{1.7}$$

can be transformed into semi-explicit form. By decomposing matrix E into

$$E = S\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}T, \tag{1.8}$$

with S, T invertible matrices and I the identity matrix with dimensions equal to the rank of E. This decomposition can be achieved by using Gaussian elimination with total pivoting. Then the variable y can be transformed into variables x, z by

$$Ty = \begin{pmatrix} x \\ z \end{pmatrix}. \tag{1.9}$$

Thus we obtain a semi explicit system where the differential and algebraic equations are defined by

$$\begin{pmatrix} f(t,z,x) \\ g(t,z,x) \end{pmatrix} = S^{-1}j(T^{-1}\begin{pmatrix} x \\ z \end{pmatrix}). \tag{1.10}$$

Therefore, any results holding for DAEs in semi-explicit form also hold for the semi linear implicit form. Initial values y_0 are consistent if it holds that $j(y_0)$ lies in the range of the matrix E.

2

Integrated Circuit Simulation

Abstract

Computer-aided design plays an increasingly important role in the development of integrated circuit. To verify the validity of the integrated circuited designed with a computer, simulations have a significant edge over prototyping. Here arises the need for mathematical models to accurate simulate the natural phenomena occurring inside the microchip. Charge-oriented modified nodal analysis is used to construct the network equations used in the simulation of the circuit. The characteristics of different components and their electrical behaviour are discussed.

In integrated circuit design, there are a significant number of design possibilities under which the internal components need to be guaranteed to work. This leads to a whole range of explorations to ensure sound functionality of the design. These explorations are performed by numerical simulations of the circuits mathematical model. To this end it is necessary to correctly describe the phenomena occurring inside a microchip with mathematical models.

This chapter discusses the mathematical properties of different types of network devices. We start by elaborating on the chosen network modelling approach and detail how each components contributes to the system of equations. The chapter ends with an overview of the different types of circuit analysis that can be performed for a given integrated circuit.

2.1 Modified Nodal Analysis

In the field of circuit simulation, a network modelling approach is used to derive mathematical models describing physical electrical circuits that can be used in computer-aided design. These mathematical models, sets of so-called network equations, are generated by the combination of network topology and characteristic equations that describe the physical behaviour of the network elements. One of these equation generating network approaches commonly used in industry is the charge-oriented modified nodal analysis (MNA), see [26], [3]. After a brief description of MNA, this section extends the standard charge-oriented MNA to include nonlinear components such as diodes and transistors. As these nonlinear network elements introduce more complexity into the network equations extra consideration is required.

2.1.1 Charge-Oriented MNA

The charge-oriented MNA describes the time behaviour of the activity of a circuit in physical quantities such as branch currents $I(t) \in \mathbb{R}^{n_I}$, branch voltages $U(t) \in \mathbb{R}^{n_I}$ and node voltages $u(t) \in \mathbb{R}^{n_u}$. More physical quantities like electrical charge $q(t) \in \mathbb{R}^{n_q}$ and magnetic fluxes $\phi(t) \in \mathbb{R}^{n_\phi}$ can be included in the set of variables, but more on this later. The derivative of a quantity with respect to time is indicated by a $\dot{}$ on top of it.

The electrical circuit is considered to be a collection of interconnected network elements and nodes. The real physical circuit elements such as resistors, capacitors and especially semiconductor devices, modelled by companion models, which idealise their behaviour. This is done by associating each element with one or more characteristic equations. The composition of these elements is governed by Kirchhoff's voltage law (KVL) and Kirchhoff's current law (KCL).

Definition *2.1.1*. Kirchhoff's voltage law The directed sum of the potential differences, voltages, around any closed loop is zero.

$$\sum_{k=1}^{n} v_k = 0. \tag{2.1}$$

Where n is the number of voltages in the closed loop. See Figure 2.1 right.

The KVL states that the algebraic sum of voltages along each loop of the network must be equal to zero at every instant of time. This law is

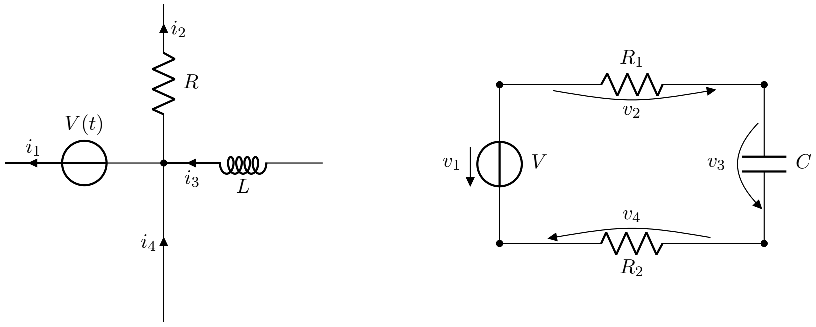

Figure 2.1: *The current entering any junction is equal to the current leaving that junction (left). The sum of all the voltages around a loop is equal to zero (right).*

used to define the relationship between branch voltages and node voltages

$$A^{\top} \cdot u(t) = U(t), \tag{2.2}$$

with an incidence matrix $A \in \{-1, 0, 1\}^{n_u \times n_I}$ that describes the directed graph that represents the network.

Definition *2.1.2*. Kirchhoff's current law The algebraic sum of currents in a network of conductors meeting at a point is zero.

$$\sum_{k=1}^{n} i_k = 0. \tag{2.3}$$

Where n is the number of branches with currents flowing towards or away from the node. See Figure 2.1 left.

The KCL states that the algebraic sum of currents traversing each cutset of the network must be equal to zero at every instant of time. Therefore, the sum of currents leaving any circuit node is zero

$$A \cdot I(t) = 0. \tag{2.4}$$

By applying the KCL to the terminals of an element and integrating over time, the KCL provides the charge neutrality requirement. This means that the sum of charges q_k over all terminals k of each element must be constant. Subsequently the constant can be set to zero without loss of generality.

With the purely topological relations defined, additional equations are provided by the characteristic equations related to the physical behaviour of each network element. The formulas used for the basic linear network elements shown in Figure 2.2, are given in Table 2.1. Interconnects and semiconductor devices such as transistors are modelled by multi-terminal elements, which will be discussed later sections for each relevant type.

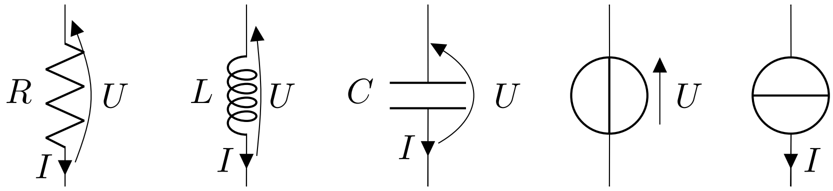

Figure 2.2: *The five basic network elements: resistor, inductor, capacitor, voltage source and current source.*

Element	Linear	General
Resistor	$I = \frac{1}{R} \cdot U$	$I = \mathcal{R}(U)$
Inductor	$U = L \cdot \dot{I}$	$U = \dot{\phi}$ with $\phi = \phi_L(I)$
Capacitor	$I = C \cdot \dot{U}$	$I = \dot{q}$ with $q = q_c(U)$
Sources	Independent	Controlled
Voltage	$U = v(t)$	$U = v(U_{\text{ctrl}}, I_{\text{ctrl}}, t)$
Current	$I = i(t)$	$I = i(U_{\text{ctrl}}, I_{\text{ctrl}}, t)$

Table 2.1: *Characteristic equations of the basic network elements.*

To set up the MNA network equations, the KCL is applied to each node, except the node that is considered the ground node. The incidence matrix is then defined as a collection of incidence matrices related to each different type of element,

$$A = \{A_R, A_L, A_C, A_V, A_I\},$$

with $A_\Omega \in \{0, +1, -1\}^{n_u \times n_\Omega}$, where n_Ω is the cardinality the set of each type of network element. Using these incidence matrices, we can relate the branch voltages in a loop and the currents accumulating in a node by applying KCL and KVL to each node, resulting in

$$A_C \dot{q} + A_R \mathcal{R}(A_R^\top u, t) + A_L \jmath_L + A_V \jmath_V + A_I i(t) = 0, \tag{2.5a}$$

$$\dot{\phi} - A_L^\top u = 0, \tag{2.5b}$$

$$v(t) - A_V^\top u = 0, \tag{2.5c}$$

$$q - q_C(A_C^\top u) = 0, \tag{2.5d}$$

$$\phi - \phi_L(\jmath_L) = 0. \tag{2.5e}$$

The unknowns q, ϕ, u, $\jmath_L$ and $\jmath_V$ are the charges, fluxes, node voltages, inductor currents and voltage source currents, respectively. All these quantities are time dependent, and are combined into one state vector $x(t) \in \mathbb{R}^m$ of unknowns. The dimension of which is given by the cumulative dimensions of the quantities. The network equations can now be stated in compact form

$$\dot{q}(x(t)) + j(x(t)) + Br(t) = 0, \tag{2.6}$$

where q and j are mappings from $\mathbb{R}^m$ to $\mathbb{R}^m$ related to the network elements, while the source term vector $r(t) \in \mathbb{R}^n$, $n = n_V + n_I$, combines the current- and voltage-sources. These are mapped to the corresponding nodes and branches by matrix $B \in \mathbb{R}^{m \times n}$.

Under the assumption that the underlying circuit only contains linear capacitors and inductors, the charge/flux considerations can be simplified. The network equations can then be written in the following matrix form

$$E\dot{x}(t) + Ax(t) + p(x(t)) + Br(t) = 0. \tag{2.7}$$

with

$$E = \begin{pmatrix} A_C \mathcal{C} A_C^\top & 0 & 0 \\ 0 & \mathcal{L} & 0 \\ 0 & 0 & 0 \end{pmatrix}, A = \begin{pmatrix} A_R \mathcal{G} A_R^\top & A_L & A_V \\ -A_L^\top & 0 & 0 \\ -A_V^\top & 0 & 0 \end{pmatrix}, B = \begin{pmatrix} A_I & 0 \\ 0 & 0 \\ 0 & I_{n_V} \end{pmatrix}.$$

Where the term $j(x(t))$ is split into a linear term $Ax(t)$ and a nonlinear term $p(x(t))$. The nonlinear term $p(x(t))$ is introduced to encapsulate nonlinear behaviour of semiconductor devices. The matrices $\mathcal{G}$, $\mathcal{L}$ and $\mathcal{C}$ are diagonal matrices containing the individual conductances, inductances and capacitances of the network elements and I_{n_V} is a n_V by n_V identity matrix.

In the case of circuit simulation there is a relation between the topology of a circuit and the index of the DAEs describing the circuit. To define this relation let us first introduce the following definitions.

Definition *2.1.3* (Loops and cutsets)**.** Given a circuit let the following structures be defined:

- A *LI*-cutset is a cutset consisting of inductors or both inductors and current sources.
- A *CV*-loop is a loop consisting of both capacitors and voltage sources.

Now by using these definitions, the following statements can be made concerning the differential index of the DAEs describing the circuit that do or do not contain these special structures.

- If a circuit contains neither CV-loops nor LI-cutsets, then the differential index of the DAEs describing this circuit is equal to 1.

- If a circuit contains either CV-loops, LI-cutsets or both CV- loops and LI-cutsets, then the differential index is equal to 2.

An example RLC-circuit consisting solely of resistor, capacitor and inductor network elements is shown in Figure 2.3. Notice that, as this circuit does not contain nonlinear elements, it can be described by Equation (2.7) with the nonlinear term set to $p(x(t)) = 0$. The network equations are then given by

$$0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & C_1 & -C_1 & 0 & 0 & 0 \\ 0 & -C_1 & C_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{u_1} \\ \dot{u_2} \\ \dot{u_3} \\ \dot{u_4} \\ \dot{I_L} \\ \dot{I_V} \end{pmatrix} + \begin{pmatrix} \frac{1}{R} & -\frac{1}{R} & 0 & 0 & 0 & 1 \\ -\frac{1}{R} & \frac{1}{R} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{R} & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ I_L \\ I_V \end{pmatrix},$$

$$+ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -V \end{pmatrix}.$$

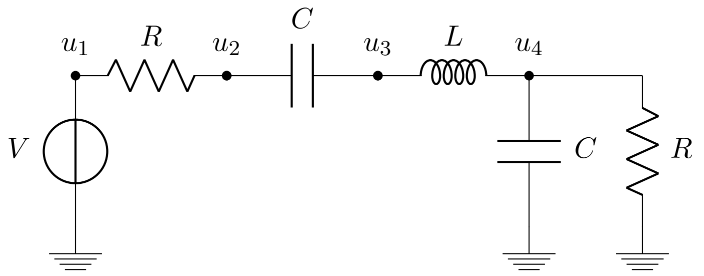

Figure 2.3: *RLC circuit.*

2.1.2 Extension of Basic MNA

Besides the linear network elements seen previously, there is also a need to model nonlinear network elements. The behaviour of each of these nonlinear elements is modelled by idealised companion models. For each of the to be considered nonlinear components, the companion model is discussed below.

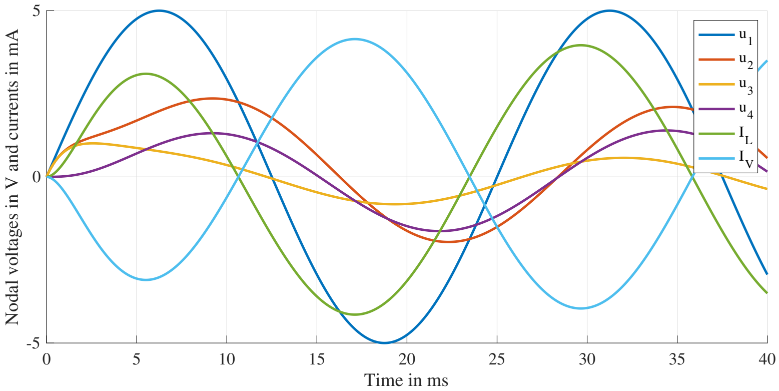

Figure 2.4: *Transient analysis of the RLC circuit for time interval* $[0, 40 \text{ ms}]$.

Shockley Diode

The characteristic function for diode network elements is modelled by the Shockley diode equation. This relates the current I of a p-n junction diode to the voltage drop over the diode V_D. This current is given by

$$I = I_S\left(e^{\frac{V_D}{nV_T}} - 1\right), \tag{2.8}$$

where I_S is the saturation current of the diode, which in most cases has an order of magnitude of 10^{-12} A. The thermal voltage is given by $V_T = kT/q \approx 26$ mV, for normal temperatures, and n is in this case the diode ideality factor, which is set to 1 for silicon diodes. As Equation (2.8) has a nonlinear dependency it is added to nonlinear vector term $p(x(t))$ on the indices of the connected nodes. $A^\top A^\top$

To illustrate the equations arising from incorporating diodes into the network elements consider the circuit shown in Figure 2.5. Applying MNA

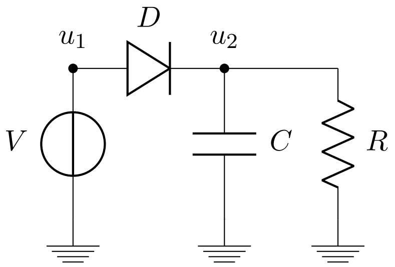

Figure 2.5: *Nonlinear time dependent circuit with a diode.*

to this circuit results in a state vector $x(t) = (u_1, u_2, I_V)^\top$, which is a combination of the nodal voltages u_1 and u_2 and the current I_V through the voltage source V. The network equations are then given by

$$0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{I}_V \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & \frac{1}{R} & 0 \\ -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ I_V \end{pmatrix},$$

$$+ \begin{pmatrix} I_S(e^{\frac{u_1-u_2}{nV_T}} - 1) \\ -I_S(e^{\frac{u_1-u_2}{nV_T}} - 1) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -V \end{pmatrix}.$$

Where $C = 10^{-6}$ F, $R = 10^3\ \Omega$, $I_S = 10^{-12}$ A and under normal temperatures with ideality factor $n = 1$. For a transient analysis of this circuit, the system of DAEs is solved on time interval [040 ms] with initial conditions $u(0) = (0, 0, 0)^\top$. The voltage source V supplies is modelled by a pulse input of $v = 5 \cdot \sin 40t$. This yields the response depicted in Figure 2.6.

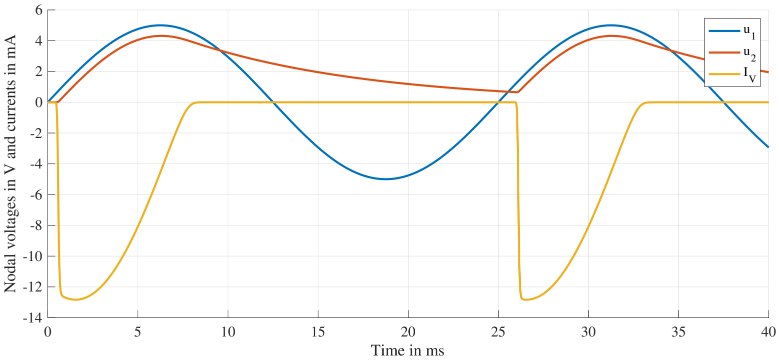

Figure 2.6: *Transient analysis of the diode circuit for the time interval* [0, 40 ms].

MOSFET

Another type of transistor is the metal-oxide-semiconductor field-effect transistor (MOSFET), which is a type of insulated-gate field-effect transistor (IGFET). Like the BJT transistor, a MOSFET can be either a p-type or an n-type, see Figure 2.7. The characteristics of the operation of a MOSFET depend on the voltages at each of the terminals, if the MOSFET is in enhancement or depletion mode and the type of MOSFET, [43], [9]. In the following discussion, the MOSFETS are considered to be enhancement-

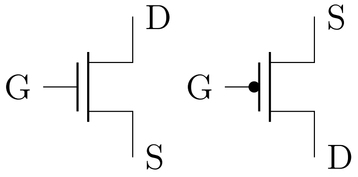

Figure 2.7: *The n-type and p-type MOSFET symbols respectively, with the terminals: gate (G), drain (D) and source (S).*

mode, n-type MOSFET. Then there are three distinct operational modes, depending on the terminal voltages.

- *Weak-inversion mode* - In this mode it holds that $V_{GS} < V_T$, thus the voltage between G and S is less than a threshold voltage V_T. In a very idealised model, the transistor is turned off and there is no conduction between D and S, thus $I_D = 0$. However, in the weak-inversion case, there is a current that is exponentially defined by the voltage V_{GS} and is approximated by

$$I_D \approx I_{D0} e^{\frac{V_{GS}-V_T}{V_T}}. \tag{2.9}$$

  With I_{D0} the current at $V_{GS} = V_T$, and the thermal voltage $V_T = kT/q$ defined by the temperature, Boltzmann constant and the specific resistance.

- *Triode mode* - When $V_{GS} \geq V_T$ but $V_{DS} < V_{GS} - V_T$ then the MOSFET is in triode mode. The transistor is switched on and operates like a resistor, which is controlled by the gate voltage. The current from the drain to the source is then modelled by

$$I_D = 2K\left[(V_{GS} - V_T) * V_{DS} - \frac{V_{DS}^2}{2}\right](1 + \lambda V_{DS}). \tag{2.10}$$

  Here $K = \frac{\mu C}{2}\frac{W}{L}$, where μ is the carrier mobility, C is the capacitance per unit area of gate and W and L the width and length of the gate respectively. The parameter λ is the channel-length modulation parameter. To simplify circuit analysis this can be set to 0, but may yield unrealistic results.

- *Saturation mode* - When $V_{GS} \geq V_T$ and $V_{DS} \geq V_{GS} - V_T$ then the MOSFET is saturated, or in active mode. The switch has been turned on, and a channel has been created. The current is now only weakly dependent upon the drain voltage and is modelled by

$$I_D = K[V_{GS} - V_T]^2[1 + \lambda(V_{DS} - V_{GS} - V_T)]. \tag{2.11}$$

A very common usage of MOSFETs is to invert a signal. To illustrate the behaviour consider the circuit of which the schematic is given in Figure 2.8, and the correponding network equations in Equation (2.12).

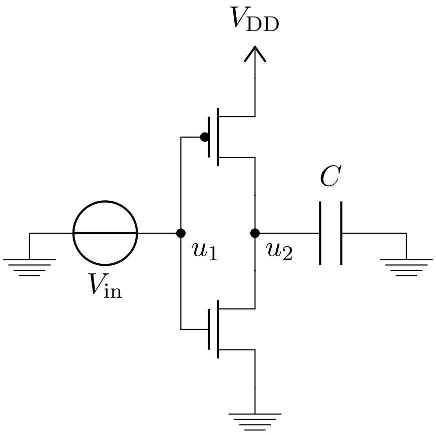

Figure 2.8: *A simple inverter circuit with n-type and p-type MOSFETs.*

$$\begin{aligned} 0 = &\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & C & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_{V_{\mathrm{dd}}} \\ \dot{I}_{V_{\mathrm{in}}} \\ \dot{I}_{V_{\mathrm{dd}}} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_{V_{\mathrm{dd}}} \\ I_{V_{\mathrm{in}}} \\ I_{V_{\mathrm{dd}}} \end{pmatrix}, \\ &+ \begin{pmatrix} 0 \\ f_{\mathrm{MOSFET,n}}(u_1, u_2, 0) - f_{\mathrm{MOSFET,p}}(u_1, u_2, u_{V_{\mathrm{dd}}}) \\ f_{\mathrm{MOSFET,p}}(u_1, u_2, u_{V_{\mathrm{dd}}}) \\ 0 \\ 0 \end{pmatrix}, \\ &+ \begin{pmatrix} 0 \\ 0 \\ 0 \\ -V_{\mathrm{in}} \\ -V_{\mathrm{DD}} \end{pmatrix}. \end{aligned} \tag{2.12}$$

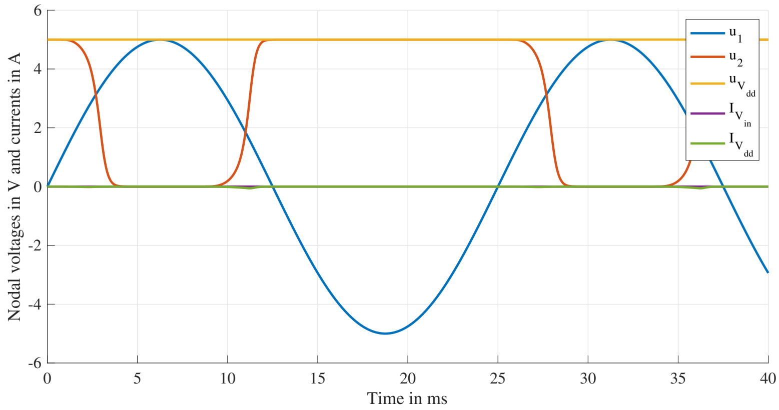

Figure 2.9: *Transient analysis of the inverter circuit for time interval* $[0, 40\text{ ms}]$.

Where $f_{\text{MOSFET},\{\text{n,p}\}}$ gives the drain to source current according to the state of the MOSFET, dependent on the voltages of terminals.

2.2 Transient Analysis

The mathematical model of the electrical circuit derived by the MNA can be used to analyse the circuit. There are several different types of circuit analysis each with their own merit and applications. If one is interested in the behaviour of the full circuit on a time interval $[0, T]$ one uses the so called transient analysis. For the transient analysis, the mathematical model of (2.6) is extended with an initial condition, $x(0) = x_0$, and transformed into an initial value problem (IVP)

$$\begin{cases} \dot{q}(x(t)) + j(x(t)) + Br(t) = 0, \\ x(0) = x_0. \end{cases} \tag{2.13}$$

The initial condition x_0 is usually known, or can be obtained by a DC analysis or integrating the system with very small time steps starting from $x_0 = 0$. If one increases the voltage supplies to their values for the initial time one obtains the steady state solution at time t_0.

As the underlying mathematical model consists of DAEs this initial condition is of the utmost importance as DAEs need consistent initial conditions and not all initial states satisfy this constraint. The solution $x(t)$

of this IVP gives the time evolution of each of the states of the dynamical system as can be seen in the Figures of the previous section. These IVPs can be solved by applying a variety of Runge Kutta or linear multistep methods. This thesis focuses solely on the transient analysis of circuits and their time integration.

3
Model Order Reduction

Abstract

Model order reduction has been a long standing method in industry for reducing numerical complexity, [1] [41]. However, most of these techniques only cater to linear time invariant models, circumventing nonlinear model order reduction. In this chapter a tried and tested nonlinear model order reduction technique is described as a frame of reference for a novel approach. The maximum entropy snapshot sampling method as developed by Kasolis, [31], is discussed and a novel parameter estimation approach is presented. The chapter concludes with a comparison between the two nonlinear model order reduction methods.

Introduction

In the previous chapter we have seen the automated approach of construction of a mathematical model describing the equations needed for integrated circuit simulation. As stated this automated approach gives rise to a sub-optimal numerical model in terms of differential-algebraic equations, which are inherently more difficult to solve. A key aspect in this modelling approach is the fact that these models may contain redundant equations. It is possible to reduce a redundant continuous DAE model directly by applying techniques such as model order reduction (MOR) and multirate integration schemes.

In this chapter we focus on removing redundancy of the continuous models by applying MOR techniques. Since there are a vast amount of different approaches to use MOR this chapter covers only a subset of these techniques. The main scope of this thesis is related to nonlinear MOR techniques and as such these will be the main topic of this Chapter. Especially techniques which use reduction by Galerkin projection.

We cover the traditional nonlinear MOR technique, the proper orthogonal decomposition (POD) method, [51], [13], [18], [19], and a novel MOR technique, the maximum entropy snapshot sampling (MESS) method, [31]. This chapter also introduces a novel method for the estimation of the reduction parameter ϵ used in the maximum entropy snapshot sampling method. Based on the approach presented by Takens in [44], the method is applied to a circuit test case. The results of this method have been published in [8].

Besides redundancy reduction our nonlinear MOR approach also incorporates hyper-reduction techniques. This is done to even further reduce the costs of the nonlinear function evaluations that are the computational bottleneck of nonlinear MOR. The hyper-reduction techniques covered in this Chapter are the Discrete Empirical Interpolation Method (DEIM), with and without the a QR extension (Q-DEIM) [18], and gappy Proper Orthogonal Decomposition, [53], [22], [55]. Both of them are used in the Galerkin projection MOR settings.

First the POD and MESS methods are defined in a general ODE setting since they are not suitable for direct application to DAEs. Then, by using the Gauss-Newton with approximated tensors (GNAT) method a significant reduction achieved for the more complex DAE setting.

3.1 Reduction by Galerkin Projection

In the scope of integrated circuit simulation, one has to regularly deal with large-scale ODE systems, they arise for instance in the semidiscretisation of parameter dependent partial differential equations (PDEs). Consider the finite difference discretisation of a nonlinear PDE with one spatial dimension, which is given by the following system of nonlinear ODEs

$$\dot{x}(t) = Ax(t) + p(x(t)), \tag{3.1}$$

with initial conditions given by $x(t_0) = x_0$. The time parameter $t \in I$ and we have that $x : I \to \mathbb{R}^m$ with $x(t) = [x_1(t), ..., x_m(t)]^\top$, componentwise nonlinear function $p : \mathbb{R}^m \to \mathbb{R}^m$. Define $p(x(t))$ as a general nonlinear function, i.e. the array function $p : \mathbb{R}^m \to \mathbb{R}^m$ is defined componentwise by

$$[p(x(t))]_i = p_i(x(t)). \tag{3.2}$$

with $p_i : \mathbb{R}^m \to \mathbb{R}$.

The dimension m can become extremely large in the case of finite difference discretisation and could thus lead to computationally intensive or even infeasible systems. Therefore, approximate models with a dimension r, where $r \ll m$, are needed to maintain computational efficiency.

To this end projection based MOR techniques are commonly used in the nonlinear case to generate a reduced order system. The original system is approximated from a subspace spanned by a reduced basis in that has the reduced dimension r. A Galerkin projection is used to obtain this system with a reduced dimension.

Consider a reduced basis $V_\mathrm{r} \in \mathbb{R}^{m \times r}$ which is a matrix with orthonormal columns. For now the reduced basis is assumed but in the next sections we elaborate on the construction of such a reduced basis. Replace the state vector $x(t)$ by a reduced state vector $\tilde{x}(t) \in \mathbb{R}^r$ with $x(t) \approx V_\mathrm{r}\tilde{x}(t)$. Then, by plugging this projection into (3.1), we obtain the reduced system

$$\dot{\tilde{x}}(t) = \underbrace{V_\mathrm{r}^\top A V_\mathrm{r}}_{\tilde{A}} \tilde{x}(t) + V_\mathrm{r}^\top p(V_\mathrm{r}\tilde{x}(t)). \tag{3.3}$$

Where we have that $\tilde{A} = V_\mathrm{r}^\top A V_\mathrm{r} \in \mathbb{R}^{r \times r}$. Now there are several choices of the reduced basis and these directly influence the accuracy of the approximation. First we consider an industry standard, the POD method.

3.1.1 Proper Orthogonal Decomposition

For the POD method it is assumed that for the system which we want to reduce it holds that the solution space is attracted to a low-dimensional manifold. Therefore, the POD method uses a set of basis functions obtained from a singular value decomposition (SVD) of a snapshot matrix. This snapshot matrix is a discrete sample of trajectories associated with a specific set of initial conditions and parameters on the time interval. When a reduced model is constructed based upon a reduced basis obtained from this sample, it will approximate the solutions of the full order model for a variety of initial conditions and parameters that are sufficiently close to those of the sample. The POD method is optimal in least-squares sense and thus the space spanned by the POD obtained basis minimises the approximation error.

Let m and n be positive integers and $m \gg n > 1$. We define a finite sequence $\{x_1, x_2, ..., x_n\}$ of system states $x_j \in \mathbb{R}^m$ at time instances $t_j \in \mathbb{R}$, with $j \in \{1, 2, ..., n\}$, of a dynamical system which we want to reduce. Then a POD basis of dimension r is a set of orthonormal vectors $\{\phi\}_{i=1}^r \in \mathbb{R}^m$ for whist the linear span optimally approximates the space $\mathcal{X} = \text{span}(x_1, ..., x_n)$. The basis set $\{\phi\}_{i=1}^r$ thus solves the minimisation problem

$$\min_{\{\phi\}_{i=1}^r} \sum_{j=1}^{n} \left\| x_j - \sum_{i=1}^{r} (x_j^\top \phi_i)\phi_i \right\|_2^2, \tag{3.4}$$

where

$$\phi_i^\top \phi_j = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \quad i, j = 1, ..., r. \tag{3.5}$$

The solution to this minimisation problem is given by the set of left singular vectors of the snapshot matrix $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m \times n}$. These vectors are obtained by performing a SVD on the snapshot matrix X.

$$X = V\Sigma W^\top, \tag{3.6}$$

where $V = [v_1, v_2, ..., v_k] \in \mathbb{R}^{m \times k}$ and $W = [w_1, w_2, ..., w_k] \in \mathbb{R}^{n \times k}$ and $\Sigma = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_k) \in \mathbb{R}^{k \times k}$, with $k = \min(m, n)$ and $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k$ are the singular values. Now the POD basis $\{v\}_{i=1}^r$ is the optimal solution to (3.4). The minimum Euclidean error induced by the approximation o the snapshots using this basis is then given by

$$\sum_{j=1}^{n} \left\| x_j - \sum_{i=1}^{r} (x_j^\top v_i)v_i \right\|_2^2 = \sum_{i=r+1}^{k} \sigma_i^2. \tag{3.7}$$

3.1.2 *Maximum Entropy Snapshot Sampling*

Besides the previously seen POD method for construction a reduced basis there are plenty of other possible methods to achieve this. One of these methods is the maximum entropy snapshot sampling method. First introduced by Kasolis and Clemens in [31], which is the key reference for this section. The MESS method uses a very different approach to achieve reduced basis construction compared to the traditional POD method.

The POD method uses the SVD which is an established data analysis factorisation, however it is also inherently linear and optimal in the least-squares sense. This makes the SVD especially robust for the analysis of samples that are obtained from linear problems. As the scope of this thesis is aimed at nonlinear model order reduction, this poses an inefficiency. When the SVD is used for the construction of a reduced basis for a nonlinear problem it removes high-frequency components that are very relevant to the evolution of the nonlinear dynamical system.

Instead of using the SVD to construct a basis, the MESS method is based on the asymptotic properties of measure-preserving transformations. A reduced basis for a nonlinear dynamical system is then obtained by a method based on a variant of the so-called Grassberger-Prociaccia correlation sum, on recurrence quantification analysis, and on an estimate of the invariant Kolmogorov-Sinai entropy. In short, the reduction is achieved by constraining the entropy estimate to be a strictly increasing function related to the time index and then applying any orthonormalization process on the reduced sample of snapshots.

The following considerations are taken from [8], page 1 in combination with [31]. Let m and n be positive integers such that $m \gg n > 1$. Define a finite sequence

$$X = (x_1, x_2, ..., x_n), \tag{3.8}$$

of numerically obtained states $x_j \in \mathbb{R}^m$ at time instances $t_j \in \mathbb{R}$, with $j \in \{1, 2, ..., n\}$, of a dynamical system governed by either ODEs or DAEs. Provided a probability distribution p of the states of the system, the second-order Rényi entropy of the sample X is

$$H_p^{(2)}(X) = -\log \sum_{j=1}^{n} p(x_j)^2 = -\log \mathbb{E}(p(X)), \tag{3.9}$$

with $\mathbb{E}(p(X))$ the expected value of the probability distribution p with respect to p itself. When n is large enough, according to the law of large numbers, the average of $p_1, p_2, ..., p_n$ almost surely converges to their expected value,

$$\frac{1}{n} \sum_{j=1}^{n} p(x_j) \to \mathbb{E}(p(X)) \quad \text{as } n \to \infty, \tag{3.10}$$

thus each $p(x_j)$ can be approximated by the sample's average sojourn time or relative frequency of occurrence. To obtain this frequency of occurrence, consider a norm $\|\cdot\|$ on $\mathbb{R}^m$. Then the notion of occurrence can be translated into a proximity condition. In particular, for each $x_j \in \mathbb{R}^m$ define the open ball that is centred at x_j and whose radius is $\epsilon > 0$,

$$B_\epsilon(x) = \{y \in \mathbb{R}^m \mid \|x - y\| < \epsilon\}, \tag{3.11}$$

and introduce the characteristic function with values

$$\chi_i(x) = \begin{cases} 1, & \text{if } x \in B_\epsilon(x_i), \\ 0, & \text{if } x \notin B_\epsilon(x_i). \end{cases} \tag{3.12}$$

Under the aforementioned considerations, the entropy of X can be estimated by

$$\hat{H}_p^{(2)}(X) = -\log\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\chi_i(x_j)\right). \tag{3.13}$$

Now for the MESS method we want a reduced sequence

$$X_\mathrm{r} = (\bar{x}_{j_1}, \bar{x}_{j_2}, \dots, \bar{x}_{j_r}), \tag{3.14}$$

with $r \le n$, that is sampled from X, by requiring that the entropy of X_r is a strictly increasing function of the index $k \in \{1, 2, \dots, r\}$, [32]. To obtain this reduced sequence, provided that the limit of the evolution of $\hat{H}_p^{(2)}$ exists, for n large enough, and measures the sensitivity of the evolution of the system itself [14, §6.6], first the notion of the recurrence matrix is needed. A symmetric matrix $R_\epsilon \in \{0, 1\}^{n \times n}$ whose entries are obtained by $\chi_i(x_i)$ is commonly referred to as the recurrence matrix associated with sample X. It is a matrix which indicates whether the pairwise distance between the columns of X are below a certain threshold ϵ. Using this notion the entropy estimation of (3.13) can be rewritten as

$$\eta_\epsilon(X) = -\log\left(\frac{1}{n^2}\|R_\epsilon\|_\mathrm{F}^2\right), \tag{3.15}$$

with the so-called information potential $v_\epsilon(X)$ defined as

$$v_\epsilon(X) = \frac{1}{n^2}\|R_\epsilon\|_\mathrm{F}^2 \neq 0, \tag{3.16}$$

also know to be the recurrence rate. Here $\|\cdot\|_F$ denotes the Frobenius norm. These quantities are the key components of the MESS method and can be defined more clearly related to the snapshot matrix X.

Definition *3.1.1* (ϵ-Frobenius entropy, ϵ-Frobenius potential, ϵ-dynamical entropy). Let m and n be positive integers such that $m \gg n > 1$, and given a snapshot matrix $X \in \mathbb{R}^{m \times n}$. Let $\epsilon > 0$ and consider the associated recurrence matrix $R_\epsilon \in \{0,1\}^{n \times n}$. Then for each index j with $j \in [1, ..., n]$ we have the following definitions.

- The ϵ-Frobenius entropy η_j of $X(:, 1:j)$ is given by

$$\eta_j = \eta_\epsilon(X(:, 1:j)) = -\log\left(\frac{1}{j^2}\left\|R_\epsilon(1:j, 1:j)\right\|_{\mathrm{F}}^2\right) \tag{3.17}$$

- The ϵ-Frobenius potential v_j of $X(:, 1:j)$ is given by

$$v_j = v_\epsilon(X(:, 1:j)) = \frac{1}{j^2}\left\|R_\epsilon(1:j, 1:j)\right\|_{\mathrm{F}}^2. \tag{3.18}$$

- The ϵ-dynamical entropy h_j of X for a given time iteration j is given by

$$h_j = \eta_{j+1} - \eta_j = -\log\left(\frac{v_{j+1}}{v_j}\right). \tag{3.19}$$

The quantity h_j can be interpreted to be the information that has been delivered by the iterate x_{j+1}, relative to the information that has been delivered by each of the previous information deliveries.

For the MESS method the information gain per iteration step h_j should be greater than zero, this means that the ϵ-Frobenius entropy is a strictly increasing function dependent on j. Through this monotonicity requirement we can state that the ϵ-Frobenius potential is thus a strictly decreasing function with respect to j. By definition, the ϵ-Frobenius potential v_j are instanteneus fractions of all pairs of snapshots that are within ϵ distance. As every snapshot x_j is within its own ball $B_\epsilon(x_j)$ and at most in every other ϵ ball around each snapshot it is clear that

$$j \leq \left\|R_\epsilon(1:j, 1:j)\right\|_{\mathrm{F}}^2 \leq j^2 \iff 1/j \leq v_j \leq 1. \tag{3.20}$$

The monotonicity of the ϵ-Frobenius entropy can be linked to the structure of the recurrencxe matrix by using Theorem 3.2 from [31]. This theorem implies that only the snapshots of X whose ϵ-Frobenius entropy is an increasing function of the snapshot index need to be sampled to generate a reduced basis for the range of X. The MESS procedure is outlined in Algorithm 1. It has been shown [31] that, depending on the recurrence properties of a system, any such basis guarantees that the Euclidean reconstruction error of each snapshot is bounded from above by ϵ, while a similar bound holds true for future snapshots, up to a specific time-horizon.

Algorithm 1: Maximum Entropy Snapshot Sampling

```
input  : Snapshot matrix $X \in \mathbb{R}^{m \times n}$, tolerance $\epsilon$.
output : Reduced basis $V \in \mathbb{R}^{m \times r}$.
$P_{i,j} \leftarrow \|x_i - x_j\|, \forall i, j \in \{1, ..., n\}$;
$P \leftarrow P / \max(P)$;
$R \leftarrow P < \epsilon$;
$idx \leftarrow [1, 0, \ldots, 0] \in \{0, 1\}^{1 \times n}$;
$k \leftarrow 1$;
$c \leftarrow 1$;
for $j- = 1, 2, \ldots, n-1$ do
    $d_j = 2 \sum_{k=1}^{j} R(j+1, k) + 1$;
    if $d - (2k+1)c < 0$ then
        $idx_{j+1} \leftarrow 1$;
        $c \leftarrow (k^2 c + d)/((k+1)^2)$;
        $k \leftarrow k + 1$;
    end
end
$[V, -] \leftarrow \mathrm{qr}(X(:, idx))$;
```

The Estimation of ϵ

The open ball parameter ϵ, which is directly responsible for the degree of reduction within the MESS framework, can be chosen arbitrarily, much like the number of selected basis vectors provided by a POD approach. For a ballpark estimate of this parameter the following optimisation approach is provided [44]. The following approach and algorithms were first presented in [5], Section 3.1. The quantity within the logarithm in the entropy estimate (3.13) is often referred to as the sample's correlation sum and can be written as

$$C_\epsilon = \frac{1}{n^2} \|R_\epsilon\|_{\mathrm{F}}^2, \tag{3.21}$$

with $R_\epsilon \in \{0, 1\}^{n \times n}$ being the recurrence matrix and $\| \cdot \|_{\mathrm{F}}^2$ being the Frobenius norm. In terms of probability theory, C_ϵ is a cumulative distribution function of ϵ, and hence, its derivative $\mathrm{d}C_\epsilon / \mathrm{d}\epsilon$ is the associated probability density function of ϵ. A commonly justified hypothesis is that the correlation sum scales as ϵ^D [46, Chapter 1], with $D \geq 0$ being the so-called correlation dimension of the manifold that is formed in $\mathbb{R}^m$ by the terms of X. Under this power law assumption, the maximum likelihood estimate [48, Chapter 8] of the correlation dimension is estimated as follows.

We find a sample $\{\epsilon_i\}$, with $\epsilon_i \in [0,1]$ for all $i \in \{1, 2, \ldots, q\}$, of a random variable E that is sampled according to C_ϵ. Then, the probability of finding a sample in $(\epsilon_i, \epsilon_i + \mathrm{d}\epsilon_i)$ in a trial is

$$\prod_{i=1}^{q} D\epsilon^{D-1} \mathrm{d}\epsilon_i. \tag{3.22}$$

To calculate the ϵ value for which this expression is maximized, we take the logarithm

$$q \cdot \ln D + (D-1) \sum_{i=1}^{q} \ln \epsilon_i, \tag{3.23}$$

and note that the maximum of this expression is attained when

$$\frac{q}{D} + \sum_{i=1}^{q} \ln \epsilon_i = 0. \tag{3.24}$$

This results in the most likely value $D_* = -1/\langle \ln E \rangle$. The value for ϵ_* is then estimated by choosing the ϵ from the sample that produces a quotient that is closest to D^*. Thus ϵ can be estimated by

$$\epsilon_* = \operatorname{argmin}(|D_* - \ln C_\epsilon / \ln \epsilon|). \tag{3.25}$$

The algorithm to calculate this most likely value for a given snapshot matrix X is described in Algorithm 2.

Algorithm 2: Epsilon estimation for a given snapshot matrix X

```
input  : Snapshot matrix $X \in \mathbb{R}^{m \times n}$.
output : Estimated tolerance value $\epsilon_*$.
$P_{i,j} \leftarrow \|x_i - x_j\|, \forall i, j \in \{1, ..., n\}$;
$P \leftarrow P / max(P)$;
$\{\epsilon^i_{\text{cdf}}\} \leftarrow LinearSpace(0, 1, n_\epsilon)$;
for $i = 1, ..., n_\epsilon$ do
    $C(\epsilon^i_{\text{cdf}}) \leftarrow \frac{1}{n_\epsilon^2} \left\| R_{\epsilon^i_{\text{cdf}}} \right\|_F^2$, Equation(3.21);
end
$\{\epsilon_i\}_{i=0}^{q} \leftarrow RandomFromCDF(q, \{\epsilon^j_{\text{cdf}}\}, \{C(\epsilon^j_{\text{cdf}})\})$, for $j = 1, ..., n_\epsilon$;
$D_* \leftarrow -\frac{1}{\langle \ln \epsilon_i \rangle}$;
$\epsilon_* \leftarrow \epsilon^i_{\text{cdf}}$, for which $i = \text{argmin}_j(|D_* - \ln C_{\epsilon^j_{\text{cdf}}} / \ln \epsilon^j_{\text{cdf}}|)$;
```

3.1.3 POD Compared With MESS

To verify the previously stated claims that the MESS method would perform better than the POD method a case study is performed. Both the MESS and POD methods are applied to a diode chain and are benchmarked for comparison.

As an instance of an integrated circuit, consider the diode chain model that is depicted in Fig. 5.1 and described by the differential-algebraic system [50]

$$\begin{aligned}
&\Phi_1 - \Phi_{\text{in}}(t) &&= 0,\\
&I(\Phi_{i-1},\Phi_i) - I(\Phi_i,\Phi_{i+1}) - \frac{\Phi_i}{R} - C\frac{\mathrm{d}\Phi_i}{\mathrm{d}t} &&= 0,\\
&I(\Phi_{m-2},\Phi_{m-1}) - \frac{\Phi_{m-1}}{R} - C\frac{\mathrm{d}\Phi_{m-1}}{\mathrm{d}t} &&= 0,\\
&i_E - I(\Phi_1,\Phi_2) &&= 0,
\end{aligned} \tag{3.26}$$

where $i \in \{2,3,\ldots,m-2\}$ with integer $m > 3$, Φ_i is the voltage at the i-th node of the circuit and is measured in V, while the time is measured in ns. The current-voltage diode characteristic function $I : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is defined by

$$I(x,y) = I_{\text{s}}\left[\mathrm{e}^{\alpha\cdot(x-y)} - 1\right], \tag{3.27}$$

where $I_{\text{S}} = 10^{-14}$ A is the saturation current and α is the inverse of the thermal voltage $\Phi_{\text{T}} = 0.0256$ V. Additional model parameters are mentioned in Fig. 5.1. Further, the excitation voltage is

$$\Phi_{\text{in}} = \begin{cases} 20, & \text{if } t \leq 10,\\ 170 - 15t, & \text{if } 10 < t \leq 11, \quad \text{(in V)}\\ 5, & \text{if } t > 11. \end{cases} \tag{3.28}$$

To simulate a transient analysis of the diode chain model depicted in Fig. 5.1, system (3.26) is integrated numerically. For large m such simulations become prohibitively expensive in terms of computational time. Here, to recover computational feasibility, reduced basis model reduction techniques are exploited.

The MESS method is applied to the nonlinear diode chain model, with $m = 40002$. The transient analysis is performed in the interval $[0,70]$ ns, using an implicit Euler scheme with time step $\Delta t = 0.1$ ns. Consistent

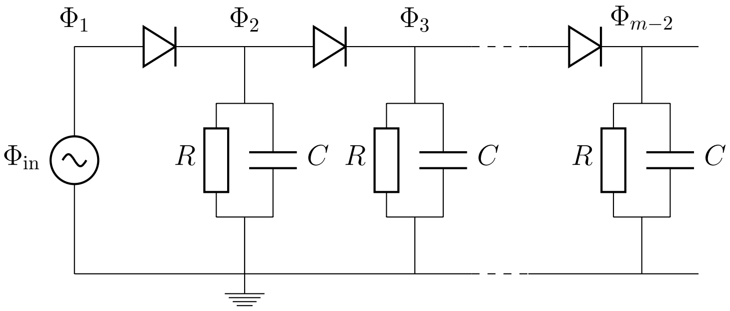

Figure 3.1: *The diode chain with $R = 10^4\ \Omega$ and $C = 10^{-12}$ F.*

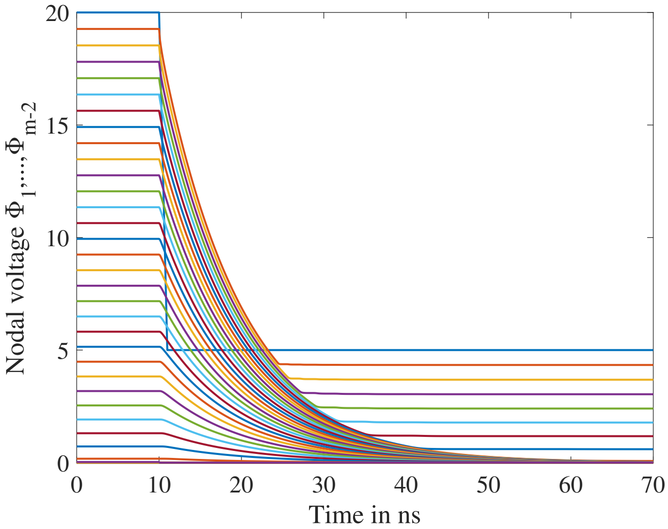

Figure 3.2: *The output of the transient analysis for all nodes.*

initial conditions are obtained through a direct current simulation using very small time steps and using a linear increasing input voltage from $\Phi_{\text{in}} = 0$ to $\Phi_{\text{in}} = 20$. The reduced bases are generated from the high-fidelity matrix $X \in \mathbb{R}^{m \times n}$, with $n = 701$, see Fig 3.2.

To benchmark the presented MESS based reduction, a comparison with the POD method is made. The number of POD modes is taken to be equal to the number of MESS-obtained basis vectors. In the Newton iterations, least squares approximations of the Jacobian matrix are employed.

Here, the estimated ϵ_* value is equal to 0.00525. However, in an attempt to maximally reduce the studied system, ϵ is manually selected close to a value that turns out to yield a numerically unstable reduced model. In Fig. 3.3, the case of the MESS reduced system for $\epsilon = 0.0325$ is depicted. There, it is shown that the solution to the MESS reduced system converges to the

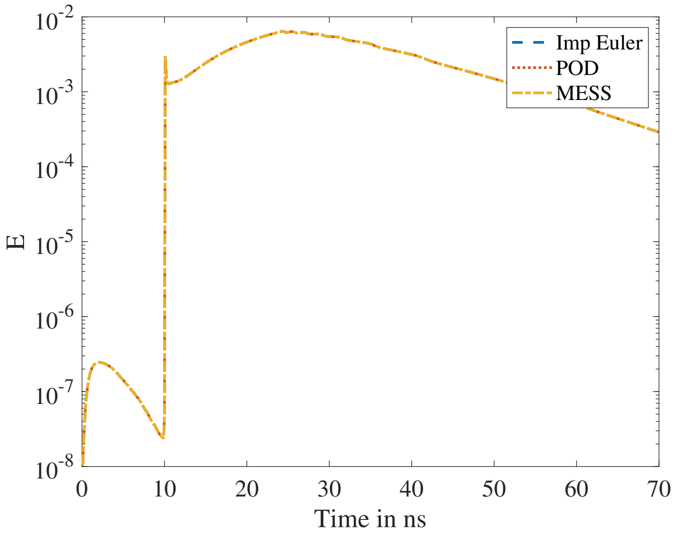

Figure 3.3: *The difference $E(t) = \|\Phi_{\mathrm{HF}} - \Phi\| / \|\Phi_{\mathrm{HF}}\|$ for the parameter value $\epsilon = 0.0325$. The subscript HF stands for "high-fidelity".*

reference solution. To illustrate that some caution is needed if ϵ is selected manually, in Fig. 3.4, a slightly higher ϵ value is chosen, when the resulting reduced model becomes unstable. In Table 3.1, the computational times that are required for generating the bases suggest that the MESS has an advantage in the offline stage. Further, for large-scale problems, the SVD becomes infeasible due to memory constraints, whereas this is not the case for MESS, since it relies on recursive evaluations.

	$\epsilon = 0.0325$		$\epsilon = 0.0425$	
	Basis generation	m	Basis generation	m
High-fidelity		40002		40002
POD	1.5400 s	31	1.3397 s	25
MESS	0.1733 s	31	0.1577 s	25

Table 3.1: *Timing MESS vs POD (time in seconds).*

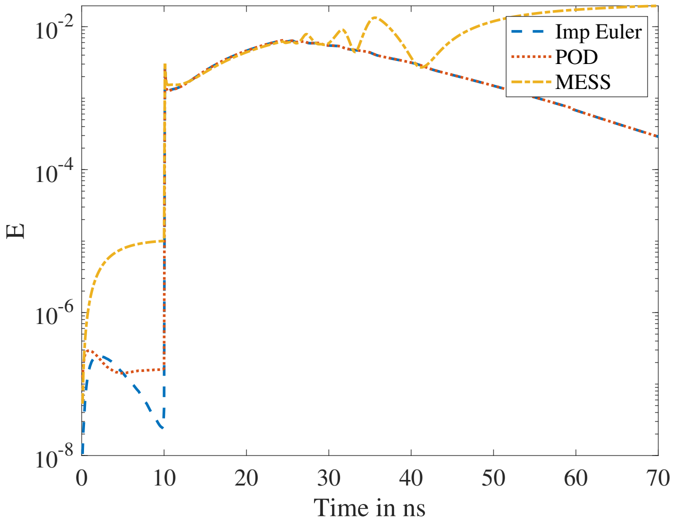

Figure 3.4: *The difference* $E(t) = \|\Phi_{\text{HF}} - \Phi\| / \|\Phi_{\text{HF}}\|$ *for the parameter value* $\epsilon = 0.0425$.

3.2 Hyper-Reduction

There are however complications when the POD is used in conjunction with a straightforward Galerkin projection, which reduces the effectiveness of the dimension reduction. This is especially the case when one considers dynamical systems that have general nonlinearities. Therefore additional efforts are required for adequate reductions.

Consider for instance the following nonlinear equation

$$\tilde{N}(\tilde{x}) := \underbrace{V_{\text{r}}^{\top}}_{r\times m} \underbrace{p(V_{\text{r}}\tilde{x}(t))}_{m\times 1}. \tag{3.29}$$

The computational complexity of the evaluation of $\tilde{N}(\tilde{x})$ depends on the full order non-reduced dimension m. To evaluate this function $2mr$ flops are needed for the matrix-vector multiplications, and further more, it requires the full evaluation of the nonlinear m dimensional vector function p. Due to this bottleneck, the complexity for solving the reduced system might be just as costly as solving the original system. Moreover, the same type of inefficiency is also present in the numerical computation of the Jacobian for each Newton iteration of the reduced-order system.

To overcome this computational bottleneck two well-established hyper-reduction methods are presented. Although they share the same type of approximation space, they are distinct in their selection criterion. The first method that we present is the discrete empirical interpolation method (DEIM), which is extended to Q-DEIM, which utilises a new selection operator compared to regular DEIM. This method enforces the reduced approximation to be interpolated from a specified number of sampling points of the original nonlinear function which are selected by a greedy sampling strategy.

The second approach is known as gappy-POD, [53] [22]. Instead of interpolation it utilises a least-squares approach to determine the reduced approximation from a number of greedy selected sampling points. However, this least-squares approach allows for significantly more flexibility in the selection of these points.

3.2.1 Discrete Empirical Interpolation Method

The DEIM approach approximates a nonlinear function by projection it onto a subspace which approximates the space generated by the nonlinear function and is spanned by a basis of dimension $g \ll m$. Let $f(\tau)$ represent the previously seen nonlinear function $p(V_{\mathrm{r}}\tilde{x}(t))$, where $\tau = t$. By using a projection approximation from $f(\tau)$ onto a subspace spanned by a basis $\{u_1, ..., u_g\} \in \mathbb{R}^m$ we have that

$$f(\tau) \approx Uc(\tau), \tag{3.30}$$

where we have the basis $U = [u_1, ..., u_g] \in \mathbb{R}^{m \times g}$ and $c(\tau)$ the corresponding coefficient vector. Determining the coefficient vector $c(\tau)$ is done by selecting g distinct rows from the overdetermined system $f(\tau) = Uc(\tau)$. This is achieved by creating a selection matrix

$$S = [e_{\mathcal{S}_i}, ..., e_{\mathcal{S}_g}] \in \mathbb{R}^{m \times g}, \tag{3.31}$$

where $e_{\mathcal{S}_i} = [0, ..., 0, 1, 0, ..., 0]^\top$ is the $\mathcal{S}_i$th column of identity matrix $I_m \in \mathbb{R}^{m \times m}$, with $i = 1, ..., g$. Then if $S^\top U$ is nonsingular, the coefficient vector $c(\tau)$ can be uniquely determined by solving

$$S^\top f(\tau) = (S^\top U)c(\tau), \tag{3.32}$$

and the projection approximation becomes

$$f(\tau) \approx Uc(\tau) = U(S^\top U)^{-1} S^\top f(\tau). \tag{3.33}$$

Evidently both a projection basis U and the interpolation indices $\{\mathcal{S}_1, ..., \mathcal{S}_g\}$ are needed for this approximation. The basis can be created by applying the POD or MESS method on the nonlinear function f, using nonlinear snapshots obtained from the original fully dimensional system. For the construction of selection matrix $\mathcal{S}$ there is a more effective method than classical DEIM, namely Q-DEIM, which computes $\mathcal{S}$ independent of a particular orthonormal basis U and enjoys a better upper bound for the condition number than classical DEIM.

As Q-DEIM outperforms classical DEIM in almost every sense, the classical DEIM algorithm is omitted in this thesis and only the Q-DEIM pseudocode according to [21] is presented in Algortihm 3.

Algorithm 3: QDEIM

```
input  : Snapshot matrix $X \in \mathbb{R}^{m \times n}$, tolerance $\epsilon$.
output : Matrix $M$.
$[U, S, V] \leftarrow SVD(X, \epsilon)$;
$[m, r] \leftarrow size(U)$;
$[Q, R, P] \leftarrow qr(U, \text{thin})$, such that $UP = QR$;
$S \leftarrow P(1 : m)$;
$M \leftarrow [eye(m); (R(:, 1 : m) \backslash R(:, m + 1 : n))']$ ;
$Pinverse(P) \leftarrow 1 : n$;
$M \leftarrow M(Pinverse, :)$;
```

3.2.2 Gappy Reconstruction

The second hyper-reduction approach is originally known as gappy-POD, [22,53]. However, as the POD part in the name only refers to the procedure of the reduced basis construction and not the vector reconstruction, the POD procedure can be interchanged with the MESS procedure. Like the gappy POD approach gappy-MESS uses a reduced basis to reconstruct gappy data. However, unlike the gappy-POD approach the basis used is now not obtained through POD but by MESS. Gappy-MESS starts by defining a mask vector n for a solution state u as

$$\begin{aligned} n_j &= 0 \text{ if } u_j \text{ is missing,} \\ n_j &= 1 \text{ if } u_j \text{ is known,} \end{aligned}$$

where j denotes the j-th element of each vector. The mask vector n is applied point-wise to a vector by $(n, u)_j = n_j u_j$. This sets all the unobserved values to 0. Then, the gappy inner product can be defined as

$(x, y)_n = ((n, x), (n, y))$, which is the inner product of the each vector masked respectively. The induced norm is then $(\|x\|_n)^2 = (x, x)_n$. Considering the reduction basis obtained by MESS $V_{\text{gap}} = \{v^i\}_{i=1}^r$, now we can construct an intermediate “repaired" full size vector $\tilde{g}$ from a reduced vector g with only r elements by

$$\tilde{g} \approx \sum_{i=1}^{r} b_i v^i, \tag{3.34}$$

where the coefficients b_i need to minimise an error E between the original and repaired vector, which is defined as

$$E = \|g - \tilde{g}\|_n^2, \tag{3.35}$$

using the gappy norm so that only the original existing data elements in g are compared. This minimisation is done by solving the linear system

$$Mb = f, \tag{3.36}$$

where

$$M_{ij} = (v^i, v^j)_n, \text{ and } f_i = (g, v^i)_n. \tag{3.37}$$

From this solution $\tilde{g}$ is constructed. Then the complete vector is reconstructed by mapping the reduced vectors elements to their original indices and filling the rest with the reconstructed values.

3.2.3 Hyper-Reduction Comparison

To illustrate the difference between the QDEIM and Gappy reconstruction approaches a numerical experiment is performed. Analogous to the case study in Subsection *3.1.3* consider the diode chain model.

Now the reduced system is integrated using backward Euler integration scheme utilising first the QDEIM hyper-reduction and another time using the Gappy reconstruction

From Figure 3.5 we see that there is indeed a difference between the two methods. However, it is also almost negligible as it is a relative difference of order 10^{-11}. As the gappy-MESS approach has a stability advantage over the QDEIM hyper-reduction approach, this will be used in subsequent numerical experiments.

Algorithm 4: Gappy reconstruction

```
input  : Snapshot matrix $X \in \mathbb{R}^{m \times n}$, tolerance $\epsilon$.
output : Matrix $M$.
$U \leftarrow MESS(X, \epsilon)$;
$[Q, R, P] \leftarrow qr(U, \text{thin})$, such that $UP = QR$;
$S \leftarrow P(1 : m)$;
for $j = 1, ..., m$ do
    if $j \in S$ then
        $n_j^{\text{mask}} = 1$;
    else
        $n_j^{\text{mask}} = 0$;
    end
end
for $i = 1, ..., m$ do
    for $j = 1, ..., m$ do
        $M_{i,j} = (U_{:,i}, U_{:,j})_{n^{\text{mask}}}$;
    end
end
```

3.3 DAE Reduction

Direct application of a Galerkin projection to reduce DAEs does not work well in practice, [33], [16]. Applying the Galerkin projection scheme directly may yield unsolvable reduced order models. The reduced Jacobian J_r may be singular, though the original Jacobian J is regular. To circumvent this problem, the Galerkin projection is applied in the numerical scheme, as opposed to the whole system of equations.

Therefore a simplified Gauß-Newton with approximated tensors (GNAT), equipped with a function-sampling-hyper-reduction scheme is used, [17], [16]. Firstly, a direct Galerkin projection may yield an unsolvable reduced system for DAEs. Secondly, the computational effort required to solve this reduced system and the full system is about the same in the nonlinear cases. This is due to the fact that the evaluation costs of the reduced system are not reduced at all because the projection basis will be a dense matrix in general.

3.3.1 Gauß-Newton with Approximated Tensors

Considering a general DAE in the form

$$\dot{\phi}(t, u) + \psi(t, u) = 0, \tag{3.38}$$

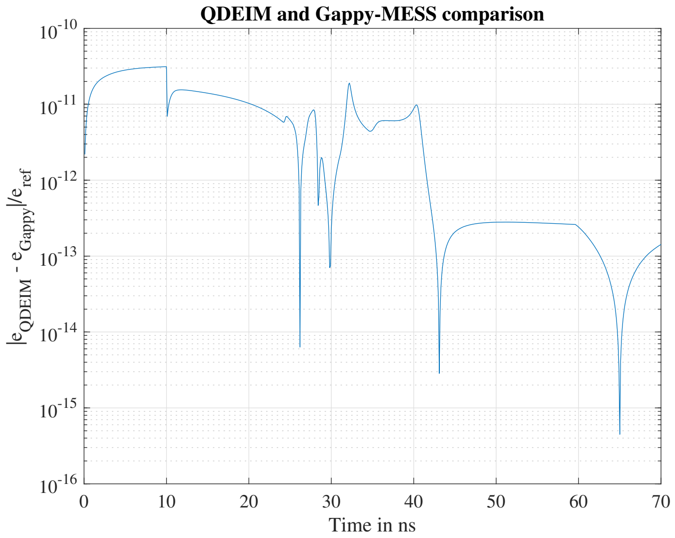

Figure 3.5: *The relative difference between the numerical approximation using QDEIM hyper-reduction and gappy-MESS hyper-reduction.*

where ϕ and ψ are functions of time t and some state vector u. In the discrete case, we assume that the numerical scheme exactly solves the following nonlinear system for each time step t_i,

$$R(u) = 0, \tag{3.39}$$

where $u \in \mathbb{R}^N$, u^0 the initial condition and the residual $R : \mathbb{R}^N \to \mathbb{R}^N$. Note that for ease of notation, the relevant time subscripts have been omitted, as this equation is solved for each individual time step. For the reduction of the dimension of Equation (3.39), a projection is used to search the approximated solution in the incremental affine trial subspace $u^0 + \mathcal{V} \subset \mathbb{R}^N$. Thus $\tilde{u}$ is given by

$$\tilde{u} = u^0 + V_u u_r, \tag{3.40}$$

where $V_u \in \mathbb{R}^{N \times n_u}$ is the n_u-dimensional projection basis for $\mathcal{V}$, and u_r denotes the reduced incremental vector of the state vector. Now deviating from the direct Galerkin projection process, Equation (3.40) is substituted into Equation (3.39). This results in an overdetermined system of N equations and n_u unknowns. Because V_u is a matrix with full column rank, it is possible to solve this system by a minimisation in least-squares sense through

$$\min_{\tilde{u}\in u^0+\mathcal{V}} \|R(\tilde{u})\|_2 . \tag{3.41}$$

This nonlinear least-squares problem is solved by the Gauß-Newton method, leading to the iterative process for $k = 1, ..., K$, solving

$$s^k = \operatorname{argmin}_{a\in\mathbb{R}^{n_u}} \left\| J^k V_u a + R^k \right\|_2 , \tag{3.42}$$

and updating the search value w_r^k with

$$w_r^{k+1} = w_r^k + s^k, \tag{3.43}$$

where K is defined through a convergence criterion, initial guess w_r^0, $R^k \equiv R(u^0 + V_u w_r^k)$ and $J^k \equiv \frac{\partial R}{\partial u}(u^0, V_u u_r^k)$. Here J^k is the full order Jacobian of the residual at each iteration step k. Since the computation of this Jacobian scales with the original full dimension of Equation (3.39) this is a computational bottleneck. This bottleneck can be circumvented by the application of the previously discussed hyper-reduction methods.

4
Numerical Integration Methods

Abstract

Multirate time integration exploits differences in the characteristic time-scales of the subsystems that comprise the whole integrated circuit model. When combining model order reduction techniques with multirate, it is important to consider which type of coupled subsystems are obtained after partitioning. This chapter provides a detailed discussion of both DAE-ODE and DAE-DAE coupled systems and the construction of reduced order multirate schemes for each case. Lastly, a numerical analysis is performed resulting in convergence results for the reduced order multirate schemes.

Introduction

From a mathematical point of view the core of every simulation software packages used in integrated circuit design is the definition of the numerical integration of the network equations. As there are a multitude of different approaches to this problem each choice of integration scheme carries its own advantages and disadvantages.

Especially in the scope of this thesis the concept of different integration schemes is something that deserves careful considerations. As reduced order multirate schemes are the central theme, the numerical mathematics behind the integration schemes are one of the two key concepts.

In this chapter we start by considering how to numerically solve systems of DAEs by introducing the concept of singular perturbation problems. Two numerical integration approaches are then presented. First, the general Runge-Kutta method and second a multistep approach by using the backward differentiation formula approach.

The next sections are dedicated to the application of these integration techniques in a multirate framework. Different types of coupling are considered and the mathematical framework is detailed from the ground up. Furthermore, the specific model order reduction approaches in each scheme are discussed.

The chapter ends with a numerical analysis related to the convergence of these newly constructed reduced order multirate schemes and results are presented. These results were first presented in [4] and [5]. The numerical verification of these results are discussed in the following chapter.

4.1 Index-1 Problems

To obtain a numerical solution for a semi-explicit DAEs one can consider the following singular perturbation problem with $\epsilon \to 0$,

$$\dot{x} = f(t, z, x), \tag{4.1a}$$

$$\epsilon \dot{z} = g(t, z, x). \tag{4.1b}$$

where x and z are the vectors of the differential and algebraic variables respectively. The functions f and g are considered to be sufficiently differentiable vector functions with dimensions according to x and z. By taking the limit $\epsilon \to 0$ a semi-explicit system of DAEs is obtained

$$\dot{x} = f(t, z, x), \tag{4.2a}$$

$$0 = g(t, z, x). \tag{4.2b}$$

which we provide with consistent initial conditions y_0 and z_0. Then under the index-1 assumption as seen in Example 1.1.1. we have that

$$\frac{\partial g}{\partial z} \text{ is nonsingular,} \tag{4.3}$$

in the neighbourhood of the solution of the semi-explicit system of DAEs. From this assumption the algebraic variable z possesses a locally unique solution by using the implicit function theorem

$$z = G(y, u). \tag{4.4}$$

For a detailed analysis of DAEs and their numerical solutions, the reader is referred to the textbooks [52] and [34].

4.1.1 Runge Kutta Methods

To solve (4.2) one approach is to apply a numerical method to the singular perturbation problem and then let $\epsilon = 0$. In this section this method is demonstrated by applying a general Runge-Kutta method. Applying this method to the singular perturbation problem the following equalities are obtained

$$X_{ni} = x_n + h\sum_{j=1}^{s} a_{ij} f(X_{nj}, Z_{nj}), \tag{4.5a}$$

$$\epsilon Z_{ni} = \epsilon z_n + h\sum_{j=1}^{s} a_{ij} g(X_{nj}, Z_{nj}), \tag{4.5b}$$

$$x_{n+1} = x_n + h\sum_{i=1}^{s} b_i f(X_{ni}, Z_{ni}), \tag{4.5c}$$

$$\epsilon z_{n+1} = \epsilon z_n + h\sum_{i=1}^{s} b_i g(X_{ni}, Z_{ni}). \tag{4.5d}$$

Now let matrix $A = (a_{ij})$ be invertible, then by defining ω_{ij} as the elements of A^{-1} z_{n+1} can be made independent of ϵ by using

$$hg(X_{ni}, Z_{ni}) = \epsilon \sum_{j=1}^{s} \omega_{ij}(Z_{nj} - z_n), \tag{4.6}$$

and inserting this into the expression for z_{n+1}. Thus obtaining

$$X_{ni} = x_n + h\sum_{j=1}^{s} a_{ij} f(X_{nj}, Z_{nj}), \tag{4.7a}$$

$$0 = g(X_{nj}, Z_{nj}), \tag{4.7b}$$

$$x_{n+1} = x_n + h\sum_{i=1}^{s} b_i f(X_{ni}, Z_{ni}), \tag{4.7c}$$

$$z_{n+1} = \left(1 - \sum_{i,j=1}^{s} b_i\omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i\omega_{ij} Z_{nj}. \tag{4.7d}$$

Although the numerical solution obtained by approach of (4.7) will not lie on the manifold $g(x, z) = 0$, this can be repaired by replacing

$$z_{n+1} = \left(1 - \sum_{i,j=1}^{s} b_i\omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i\omega_{ij} Z_{nj}. \tag{4.8}$$

by the following condition

$$0 = g(x_{n+1}, z_{n+1}). \tag{4.9}$$

We especially consider stiffly accurate methods in this thesis, this are methods for which it holds that

$$a_{si} = b_i \text{ for } i = 1, ...s. \tag{4.10}$$

Then it automatically holds that $x_{n+1} = X_{ns}$ and $z_{n+1} = Z_{ns}$.

4.1.2 Backward Differentiation Formula

For a multistep method using the Backward Differentiation Formula (BDF) again we consider a singular perturbation problem, (4.1). This method is then again applied to a DAE system by using the ϵ-embedding method. Consider a semi-explicit system with dynamical variables x and algebraic variables z, then the multistep method gives

$$\sum_{i=0}^{k} \alpha_i x_{n+i} = h \sum_{i=0}^{k} \beta_i f(x_{n+i}, z_{n+i}), \tag{4.11a}$$

$$\epsilon \sum_{i=0}^{k} \alpha_i z_{n+i} = h \sum_{i=0}^{k} \beta_i g(x_{n+i}, z_{n+i}). \tag{4.11b}$$

$$\tag{4.11c}$$

Then by putting $\epsilon = 0$ we obtain

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i f(x_{n+i}, z_{n+i}), \tag{4.12a}$$

$$0 = \sum_{i=0}^{k} \beta_i g(x_{n+i}, z_{n+i}). \tag{4.12b}$$

$$\tag{4.12c}$$

which enables us to apply this method to a semi-explicit differential algebraic system. However, we want to be able to solve implicit differential algebraic systems. Therefore, the multistep system for an implicit system of DAEs, $M\dot{x} = f(x)$, is given by

$$M \sum_{i=0}^{k} \alpha_i x_{n+i} = h \sum_{i=0}^{k} \beta_i f(x_{n+i}) \tag{4.13}$$

In general form, applying Equation (4.13) to an implicit nonlinear system of DAEs at time step t_n yields

$$f(\frac{1}{h} \sum_{i=0}^{k} \frac{\alpha_i}{\beta_i} x_{n-i}, x_n, t_n) = 0. \tag{4.14}$$

This gives that the numerical solution of the system is thus reduced to the solution of the system of nonlinear Equations (4.14). This system is solved iteratively for x_n by applying Newton's method.

4.2 Multirate Integration

The name multirate stems from the fact that these methods use multiple rates in their process of integration. In contrast to classical integration schemes, which use single rate step sizes. Multirate integration schemes integrate different parts of the complete system with different step sizes, or even with different schemes.

When a system has parts that operate on different intrinsic time-scales, such as heat compared to electricity, multirate integration can be used to exploit these different characteristics. The faster operating systems need to be integrated using a refined time-grid, whilst the slow parts do just fine on a coarser time-grid. Let the coarse time-grid be defined by step sizes $H_n = T_n - T_{n-1}$, and define the refined grid by $\{t_{n-1,q}, 1 \leq l \leq m\}$ with step sizes $h_{n,l} = t_{n,l} - t_{n,l-1}$. Here m is the so-called multirate factor and we have that the following equality holds

$$H_n = \sum_{l=1}^{m} h_{n-1,l}. \tag{4.15}$$

See Figure 4.1 for a visualisation.

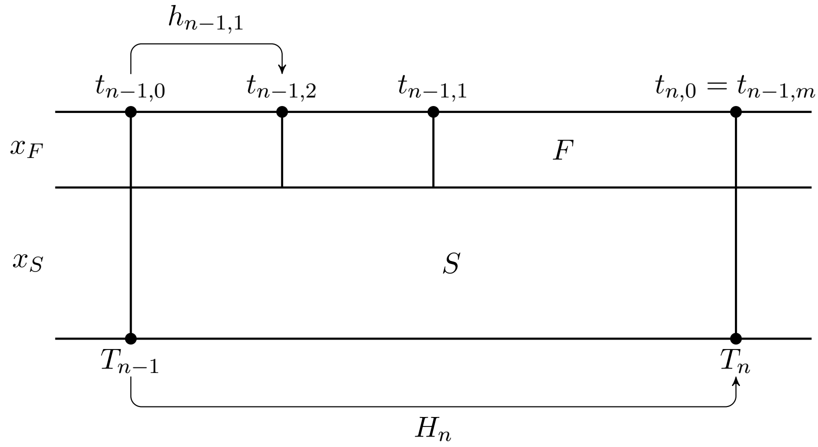

Figure 4.1: *Visual depiction of the multirate integration approach.*

4.2.1 Partitioning

From the previous section it is evident that the variables needs to be partitioned into a slow (S) and a fast (F) part, more than two parts is also possible but omitted for clarity. Consider the previously defined system

$$\frac{d}{dt}q(t,y)+j(t,y)=0, y(0)=y_0. \tag{4.16}$$

Where we consider $y \in \mathbb{R}^m$ to be the electrical state vector containing both the differential and algebraic variables $[x,z]^\top$, while the functions $q, j : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^d$ represent the charges and currents in the circuit. We have that q and j can be strongly nonlinear with respect to y and q is generally not invertible.

Let $B_\mathrm{F} \in \{0,1\}^{m_\mathrm{F} \times M}$ and $B_\mathrm{S} \in \{0,1\}^{m_\mathrm{S} \times M}$, be selection operators where $m_\mathrm{F} + m_\mathrm{S} = M$ and the following orthogonal properties: $B_\mathrm{F} B_\mathrm{F}^\top = I_\mathrm{F}$, $B_\mathrm{S} B_\mathrm{S}^\top = I_\mathrm{S}$ and $B_\mathrm{F} B_\mathrm{S}^\top = B_\mathrm{S} B_\mathrm{F}^\top = 0$, where $I_{\{\mathrm{F},\mathrm{S}\}}$ is the identity matrix with the respective subsystems dimension m_F or m_S. Then the variables and functions of each subsystem can be split into parts $y_\mathrm{F} \in \mathbb{R}^m_\mathrm{F}$, $y_\mathrm{S} \in \mathbb{R}^m_\mathrm{S}$. Then the variables and the functions can be split into the following parts:

$$\begin{aligned} y &= B_\mathrm{F}^\top y_\mathrm{F} + B_\mathrm{S}^\top y_\mathrm{S}, && \text{(4.17a)} \\ q(t,y) &= B_\mathrm{F}^\top q_\mathrm{F}(t, B_\mathrm{F} y, B_\mathrm{S} y)) + q_\mathrm{S}(t, B_\mathrm{F} y, B_\mathrm{S} y), && \text{(4.17b)} \\ j(t,y) &= B_\mathrm{F}^\top j_\mathrm{F}(t, B_\mathrm{F} y, B_\mathrm{S} y)) + j_\mathrm{S}(t, B_\mathrm{F} y, B_\mathrm{S} y). && \text{(4.17c)} \end{aligned}$$

Applying this partition to the network equations, Equation (2.6), results in the following systems

$$\begin{aligned} &\frac{d}{dt}(q_\mathrm{F}(t, B_\mathrm{F} y, B_\mathrm{S} y)) + j_\mathrm{F}(t, B_\mathrm{F} y, B_\mathrm{S} y) = 0, \qquad y_\mathrm{F}(0) = y_{\mathrm{F},0}, && \text{(4.18a)} \\ &\frac{d}{dt}(q_\mathrm{S}(t, B_\mathrm{F} y, B_\mathrm{S} y)) + j_\mathrm{S}(t, B_\mathrm{F} y, B_\mathrm{S} y), = 0, \qquad y_\mathrm{S}(0) = y_{\mathrm{S},0}. && \text{(4.18b)} \end{aligned}$$

Note however, that although the integration methods used for the integration of each sub-circuit can be A-stable, this does not hold automatically for the multirate variant. This is due to the fact that the results of the subsystems depend on the interpolated or extrapolated values of the other subsystems. Therefore, we have that the stability of multirate methods is heavily dependent on the partitioning and on the coupling between the subsystems.

4.2.2 Different Multirate Approaches

As stated in the previous sections, multirate time-integration schemes are characterised by the utilisation of different integration time-steps. The slowest subsystem is integrated by large macro-steps, with size H, whilst the faster subsystem is integrated with smaller micro-steps, with size h. Although multirate methods are independent of the integration method, here we use identical methods. Naturally two distinct approaches come to mind, slowest-first and fastest-first.

- In the slowest-first approach, first the slow subsystem is integrated for the large macro-step. Subsequently, the faster subsystem is integrated utilising the smaller micro-steps until the synchronisation point is reached.
- For fastest-first multirate, first the subsystem with the fastest behaviour is integrated. Then the subsystems with decreasingly faster dynamics are integrated.

Besides these two decoupled approaches where for the interpolated values needed for the integration of the first system constant interpolation is used, there are also coupled approaches. Since the systems used in circuit simulation are often coupled, it makes sense to use a coupled-first multirate approach, in which the first step is performed for the whole coupled system.

For both the coupled-fastest-first and coupled-slowest first approaches, the step size H is decided by the stability of the whole system for that step size. In the slowest-first approach this can be immediately checked in the first macro-step. However, for the fastest-first approach, first the fast subsystem would be integrated for m steps. Only after these m steps, the whole system would be integrated. Should this fail, it would have cost m more fast steps for the fastest-first approach opposed to the slowest-first approach.

Coupled-Slowest-First Multirate

The algebraic constraints are partitioned into the fast subsystem. This type of coupling lets us consider electrical circuits with a differential index up to 1, coupled with slower ODE systems. The total index-1 system can be integrated with the stiffly accurate implicit Euler method. To exploit the assumed different time scales, a multirate integration method is proposed. This approach is analogous to [28] but with the algebraic constraint in the fast subsystem.

Consider the following coupled system

$$\begin{aligned}
\dot{x}_{\mathrm{F}} &= f_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S}}), \quad x_{\mathrm{F}}(0) = x_{\mathrm{F},0}, && (4.19a)\\
\dot{x}_{\mathrm{S}} &= f_{\mathrm{S}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S}}), \quad x_{\mathrm{S}}(0) = x_{\mathrm{S},0}, && (4.19b)\\
0 &= g_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S}}), \quad z_{\mathrm{F}}(0) = z_{\mathrm{F},0}. && (4.19c)
\end{aligned}$$

The integration of the coupled system, Equations (4.26a-4.26c), for one macro-step $t_n \to t_{n+1} = t_n + H$ is defined as

$$x_{\mathrm{F},n+(l+1)/m} = x_{\mathrm{F},n+l/m} + h f_{\mathrm{F}}(x_{\mathrm{F},n+(l+1)/m}, z_{\mathrm{F},n+(l+1)/m}, \bar{x}_{\mathrm{S},n+(l+1)/m}), \tag{4.20a}$$

$$\begin{aligned}
x_{\mathrm{S},n+1} &= x_{\mathrm{S},n} + H f_{\mathrm{S}}(\bar{x}_{\mathrm{F},n+1}, \bar{z}_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}), && (4.20b)\\
0 &= g_{\mathrm{F}}(x_{\mathrm{F},n+(l+1)/m}, z_{\mathrm{F},n+(l+1)/m}, \bar{x}_{\mathrm{S},n+(l+1)/m}). && (4.20c)
\end{aligned}$$

With $l = 0, \ldots, m-1$ for the micro grid and the coupling variables denoted by $\bar{x}_{\mathrm{F}}$, $\bar{z}_{\mathrm{F}}$, $\bar{x}_{\mathrm{S}}$. The coupling strategy is chosen to be the coupled-slowest-first approach as this is shown to have a consistency of order 1 for the problem posed in [28]. First the whole system is solved for the macro-step.

$$\begin{aligned}
x^*_{\mathrm{F},n+1} &= x_{\mathrm{F},n} + H f_{\mathrm{F}}(x^*_{\mathrm{F},n+1}, z^*_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}), && (4.21a)\\
x_{\mathrm{S},n+1} &= x_{\mathrm{S},n} + H f_{\mathrm{S}}(x^*_{\mathrm{F},n+1}, z^*_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}), && (4.21b)\\
0 &= g_{\mathrm{F}}(x^*_{\mathrm{F},n+1}, z^*_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}). && (4.21c)
\end{aligned}$$

Where the step size H is chosen according to the slow dynamics. From this it follows that the fast solutions, $x^*_{\mathrm{F},n+1}$ and $z^*_{\mathrm{F},n+1}$, are not accurate and discarded. Following the micro-step integration the fast solutions are computed for $l = 0, ..., m-1$, using linearly interpolated values for the slow variables.

4.3 Reduced Order Multirate

In order to create a reduced order multirate scheme, we combine the techniques of all of the previous sections. This section shows the complete mathematical setup of how to combine model order reduction and multirate for a DAE-ODE coupled system and for DAE-DAE coupled systems where the latter of the two is the subsystem that is to be reduced. The considerations in these sections closely follow [4] and [5] were they were first published.

4.3.1 DAE-ODE Coupling

The dynamics of an electrical circuit are described by a system of DAEs or network equations, which are analytically equivalent to a semi-explicit system of DAEs of the form, see [26],

$$\dot{y} = \tilde{f}(y, z, t) := f(y, z, u), \quad y(t_0) = y_0, \quad u(t_0) = u_0, \tag{4.22a}$$
$$0 = \tilde{g}(y, z, t) := g(y, z, u), \quad z(t_0) = z_0, \tag{4.22b}$$

with functions $\tilde{f}\colon \mathbb{R}^n \times \mathbb{R}^m \times I \to \mathbb{R}^n$ and $\tilde{g}\colon \mathbb{R}^n \times \mathbb{R}^m \times I \to \mathbb{R}^m$ for the differential and algebraic part, respectively. The quantities $y\colon I \to \mathbb{R}^n$ and $z\colon I \to \mathbb{R}^m$ denote the differential and algebraic variables defined on the time-interval $[t_0, t_1]$, and u couples the network to the PDE model described below. Furthermore y_0, z_0 and u_0 need to be consistent initial conditions: for the index-1 system assumed in the following, this reads $0 = g(y_0, z_0, u_0)$. Secondly, other phenomena requiring to describe spatial effects can be included via PDEs, which are denoted in general form, given by $\dot{u} = \mathcal{L}(u)$. Here $\mathcal{L}$ is a differential operator acting on $u = u(t, x)$, which maps $t \in \mathbb{R}_0^+$ and $x \in D$, where $D \subset \mathbb{R}^d$ with $d \in \{1, 2, 3\}$ denotes the dimension of the spatial domain, into $\mathbb{R}^m$. This PDE system is coupled to the system of DAEs above via boundary conditions, source terms and/or parameters. After applying a suitable space discretization to the PDE system, an initial value problem of semi-explicit system of DAEs is obtained:

$$\dot{y} = f(y, z, u), \quad y(t_0) = y_0, \tag{4.23a}$$
$$0 = g(y, z, u), \quad z(t_0) = z_0, \tag{4.23b}$$
$$\dot{u} = h(y, z, u), \quad u(t_0) = u_0. \tag{4.23c}$$

Remark. Transient sources have been omitted for notational convenience.

This system is guaranteed to be of index-1 by the assumption that the Jacobian

$$\frac{\partial g(y, z, u)}{\partial z} \text{ is invertible,} \tag{4.24}$$

in a neighbourhood of the solution of the system, Equations (4.23a-4.23c), see Chapter 1. From this assumption the algebraic variable z can be solved locally by using the implicit function theorem

$$z = G(y, u). \tag{4.25}$$

Since the coupled system, Equations (4.23a-4.23c), is constructed by the combination of two different processes it can be assumed that they act within different time scales. To exploit this characteristic, the total system is partitioned into fast ($x_{\mathrm{F}} = y$ and $z_{\mathrm{F}} = z$) and slow ($x_{\mathrm{S}} = u$) subsystems,

$$\dot{x}_{\mathrm{F}} = f_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S}}), \quad x_{\mathrm{F}}(0) = x_{\mathrm{F},0}, \tag{4.26a}$$

$$\dot{x}_{\mathrm{S}} = f_{\mathrm{S}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S}}), \quad x_{\mathrm{S}}(0) = x_{\mathrm{S},0}, \tag{4.26b}$$

$$0 = g_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S}}), \quad z_{\mathrm{F}}(0) = z_{\mathrm{F},0}, \tag{4.26c}$$

with $f_{\mathrm{F}} := f, f_{\mathrm{S}} := h$ and $g_f := g$. The subscripts $\{\mathrm{F}, \mathrm{S}\}$ in the differential variables $x_{\mathrm{F}} \in \mathbb{R}^{n_{\mathrm{F}}}(n_{\mathrm{F}} := n)$, $x_{\mathrm{S}} \in \mathbb{R}^{n_{\mathrm{S}}}$ and algebraic variables $z_{\mathrm{F}} \in \mathbb{R}^{n_Z}(n_Z := m)$ indicate fast or slow dynamics, for $t \in [t_0, t_1]$ with consistent initial conditions. The algebraic constraints are assumed to be fast, as the whole dynamics of the DAE system is fast. This type of coupling lets us consider electrical circuits with a differential index up to 1, coupled to slower ODE systems.

Reducing The System

Applying a space discretization to the PDE can result in large nonlinear ODE systems. To reduce the computational effort needed to solve this system in each time step MOR techniques are used. Due to the nonlinearity of the ODE most conventional MOR techniques can be discarded as they are tailored to linear systems. Hence the chosen method for this system is a reduction by a Galerkin projection constructed by Proper Orthogonal Decomposition (POD), [13]. This is then extended by the application of the Discrete Empirical Interpolation Method (DEIM), [18], exploiting a QR selection procedure (Q-DEIM), [21]. By using a Galerkin projection a reduced model is constructed, [21], which guarantees that the reduced system is again index 1. Let $V \in \mathbb{R}^{n_{\mathrm{S}} \times r}$ be a non-square matrix with independent and orthonormal columns, with $n_{\mathrm{S}} \gg r$, then $\mathcal{V}_r$ denotes an r-dimensional subspace spanned by these columns. The full state of the slow subsystem x_{S} is then approximated by $x_{\mathrm{S}} \approx V x_{\mathrm{S},r}$ using the model reduction basis V. The reduced model in the fast unknowns x_{F}, z_{F} and slow unknowns $x_{\mathrm{S},r}$ is then defined by

$$\begin{aligned} \dot{x}_{\mathrm{F}} &= f_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, Vx_{\mathrm{S},r}), & x_{\mathrm{F}}(0) &= x_{\mathrm{F},0}, \qquad (4.27a)\\ \dot{x}_{\mathrm{S},r} &= f_{\mathrm{S},r}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S},r}), & x_{\mathrm{S},r}(0) &= x_{\mathrm{S},r,0}, \qquad (4.27b)\\ 0 &= g_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, Vx_{\mathrm{S},r}), & z_{\mathrm{F}}(0) &= z_{\mathrm{F},0} \qquad (4.27c) \end{aligned}$$

with $f_{\mathrm{S},r}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S},r}) = V^T f_{\mathrm{S}}(x_{\mathrm{F}}, z_{\mathrm{F}}, Vx_{\mathrm{S},r})$, where the full state is needed for the coupling. The reduction basis V is constructed through POD. note that there is still a problem with the Galerkin projection of the reduced term, $f_{\mathrm{S},r}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S},r})$, causing computational inefficiencies. his term has a computational complexity that depends on the non-reduced full order size n_{S}. To reduce the computational complexity Q-DEIM is applied, [18].

Consider the nonlinear function $f_{\mathrm{S}} : \mathcal{T} \to \mathbb{R}^{n_{\mathrm{S}}}$ with $\mathcal{T} \subset \mathbb{R}^{n_{\mathrm{S}}}$, and matrix $U \in \mathbb{R}^{n_{\mathrm{S}} \times m}$ of rank m. Then the DEIM approximation of f_{S} is defined by [18, Definition 3.1],

$$\hat{f}_{\mathrm{S}}(\tau) := U(\mathbb{S}U)^{-1}\mathbb{S}^T f_{\mathrm{S}}(x_{\mathrm{F}}(\tau), z_{\mathrm{F}}(\tau), x_{\mathrm{S}}(\tau)), \qquad (4.28)$$

where $\mathbb{S}$ is a selection matrix of size $n_{\mathrm{S}} \times m$ by selecting columns of identity matrix $\mathbb{I}$ of size $n_{\mathrm{S}} \times n_{\mathrm{S}}$. Then the reduced nonlinear function $f_{\mathrm{S},r}$ is approximated with the Q-DEIM, we replace f_{S} by $f_{\mathrm{S},r}$ (the coupled terms x_{F} and z_{F} are dropped here from the notation as these are not reduced)

$$f_{\mathrm{S},r}(x_{\mathrm{S},r}) \approx V^T U(\mathbb{S}U)^{-1}\mathbb{S}^T f_{\mathrm{S}}(Vx_{\mathrm{S},r}). \qquad (4.29)$$

Using the interpolation of general nonlinear functions, outlined in [18, Section 3.5], a general nonlinear function can be represented as

$$\big[F(y)\big]_i = F_i(y) = F_i(y_{j_1^i}, y_{j_2^i}, \ldots, y_{j_{n_i}^i}) = F_i\big(y(j_i)\big), \qquad (4.30)$$

where $F_i : \mathcal{Y}_i \to \mathbb{R}$, $\mathcal{Y}_i \subset \mathbb{R}^{n_i}$, with integer vector $j_i = [j_1^i, j_2^i, \ldots, j_{n_i}^i]$ denoting the indices of the components required to evaluate F_i. The numerical implementation of this allows to compute Equation (4.29) without the full evaluation of f_{S}.

Reduced Order Multirate

The integration of the coupled system, Equations (4.27a)-(4.27c), for one macro-step $t_n \to t_{n+1} = t_n + H$ is defined as

$$x_{\mathrm{F},n+(l+1)/m} = x_{\mathrm{F},n+l/m} + h f_{\mathrm{F}}(x_{\mathrm{F},n+(l+1)/m}, z_{\mathrm{F},n+(l+1)/m}, \bar{x}_{\mathrm{S},n+(l+1)/m}), \tag{4.31a}$$

$$x_{\mathrm{S},n+1} = x_{\mathrm{S},n} + H f_{\mathrm{S},r}(\bar{x}_{\mathrm{F},n+1}, \bar{z}_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}), \tag{4.31b}$$

$$0 = g_{\mathrm{F}}(x_{\mathrm{F},n+(l+1)/m}, z_{\mathrm{F},n+(l+1)/m}, \bar{x}_{\mathrm{S},n+(l+1)/m}), \tag{4.31c}$$

with $l = 0, \ldots, m-1$, counting the micro step approximations at the micro grid, where $h = H/m$. The coupling variables are denoted by $\bar{x}_{\mathrm{F}}$, $\bar{z}_{\mathrm{F}}$, $\bar{x}_s$. The coupling strategy is chosen to be the coupled-slowest-first approach as this is shown to have a consistency of order 1 for the problem posed in [28]. First the whole system is solved for the macro-step.

$$x^*_{\mathrm{F},n+1} = x_{\mathrm{F},n} + H f_{\mathrm{F}}(x^*_{\mathrm{F},n+1}, z^*_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}), \tag{4.32a}$$

$$x_{\mathrm{S},n+1} = x_{\mathrm{S},n} + H f_{\mathrm{S},r}(x^*_{\mathrm{F},n+1}, z^*_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}), \tag{4.32b}$$

$$0 = g_{\mathrm{F}}(x^*_{\mathrm{F},n+1}, z^*_{\mathrm{F},n+1}, x_{\mathrm{S},n+1}). \tag{4.32c}$$

The step size H is chosen according to the slow dynamics, whilst the full system remains solvable. From this it follows that the fast solutions, $x^*_{\mathrm{F},n+1}$ and $z^*_{\mathrm{F},n+1}$, are not accurate enough and can be discarded, as they will be computed in the last micro step. In a second step, the fast solutions are computed for the micro steps $l = 0, \ldots, m-1$, using for the interpolated values $\bar{x}_{\mathrm{S},n+(l+1)/m}$ linear interpolation based on the available information $x_{\mathrm{S},n}$ and $x_{\mathrm{S},n+1}$ in time values for the slow variables.

4.3.2 DAE-DAE Coupling

Consider the following coupled system of two semi explicit system of DAEs, where the subscripts $\{F, S\}$ indicate a fast or slow time-scale, respectively, and independent transient sources have been omitted for notational convenience:

$$\frac{d}{dt} y_{\mathrm{F}} = f_{\mathrm{F}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad y_{\mathrm{F}}(t_0) = y_{F_0}, \tag{4.33a}$$

$$0 = g_{\mathrm{F}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad z_{\mathrm{F}}(t_0) = z_{F_0}, \tag{4.33b}$$

$$\frac{d}{dt} y_{\mathrm{S}} = f_{\mathrm{S}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad y_{\mathrm{S}}(t_0) = y_{S_0}, \tag{4.33c}$$

$$0 = g_{\mathrm{S}}(t, y_{\mathrm{F}}, z_{\mathrm{F}}, y_{\mathrm{S}}, z_{\mathrm{S}}), \quad z_{\mathrm{S}}(t_0) = z_{S_0}, \tag{4.33d}$$

with the functions $f_A : \mathbb{R}\times\mathbb{R}^a\times\mathbb{R}^b\times\mathbb{R}^c\times\mathbb{R}^d \to \mathbb{R}^a$, with $A \in \{\mathrm{F},\mathrm{S}\}$, where $\{a,b,c,d\} \in \mathbb{N}$ are the respective dimensions, and equivalent definitions for g_A. Consistent initial conditions are assumed, which means that Equations (4.33b) and (4.33d) are satisfied at initial time t_0. The quantities $y_{\{F,S\}} : I \to \mathbb{R}^{\{a,b\}}$ and $z_{\{F,S\}} : I \to \mathbb{R}^{\{c,d\}}$ denote the differential and algebraic variables defined on the time interval $[t_0,t_1] = I$. Both subsystems and the joint system are guaranteed to be index-1 by the assumption that the Jacobians

$$\frac{\partial g_\mathrm{F}}{\partial z_\mathrm{F}},\ \frac{\partial g_\mathrm{S}}{\partial z_\mathrm{S}} \text{ and } \begin{pmatrix} \frac{\partial g_\mathrm{F}}{\partial z_\mathrm{F}} & \frac{\partial g_\mathrm{F}}{\partial z_\mathrm{S}} \\ \frac{\partial g_\mathrm{S}}{\partial z_\mathrm{F}} & \frac{\partial g_\mathrm{S}}{\partial z_\mathrm{S}} \end{pmatrix} \text{ are invertible} \tag{4.34}$$

in the neighbourhood of the solution of the system. From this assumption the algebraic variables $z_{\{\mathrm{F},\mathrm{S}\}}$ can be solved locally by using the implicit function theorem

$$z_F = G_{t,\mathrm{F}}(y_\mathrm{F}, z_S, y_\mathrm{S}), \tag{4.35a}$$
$$z_S = G_{t,\mathrm{S}}(y_\mathrm{F}, z_F, y_\mathrm{S}), \tag{4.35b}$$

where the second z subscript is the opposite of the first z subscript. The partition of the system into subsystems can originate from different physical systems, such as temperature diffusion and electric currents. However, differences in time scale can also be identified by different orders of time derivatives. Here the partition is considered to be fixed during the time integration.

Reducing The System

To incorporate the previous two sections into the partitioned system of DAEs, Equations (4.33a)-(4.33d), we first rewrite Equations (4.33c)-(4.33d) in a more general DAE form, to have the slow subsystem encapsulated into one equation.

$$\frac{d}{dt} y_\mathrm{F} = f_\mathrm{F}(t, y_\mathrm{F}, z_\mathrm{F}, u_\mathrm{S}), \quad y_\mathrm{F}(t_0) = y_{\mathrm{F}_0}, \tag{4.36a}$$
$$0 = g_\mathrm{F}(t, y_\mathrm{F}, z_\mathrm{F}, u_\mathrm{S}), \quad z_\mathrm{F}(t_0) = z_{\mathrm{F}_0}, \tag{4.36b}$$
$$\frac{d}{dt}\phi(u_\mathrm{S}) = F_\mathrm{S}(t, y_\mathrm{F}, z_\mathrm{F}, u_\mathrm{S}), \quad u_\mathrm{S}(t_0) = (y_{\mathrm{S}_0}, z_{\mathrm{S}_0})^\top, \tag{4.36c}$$

where $F_\mathrm{S} : \mathbb{R}\times\mathbb{R}^a\times\mathbb{R}^b\times\mathbb{R}^{m_\mathrm{S}} \to \mathbb{R}^{m_\mathrm{S}}$ and $u_\mathrm{S} = (y_\mathrm{S}, z_\mathrm{S})^\top$. Into these equations we incorporate the back projected reduced state $\tilde{u}_{\mathrm{S}_\mathrm{r}} = u_\mathrm{S}^0 + V_u u_{\mathrm{S}_\mathrm{r}}$

$$\frac{d}{dt}y_{\mathrm{F_r}} = f_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{4.37a}$$

$$0 = g_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{4.37b}$$

$$\frac{d}{dt}\phi(\tilde{u}_{\mathrm{S_r}}) = F_{\mathrm{S}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}). \tag{4.37c}$$

and then, with the Gappy MESS complexity reduction incorporated we obtain

$$\frac{d}{dt}y_{\mathrm{F_r}} = f_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{4.38a}$$

$$0 = g_{\mathrm{F}}(t, y_{\mathrm{F_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}), \tag{4.38b}$$

$$\frac{d}{dt}\phi(\tilde{u}_{\mathrm{S_r}}) = F_{\mathrm{S_r}}(t, y_{\mathrm{S_r}}, z_{\mathrm{F_r}}, \tilde{u}_{\mathrm{S_r}}). \tag{4.38c}$$

Where $F_{\mathrm{S_r}}$ denotes the function F_{S} solved by the reduced least squares approach. Note that the subscript r denotes a reduction, and not the reduction factor.

Reduced Order Multirate

The overall index-1 system, Equations (4.38a)-(4.38c), can be integrated with the stiffly accurate implicit Euler scheme, which automatically assures that also for $t > 0$ the algebraic constraints will remain consistent. To exploit the assumed different time scales, a multirate integration scheme is proposed. This approach is analogous to [4], but with a slow subsystem consisting of DAEs. Here $h = H/m$ and the coupling variables are denoted by $\bar{y}_{\mathrm{F}}$, $\bar{z}_{\mathrm{F}}$, $\bar{\tilde{u}}_{\mathrm{S_r}}$. The coupling strategy is chosen to be the coupled-slowest-first approach as this is shown to have a consistency of order 1 for the problem posed in [28]: First the whole system is solved for the macro-step, $t_n \to t_{n+1} = t_n + H$

$$y^*_{\mathrm{F_r},n+1} = y_{\mathrm{F_r}n} + Hf_{\mathrm{F}}(y^*_{\mathrm{F_r},n+1}, z^*_{\mathrm{F_r},n+1}, \tilde{u}_{\mathrm{S_r},n+1}), \tag{4.39a}$$

$$0 = g_{\mathrm{F}}(y^*_{\mathrm{F_r},n+1}, z^*_{\mathrm{F_r},n+1}, \tilde{u}_{\mathrm{S_r},n+1}), \tag{4.39b}$$

$$\phi(\tilde{u}_{\mathrm{S_r},n+1}) = \phi(\tilde{u}_{\mathrm{S_r},n}) + HF_{\mathrm{S_r}}(y^*_{\mathrm{F_r},n+1}, z^*_{\mathrm{F_r},n+1}, \tilde{u}_{\mathrm{S_r},n+1}). \tag{4.39c}$$

The step size H is chosen according to the slow dynamics, whilst the full system remains solvable. From this it follows that the fast solutions, $y^*_{\mathrm{F_r},n+1}$ and $z^*_{\mathrm{F_r},n+1}$, are not accurate enough and can be discarded, as they will be computed in the last micro step. In a second step, the fast solutions are computed for the micro steps $l = 0, \dots, m-1$,

$$\begin{aligned} y_{\mathrm{F_r},n+(l+1)/m} &= y_{\mathrm{F_r},n+l/m} \\ &\quad + h f_{\mathrm{F}}(y_{\mathrm{F_r},n+(l+1)/m}, z_{\mathrm{F_r},n+(l+1)/m}, \bar{\tilde{u}}_{\mathrm{S_r},n+(l+1)/m}), \end{aligned} \tag{4.40a}$$

$$0 = g_{\mathrm{F}}(y_{\mathrm{F_r},n+(l+1)/m}, z_{\mathrm{F_r},n+(l+1)/m}, \bar{\tilde{u}}_{\mathrm{S_r},n+(l+1)/m}), \tag{4.40b}$$

$$\begin{aligned} \phi(\tilde{u}_{\mathrm{S_r},n+(l+1)/m}) &= \phi(\tilde{u}_{\mathrm{S_r},n+l/m}) \\ &\quad + h F_{\mathrm{S_r}}(\bar{y}_{\mathrm{F_r},n+(l+1)/m}, \bar{z}_{\mathrm{F_r},n+(l+1)/m}, \tilde{u}_{\mathrm{S_r},n+(l+1)/m}). \end{aligned} \tag{4.40c}$$

For stability reasons, the interpolated values $\bar{\tilde{u}}_{\mathrm{S_r},n+(l+1)/m}$ are obtained by constant interpolation based on $\tilde{u}_{\mathrm{S_r},n+1}$, then the coupled-slowest-first Euler approach is unconditionally A-stable.

4.4 Numerical Analysis

In this section the convergence of the different reduced order multirate approaches are studied and proofs are given. In the subsequent chapter, these analytical results are verified using numerical experiments. This analysis was first presented in [4].

4.4.1 DAE-ODE Coupling

Let $y_{\{\text{full,red}\}} : [t_0, t_{\text{end}}] \to \mathbb{R}^k$ denote all the variables of the full system, Equations (4.26a)–(4.26c), and of the reduced system, Equations (4.27a)–(4.27c), respectively. Let the exact values be denoted with $y(t)$ for t in the interval $[t_0, t_n]$. Furthermore, let y_n denote the numerical approximation after n macro-steps. Let $|| \cdot ||$ denote the 2-norm in Euclidean space. To numerically integrate the reduced order system, Equations (4.27a)-(4.27c), with a reduced order multirate scheme we are interested in the difference between the numerical approximation computed by the reduced multirate scheme and the exact solution of the full order model.

$$\left\| y_{\text{full}}(t_{n+1}) - \tilde{V} y_{\text{red},n+1} \right\|, \tag{4.41}$$

where $\tilde{V}$ is the orthonormal projection matrix for the full system, defined as

$$\begin{pmatrix} I & 0 & 0 \\ 0 & V & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{4.42}$$

Here V is the projection of the slow part, the other quantities are not reduced. We split the approximation errors into two parts, the reduction error and the numerical error.

$$\begin{aligned} \left\| y_{\text{full}}(t_{n+1}) - \tilde{V} y_{\text{red},n+1} \right\| \leq & \left\| y_{\text{full}}(t_{n+1}) - \tilde{V} y_{\text{red}}(t_{n+1}) \right\| && \text{(4.43a)} \\ & + \left\| \tilde{V} y_{\text{red}}(t_{n+1}) - \tilde{V} y_{\text{red},n+1} \right\|. && \text{(4.43b)} \end{aligned}$$

By using [19, Theorem 4.2], we have that

$$\int_0^{t_{n+1}} \left\| y_{\text{full}}(t) - \tilde{V} y_{\text{red}}(t) \right\|^2 dt \leq C(t_{n+1})(\mathcal{E}_y + \mathcal{E}_F), \tag{4.44a}$$

where $C(t_{n+1})$ is the magnification factor, and $\mathcal{E}_y$ and $\mathcal{E}_F$ are the 2-norm errors from approximating the solution $y_{\text{full}}(t)$ with POD and the nonlinear function with Q-DEIM. Now we have that $\left\| y_{\text{full}}(t) - \tilde{V} y_{\text{red}}(t) \right\|^2$ is a Riemann integrable function and thus it holds that

$$\left\| y_{\text{full}}(t_{n+1}) - \tilde{V} y_{\text{red}}(t_{n+1}) \right\|^2 \leq \tilde{C}(t_{n+1})(\mathcal{E}_y + \mathcal{E}_F), \tag{4.45a}$$

where $\tilde{C}(t_{n+1}) = \frac{1}{t_n} C(t_{n+1})$. This follows from the upper and lower bound inequality of Riemann integrals, and therefore we have obtained a bound for the reduction error. For the bound of the numerical error from Equation (4.43b) we first observe that obtaining the fast algebraic approximates via the implicit function theorem and via the implicit Euler method are the same, as the implicit Euler method is stiffly accurate and thus automatically consistent. Therefore the numerical approximation error in the algebraic variable only depends on y_F and y_{S_r}:

$$\begin{aligned} z_{F,n+\frac{l}{m}} - z_F(t_{n+\frac{l}{m}}) = & G(y_{F,n+\frac{l}{m}}, y_{S_r,n+\frac{l}{m}}) && \text{(4.46a)} \\ & - G(y_F(t_{n+\frac{l}{m}}), y_{S_r}(t_{n+\frac{l}{m}})). && \text{(4.46b)} \end{aligned}$$

Remark. As in the case of classical numerical integration schemes for index-1 DAE systems, these nonlinear equations are not solved exactly, but with high accuracy via Newton's method. However, this additional error is much lower than the leading numerical discretization error of the integration scheme.

This is equal to the case c in, [28, Lemma 2]. Therefore, [28, Theorem 2], can be applied which gives that the numerical error

$$\left\|\tilde{V}y_{\mathrm{red}}(t_{n+1})-\tilde{V}y_{\mathrm{red},n+1}\right\|\approx\mathcal{O}(H), \tag{4.47}$$

where H is the macro step-size of the mrIRK-DAE1 scheme. From this it follows that the reduced order multirate global approximation error is given by

$$\left\|y_{\mathrm{full}}(t_{n+1})-\tilde{V}y_{\mathrm{red},n+1}\right\|\approx\sqrt{\tilde{C}(t_{n+1})(\mathcal{E}_y+\mathcal{E}_F)}+\mathcal{O}(H). \tag{4.48}$$

Thus if the reduction error is chosen such that

$$\sqrt{\tilde{C}(t_{n+1})(\mathcal{E}_y+\mathcal{E}_F)}=\mathcal{O}(H) \tag{4.49}$$

the method converges with order 1 in H to a solution which is within $\mathcal{O}(H)$ distance of the full solution. This small choice is always possible as the bound can be made arbitrarily small. One caveat could be that if the system is unsuitable for reduction, the reduction error will only be small for a small dimension reduction. Then this will not lead to improved computational times as the system will not be sufficiently reduced and the large projections can increase the computational effort. However, our thermal-electrical benchmark perfectly allows for reduction.

4.4.2 DAE-DAE Coupling

In this section, the local truncation error induced by the reduced order multirate scheme from one macro-step $t_n\to t_{n+1}=t_n+H$ is estimated. This analysis was first presented in [5].

We define the error in each variable class as

$$\left\|y_{\mathrm{F}}(t_{n+1})-y_{\mathrm{F_r},n+1}\right\|, \tag{4.50a}$$
$$\left\|z_{\mathrm{F}}(t_{n+1})-z_{\mathrm{F_r},n+1}\right\|, \tag{4.50b}$$
$$\left\|u_{\mathrm{S}}(t_{n+1})-\tilde{u}_{\mathrm{S_r},n+1}\right\|. \tag{4.50c}$$

Here $\|\cdot\|$ is the 2-norm in Euclidean space. The analytical solutions of a state variable is notated by with a parenthesised time argument whilst the numerical approximation is noted with subscript, e.g. $u_{\mathrm{S}}(t_n)$ and $u_{\mathrm{S},n}$. To analyse the local truncation error, it is split into two parts, the numerical approximation error and the discrete reduction error.

$$\|y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F_r},n+1}\| \leq \|y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F},n+1}\| + \|y_{\mathrm{F},n+1} - y_{\mathrm{F_r},n+1}\|\,, \tag{4.51a}$$

$$\|z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F_r},n+1}\| \leq \|z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F},n+1}\| + \|z_{\mathrm{F},n+1} - z_{\mathrm{F_r},n+1}\|\,, \tag{4.51b}$$

$$\|u_{\mathrm{S}}(t_{n+1}) - \tilde{u}_{\mathrm{S_r},n+1}\| \leq \|u_{\mathrm{S}}(t_{n+1}) - u_{\mathrm{S},n+1}\| + \|u_{\mathrm{S},n+1} - \tilde{u}_{\mathrm{S_r},n+1}\|\,. \tag{4.51c}$$

The first error term on the right-hand side of the inequality can be identified to be the error induced by a non-reduced order implicit multirate scheme. This error E_{MR} is $\mathcal{O}(H^2)$, following [28, Theorem 2]. The second error on the right-hand side, the error induced by the GNAT and hyper-reduction method, can be analysed in the following manner. For the slow subsystem, we only have to consider the macro-step and thus it holds that

$$\|u_{\mathrm{S},n+1} - \tilde{u}_{\mathrm{S_r},n+1}\| \leq E_{\mathrm{macro}} \tag{4.52}$$

where E_{macro} is the error bound of [17, Proposition 4.1]. This, due to the fact that the macro-step of the reduced order multirate scheme is an implicit Euler step using GNAT and hyper-reduction, identical to the prerequisites of the proposition, using the fact that the algebraic variable values are directly obtained through the implicit function theorem. The only error that now needs to be bounded such that the whole reduced order multirate induced error is bounded, is the micro-step error. For each micro-step, again using the fact that the algebraic values are solved locally by the implicit function theorem, we only have to analyse the error in the fast dynamical variables y_{F}

$$R^n(y_{\mathrm{F},n+(l+1)/m}) = y_{\mathrm{F},n+(l+1)/m} - y_{\mathrm{F},n+l/m} - hf_{\mathrm{F}}(y_{\mathrm{full},n+(l+1)/m}), \tag{4.53}$$

and

$$\tilde{R}^n(\tilde{y}_{\mathrm{F_r},n+(l+1)/m}) = y_{\mathrm{F_r},n+(l+1)/m} - y_{\mathrm{F_r},n+l/m} - hf_{\mathrm{F}}(\tilde{y}_{\mathrm{full},n+(l+1)/m}). \tag{4.54}$$

Here $y_{\mathrm{full},n}$ is a shorthand notation for the full state $(y_{\mathrm{F},n}, z_{\mathrm{F},n}, u_{\mathrm{S},n})$ for f.

Using $\zeta : (x) \to x - h f_{\mathrm{F}}(x)$ and the inverse Lipschitz constant

$$\mathcal{L}_n \equiv \sup_{x \neq y} \frac{\|x - y\|}{\|\zeta(x) - \zeta(y)\|}. \tag{4.55}$$

we obtain a bound for the local micro-step approximation error

$$\left\| y_{\mathrm{F_r},n+(l+1)/m} - \tilde{y}_{\mathrm{F_r},n+(l+1)/m} \right\| \leq \mathcal{L}_n (\epsilon_{\mathrm{Newton}} + \tag{4.56a}$$

$$\left\| \tilde{R}^n (\tilde{y}_{\mathrm{F_r},n+(l+1)/m}) \right\| + \tag{4.56b}$$

$$\left\| y_{\mathrm{F},n+l/m} - \tilde{y}_{\mathrm{F_r},n+l/m} \right\|). \tag{4.56c}$$

This then results in

$$\| y_{\mathrm{F_r},n+1} - \tilde{y}_{\mathrm{F_r},n+1} \| \leq \sum_{k=0}^{m-1} a^k (\epsilon_{\mathrm{Newton}} + \tag{4.57a}$$

$$\left\| \tilde{R}^k (\tilde{y}_{\mathrm{F_r},n+(k+1)/m}) \right\| + \tag{4.57b}$$

$$\left\| y_{\mathrm{F},n+k/m} - \tilde{y}_{\mathrm{F_r},n+k/m}) \right\|), \tag{4.57c}$$

where $a = \mathcal{L} \equiv \sup_{k \in \{0,\ldots,m-1\}} \mathcal{L}_k$. For h small enough, it follows that

$$\| y_{\mathrm{F},n+1} - y_{\mathrm{F_r},n+1} \| \leq E_{\mathrm{micro}}. \tag{4.58}$$

Thus it has been shown that the cumulative micro-step error is bounded as well. Now we assume that the reduction induced error bound $E_{\{\mathrm{macro,micro}\}} \ll E_{\mathrm{MR}}$, which should always be the case as model order reduction should only be used if the reduced model is able to accurately capture the full order dynamics. So for a macro-step, the following holds,

$$\| y_{\mathrm{F}}(t_{n+1}) - y_{\mathrm{F_r},n+1} \| \leq E_{\mathrm{MR}} + E_{\mathrm{micro}} \approx \mathcal{O}(H^2), \tag{4.59a}$$

$$\| z_{\mathrm{F}}(t_{n+1}) - z_{\mathrm{F_r},n+1} \| \leq E_{\mathrm{MR}} + E_{\mathrm{micro}} \approx \mathcal{O}(H^2), \tag{4.59b}$$

$$\| u_{\mathrm{S}}(t_{n+1}) - \tilde{u}_{\mathrm{S_r},n+1} \| \leq E_{\mathrm{MR}} + E_{\mathrm{macro}} \approx \mathcal{O}(H^2). \tag{4.59c}$$

Then, for the error propagation over several macro-steps we obtain by using [28, Theorem 2] that the global error is $\mathcal{O}(H)$.

5

Applications

Abstract

Previous chapters have introduced the novel reduced order multirate schemes. This chapter is presents several numerical experiments that verify the theoretical results. Both academic test cases have been used, especially in early stages, and then an industrial test-case as provide by STMicroelectronics has been investigated. Results show the positive impact that the reduced order multirate schemes have on the computation time.

Besides obtaining only theoretical results, numerical results are just as important. Due to the large size of the systems that need to be simulated, theoretical results might not be achievable or feasible in reality. Therefore, verification of the reduced order multirate method, utilising the discussed methods, is done through numerical experiments. For the first two experiments we consider two DAE-DAE type academic circuit consisting of resistors, capacitors and diodes. Then real world industrial circuits are considered for both DAE-ODE and DAE-DAE coupling, with the first experiment originating from [10] and the second experiment as provided by STMicroelectronics. Although these test examples come from the industry partner, some simplification assumptions have been made about the parameters of the underlying network components.

5.1 Academic Experiments

A transient analysis is performed for an academic test case and the convergence of the error is investigated. Furthermore, computational times are compared to verify the efficiency of the reduced order multirate scheme.

5.1.1 The Diode Chain Model

Following the results presented in [5], the underlying test case consists of a large diode chain model, [11], that is very suitable for reduction due to internal redundancy, coupled to a two dimensional oscillating DAE system. This second subsystem is dependent on the large diode chain through the voltage at Φ_2.

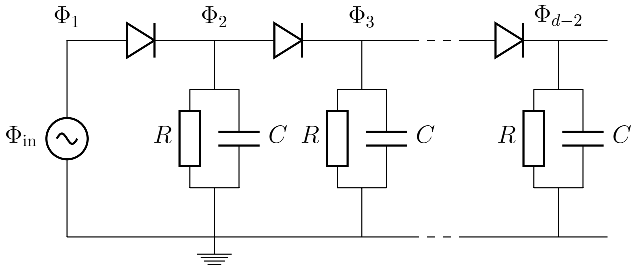

Figure 5.1: *The diode chain*

The diode chain model is described by the following differential-algebraic equations

$$\begin{aligned}
\Phi_1 - \Phi_{\text{in}} &= 0,\\
I(\Phi_1, \Phi_2) - I(\Phi_2, \Phi_3) - C\frac{\mathrm{d}\Phi_2}{\mathrm{d}t} - \frac{1}{R}\Phi_2 &= 0,\\
I(\Phi_{i-1}, \Phi_i) - I(\Phi_i, \Phi_{i+1}) - C\frac{\mathrm{d}\Phi_i}{\mathrm{d}t} - \frac{1}{R}\Phi_i &= 0,\\
I(\Phi_i, \Phi_{i+1}) - C\frac{\mathrm{d}\Phi_{i+1}}{\mathrm{d}t} - \frac{1}{R}\Phi_{i+1} &= 0,\\
i_{\text{E}} - I(\Phi_1, \Phi_2) &= 0.
\end{aligned}$$

$$I(x,y) = I_{\text{s}}\left[\mathrm{e}^{(x-y)/0.0256} - 1\right], \quad \Phi_{\text{in}} = 8\sin(7\cdot 10^8 \cdot \frac{t}{2\pi}).$$

Where $C = 10^{-11}$ and R is a coupled resistance term. The saturation current I_S is set to $10^{-12}\tilde{A}$. Through this term, the variables of the slow subsystems depend only weakly on the variables variables of the fast subsystem, this coupling is given by $R = R_0 + y_1 \cdot 10^2$, where y_1 is defined by a fast oscillating academic DAE system that is dependent on the nodal voltage Φ_2.

$$\begin{aligned} 0 &= C_\mathrm{A}\frac{\mathrm{d}y_1}{\mathrm{d}t} - y_2 - \frac{1}{R}\Phi_2 \\ 0 &= y_2 - \sin(7 \cdot 10^8 t). \end{aligned}$$

Below, in Figure 5.2 and 5.3, the results of a transient analysis for a time interval from 0 to 37.5 ns is shown. The diode chain parameters are $R_0 = 10000\ \Omega$.

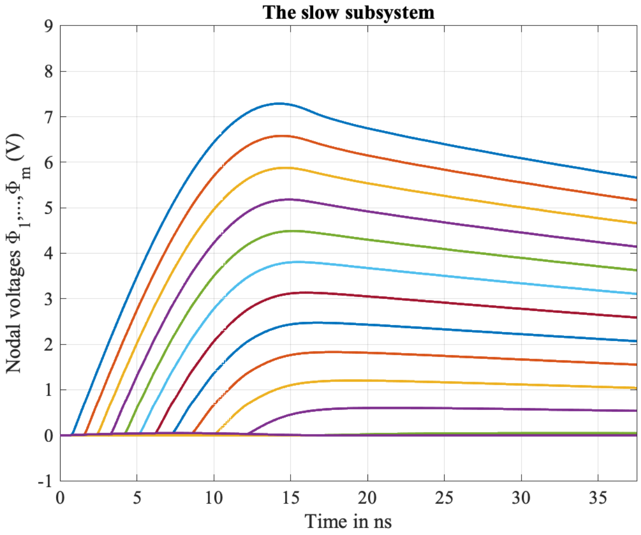

Figure 5.2: *The evolution of the slow subsystems from the transient analysis. From top to bottom we have $\Phi_1, ..., \Phi_m$.*

For the fast subsystem the resistance is taken to be equal to that of the diode chains resistors, and the capacitance is set to $C_\mathrm{A} = 10^{-10}$. The dimension parameter of the diode chain is set to $d = 1000$. The snapshot matrix is provided by a high accuracy integration of the full system and snapshots are taken with $\Delta t_\mathrm{HF} = 0.0375$ ns. By applying the ϵ estimation procedure we obtain $\epsilon_* = 0.1928$ and this results in the reduced system size $r = 14$. The same reduced basis size is used for the gappy reconstruction. The multirate integration factor $m = 20$. The initial conditions for the slow subsystem are 0 for each node, and for the fast subsystem $(-1, 0)^\top$.

Regarding the convergence of the ROMR integration scheme, Figure 5.8

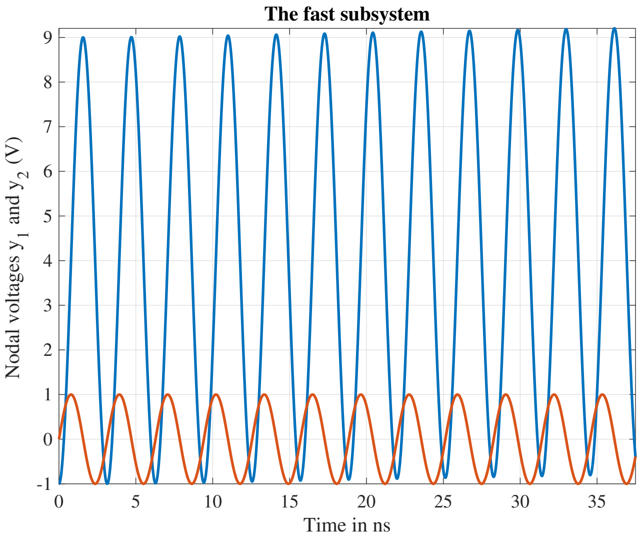

Figure 5.3: *The evolution of the fast subsystems from the transient analysis. In blue we have y_1 and in red y_2.*

illustrates the order 1 convergence rate. We see that the ROMR accuracy is nearly identical to that of the full order solutions. Furthermore, in Figure 5.5 it shows that this accuracy is achieved with a significant reduction in computational time. The computational effort is almost a order of magnitude lower for the reduced schemes, while the precision is maintained. The positive effects of model order reduction, multirate time integration and the combination of both is evident.

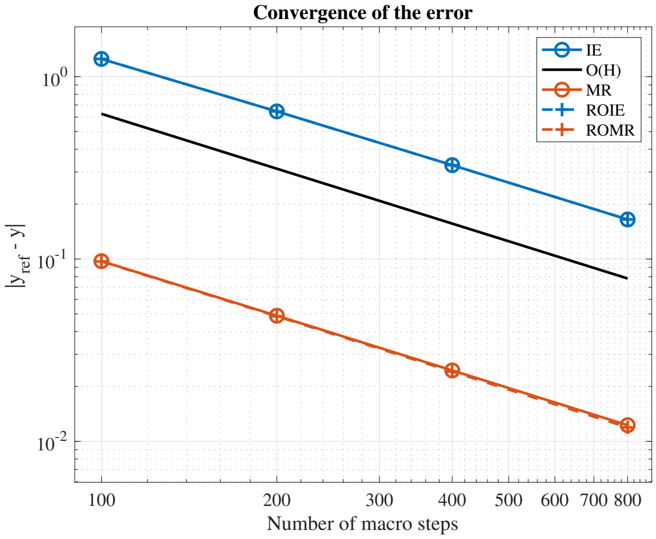

Figure 5.4: *The order 1 convergence of the computational error descending parallel to the black reference convergence.*

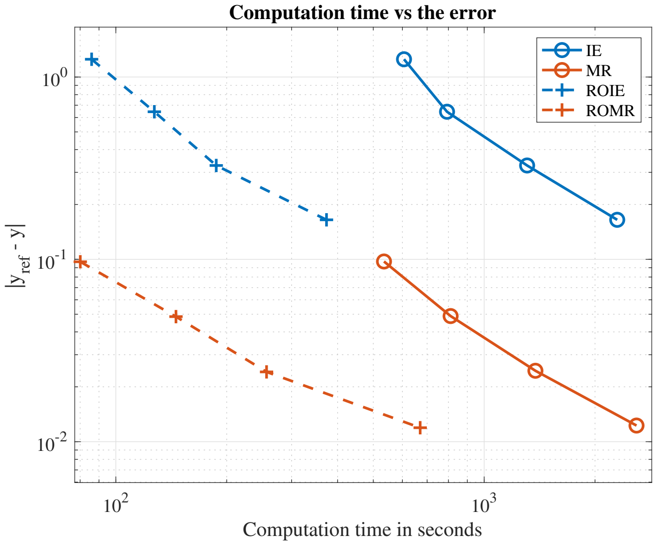

Figure 5.5: *The effect of the different numerical methods on the computational time and accuracy.*

5.1.2 System of Coupled DAEs

The following experiment was first presented in [7][Section 4.1]. The academic circuit shown in Figure 5.6 is a combination of a short diode chain and then a long ladder of diodes and resistors. Similar to a standard diode chain model [50], this model contains sufficient redundancy to make it eligible for model order reduction. Furthermore, increasing the resistance for each R_i with $R < R_i < R_{i+1}$ makes that the ladder part of the circuits behaves on a slower timescale. This makes the circuit excellent for a time integration with the reduced order multirate approach. The simulation parameters are given in box below.

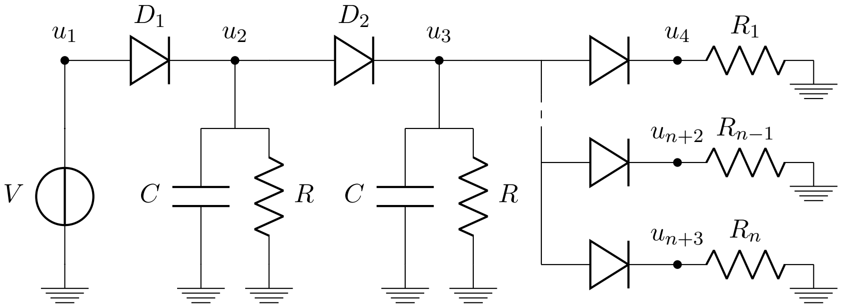

Figure 5.6: *The academic diode chain test model with redundancy.*

Starting time	t_0	0 s
Ending time	t_N	0.004 s
Number of steps	N	[100 200 400 800]
Multirate factor	m	20
Newton tolerance	tol	10^{-8}
Reduced dimension	r	2
Hyper-reduction factor	g	23
Voltage source	V	$5\sin(40 \cdot 2\pi t)$ V
Resistance	R	1000 Ω
Resistance	R_i	$i \cdot 1000$ Ω
Capacitance	C	10 μF
Diode saturation current	I_S	10^{-12} A

Table 5.1: *Simulation parameters of the academic model*

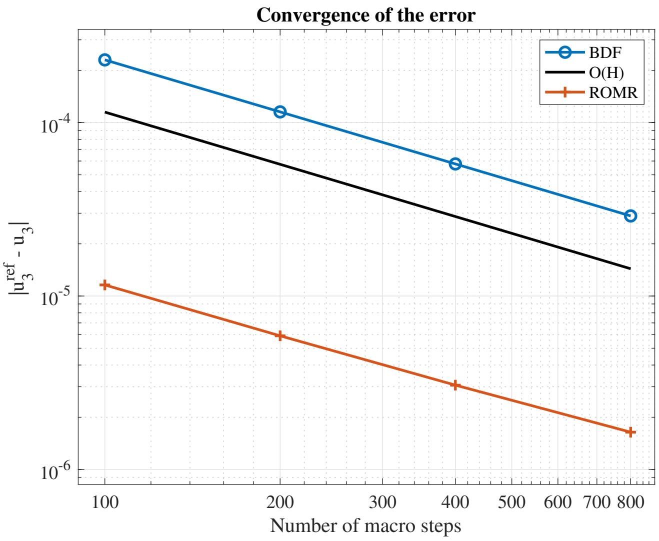

Figure 5.7: *Convergence of the numerical schemes, where the error is plotted against the number of macro steps. The MR error is omitted in the figure as the difference introduced by the reduction is negligible.*

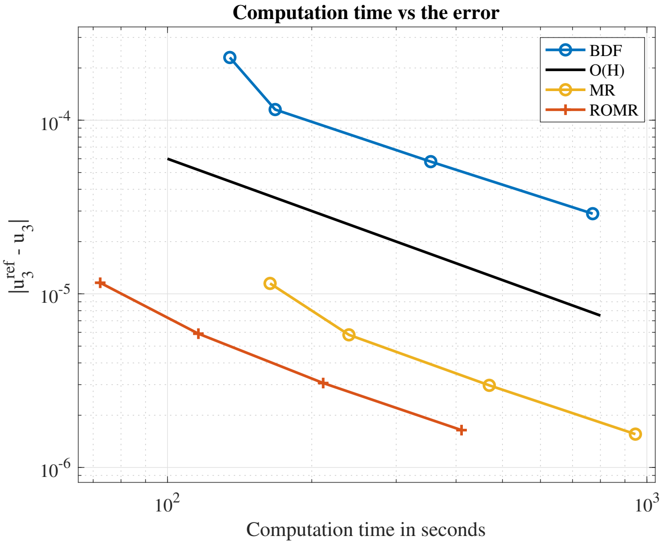

Figure 5.8: *Computational effort of the numerical schemes, where the error is plotted against the computation time in seconds. The error is defined as the absolute value between the computed voltage and reference voltage for the output node.*

5.2 Industrial Applications

Besides the academic test models, this section presents benchmark cases from industry. The first benchmark case considers a thermal resistor in a small circuit. The different inherent time scales are rooted in the distinction between the electrical part and the thermal part of the full system.

5.2.1 The Thermal-Electric Circuit

Following the experiments as described in [8][Section 3.B]. As the reduced order multirate case study circuit needs to contain both coupling and different intrinsic time scales, a thermal-electric test circuit is used [10]. This test circuit originates from Silicon-On-Insulator technology, which facilitates lower switching capacities by a floating body. The system is modelled by using a medium-grained approach for uni-dimensional and lumped elements. This circuit consists of an operational amplifier, two resistors, a diode, and a capacitor. The thermal resistor $R(T)$ is modelled by a structure of length $d = 0.03$ m which produces and transports heat along a uni-dimensional wire between the amplifier and the diode, it has a variable diameter

$$a(x) = a_0/[1 + b(d - x)x], \tag{5.1}$$

with $x \in [0, d]$, while the material parameters are those of a copper wire. The local resistance

$$\rho(T) = r_0(1 + \alpha(T - T_{\text{meas}}) + \beta(T - T_{\text{meas}})^2) \tag{5.2}$$

exhibits quadratic dependence on the temperature. The local resistance per unit cross-section is thus expressed in Ωm. Using this expression, the total resistance of the wire is

$$R(T) = \int_0^d \frac{\rho(\xi, T(t, \xi)}{a(\xi)} \mathrm{d}\xi. \tag{5.3}$$

The amplifier is modelled as a heat source and the diode has a temperature dependent characteristic $i_{\text{di}}(u_{\text{di}}, T_{\text{di}})$ curve

$$i_{\text{di}}(u_{\text{di}}, T_{\text{di}}) = \hat{I}_S(T_{\text{di}}) \left[\mathrm{e}^{\frac{u_{\text{di}}}{v_T}} - 1 \right], \tag{5.4}$$

$$\hat{I}_S(T_{\text{di}}) = 10^{-12} \left(\frac{T_{\text{di}}}{300\text{K}} \right)^3 \mathrm{e}^{\frac{-qE_g(300\text{K})}{k_{\text{B}} T_{\text{di}}} (1 - \frac{T_{\text{di}}}{300\text{K}})}. \tag{5.5}$$

Material	Cu (copper)
Specific resistance	$r_0 = 1.7\ \mu\Omega \cdot \mathrm{m}$
Reference temperature	$T_{\mathrm{meas}} = 291$ K
Length	$d = 0.03$ m
Cross section	$a_0 = 540$ m
Profile	$b = (2/d)^2\ \mathrm{m}^2$
1st thermal coefficient	$\alpha = 1/(273\ \mathrm{K})$
2nd thermal coefficient	$\beta = 1/(273\ \mathrm{K})^2$

Table 5.2: *Electrical parameters of one-dimensional resistor*

Density	$d_{\mathrm{w}} = 8.98 \cdot 10^3\ \mathrm{kg/m^3}$
Heat conductivity	$\lambda_{\mathrm{w}} = 390\ \mathrm{W/(mK)}$
Specific heat	$c_{\mathrm{w}} = 385\ \mathrm{J/(kgK)}$
Transition coefficient	$\gamma = 1.0\ \mathrm{W/(m^2K}$
Thermal mass	$M'_{\mathrm{w},i} = a(x_i)d_{\mathrm{w}}c_{\mathrm{w}}$ J/K
Cooling surface	$S'_{\mathrm{w},i}(x) = 2\sqrt{\pi a(x)}$

Table 5.3: *Thermal parameters of the one-dimensional resistor*

The electric behaviour of the circuit is modelled by modified nodal analysis based on Kirchhoff's laws. The thermal model is nonlinear due to the coupling terms, where the local self-heating term, P_{w}, introduces the nonlinear terms. After discretising in space, the following thermal-electric system is obtained.

Specific resistance	$q = 1.602 \cdot 10^{-19}$ C
Energy gap	$E_g(300\mathrm{K}) = 1.11$ V
Boltzmann constant	$k_{\mathrm{B}} = 1.381 \cdot 10^{-23}$ J/K
Thermal voltage	$v_T = k_{\mathrm{B}} \cdot T_{\mathrm{di}}/q$ V
Operational power	$v_{\mathrm{op}} = 15$ V
Amplification	$A = 20000$
Load resistance	$R_{\mathrm{L}} = 0.3\ \mathrm{k}\Omega$
Capacitance	$C = 500$ nF

Table 5.4: *Electrical parameters of the zero-dimensional elements*

Electric network

$$\begin{aligned} 0 &= (Av(t) - u_3)/R(T) + i_{di}(u_3 - u_4, T_{\mathrm{di}}), \\ C\dot{u}_4 &= i_{\mathrm{di}}(u_3 - u_4, T_{\mathrm{di}}) - u_4/R_{\mathrm{L}}, \end{aligned}$$

Coupling interfaces

$$P_{\mathrm{op}} = |(v_{\mathrm{op}} - |v(t)|) \cdot (Av(t) - u_3)/R|, \quad P_{\mathrm{w}} = (Av(t) - u_3)^2/R,$$

$$R(T) = \left(\frac{1}{2}(\rho(0, T_0) + \sum_{i=1}^{N-1} \rho(X_i, T_i) + \frac{1}{2}\rho(l, T_N) \right) \cdot h,$$

Heat equation

$$\begin{aligned}
M'_{\mathrm{w},i} h \dot{T}_i, &= \Lambda \frac{T_{i+1} - 2T_i + T_{i-1}}{h} + P_{\mathrm{w}} \frac{\tilde{\rho}(X_i, T_i)}{R} h \\
&\quad - \gamma S'_{\mathrm{w},i} h (T_i - T_{\mathrm{env}}), (i = 1, ..., N-1), \\
(M'_{\mathrm{w},0} \cdot \frac{h}{2} + M_{\mathrm{op}}) \dot{T}_0 &= \Lambda \frac{T_1 - T_0}{h} + P_{\mathrm{w}} \frac{\tilde{\rho}(0, T_0)}{R} \frac{h}{2} \\
&\quad - \gamma (S'_{\mathrm{w},0} \frac{h}{2} + S_{\mathrm{op}}) \cdot (T_0 - T_{\mathrm{env}}) + P_{\mathrm{op}}, \\
(M'_{\mathrm{w},N} \cdot \frac{h}{2} + M_{\mathrm{di}}) \dot{T}_N &= \Lambda \frac{T_{N-1} - T_N}{h} + P_{\mathrm{w}} \frac{\tilde{\rho}(X_N, T_N)}{R} \frac{h}{2} \\
&\quad - \gamma (S'_{\mathrm{w},N} \frac{h}{2} + S_{\mathrm{di}}) \cdot (T_N - T_{\mathrm{env}})
\end{aligned}$$

Amplifier	cubic
Material	Al (aluminium)
Size	$e_{\mathrm{op}} = 0.5$ mm
Heat capacity	$c_{\mathrm{Al}} = 449$ J/(kgK)
Density	$d_{\mathrm{al}} = 2.7 \cdot 10^3 \mathrm{kg/m}^3$
Cooling surface	$S_{\mathrm{op}} = 6 \cdot e_{\mathrm{op}}^2$ mm^2
Diode	cubic
Material	Si (silicon)
Size	$e_{\mathrm{di}} = 0.167$ mm
Heat capacity	$c_{\mathrm{Al}} = 700$ J/(kgK)
Density	$d_{\mathrm{si}} = 2.33 \cdot 10^3 \mathrm{kg/m}^3$
Cooling surface	$S_{\mathrm{di}} = 6 \cdot e_{\mathrm{di}}^2$ mm^2

Table 5.5: *Extension parameters of the zero-dimensional elements*

To illustrate the behaviour of the thermal resister the voltages at nodes u_3 and u_4 have been plotted in Figure 5.9. The heat development through time in the wire is shown in Figure 5.10

The computational cost of coupled network simulations are reduced by applying the reduced order multirate scheme. Here, multirate integration and the maximum entropy snapshot sampling method are employed for solving the equations that govern the thermal-electric circuit that is depicted

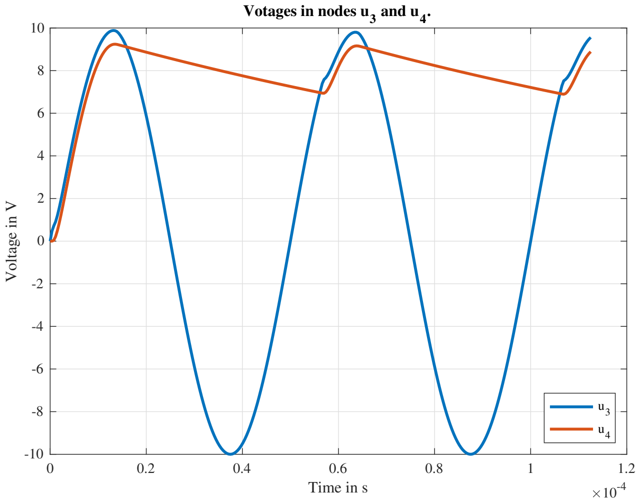

Figure 5.9: *Output of the thermal-electric circuit at nodes u_3 and u_4.*

in Fig. 5.11. After partitioning the slow and fast varying time-scales, the system of Equations (4.27a)–(4.27c) becomes

$$\begin{aligned} \dot{x}_{\mathrm{F}} &= f_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, Vx_{\mathrm{S,r}}), & x_{\mathrm{F}}(0) &= x_{\mathrm{F},0}, \\ \dot{x}_{\mathrm{S,r}} &= f_{\mathrm{S,r}}(x_{\mathrm{F}}, z_{\mathrm{F}}, x_{\mathrm{S,r}}), & x_{\mathrm{S,r}}(0) &= x_{\mathrm{S,r},0}, \\ 0 &= g_{\mathrm{F}}(x_{\mathrm{F}}, z_{\mathrm{F}}, Vx_{\mathrm{S,r}}), & z_{\mathrm{F}}(0) &= z_{\mathrm{F},0}. \end{aligned} \tag{5.6}$$

Here, $x_{\mathrm{F}} = u_4$, $z_{\mathrm{F}} = u_3$, and $x_{\mathrm{S,r}} = V^{\top}x_{\mathrm{S}}$, with $x_{\mathrm{S}} \in \mathbb{R}^m$ being the discretized temperature in the thermal resistor.

The system of Equations (5.6) is integrated with a reduced order multirate method, and the parameter ϵ is computed by (3.25), see Figure 5.13. To verify the performance of the reduced order multirate scheme, maximum entropy snapshot sampling method and estimated epsilon, a transient analysis for the output u_4 is performed. A reference solution is obtained with a standard multirate scheme of five fine-grid steps for a problem with $(m, n) = (10^4, 500)$. Then, the reduced order multirate scheme is used, once with the maximum entropy snapshot sampling method with ϵ_* and once with the proper orthogonal decomposition. In Fig. 5.12 the correlation sum is shown, and in Figure 5.13 the accuracy plot for maximum entropy snapshot sampling with $\epsilon_* = 0.0816$ is depicted. The accuracy result for the proper orthogonal decomposition is indistinguishable from the one depicted in Fig. 5.13, and hence, it is omitted. The degrees

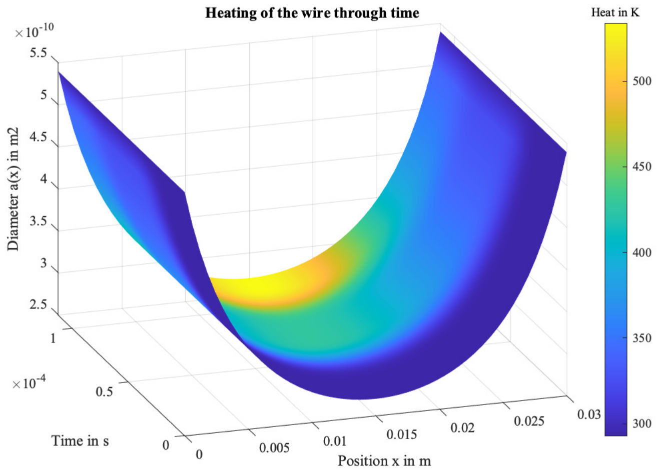

Figure 5.10: *The heat flow in time through the resistor*

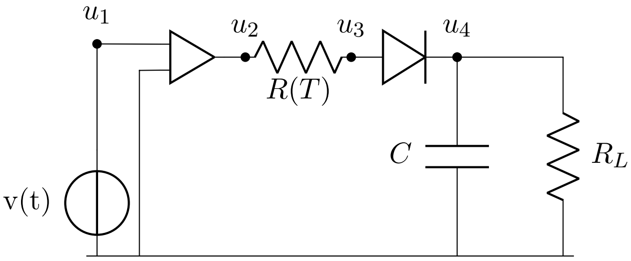

Figure 5.11: *The circuit used for the numerical experiments.*

of freedom are reduced from 10^4 to 13, while the optimal ϵ_* is estimated in 2.9 s and the maximum entropy snapshot sampling base is constructed in 0.16 s, in contrast to a total of 6.33 s that is required by the POD.

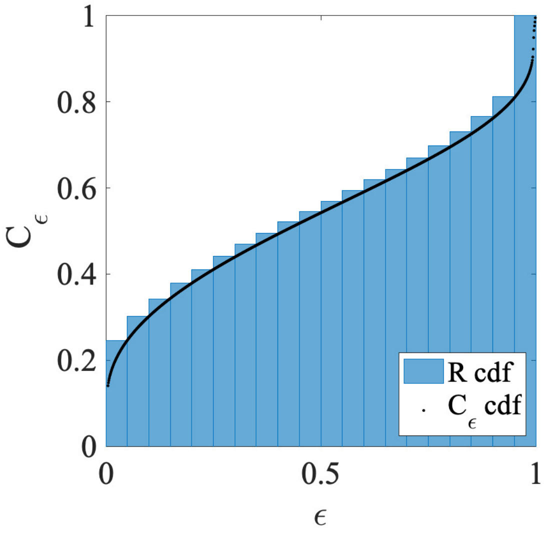

Figure 5.12: *The cumulative distribution of R and C_ϵ.*

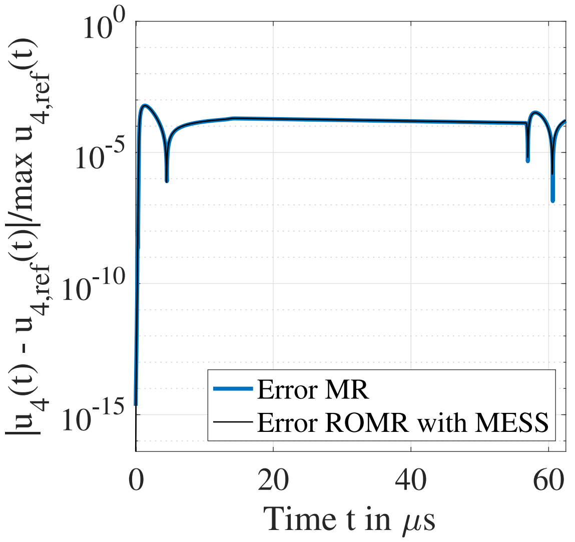

Figure 5.13: *The relative difference between standard multirate (MR) and reduced order multirate (ROMR) with (MESS, 0.0816).*

5.2.2 Solar Panel Simulation

As final industrial application STMicroelectronics has provided a test case from their development team. These results have been published in [7][Section 4.2]. The test case concerns a transient analysis of a solar panel circuit. In this case we consider a silicon photovoltaic cell that is composed of two layers of semiconducting material, e.g. silicon or gallium arsenide, with different doping. The process of converting solar radiation into electricity is based on the photovoltaic effect.

To model the photovoltaic effects taking place in the solar panel, a mathematical model is used based on the construction of an equivalent circuit. When the cell is lit, the generated power can be modelled as a current source. Therefor, an equivalent circuit of the photovoltaic cell is shown in Figure 5.14.

The full solar panel circuit is a conglomeration of photovoltaic cells, Figure 5.14, that are linked in series and parallel. This grid of photovoltaic cells is then connected to a DC-DC buck converter to stabilise the output. The full schematic of the solar panel circuit is shown in Figure 5.15.

As the voltages and currents generated by the solar panel vary significantly slower than the operating voltages in the buck converter there is an inherent multirate advantage. Due to the replication in the structure of the solar panel, model order reduction can be used to significantly reduce the complexity of the panel simulation.

The full test case for the silicon photovoltaic solar panel is run for a series and parallel grid that is 35 by 35, which makes the full dimensional system consist of 2422 equations.

The reduction parameter selection procedure provides us with $\epsilon = 0.43$ which leads to quite a significant reduction. As shown in Table 5.6 the dimension of the slow subsystem of photovoltaic cells is decreased from 2416 to only 6 equations, and the hyper-reduced dimension of the nonlinear function is reduced to only 8 equations.

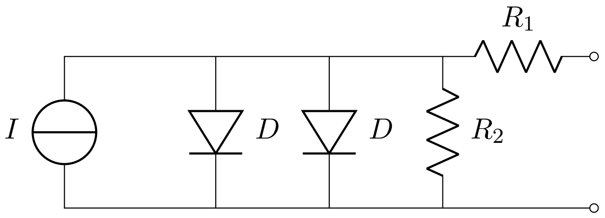

Figure 5.14: *The equivalent circuit of a photovoltaic cell.*

Starting time	t_0	0 s
Ending time	t_N	100 ms
Number of steps	N	[75 150 300 600 1200]
Multirate factor	m	20
Newton tolerance	tol	10^{-8}
Voltage source	V	$10\sin(100 \cdot 2\pi t)$ V
Resistance	R_1	3.5 mΩ
Resistance	R_2	150 Ω
Resistance	R_3	2 $\mu\Omega$
Resistance	R_4	2.4 Ω
Current source	I	0.9 A
Diode	I_S	10^{-12} A
Original dimension	d	2422
Reduced dimension	r	6
Hyper-reduction factor	g	8

Table 5.6: *Simulation parameters of the industrial model*

From Figures 5.16 and 5.17, we see that the academically achieved results are replicated for an industrial test case. Due to the model order reduction there is a substantial decrease in computation time, with the reduced order multirate scheme being roughly 6.35 times faster than the multirate integration scheme, without losing accuracy.

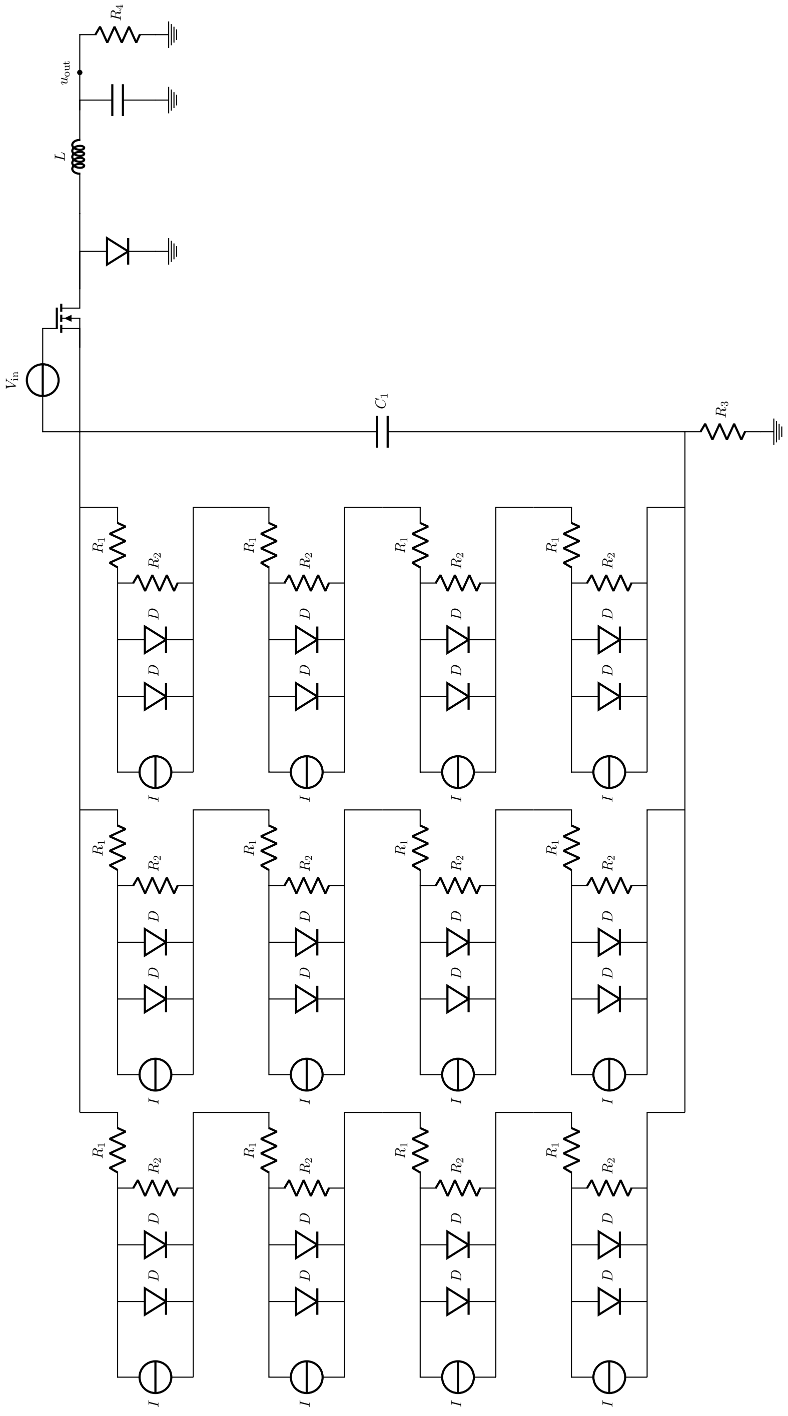

Figure 5.15: *The industrial solar panel test model with redundancy. Here a 4 by 3 grid of photovoltaic cells are linked together in series and parallel and connected to a DC/DC converter.*

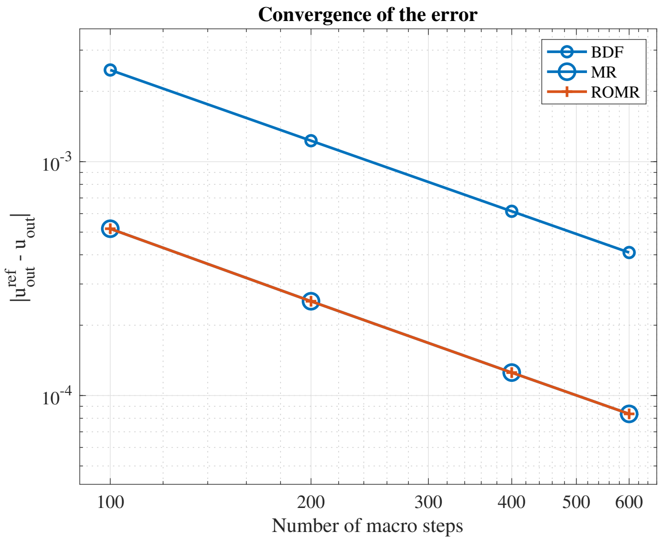

Figure 5.16: *Convergence of the numerical schemes, where the error is plotted against the number of macro steps.*

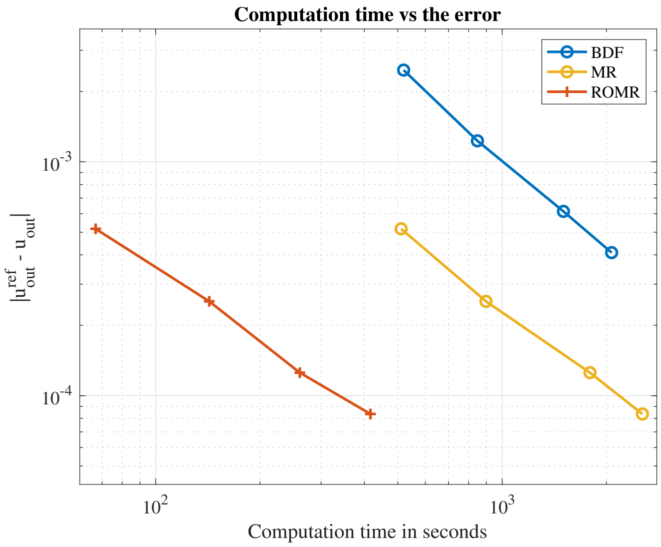

Figure 5.17: *Computational effort of the numerical schemes, where the error is plotted against the computation time in seconds.*

Conclusions and Outlook

In this thesis we have extensively described the development of reduced order multirate schemes in a industrial circuit simulation setting. We have shown that the combination of model order reduction techniques and multirate integration schemes provides possibilities for significant decreases in computational effort, whilst the accuracy of the schemes is preserved.

The research started out by coupling a system of differential-algebraic equations to a system of ordinary differential equations derives from a discretisation of a system of partial differential equations. A significant reduction in computation time has been achieved by using a combination of using the Galerkin projection model order reduction with a basis provided by proper orthogonal decomposition. The nonlinear bottleneck was circumvented using a QDEIM hyper-reduction.

In a subsequent paper, this approach has been proven to converge by numerical analysis, given that the some accuracy criteria related to the model order reduction approximation have been satisfied. However, in an effort to further generalise the theory of reduced order multirate schemes now a double differential-algebraic coupled system was considered.

As direct Galerkin projection model order reduction introduced a lot of complications and instability when applied to differential-algebraic systems another approach was considered. By using the Gauß-Newton with approximated tensors method it was possible to reduce these differential-algebraic systems. Furthermore, a novel reduced basis construction method, the maximum entropy snapshot sampling method, was utilised instead of the proper orthogonal decomposition approach.

This type of reduced order multirate schemes has also been shown to converge. This was verified by numerical simulation of academic circuits. Besides only showing results related to the convergence of reduced simulation schemes, also a method for selecting the reduction parameter of the maximum entropy snapshot sampling method has been presented. In joint work with the chair of electromagnetic theory at the Bergische Universität Wuppertal an algorithm for obtaining this parameter has been presented.

This first exploratory research project into the combination of reduced order multirate schemes established a point of reference which future endeavours can use as a reference. The reduced order multirate schemes presented in this thesis show a huge potential for application in industrial circuit simulation. However, there is still a lot to be done. Construction a software suite capable of handling all different types of coupling can be quite a challenge. Especially incorporating various different flavours of model order reduction approaches, hyper-reduction methods and multirate integration schemes. In contrast, the upside advantages of these schemes may save a substantial amount of time in the verification process of designing microchips.

Bibliography

[1] A. C. Antoulas. *Approximation of large-scale dynamical systems.* SIAM, 2005.

[2] U. M. Ascher and L. R. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*, volume 61. Siam, 1998.

[3] S. Bächle. *Numerical solution of differential-algebraic systems arising in circuit simulation.* PhD thesis, 2007.

[4] M. W. F. M. Bannenberg, A. Ciccazzo, and M. Günther. Coupling of model order reduction and multirate techniques for coupled dynamical systems. *Applied Mathematics Letters*, 112:106780, 2020.

[5] M. W. F. M. Bannenberg, A. Ciccazzo, and M. Günther. Reduced order multirate schemes for coupled differential-algebraic systems. *Applied Numerical Mathematics*, 2021.

[6] M. W. F. M. Bannenberg, M. Günther, et al. Benchmark cases, deliverable 5.3. 2020.

[7] Marcus WFM Bannenberg, Angelo Ciccazzo, and Michael Günther. Reduced order multirate schemes in industrial circuit simulation. *Journal of Mathematics in Industry*, 12(1):1–13, 2022.

[8] Marcus WFM Bannenberg, Fotios Kasolis, Michael Günther, and Markus Clemens. Maximum entropy snapshot sampling for reduced basis modelling. *COMPEL-The international journal for computation and mathematics in electrical and electronic engineering*, 2021.

[9] A. Bartel and M. Günther. A multirate w-method for electrical networks in state–space formulation. *Journal of Computational and Applied Mathematics*, 147(2):411–425, 2002.

[10] A. Bartel, M. Günther, and M. Schulz. Modeling and discretization of a thermal-electric test circuit. In *K. Antreich, R. Bulirsch, A. Gilg, P. Rentrop (Eds.): Modeling, Simulation, and Optimization of Integrated Circuits*, pages 187–201. Springer, 2003.

[11] T. Bechtold, A. Verhoeven, E. J. W. Ter Maten, and T. Voss. Model order reduction: an advanced, efficient and automated computational tool for microsystems. In *Applied And Industrial Mathematics In Italy II*, pages 113–124. World Scientific, 2007.

[12] P. Benner, M. Hossain, and T. Stykel. Model reduction of periodic descriptor systems using balanced truncation. In *Model Reduction for Circuit Simulation*, pages 193–206. Springer, 2011.

[13] G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Ann. Rev. Fluid Mech.*, 25(1):539–575, 1993.

[14] H. Broer and F. Takens. *Dynamical systems and chaos.* Springer-Verlag New York, 2011.

[15] L. A. Bunimovich, I. P. Cornfeld, R. L. Dobrushin, N. B. Maslova, Y. B. Pesin, and A. M. Vershik. *Dynamical Systems II: Ergodic Theory with Applications to Dynamical Systems and Statistical Mechanics*, volume 2. Springer Science & Business Media, 2013.

[16] K. Carlberg, C. Bou-Mosleh, and C. Farhat. Efficient non-linear model reduction via a least-squares petrov–galerkin projection and compressive tensor approximations. *International Journal for numerical methods in engineering*, 86(2):155–181, 2011.

[17] K. Carlberg, C. Farhat, J. Cortial, and D. Amsallem. The GNAT method for nonlinear model reduction: effective implementation and application to computational fluid dynamics and turbulent flows. *Journal of Computational Physics*, 242:623–647, 2013.

[18] S. Chaturantabut and D. C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, 2010.

[19] S. Chaturantabut and D. C. Sorensen. A state space error estimate for POD-DEIM nonlinear model reduction. *SIAM J. Numer. Anal.*, 50(1):46–63, 2012.

[20] C. F. Curtiss and J. O Hirschfelder. Integration of stiff equations. *Proceedings of the National Academy of Sciences of the United States of America*, 38(3):235, 1952.

[21] Z. Drmac and S. Gugercin. A new selection operator for the Discrete Empirical Interpolation Method – improved a priori error bound and extensions. *SIAM J. Sci. Comput.*, 38(2):A631–A648, 2016.

[22] R. Everson and L. Sirovich. Karhunen–loeve procedure for gappy data. *JOSA A*, 12(8):1657–1664, 1995.

[23] C. W. Gear and D. R. Wells. Multirate linear multistep methods. *BIT Numerical Mathematics*, 24(4):484–502, 1984.

[24] P. Grassberger and I. Procaccia. Estimation of the kolmogorov entropy from a chaotic signal. *Physical review A*, 28(4):2591, 1983.

[25] S. Gugercin and A. C. Antoulas. A survey of model reduction by balanced truncation and some new results. *International Journal of Control*, 77(8):748–766, 2004.

[26] M. Günther, U. Feldmann, and J. ter Maten. Modelling and discretization of circuit problems. *HU. Langer, W. Amrhein, W. Zulehner (Eds.), Scientific Computing in Electrical Engineering - SCEE 2016. Book Series: Mathematics in Industry*, 28:91–100, 2018.

[27] M. Günther, A. Kvaernø, and P. Rentrop. Multirate partitioned Runge-Kutta methods. *BIT Numerical Mathematics*, 41(3):504–514, 2001.

[28] C. Hachtel, A. Bartel, M. Günther, and A. Sandu. Multirate implicit Euler schemes for a class of differential–algebraic equations of index-1. *J. Comput. Appl. Math.*, page 112499, 2019.

[29] P. Heres and W. H. A. Schilders. Krylov subspace methods in the electronic industry. In *Progress in Industrial Mathematics at ECMI 2004*, pages 139–143. Springer, 2006.

[30] P. J. Heres. Robust and efficient krylov subspace methods for model order reduction. 2005.

[31] F. Kasolis and M. Clemens. Maximum entropy snapshot sampling for reduced basis generation. *arXiv preprint arXiv:2005.01280*, 2020.

[32] F. Kasolis, D. Zhang, and M. Clemens. Recurrent quantification analysis for model reduction of nonlinear transient electro-quasistatic field problems. In *International Conference on Electromagnetics in Advanced Applications (ICEAA 2019)*, pages 14–17, 2019.

[33] B. Kramer and K. Willcox. Nonlinear model order reduction via lifting transformations and proper orthogonal decomposition. *AIAA Journal*, 57(6):2297–2307, 2019.

[34] P. Kunkel and V. Mehrmann. *Differential-algebraic equations: analysis and numerical solution*, volume 2. European Mathematical Society, 2006.

[35] J. Liesen and Z. Strakos. *Krylov subspace methods: principles and analysis.* Oxford University Press, 2013.

[36] M. Rathinam and L. R. Petzold. A new look at proper orthogonal decomposition. *SIAM Journal on Numerical Analysis*, 41(5):1893–1925, 2003.

[37] J. R. Rice. Split runge-kutta method for simultaneous. *Journal of Research of the National Bureau of Standards: Mathematics and mathematical physics. B*, 64:151, 1960.

[38] J. Rommes. *Methods for eigenvalue problems with applications in model order reduction.* PhD thesis, 2007.

[39] Y. Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of computation*, 37(155):105–126, 1981.

[40] Y. Saad. Krylov subspace methods on supercomputers. *SIAM Journal on Scientific and Statistical Computing*, 10(6):1200–1232, 1989.

[41] W. H. A. Schilders, H. A. Van der Vorst, and J. Rommes. *Model order reduction: theory, research aspects and applications*, volume 13. Springer, 2008.

[42] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[43] H. Shichman and D. A. Hodges. Modeling and simulation of insulated-gate field-effect transistor switching circuits. *IEEE Journal of Solid-State Circuits*, 3(3):285–289, 1968.

[44] F. Takens. On the numerical determination of the dimension of an attractor. In *Dynamical systems and bifurcations*, pages 99–106. Springer, 1985.

[45] F. Takens and E. Verbitski. Generalized entropies: Rényi and correlation integral approach. *Nonlinearity*, 11(4):771, 1998.

[46] H. A. M. Tong. *Dimension estimation and models*, volume 1. World Scientific, 1993.

[47] M. van Beurden, N. V. Budko, and W. H. A. Schilders. *Scientific Computing in Electrical Engineering: SCEE 2020, Eindhoven, The Netherlands, February 2020*, volume 36. Springer Nature, 2021.

[48] B. L. van der Waerden. *Mathematical Statistics.* Springer, 1969.

[49] A. Verhoeven. *Redundancy reduction of IC models : by multirate time-integration and model order reduction.* PhD thesis, Department of Mathematics and Computer Science, 2008.

[50] A. Verhoeven, J. Ter Maten, M. Striebel, and R. Mattheij. Model order reduction for nonlinear ic models. In *IFIP Conference on System Modeling and Optimization*, pages 476–491. Springer, 2007.

[51] S. Volkwein. Model reduction using proper orthogonal decomposition. *Lecture Notes, Institute of Mathematics and Scientific Computing, University of Graz*, 1025, 2011.

[52] G. Wanner and E. Hairer. *Solving ordinary differential equations II*, volume 375. Springer Berlin Heidelberg, 1996.

[53] K. Willcox. Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition. *Computers & fluids*, 35(2):208–226, 2006.

[54] D. Wirtz. *Model reduction for nonlinear systems: kernel methods and error estimation.* epubli, 2014.

[55] R. Zimmermann and K. Willcox. An accelerated greedy missing point estimation procedure. *SIAM Journal on Scientific Computing*, 38(5):A2827–A2850, 2016.