Chapter 3
Simulation of Coupled PDAEs: Dynamic Iteration and Multirate Simulation

Giuseppe Alì, Andreas Bartel, Michael Günther, Vittorio Romano, and Sebastian Schöps

Abstract This chapter investigates the error transport in dynamic iteration schemes for coupled DAE systems. The essential theory is developed in detail. Then the results are applied to various coupled systems stemming from applications in electrical engineering.

3.1 Aim and Outline

In practice, we often have to deal with multiphysical descriptions of mathematical models and as well with systems which exhibit widely separated time scales. A common approach for multiphysical systems is the application of dynamic iteration (or co-simulation), which allows to treat each subsystem with a dedicated solver, and also an according discretization. Furthermore, so-called multirate techniques can be applied to specifically exploit different time scales.

G. Alì
Department of Physics, University of Calabria via Pietro Bucci 30/B, 87036 Arcavacata di Rende, Cosenza, Italy
e-mail: giuseppe.ali@unical.it

A. Bartel (✉) • M. Günther
Chair of Applied Mathematics/Numerical Analysis, Bergische Universität Wuppertal, Gaußstraße 20, D-42119 Wuppertal, Germany
e-mail: bartel@math.uni-wuppertal.de; guenther@math.uni-wuppertal.de

S. Schöps
TU Darmstadt, Graduate School of Excellence Computational Engineering, Dolivostraße 15, 64293 Darmstadt, Germany
e-mail: schoeps@gsc.tu-darmstadt.de

V. Romano
Università di Catania, Dipartimento di Matematica e informatica, Viale A. Doria no, 95125 Catania, Italy
e-mail: romano@dmi.unict.it

M. Günther (ed.), *Coupled Multiscale Simulation and Optimization in Nanoelectronics*, Mathematics in Industry 21,
DOI 10.1007/978-3-662-46672-8_3

To reflect this, the aim of this chapter is twofold. First we address dynamic iteration of spatially discretized PDAE systems, which are in fact coupled DAE systems. We demonstrate the crucial differences between coupled DAE and coupled ODE systems by investigating the splitting error of these coupled systems theoretically. Then we apply the obtained knowledge to coupled systems from Chap. 2. Secondly, a multirate strategy is discussed and studied numerically.

To this end, this chapter is organized as follows. It starts with the detailed theory of dynamic iteration schemes for coupled DAEs. First we consider a single window and proof an error recursion for any investigated window. Then we treat multiple windows and generalize the results. In the following section, we apply our results to some of the DAE models introduced in Chap. 2: refined network models, electric networks and Maxwell's magnetostatic equations. Finally, a multirate method for the coupled simulation of thermal effects in silicon devices is investigated.

3.2 Theory of Dynamic Iteration Schemes for Coupled DAEs

Here we address the time-domain solution of PDAEs by means of dynamical iteration schemes. To explain the basic concept, let us suppose that we want to solve an initial value problem for a system of PDAEs, on a time interval $[0, t_e]$. To this end, the time interval $[0, t_e]$ is split in windows $[t_n, t_{n+1}]$ with so-called synchronization points t_n, which satisfy: $0 = t_0 < t_1 < \cdots < t_N = t_e$. The windows are treated sequentially and in each window the subsystems are solved iteratively. Mathematically speaking, this leads to apply a dynamic iteration scheme.

Coupled systems as our PDAEs, see Chap. 2, can be treated with coupled simulators, each designed and tailored to the respective subsystem's structure. This is called simulator-coupling, co-simulation or distributed (time-)integration. Compared to monolithic approaches, where the overall system is treated by any standard integration the distributed computation offers potential w.r.t. parallelization and incorporates adapted step sizes and orders to every subsystem automatically.

Although we have in mind applications to PDAEs, we will develop the theory of dynamic iteration schemes for DAEs. For practical applications, all the results presented in this Chapter can be extended to PDAE after performing suitable spatial discretizations. A detailed example of this approach is given for PDAEs arising in refined network modeling.

Iteration schemes were first applied to coupled ODE systems, including multirate, multi-order, multi-method and dynamic iteration. For the latter, which is our focus, convergence is unconditional (see [10]) if the windowing technique is applied. However, the situation changes, when this methods are applied to DAEs. Here instabilities may occur and solutions can explode even if a windowing technique is in use. Here convergence, that is, contraction of the corresponding fixed point operator, can be guaranteed by fulfilling additional stability constraints. This dates back to Lelarasmee [24] and was applied for single window convergence [3, 22] and specially coupled systems for multiple windows in [1, 4]. We note that

the stability restrictions where also encountered in the numerical analysis of DAEs, see [16, 21] and [1, 23].

Here we follow [4] with some more details to derive a general representation of the error recursion for coupled systems. The preceding steps, e.g. for convergence result, are as in [1]. Furthermore we aim at extracting the underlying principle: algebraic to algebraic coupling is to be excluded or damped.

3.2.1 Description of Coupled Systems

After applying a suitable space discretization to the PDAE problems discussed in the first chapter, we are faced with the following simulation problem: solve an initial-value problems of semi-explicit differential-algebraic equations

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}), \tag{3.1a}$$

$$0 = \mathbf{g}(\mathbf{y}, \mathbf{z}), \tag{3.1b}$$

where the dot denotes differentiation with respect to time. In this formulation we do not distinguish between different subsystems, but all subsystems are comprised within one system. As we will see, this is enough to treat dynamic iteration schemes. It is specially well-applicable for linear PDE-parts, where space and time discretization can be easily separated. Also a non-autonomous system can be casted in this form, by introducing an additional equation: $\dot{t} = 1$. We assume that this problem, equipped with initial values

$$\mathbf{y}(0) = \mathbf{y}_0\,, \quad \mathbf{z}(0) = \mathbf{z}_0, \tag{3.2}$$

has a unique solution $\mathbf{y} : [0, t_e] \to \mathbb{R}^{n_y}$, $\mathbf{z} : [0, t_e] \to \mathbb{R}^{n_z}$ on the finite time interval $[0, t_e]$. In a neighborhood of this solution the functions $\mathbf{f}$ and $\mathbf{g}$ are supposed to be sufficiently often differentiable. Furthermore, it is supposed that

$$\text{the Jacobian } \partial\mathbf{g}/\partial\mathbf{z} \text{ is non-singular,} \tag{3.3}$$

in the neighborhood of the solution. Hence system (3.1) has index-1. Moreover, the initial values (3.2) have to be consistent, that is for our semi-explicit index-1 system (3.1), the explicit algebraic constraint (3.1b) is fulfilled for the initial data.

Next we discuss the representation of coupled systems. In multiphysics problems, system (3.1) is often directly given as a coupled system of r DAE subsystems

$$\dot{\mathbf{y}}_i = \mathbf{f}_i(\mathbf{y}, \mathbf{z}), \tag{3.4a}$$

$$0 = \mathbf{g}_i(\mathbf{y}, \mathbf{z}) \tag{3.4b}$$

for $i = 1, \ldots, r$, with $\mathbf{y}^\top = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_r^\top)$, $\mathbf{z}^\top = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_r^\top)$, $\mathbf{f}^\top = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_r^\top)$, $\mathbf{g}^\top = (\mathbf{g}_1^\top, \ldots, \mathbf{g}_r^\top)$. In addition to the index-one assumption (3.3) for the whole

system (3.1), we now assume that

$$\partial \mathbf{g}_i / \partial \mathbf{z}_i \text{ is non-singular for all } i = 1, \ldots, r, \tag{3.5}$$

so that the equations $\mathbf{g}_i(\mathbf{y}, \mathbf{z}) = 0$ are locally uniquely solvable with respect to $\mathbf{z}_i$, with other words: system (3.4) defines an index-1 system for unknown functions $\mathbf{y}_i$, $\mathbf{z}_i$ assuming that all other variables $\mathbf{y}_j$, $\mathbf{z}_j$ $(j \neq i)$ are given as time-dependent functions.

Sometimes system (3.1) may be given as r coupled ODE systems linked to only one algebraic equation:

$$\dot{\mathbf{y}}_i = \mathbf{f}_i(\mathbf{y}, \mathbf{z}), \tag{3.6a}$$

$$0 = \mathbf{g}(\mathbf{y}, \mathbf{z}), \tag{3.6b}$$

for $i = 1, \ldots, r$. The index-1 assumption now again reads as in (3.3), that is, we assume that $\partial \mathbf{g} / \partial \mathbf{z}$ is non-singular in a neighborhood of the solution.

Sometimes a separation in subsystems is not a priori fixed by a simple partition (e.g. (3.6)). This leads to the following notation, where some quantities are assigned to several subsystems.

Overlapping modeling The structure (3.6) gives more freedom in a dynamic iteration scheme by applying appropriate overlapping strategies [2]. For such a strategy, the system is replaced by a number of overlapping subsystems, defined by means of splitting matrices. As splitting matrices we introduce $\mathbf{P}_i \in \mathbb{R}^{n_z \times l_i}$ with $1 \leq l_i \leq n_z$ and $\operatorname{rank}(P_i) = l_i$ for $i = 1, \ldots, r$, such that the matrix

$$(\mathbf{P}_1 \ldots \mathbf{P}_r) \in \mathbb{R}^{n_z \times (\sum_i l_i)} \text{ has full rank } n_z \tag{3.7}$$

(thus we implicitly require $\sum_i l_i \geq n_z$). In this way, arbitrary parts $\mathbf{P}_i^\top \mathbf{g}$ of the algebraic equation (3.6b) can be extracted, since it holds:

$$(\mathbf{P}_1, \ldots, \mathbf{P}_r)^\top \mathbf{g} = 0 \quad \text{if and only if} \quad \mathbf{g} = 0.$$

Next, we assign the extracted components to the i-th ODE subsystem to define r overlapping DAE systems:

$$\dot{\mathbf{y}}_i = \mathbf{f}_i(\mathbf{y}, \mathbf{w}_i), \tag{3.8a}$$

$$0 = \mathbf{P}_i^\top \mathbf{g}(\mathbf{y}, \mathbf{w}_i), \tag{3.8b}$$

substituting $\mathbf{z}$ by $\mathbf{w}_i$ (for $i = 1, \ldots, r$). Also $\mathbf{z}$ is split into further components $\bar{\mathbf{z}}_i := \mathbf{P}_i^\top \mathbf{z}$, such that it holds

$$\mathbf{w}_i = \mathbf{z} = (\mathbf{I} - \mathbf{P}_i \mathbf{P}_i^\top)\mathbf{z} + \mathbf{P}_i \bar{\mathbf{z}}_i. \tag{3.9}$$

This splitting is crucial for any modular time integration to come. Adding the coupling equation (3.9) to the rth system, we obtain in fact:

$$\dot{\mathbf{y}}_i = \tilde{\mathbf{f}}_i(\mathbf{y}, \mathbf{z}, \bar{\mathbf{z}}_i), \tag{3.10a}$$

$$0 = \tilde{\mathbf{g}}_i(\mathbf{y}, \mathbf{z}, \bar{\mathbf{z}}_i), \tag{3.10b}$$

for $i = 1, \ldots, r$, with

$$\begin{aligned}
\tilde{\mathbf{f}}_i(\mathbf{y}, \mathbf{z}, \bar{\mathbf{z}}_i) &:= \mathbf{f}_i(\mathbf{y}, (\mathbf{I} - \mathbf{P}_i\mathbf{P}_i^\top)\mathbf{z} + \mathbf{P}_i\bar{\mathbf{z}}_i), \quad i = 1, \ldots, r, \\
\tilde{\mathbf{g}}_i(\mathbf{y}, \mathbf{z}, \bar{\mathbf{z}}_i) &:= \mathbf{P}_i^\top \mathbf{g}(\mathbf{y}, (\mathbf{I} - \mathbf{P}_i\mathbf{P}_i^\top)\mathbf{z} + \mathbf{P}_i\bar{\mathbf{z}}_i) \quad i = 1, \ldots, r-1, \\
\tilde{\mathbf{g}}_r(\mathbf{y}, \mathbf{z}, \bar{\mathbf{z}}) &:= \begin{pmatrix} \mathbf{P}_r^\top \mathbf{g}(\mathbf{y}, (\mathbf{I} - \mathbf{P}_r\mathbf{P}_r^\top)\mathbf{z} + \mathbf{P}_r\bar{\mathbf{z}}_r) \\ \mathbf{z} - \left[(\mathbf{I} - \sum_{j=1}^{r} \mathbf{P}_j\mathbf{P}_j^\top)\mathbf{z} + \sum_{j=1}^{r} \mathbf{P}_j\bar{\mathbf{z}}_j \right] \end{pmatrix}.
\end{aligned}$$

If the original system (3.6) has index-1, then also system (3.10) has index-1. In fact, the index-1 conditions for system (3.10) are:

$$\mathbf{P}_i^\top (\partial \mathbf{g}/\partial \mathbf{z})\mathbf{P}_i \text{ regular},$$

$$\sum_{j=1}^{r} \mathbf{P}_j\mathbf{P}_j^\top \text{ regular},$$

which are ensured by the index-1 condition (3.3), and by the definition of our matrices $\mathbf{P}_j$, which satisfy condition (3.7).

Lastly, we notice: (a) according to our system (3.6), we have only overlapping in the algebraic system; of course, more general situations are conceivable; (b) the case of additional coupling equations can be also retrieved within the above discussed case.

Next, we discuss several types of iteration schemes, which we can identify with splitting functions.

3.2.2 *Iteration Schemes for Coupled DAE Systems*

The idea of our dynamic iteration schemes is now to work directly on the splitting structure of system (3.1) given by either (3.4) or (3.6) to exploit the varying properties of the subsystems via multirate and multimethod approaches.

Before going into the details of exploiting the special structure, we define a generic dynamic iteration scheme in the following. In a first step we split the whole integration interval $[0, t_e]$ into windows $[t_n, t_{n+1}] \subset [0, t_e]$ ($n = 0, 1, \ldots, N-1$ with $t_0 = 0$ and $t_N = t_e$), of size $H_n := t_{n+1} - t_n$. As already mentioned, this windowing

technique guarantees convergence in the case of purely coupled ODE systems and for DAE systems additional stability restrictions to be discussed play an important role (for convergence and fast numerical computation of solutions).

Let us now consider a window $[t_n, t_{n+1}]$ and suppose that the numerical solution

$$(\tilde{\mathbf{y}}, \tilde{\mathbf{z}})^\top : [0, t_e] \to \mathbb{R}^{n_y} \times \mathbb{R}^{n_z}$$

has already been computed for $t \in [0, t_n]$. To get a numerical approximation in the next window $[t_n, t_{n+1}]$,

$$\tilde{\mathbf{y}}|_{(t_n, t_{n+1}]}, \qquad \tilde{\mathbf{z}}|_{(t_n, t_{n+1}]},$$

we proceed as follows:

- *Extrapolation step*: the iteration starts with

$$\begin{pmatrix} \tilde{\mathbf{y}}_n^{(0)} \\ \tilde{\mathbf{z}}_n^{(0)} \end{pmatrix} := \Phi_n \begin{pmatrix} \tilde{\mathbf{y}}|_{(t_{n-1}, t_n]} \\ \tilde{\mathbf{z}}|_{(t_{n-1}, t_n]} \end{pmatrix} \qquad \text{with } \Phi_n = \begin{pmatrix} \Phi_{\mathbf{y},n} \\ \Phi_{\mathbf{z},n} \end{pmatrix}, \tag{3.11}$$

  where $\Phi_n : \bar{C}_{n-1}^{1,0} \to C_n^{1,0}$ denotes an operator that extrapolates $(\tilde{\mathbf{y}}, \tilde{\mathbf{z}})$ continuously from $(t_{n-1}, t_n]$ to $[t_n, t_{n+1}]$ with corresponding spaces

$$\bar{C}_n^{1,0} := \left\{ (\mathbf{y}, \mathbf{z})|_{(t_n, t_{n+1}]} : (\mathbf{y}, \mathbf{z}) \in C_n^{1,0} \right\},$$
$$C_n^{1,0} := C^1([t_n, t_{n+1}], \mathbb{R}^{n_y}) \times C([t_n, t_{n+1}], \mathbb{R}^{n_z}).$$

  The most simple initial guesses are constant functions

$$\tilde{\mathbf{y}}_n^{(0)}(t) = \tilde{\mathbf{y}}(t_n), \quad \mathbf{z}_n^{(0)}(t) = \tilde{\mathbf{z}}(t_n) \quad (\text{f.a. } t \in [t_n, t_{n+1}])$$

  which results in approximation errors proportional to the window size H_n. Approximations of higher order may be obtained by using higher degree polynomials. In any case, these extrapolation operators satisfy uniform Lipschitz conditions independent of the window size (see [1]).
- *Iteration step*: the k-th iteration step in the dynamic iteration scheme (with $k = 1, \ldots, k_n$) defines a mapping

$$\begin{pmatrix} \tilde{\mathbf{y}}_n^{(k-1)} \\ \tilde{\mathbf{z}}_n^{(k-1)} \end{pmatrix} \to \begin{pmatrix} \tilde{\mathbf{y}}_n^{(k)} \\ \tilde{\mathbf{z}}_n^{(k)} \end{pmatrix} := \Psi_n \begin{pmatrix} \tilde{\mathbf{y}}_n^{(k-1)} \\ \tilde{\mathbf{z}}_n^{(k-1)} \end{pmatrix} \qquad \text{with } \Psi_n = \begin{pmatrix} \Psi_{\mathbf{y},n} \\ \Psi_{\mathbf{z},n} \end{pmatrix}, \tag{3.12}$$

  $\Psi_n : C_n^{1,0} \to C_n^{1,0}$. Here we assume k_n to denote the finite number of iterations to be performed in the n-th window ($[t_n, t_{n+1}]$). Regarding the general setting (3.1), the iteration operator Ψ_n is implicitly defined via splitting functions $\mathbf{F}$ and $\mathbf{G}$ by

solving the initial value problem

$$\dot{\tilde{\mathbf{y}}}_n^{(k)} = \mathbf{F}(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}) \tag{3.13a}$$

$$0 = \mathbf{G}(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}) \tag{3.13b}$$

with initial value

$$\tilde{\mathbf{y}}_n^{(k)}(t_n) = \tilde{\mathbf{y}}_n^{(k-1)}(t_n). \tag{3.13c}$$

The splitting functions $\mathbf{F}$ and $\mathbf{G}$ can be chosen as arbitrarily smooth functions provided that they are related to the right-hand-sides $\mathbf{f}$ and $\mathbf{g}$ of the DAE system (3.1) by the compatibility conditions

$$\mathbf{F}(\mathbf{y}, \mathbf{y}, \mathbf{z}, \mathbf{z}) = \mathbf{f}(\mathbf{y}, \mathbf{z}), \qquad \mathbf{G}(\mathbf{y}, \mathbf{y}, \mathbf{z}, \mathbf{z}) = \mathbf{g}(\mathbf{y}, \mathbf{z}). \tag{3.14}$$

As $\mathbf{f}$, $\mathbf{g}$ are assumed to be sufficiently often differentiable, this is also assumed for $\mathbf{F}$ and $\mathbf{G}$.

Remark 3.4 Notice, that the analytic solution $(\mathbf{y}, \mathbf{z})$ is a fixed-point of the iteration operator Ψ_n due to the compatibility conditions (3.14).

With these notations the dynamic iteration step for window $[t_n, t_{n+1}]$ may be written as composition of the above introduced operators:

$$\begin{pmatrix} \tilde{\mathbf{y}}|_{(t_n, t_{n+1}]} \\ \tilde{\mathbf{z}}|_{(t_n, t_{n+1}]} \end{pmatrix} = (\Psi_n^{k_n} \circ \Phi_n)(\begin{pmatrix} \tilde{\mathbf{y}}|_{(t_{n-1}, t_n]} \\ \tilde{\mathbf{z}}|_{(t_{n-1}, t_n]} \end{pmatrix}). \tag{3.15}$$

We now come back to the question how to exploit the given structure of the coupled DAE system. If the DAE system is given in partitioned form (3.4), we are looking for numerical approximations

$$\tilde{\mathbf{y}}_n = (\mathbf{y}_{1,n}, \ldots, \mathbf{y}_{r,n})^\top, \qquad \tilde{\mathbf{z}}_n = (\mathbf{z}_{1,n}, \ldots, \mathbf{z}_{r,n})^\top$$

in split form. Now the iteration operator Ψ_n should reflect this partitioning. Instead of (3.13), Ψ_n is now implicitly defined by the r initial-value problems

$$\dot{\tilde{\mathbf{y}}}_{i,n}^{(k)} = \mathbf{F}_i(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}), \tag{3.16a}$$

$$0 = \mathbf{G}_i(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}), \tag{3.16b}$$

for $i = 1, \ldots, r$, with initial value

$$\tilde{\mathbf{y}}_{i,n}^{(k)}(t_n) = \tilde{\mathbf{y}}_{i,n}^{(k-1)}(t_n). \tag{3.16c}$$

Again, all splitting functions $\mathbf{F}_i$ and $\mathbf{G}_i$ are related to the right-hand-sides $\mathbf{f}_i$ and $\mathbf{g}_i$ of the DAE system (3.4) by the compatibility conditions

$$\mathbf{F}_i(\mathbf{y},\mathbf{y},\mathbf{z},\mathbf{z}) = \mathbf{f}_i(\mathbf{y},\mathbf{z}), \qquad \mathbf{G}_i(\mathbf{y},\mathbf{y},\mathbf{z},\mathbf{z}) = \mathbf{g}_i(\mathbf{y},\mathbf{z}).$$

And it holds:

$$\mathbf{F}^\top = (\mathbf{F}_1^\top, \ldots, \mathbf{F}_r^\top) \quad \text{and} \quad \mathbf{G}^\top = (\mathbf{G}_1^\top, \ldots, \mathbf{G}_r^\top).$$

In the notation of splitting functions, the following important classes of dynamic iterations schemes for the coupled system (3.4) read as:

$$\mathbf{F}_i(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}) = \mathbf{f}_i(\mathbf{Y}_{i,n}^{(k)}, \mathbf{Z}_{i,n}^{(k)}), \tag{3.17a}$$

$$\mathbf{G}_i(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}) = \mathbf{g}_i(\mathbf{Y}_{i,n}^{(k)}, \mathbf{Z}_{i,n}^{(k)}), \tag{3.17b}$$

for $i = 1, \ldots, r$, with:

- *Picard iteration*:

$$\mathbf{Y}_{i,n}^{(k)} = \tilde{\mathbf{y}}_n^{(k-1)},$$
$$\mathbf{Z}_{i,n}^{(k)} = \tilde{\mathbf{z}}_n^{(k-1)},$$

- *Jacobi iteration:*

$$\mathbf{Y}_{i,n}^{(k)} = (\tilde{\mathbf{y}}_{1,n}^{(k-1)}, \ldots, \tilde{\mathbf{y}}_{i-1,n}^{(k-1)}, \tilde{\mathbf{y}}_{i,n}^{(k)}, \tilde{\mathbf{y}}_{i+1,n}^{(k-1)}, \ldots, \tilde{\mathbf{y}}_{r,n}^{(k-1)})^\top,$$
$$\mathbf{Z}_{i,n}^{(k)} = (\tilde{\mathbf{z}}_{1,n}^{(k-1)}, \ldots, \tilde{\mathbf{z}}_{i-1,n}^{(k-1)}, \tilde{\mathbf{z}}_{i,n}^{(k)}, \tilde{\mathbf{z}}_{i+1,n}^{(k-1)}, \ldots, \tilde{\mathbf{z}}_{r,n}^{(k-1)})^\top,$$

- *Gauss-Seidel iteration:*

$$\mathbf{Y}_{i,n}^{(k)} = (\tilde{\mathbf{y}}_{1,n}^{(k)}, \ldots, \tilde{\mathbf{y}}_{i,n}^{(k)}, \tilde{\mathbf{y}}_{i+1,n}^{(k-1)}, \ldots, \tilde{\mathbf{y}}_{r,n}^{(k-1)})^\top,$$
$$\mathbf{Z}_{i,n}^{(k)} = (\tilde{\mathbf{z}}_{1,n}^{(k)}, \ldots, \tilde{\mathbf{z}}_{i,n}^{(k)}, \tilde{\mathbf{z}}_{i+1,n}^{(k-1)}, \ldots, \tilde{\mathbf{z}}_{r,n}^{(k-1)})^\top.$$

These techniques can be applied to the system derived from overlapping (3.8). The involved multiple computation of certain quantities, enables higher flexibility with respect to stability, as we will see. In the following we discuss a variant of the Gauss-Seidel scheme.

Overlapping technique For a DAE system given in form (3.6) (with an overall algebraic equation), overlapping was introduced in (3.10) with dynamic iteration as the method of choice [2]. For a Gauss-Seidel-like scheme, this overlapping modular time integration reads as follows. First, each subsystem

$$\dot{\mathbf{y}}_i = \mathbf{f}_i(\mathbf{y}_1, \ldots, \mathbf{y}_r, \mathbf{W}_i), \tag{3.18a}$$

$$0 = \mathbf{P}_i^\top \mathbf{g}(\mathbf{y}_1, \ldots, \mathbf{y}_r, \mathbf{W}_i), \tag{3.18b}$$

for $i = 1, \ldots, r$, is equipped with the relation

$$\mathbf{W}_i = (\mathbf{I} - \mathbf{P}_i \mathbf{P}_i^\top)\mathbf{z}_n^{(k-1)} + \mathbf{P}_i \mathbf{P}_i^\top \mathbf{Z}_i^{(k)},$$

introducing an additional stage vector $\mathbf{Z}_i^{(k)}$, which serves as an intermediate approximation for components of $\mathbf{z}$. Translated into splitting functions (and adding the Gauss-Seidel scheme), this leads to system (3.16), with

$$\begin{aligned}
&\mathbf{F}_i(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}) = \mathbf{f}_i(\mathbf{Y}_{i,n}^{(k)}, \mathbf{W}_{i,n}^{(k)}), \quad i = 1, \ldots, r,\\
&\mathbf{G}_i(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)}) = \mathbf{P}_i^\top \mathbf{g}_i(\mathbf{Y}_{i,n}^{(k)}, \mathbf{W}_{i,n}^{(k)}), \quad i = 1, \ldots, r-1,\\
&\mathbf{G}_r(\tilde{\mathbf{y}}_n^{(k)}, \tilde{\mathbf{y}}_n^{(k-1)}, \tilde{\mathbf{z}}_n^{(k)}, \tilde{\mathbf{z}}_n^{(k-1)})\\
&\quad = \begin{pmatrix} \mathbf{P}_r^\top \mathbf{g}_r(\mathbf{Y}_{i,n}^{(k)}, \mathbf{W}_{i,n}^{(k)}) \\ \tilde{\mathbf{z}}_n^{(k)} - \left(\mathbf{I} - \sum_{j=1}^r \mathbf{A}_j \mathbf{P}_j \mathbf{P}_j^\top\right) \tilde{\mathbf{z}}_n^{(k-1)} - \sum_{j=1}^r \mathbf{A}_j \mathbf{P}_j \mathbf{Z}_j^{(k)} \end{pmatrix},
\end{aligned}$$

where we have posed

$$\begin{aligned}
\mathbf{Y}_{i,n}^{(k)} &= (\tilde{\mathbf{y}}_{1,n}^{(k)}, \ldots, \tilde{\mathbf{y}}_{i,n}^{(k)}, \tilde{\mathbf{y}}_{i+1,n}^{(k-1)}, \ldots, \tilde{\mathbf{y}}_{r,n}^{(k-1)})^\top,\\
\mathbf{W}_{i,n}^{(k)} &= (\mathbf{I} - \mathbf{P}_i \mathbf{P}_i^\top)\tilde{\mathbf{z}}_n^{(k-1)} + \mathbf{P}_i \mathbf{P}_i^\top \mathbf{Z}_i^{(k)}.
\end{aligned}$$

Thereby in the last algebraic constraint, we have introduced additional matrices

$$\mathbf{A}_j \in \mathbb{R}^{n_z \times n_z} \qquad (j = 1, \ldots, r)$$

as free parameters for enforcing better stability properties. Notice that the special choice $\mathbf{P}_i = e_i^\top$, $\mathbf{A}_i = \mathbf{I}$ $(i = 1, \ldots, r)$ leads back to system (3.4), solved by the Jacobi-like iteration scheme, while regarding the algebraic part only. Last, the index-1 hypothesis, leads to the assumption that

$$\text{the matrix } \sum_{j=1}^r \mathbf{A}_j \mathbf{P}_j \mathbf{P}_j^\top \text{ is regular.} \tag{3.19}$$

This is the case, if $(\mathbf{A}_1\mathbf{P}_1, \ldots, \mathbf{A}_r\mathbf{P}_r)$ has full rank.

The discussed method corresponds to a dynamic iteration for the overlapping DAE systems (3.10), with slight generalization with respect to the free parameter matrices.

Applying Gauss-Seidel, Jacobi or Picard like dynamic iteration schemes, as well as overlapping modular time integration, to coupled ODEs convergence may always be achieved using sufficiently small window sizes. In the application to coupled

differential-algebraic equations, however, two additional contractivity conditions have to be satisfied to achieve

- Convergence within one window, and
- A stable error propagation in the algebraic components z from one window to another.

This will be the topic of the next sections, where we generalize corresponding results of [1] obtained for a special coupled system to the general case of system (3.1).

3.2.3 Convergence and Stability

In the following we address the convergence of the above defined dynamic iteration schemes. That is, we want to deal with (a) the error within one window, and (b) the transport and amplification of error from window to window. To this end, we introduce the related error notations. First, we derive the error recursions for the error within one window, and prove convergence within each single window under certain stability requirements. Secondly, we treat a finite number of windows and prove the convergence under the related requirements.

We consider an analytic error recursion, thus error due to time integration are not considered explicitly, here. We follow basically [1], but put everything in a more general context as already started in[3]. Thus in fact, only Lemma 3.1 and the exact definition of α differ from the preceding work. Here we adopt a more general viewpoint, to reveal the most prominent structural properties.

3.2.3.1 Error Recursion

Following standard procedures in error analysis, e.g. [21], we define the global error $\boldsymbol{\epsilon}_{\mathbf{y},n}(t)$, $\boldsymbol{\epsilon}_{\mathbf{z},n}(t)$ on the n-th time window ($t \in [t_n, t_{n+1}]$) as the difference of the numerical approximation $\tilde{\mathbf{y}}(t)$, $\tilde{\mathbf{z}}(t)$ and the exact solutions $\mathbf{y}(t)$, $\mathbf{z}(t)$, where the unknowns and hence the errors are split into algebraic and differential components:

$$\begin{pmatrix} \boldsymbol{\epsilon}_{\mathbf{y},n} \\ \boldsymbol{\epsilon}_{\mathbf{z},n} \end{pmatrix} := \begin{pmatrix} (\tilde{\mathbf{y}} - \mathbf{y})\,|_{(t_n,t_{n+1}]} \\ (\tilde{\mathbf{z}} - \mathbf{z})\,|_{(t_n,t_{n+1}]} \end{pmatrix} = \left(\Psi_n^{k_n} \circ \Phi_n\right) \begin{pmatrix} \tilde{\mathbf{y}}|_{(t_{n-1},t_n]} \\ \tilde{\mathbf{z}}|_{(t_{n-1},t_n]} \end{pmatrix} - \begin{pmatrix} \mathbf{y}|_{[t_n,t_{n+1}]} \\ \mathbf{z}|_{[t_n,t_{n+1}]} \end{pmatrix}.$$

Here the numerical approximation on the current time window is given by an approximation on the previous time window, which is extrapolated by Φ_n and then k_n-times iterated by the dynamic iteration operator (e.g. using the Gauss-Seidel scheme).

Classically, the global error is split into contributions from previous windows due to error propagation $\mathbf{e}_{\mathbf{y},n}$, $\mathbf{e}_{\mathbf{z},n}$ and into the errors from the current window $\mathbf{d}_{\mathbf{y},n}$, $\mathbf{d}_{\mathbf{z},n}$, i.e.,

$$\begin{aligned} \boldsymbol{\epsilon}_{\mathbf{y},n} &= \mathbf{e}_{\mathbf{y},n} + \mathbf{d}_{\mathbf{y},n} \\ \boldsymbol{\epsilon}_{\mathbf{z},n} &= \mathbf{e}_{\mathbf{z},n} + \mathbf{d}_{\mathbf{z},n}, \end{aligned} \tag{3.20}$$

where the propagated errors are described by

$$\begin{pmatrix} \mathbf{e}_{\mathbf{y},n} \\ \mathbf{e}_{\mathbf{z},n} \end{pmatrix} := \left(\Psi_n^{k_n} \circ \Phi_n\right) \begin{pmatrix} \tilde{\mathbf{y}}|_{(t_n-1,t_n]} \\ \tilde{\mathbf{z}}|_{(t_n-1,t_n]} \end{pmatrix} - \left(\Psi_n^{k_n} \circ \Phi_n\right) \begin{pmatrix} \mathbf{y}|_{(t_n-1,t_n]} \\ \mathbf{z}|_{(t_n-1,t_n]} \end{pmatrix} \tag{3.21}$$

and the local error contributions by

$$\begin{pmatrix} \mathbf{d}_{\mathbf{y},n} \\ \mathbf{d}_{\mathbf{z},n} \end{pmatrix} := \left(\Psi_n^{k_n} \circ \Phi_n\right) \begin{pmatrix} \mathbf{y}|_{(t_n-1,t_n]} \\ \mathbf{z}|_{(t_n-1,t_n]} \end{pmatrix} - \Psi_n^{k_n} \begin{pmatrix} \mathbf{y}|_{[t_n,t_{n+1}]} \\ \mathbf{z}|_{[t_n,t_{n+1}]} \end{pmatrix}. \tag{3.22}$$

The sum gives indeed global error, since the exact solution $(\mathbf{y}, \mathbf{z})$ is a fixed point of Ψ_n.

To investigate the convergence of the dynamic iteration scheme applied to system (3.1), we introduce a neighborhood $\mathscr{U}_{d,n}$ of the exact solution $\mathbf{x}|_{[t_n,t_{n+1}]} := (\mathbf{y}, \mathbf{z})|_{[t_n,t_{n+1}]}$, defined for any given $d > 0$ by

$$\mathscr{U}_{d,n} = \left\{ (\mathbf{Y}, \mathbf{Z}) \in C_n^{1,0} \,:\, \left|\left|\mathbf{Y} - \mathbf{y}|_{[t_n,t_{n+1}]}\right|\right|_{2,\infty},\ \left|\left|\mathbf{Z} - \mathbf{z}|_{[t_n,t_{n+1}]}\right|\right|_{2,\infty} \le d \right\},$$

with $||\mathbf{v}||_{2,\infty} = \max_t |\mathbf{v}(t)|$, where the maximum is taken on the interval of definition of the vector function $\mathbf{v}(t)$, and $|\cdot|$ denotes the vector 2-norm, that is, the Euclidean norm. Furthermore, we assume:

Assumption 3.1 *For our problem, there exists $d_0 > 0$ such that*

- *The splitting function $\mathbf{F}$ is Lipschitz-continuous in all its coordinates on $\mathscr{U}_{d_0,n}$ with constant $L_{\mathbf{F}} > 0$,* (3.23)
- *The splitting function $\mathbf{G}$ is totally differentiable, and its derivatives are Lipschitz-continuous on $\mathscr{U}_{d_0,n}$,* (3.24)
- *The partial derivative $\mathbf{G}_{\mathbf{z}^{(k)}}$ is invertible on $\mathscr{U}_{d_0,n}$.* (3.25)

The Lipschitz continuity means: for any fixed time t and for any set of vectors $\mathbf{Y}_i, \tilde{\mathbf{Y}}_i \in \mathbb{R}^{n_y}$, $\mathbf{Z}_i, \tilde{\mathbf{Z}}_i \in \mathbb{R}^{n_z}$, $i = 1, 2$, that satisfy $|\mathbf{Y}_i - \mathbf{y}(t)|$, $|\mathbf{Z}_i - \mathbf{z}(t)|$,

$|\tilde{\mathbf{Y}}_i - \mathbf{y}(t)|$, $|\tilde{\mathbf{Z}}_i - \mathbf{z}(t)| \le d_0$, it holds

$$\begin{aligned}&\left|\mathbf{F}(\mathbf{Y}_1, \tilde{\mathbf{Y}}_1, \mathbf{Z}_1, \tilde{\mathbf{Z}}_1) - \mathbf{F}(\mathbf{Y}_2, \tilde{\mathbf{Y}}_2, \mathbf{Z}_2, \tilde{\mathbf{Z}}_2)\right| \\ &\qquad \le L_{\mathbf{F}}(|\mathbf{Y}_1 - \mathbf{Y}_2| + |\tilde{\mathbf{Y}}_1 - \tilde{\mathbf{Y}}_2| + |\mathbf{Z}_1 - \mathbf{Z}_2| + |\tilde{\mathbf{Z}}_1 - \tilde{\mathbf{Z}}_2|)\end{aligned}$$

To have a well-defined solution to (3.13), we have the second and third assumption; it is analogous to the index-1 condition.

For $0 < d < d_0$, let us consider arbitrary functions $\mathbf{X} := (\mathbf{Y}, \mathbf{Z})^\top$ and $\tilde{\mathbf{X}} := (\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}})^\top \in \mathscr{U}_{d,n}$, and denote their image after k dynamic iterations by

$$\begin{aligned}\mathbf{Y}_n^k &:= \Psi_{\mathbf{y},n}^k \mathbf{X}, \quad \mathbf{Z}_n^k := \Psi_{\mathbf{z},n}^k \mathbf{X}, \\ \tilde{\mathbf{Y}}_n^k &:= \Psi_{\mathbf{y},n}^k \tilde{\mathbf{X}}, \quad \tilde{\mathbf{Z}}_n^k := \Psi_{\mathbf{z},n}^k \tilde{\mathbf{X}}.\end{aligned} \tag{3.26}$$

Do not confuse the above definition (3.26) with the notation in (3.17).

Let us denote distances of the y-component after k dynamic iteration by

$$\begin{aligned}\Delta_{\mathbf{y},n}^k(\mathbf{X}, \tilde{\mathbf{X}})(t) &:= \mathbf{Y}_n^k(t) - \tilde{\mathbf{Y}}_n^k(t), \\ \Delta_{\mathbf{z},n}^k(\mathbf{X}, \tilde{\mathbf{X}})(t) &:= \mathbf{Z}_n^k(t) - \tilde{\mathbf{Z}}_n^k(t), \\ \delta_{\mathbf{y},n}^k(\mathbf{X}, \tilde{\mathbf{X}}) &:= ||\Delta_{\mathbf{y},n}^k(\mathbf{X}, \tilde{\mathbf{X}})||_{2,\infty}, \\ \delta_{\mathbf{y},n}^k(\mathbf{X}, \tilde{\mathbf{X}}) &:= ||\Delta_{\mathbf{y},n}^k(\mathbf{X}, \tilde{\mathbf{X}})||_{2,\infty}.\end{aligned} \tag{3.27}$$

Now, we deduce an estimate for the error when the dynamic iteration is applied to the functions in $\mathscr{U}_{d,n}$. As in [3], we have

Lemma 3.1 (Error recursion) *Given a DAE (3.1) – with initial conditions (3.2) – and a dynamic iteration (3.13) with consistent splitting functions* $\mathbf{F}$, $\mathbf{G}$. *For the current time window* $[t_n, t_{n+1}]$ *let Assumption 3.1 hold true. Then there are constants* $C, \tilde{c} > 0$, *such that for* $d < \min\{d_0/C,\ 1/(2\tilde{c})\}$, $H < H_0 := 1/C$, *and*

$$\Psi_n^{k-1}\mathbf{X},\ \Psi_n^{k-1}\tilde{\mathbf{X}} \in \mathscr{U}_{d,n}$$

implies

$$\begin{pmatrix}\delta_{\mathbf{y},n}^k(\mathbf{X}, \tilde{\mathbf{X}}) \\ \delta_{\mathbf{z},n}^k(\mathbf{X}, \tilde{\mathbf{X}})\end{pmatrix} \le \mathbf{K} \begin{pmatrix}\delta_{\mathbf{y},n}^{k-1}(\mathbf{X}, \tilde{\mathbf{X}}) \\ \delta_{\mathbf{z},n}^{k-1}(\mathbf{X}, \tilde{\mathbf{X}})\end{pmatrix} + \begin{pmatrix}1 \\ 0\end{pmatrix} |\Delta_{\mathbf{y},n}^{k-1}(\mathbf{X}, \tilde{\mathbf{X}})(t_n)| \tag{3.28}$$

with

$$\mathbf{K} := \begin{pmatrix}CH & CH \\ C & CH + \alpha_n\end{pmatrix}, \tag{3.29}$$

$$\alpha_n := (1 + \tilde{c}\, d)\, ||\mathbf{G}_{\mathbf{z}^{(k)}}^{-1}\, \mathbf{G}_{\mathbf{z}^{(k-1)}}||_{2,\infty} + Cd. \tag{3.30}$$

Notice $\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n) = \Delta^{0}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n)$ denotes the offset due to differing initial values at the beginning of the n-th time window.

Proof We apply the technique used in [1, 3]. First we show

$$\Psi_n^{k-1}\mathbf{X},\ \Psi_n^{k-1}\tilde{\mathbf{X}} \in \mathscr{U}_{d,n} \implies \delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}),\ \delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \le Cd \tag{3.31}$$

thus $\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}),\ \delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \in \mathscr{U}_{d_0,n}$. On the one hand, we investigate the differential part (3.13a). To this end, we write this equation for any two sets of functions $\tilde{\mathbf{X}} = (\tilde{\mathbf{Y}},\tilde{\mathbf{Z}})^\top$ and $\mathbf{X} = (\mathbf{Y},\mathbf{Z})^\top$ from $\mathscr{U}_{d,n}$, which approximate the solution at the start of the dynamic iteration. Here we take the difference, and time integrate over the interval $[t_n,\tau]$, with $t_n < \tau \le t_{n+1}$. This gives for the k-th iterate, with $k > 0$,

$$\begin{aligned}|\Delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(\tau)| \le\ & |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n)| \\ & + L_F \int_{t_n}^{\tau} \big\{ |\Delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})| + |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})| \\ & \qquad + |\Delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})| + |\Delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})| \big\}\,\mathrm{d}t,\end{aligned} \tag{3.32}$$

using Lipschitz-continuity and consistency of $\mathbf{F}$, and observing that the initial value does not change in the iterations

$$\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n) = \Delta^{k}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n).$$

On the other hand, the algebraic part (3.13b) can be solved for variable $\mathbf{Z}^{(k)} = \hat{\boldsymbol{\phi}}(\mathbf{Y}^{(k)},\mathbf{Y}^{(k-1)},\mathbf{Z}^{(k-1)})$ due to Assumption 3.1. The Lipschitz continuity of $\hat{\boldsymbol{\phi}}$ (due to the implicit function theorem on $\mathscr{U}_{d_0,n}$) leads to

$$\begin{aligned}|\Delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})| &= |\hat{\boldsymbol{\phi}}(\mathbf{Y}^{(k)},\mathbf{Y}^{(k-1)},\mathbf{Z}^{(k-1)}) - \hat{\boldsymbol{\phi}}(\tilde{\mathbf{Y}}^{(k)},\tilde{\mathbf{Y}}^{(k-1)},\tilde{\mathbf{Z}}^{(k-1)})| \\ &\le L_{\hat{\phi}} \left(|\Delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})| + |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})| + |\Delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})| \right)\end{aligned} \tag{3.33}$$

for some $L_{\hat{\phi}} > 0$. Plugging this estimate into (3.32), we obtain

$$\begin{aligned}\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) \le\ & |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n)| \\ & + L_0\, H \left(\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) + \delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) + \delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \right),\end{aligned}$$

where $L_0 := L_F(1 + L_{\hat{\phi}})$. Now solving for $\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})$ gives

$$\begin{aligned}\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) \le\ & \left(1 + \frac{L_0}{1 - L_0\, H}\, H\right) |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n)| \\ & + \frac{L_0}{1 - L_0\, H}\, H \left(\delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) + \delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \right).\end{aligned}$$

The smallness of H, i.e., $H < H_0 = C^{-1}$, implies for $C > L_0$

$$H\, L_0 < H_0\, L_0 < 1.$$

This motivates the definition $c_y := 2L_0/(1 - L_0 H_0)$ from which follows

$$\begin{aligned}\delta^k_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) &\le \Big(1 + \frac{c_y}{2} H\Big)\, |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})(t_n)| \\ &\qquad + \frac{c_y}{2} H \big(\delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) + \delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}})\big) \\ &\le |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})(t_n)| + c_y H \big(\delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) + \delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}})\big), \end{aligned} \tag{3.34}$$

because the initial error at time t_n is smaller than the maximal error on the whole interval

$$|\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})(t_n)| \le \delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}).$$

Estimate (3.34) controls the error propagation for the differential variables, and it is the first line of the estimate (3.28) with the global constant $C > \max\{c_y, L_0\} = c_y$ (so far).

From the estimates (3.34) and (3.33), it is immediate to prove (3.31). In fact, by hypothesis, the $(k-1)$th iterates differ at most by $2d$, so we have

$$\begin{aligned}\delta^k_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) &\le 2(1 + 2c_y H_0)\, d, \\ \delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) &\le 2L_{\hat{\phi}}(3 + 2c_y H_0)\, d. \end{aligned} \tag{3.35}$$

Thus (3.31) holds with

$$C > \max\Big\{c_y,\ 2(1 + 2c_y H_0)d,\ 2L_{\hat{\phi}}(3 + 2c_y H_0)\, d\Big\}.$$

The error recursion estimate for the algebraic component, in the second line of estimate (3.28), can be deduced from the following homotopy of the kth iterates: let $\theta \in [0, 1]$, and let us put

$$\begin{aligned}\mathbf{Y}^{(k),\theta}(t) &:= \theta \tilde{\mathbf{Y}}^k_n(t) + (1 - \theta)\mathbf{Y}^k_n(t). \\ \mathbf{Z}^{(k),\theta}(t) &:= \theta \tilde{\mathbf{Z}}^k_n(t) + (1 - \theta)\mathbf{Z}^k_n(t). \end{aligned}$$

For the splitting function of the algebraic part, we use the short notation

$$\mathbf{G}(\theta) := \mathbf{G}\big(\mathbf{Y}^{(k),\theta},\, \mathbf{Y}^{(k-1),\theta},\, \mathbf{Z}^{(k),\theta},\, \mathbf{Z}^{(k-1),\theta}\big) \qquad \text{and} \qquad \mathbf{G}_{\mathbf{u}}(\theta) := \frac{\partial \mathbf{G}}{\partial \mathbf{u}}(\theta),$$

for any argument $\mathbf{u}$ of $\mathbf{G}$. Notice that $\mathbf{G}(0) = \mathbf{G}(1) = 0$. Thus a version of the fundamental theorem of calculus yields:

$$\begin{aligned}
0 &= \mathbf{G}(1) - \mathbf{G}(0) \\
&= \int_0^1 \Big(\mathbf{G}_{\mathbf{y}^{(k)}}(\theta) \Delta^k_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathbf{G}_{\mathbf{y}^{(k-1)}}(\theta) \Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) \\
&\qquad + \mathbf{G}_{\mathbf{z}^{(k)}}(\theta)\, \Delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathbf{G}_{\mathbf{z}^{(k-1)}}(\theta)\, \Delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) \Big) \mathrm{d}\theta,
\end{aligned} \tag{3.36}$$

since $\frac{\partial}{\partial \theta} \mathbf{Y}^{(k),\theta} = \Delta^k_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})$, and so forth. The upper bound of d in terms of d_0, i.e.,

$$C d \le d_0$$

allows us to use the Lipschitz continuity of $\mathbf{G}_{\mathbf{z}^{(k)}}$ on $\mathscr{U}_{d_0,n}$ (inside the integral of (3.36)). We denote the corresponding constant by L'_G. Together with the above estimate (3.31), we obtain for any time $t \in [t_n, t_{n+1}]$

$$\begin{aligned}
|\mathbf{G}_{\mathbf{u}}(\theta) - \mathbf{G}_{\mathbf{u}}(\hat{\theta})| \;\le\; L_{G'} \Big(& \big| \theta \tilde{\mathbf{Y}}^k_n(t) + (1-\theta) \mathbf{Y}^k_n(t) \\
& - \hat{\theta} \tilde{\mathbf{Y}}^k_n(t) - (1-\hat{\theta}) \mathbf{Y}^k_n(t) \big| \\
& + \cdots + \big| \theta \tilde{\mathbf{Z}}^{k-1}_n(t) + (1-\theta) \mathbf{Z}^{k-1}_n(t) \\
& \qquad - \hat{\theta} \tilde{\mathbf{Z}}^{k-1}_n(t) - (1-\hat{\theta}) \mathbf{Z}^{k-1}_n(t) \big| \Big) \\
= \; L_{G'} |\theta - \hat{\theta}| \Big(& |\Delta^k_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})| + |\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})| \\
& + |\Delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}})| + |\Delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}})| \Big) \le c_g d.
\end{aligned} \tag{3.37}$$

(This defines c_g in the obvious way.) The operator $\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}(0)$ exists due to Assumption 3.1. Left-multiplication of (3.36) by $\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}(0)$ yields:

$$\begin{aligned}
0 = \int_0^1 \mathbf{G}^{-1}_{\mathbf{z}^{(k)}}(0) \Big(& \big(\mathbf{G}_{\mathbf{z}^{(k)}}(0) + \big[\mathbf{G}_{\mathbf{z}^{(k)}}(\theta) - \mathbf{G}_{\mathbf{z}^{(k)}}(0) \big] \big) \Delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) \\
& + \big(\mathbf{G}_{\mathbf{z}^{(k-1)}}(0) + \big[\mathbf{G}_{\mathbf{z}^{(k-1)}}(\theta) - \mathbf{G}_{\mathbf{z}^{(k-1)}}(0) \big] \big) \Delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) \\
& + \big(\mathbf{G}_{\mathbf{y}^{(k)}}(0) + \big[\mathbf{G}_{\mathbf{y}^{(k)}}(\theta) - \mathbf{G}_{\mathbf{y}^{(k)}}(0) \big] \big) \Delta^k_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) \\
& + \big(\mathbf{G}_{\mathbf{y}^{(k-1)}}(0) + \big[\mathbf{G}_{\mathbf{y}^{(k-1)}}(\theta) - \mathbf{G}_{\mathbf{y}^{(k-1)}}(0) \big] \big) \Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) \Big) \, \mathrm{d}\theta.
\end{aligned}$$

The matrices $\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}$, $\mathbf{G}_{\mathbf{z}^{(k-1)}}$, $\mathbf{G}_{\mathbf{y}^{(k)}}$, $\mathbf{G}_{\mathbf{y}^{(k-1)}}$ are uniformly bounded on $\mathscr{U}_{d_0,n}$. Let the constant be denoted by c'_g. Now, this equation is (partially) solved for the first bit of

$\Delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})$. Using

$$\begin{aligned}||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\,\mathbf{G}_{\mathbf{z}^{(k-1)}}||_2 &= ||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\,\mathbf{G}_{\mathbf{z}^{(k-1)}}||_2(0)\\ &= \left|\left|\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\,\mathbf{G}_{\mathbf{z}^{(k-1)}}\right|\right|_2\left(\mathbf{Y}^k_n(t),\mathbf{Y}^k_n(t),\mathbf{Z}^k_n(t),\mathbf{Z}^k_n(t)\right)\end{aligned}$$

and applying the maximum norm in time as well as (3.37) gives

$$\begin{aligned}\delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \le &\left(||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\,\mathbf{G}_{\mathbf{z}^{(k-1)}}||_{2,\infty}+\frac{\tilde{c}}{2}d\right)\delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})\\ &+\frac{\tilde{c}}{2}\,d\,\delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})+c_h\left(\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})+\delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})\right)\end{aligned}$$

with $c_h := (c_g d + c'_g)\,c'_g$ and $\tilde{c} := 2\,c_g c'_g$. Inserting the estimate for $\delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})$ (3.34), we deduce, having H and d small enough, such that $d < 1/(2\tilde{c})$, the estimate

$$\begin{aligned}\delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) &\le (1+\tilde{c}d)c_h\left(|\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n)|+(1+c_yH)\delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})\right) \qquad (3.38)\\ &\quad+(1+\tilde{c}d)\left(c_hc_yH+||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\,\mathbf{G}_{\mathbf{z}^{(k-1)}}||_{2,\infty}+\frac{\tilde{c}}{2}d\right)\delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}})\\ &\le (1+\tilde{c}d)c_h(2+c_yH_0)\delta^{k-1}_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) \qquad (3.39)\\ &\quad+(1+\tilde{c}d)\left(c_hc_yH+||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\,\mathbf{G}_{\mathbf{z}^{(k-1)}}||_{2,\infty}+\frac{\tilde{c}}{2}d\right)\delta^{k-1}_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}),\end{aligned}$$

($H < H_0$). Finally, summing up, the global constant C should be large enough to state (3.31) from (3.34), (3.35) and to obtain from estimate (3.39) the claim (3.28) with (3.29). Hence we conclude

$$\begin{aligned}C > \max\Big\{&c_y,\ 2(1+2c_yH_0)\,d,\ 2L_{\hat{\phi}}(3+2c_yH_0)\,d\\ &(1+\tilde{c}d)c_h(2+c_yH_0),(1+\tilde{c}d)c_hc_y,\frac{\tilde{c}}{2}\Big\}.\end{aligned}$$

Then (3.34) and (3.39) yield the recursion (3.28), our claim. □

When iteratively applying Lemma 3.1, one can deduce the following rather technical result, which is proven for an analogous recursion in [1]:

Proposition 3.1 (Recursion Estimate) *Let the splitting functions fulfill the assumptions of Lemma 3.1 and $\alpha_n < 1$, $C > \alpha_n$, then there is a constant C_0*

such that

$$\begin{pmatrix} \delta^k_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) \\ \delta^k_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \end{pmatrix} \leq \begin{pmatrix} C(4C+1)H\mu_n^{\max(0,k-2)} & 4CH\mu_n^{k-2} \\ 4C\mu_n^{k-1} & \mu_n^k + (\mu_n - \alpha_n)^k \end{pmatrix} \begin{pmatrix} \delta^0_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}}) \\ \delta^0_{\mathbf{z},n}(\mathbf{X},\tilde{\mathbf{X}}) \end{pmatrix} + \begin{pmatrix} 1 + C_0 H \\ C_0 \end{pmatrix} \cdot \delta^0_{\mathbf{y},n}(\mathbf{X},\tilde{\mathbf{X}})(t_n) \tag{3.40}$$

with

$$\mu_n = \mu(\alpha_n, H) := \alpha_n + \frac{2CH}{\frac{\alpha_n}{2C} + \sqrt{H}} \tag{3.41}$$

is satisfied for all $k \geq 1$ *and for all* $H \leq H_0$.

This result is proven for a similar setting in [1]. It is established using the same arguments as in the proof of Theorem 3.1 for the local error: the iteration error is determined by the powers of its matrix $\mathbf{K}$ as given in (3.29) and the computation of the eigenvalues and eigenvectors as in (3.44) proofs the claim.

Next, we will employ the above estimates to show that the mapping is indeed a fixed-point operator.

3.2.3.2 Contraction and Local Error

We consider in this section the local error as defined in Eq. (3.22) only, where the error of a single iteration starting from exact data is analyzed

$$\mathbf{d}_{\mathbf{y},n} = \Delta^{k_n}_{\mathbf{y},n}(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n \mathbf{x}|_{(t_{n-1},t_n]}),$$

and analogously for $\mathbf{d}_{\mathbf{z},n}$ with $\mathbf{x} := (\mathbf{y},\mathbf{z})^\top$ in both cases. We follow [3] and the strategy from [1] to proof the following result, that is already predicted in [19]. It shows that the crucial point in the coupling lies in the algebraic-to-algebraic coupling, which is represented by the additional DAE-contraction factor α.

Theorem 3.1 (Contraction) *The splitting functions shall fulfill the assumptions of Lemma 3.1 including the index-1 assumption. Furthermore, let* $\mathbf{x}$ *denote our exact solution. Then for* d *and* $H < H_0$ *small enough the map*

$$\delta_n^{k-1}(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n \mathbf{x}|_{(t_{n-1},t_n]}) \mapsto \delta_n^k(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n \mathbf{x}|_{(t_{n-1},t_n]}) \tag{3.42}$$

is strongly contractive for all k *provided that*

$$||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\, \mathbf{G}_{\mathbf{z}^{(k-1)}}||_{2,\infty} < 1. \tag{3.43}$$

Proof We show contractivity for the constant extrapolation with $\tilde{y}_n^{(0)} = \tilde{y}(t_n)$, $z_n^{(0)} = \tilde{z}(t_n)$, from which the contraction for any higher order polynomial extrapolation follows automatically.

By induction we setup the error recursion (3.28) in $\mathscr{U}_{d,n}$: as induction basis, we have for $k = 0$ and $\tau \in [t_n, t_{n+1}]$

$$|\Delta^0_{\mathbf{y},n}(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n\mathbf{x}|_{(t_{n-1},t_n]})|(\tau) = \Big|\int_{t_n}^{\tau} \mathbf{f}(\mathbf{y},\, \mathbf{z})\, \mathrm{d}t\, \Big| \leq\, c_f\, H,$$

where $c_f := ||\mathbf{f}(\mathbf{y}, \mathbf{z})||_{2,\infty}$. Then the index-1 assumption implies for $\mathbf{z}$

$$\begin{aligned}\left|\Delta^0_{\mathbf{z},n}(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n\mathbf{x}|_{(t_{n-1},t_n]})\right|(\tau) \;&\leq \left|\boldsymbol{\phi}(\Phi_{\mathbf{y},n}\mathbf{x}|_{(t_{n-1},t_n]}) - \boldsymbol{\phi}(\mathbf{y})\right| \\ &\leq\; L_\phi \left|\Phi_{\mathbf{y},n}\mathbf{x}|_{(t_{n-1},t_n]} - \mathbf{y}\right| \\ &\leq\; c_f L_\phi\, H;\end{aligned}$$

thus choosing H sufficient small, such that $c_f\,(L_\phi + 1)\, H_0 < 1$ (and $H < H_0$), we obtain an extrapolation, which lies in the neighborhood of the solution: $\Phi_n\mathbf{x} \in \mathscr{U}_{d,n}$.

Recall the definition of the matrix $\mathbf{K}$ (3.29), which denotes an upper bound on the error recursion. Now, the mapping (3.42) is contractive if the spectral radius $\rho(\mathbf{K}) < 1$. The eigenvalues of $\mathbf{K}$ are

$$\lambda_{1,2}(\mathbf{K}) = \frac{1}{2}\Big(\alpha_n + 2\,CH \pm \sqrt{\alpha_n^2 + 4\,C^2 H}\,\Big), \tag{3.44}$$

Therefore $\alpha_n < 1$ is sufficient for contraction provided that d and H_0 are small enough. Inspecting (3.30), this translates into:

$$||\mathbf{G}^{-1}_{\mathbf{z}^{(k)}}\, \mathbf{G}_{\mathbf{z}^{(k-1)}}||_{2,\infty} < 1.$$

Eventually applying Lemma 3.1 iteratively and using

$$\delta^0_{\mathbf{y},n}(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n\mathbf{x}|_{(t_{n-1},t_n]})(t_n) = 0$$

concludes the proof. □

Remark 3.5 (Convergence Order of Iteration) The eigenvalues of $\mathbf{K}$, defined in (3.29), suggest a certain order of convergence (i.e., for the asymptotics as $H \to 0$) for the dynamic iteration (3.13): For the rate of convergence, we use Taylor expansion of the square root term in $\lambda(\mathbf{K})$ (3.44) and find

$$\sqrt{\alpha_n^2 + 4\,C^2 H} = \alpha_n\,(1 + 2\,C^2 H/\alpha_n^2) + \mathscr{O}(H^2).$$

This suggests a order of $\alpha_n + \mathscr{O}(H)$, if α_n does not vanish and $4\,C^2 H < \alpha_n^2$. For $\alpha_n = 0$, we have order $\sqrt{H}$.

We notice: convergence of the DAE-distributed time integration depends on the stability of the algebraic-to-algebraic component coupling (3.43) (and, of course, depends on the mentioned hypothesis). Thus modeling coupling is important for DAEs and should be organized if possible in a way, s.t. contractivity (stability) is directly given. The following important special case avoids these kinds of dependencies:

Corollary 3.1 (Simple Coupling) *Let the hypothesis of Lemma 3.1 be fulfilled.*

(i) If no algebraic constraint depends on an old algebraic variable, i.e.,

$$\mathbf{G}_{\mathbf{z}^{(k-1)}} = 0 \tag{3.45}$$

then contraction is archived with $\alpha_n = 0$.

(ii) If no algebraic constraint depends on an old algebraic or a differential variable, i.e.,

$$\mathbf{G}_{\mathbf{z}^{(k-1)}} = 0 \quad \textit{and} \quad \mathbf{G}_{\mathbf{y}^{(k-1)}} = 0 \tag{3.46}$$

then the contraction is archived with convergence order H.

Proof We discuss (3.36) for the special cases in which the given partial derivatives vanish.

(i) The assumption $\mathbf{G}_{\mathbf{z}^{(k-1)}} = 0$ gives the following estimate for the algebraic part replacing (3.39)

$$\delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) \leq \leq C\, \delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) + CH\, \delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}).$$

This is (3.28) with $\alpha_n = 0$. This result is in the spirit of the numerical DAE-theory (cf. [21]).

(ii) Analogously $\mathbf{G}_{\mathbf{y}^{(k-1)}} = \mathbf{G}_{\mathbf{z}^{(k-1)}} = 0$ yields

$$\begin{aligned}\delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) \leq & (1 + \tilde{c}d)c_h\Big(|\Delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})(t_n)| + c_y H \delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})\Big) \\ & + (1 + \tilde{c}d)c_h c_y H\, \delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}})\end{aligned}$$

replacing (3.38). This give the error recursion

$$\delta^k_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) \leq CH\, \delta^{k-1}_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}}) + CH\, \delta^{k-1}_{\mathbf{z},n}(\mathbf{X}, \tilde{\mathbf{X}}) + C\,|\Delta^0_{\mathbf{y},n}(\mathbf{X}, \tilde{\mathbf{X}})(t_n)|$$

which unveils a contraction operator $\mathbf{K} = \mathcal{O}(H)$ and hence implies a convergence order of H, cf. Remark 3.5. Only the initial offset cannot be improved.

□

Now still following the strategy from [1], we prove estimates for the local and propagated errors, and conclude from those results the overall stability and convergence of the method for the nth time window.

Proposition 3.2 (Local Error) *Let the assumptions of Lemma 3.1 be fulfilled, then the recursion (3.40) with μ_n (3.41) of that Lemma holds. Moreover, then there is for a sufficiently small $H < H_0$ a constant $C_{\mathbf{d}^\star}$, independent of H, α_n and k_n, such that the local error is bounded by*

$$||\mathbf{d}_{\mathbf{y},n}|| + H||\mathbf{d}_{\mathbf{z},n}|| \leq C_{\mathbf{d}^\star} H \delta_n^0$$

where the right-hand-side is given in terms of the extrapolation errors

$$\begin{aligned}\delta_n^0 :=& \mu_n^{\max(0,k_n-2)} \delta_{\mathbf{y},n}^0(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n\mathbf{x}|_{(t_{n-1},t_n]}) \\ &+ \mu_n^{k_n-1} \delta_{\mathbf{z},n}^0(\mathbf{x}|_{[t_n,t_{n+1}]}, \Phi_n\mathbf{x}|_{(t_{n-1},t_n]})\end{aligned}$$

Proof The proof of Theorem 3.1 showed that $\Phi_n\mathbf{x}|_{(t_{n-1},t_n]} \in \mathscr{U}_{d,n}$ for H sufficiently small. Therefore applying Proposition 3.1 to the specific functions

$$\mathbf{X} := \mathbf{x}|_{(t_n,t_{n+1}]} \quad \text{and} \quad \tilde{\mathbf{X}} := \Phi_n\mathbf{x}|_{(t_{n-1},t_n]}$$

where $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ is the exact solution. Notice $\delta_{\mathbf{y},n}^0(\mathbf{X}, \tilde{\mathbf{X}})(t_n) = 0$ holds, since the initial values are equal. Summation of the two equations in (3.40) yields the claimed estimate. □

This proves convergence for one window (for $k_n \to \infty$), since $\mu_n < 1$ for H sufficiently small. Next, we have to address the error transport, since the iteration is stopped after a finite number of iterations and we are not performing $k_n \to \infty$ in the numerical treatment.

3.2.3.3 Stability and Convergence for Windowing Technique

To obtain convergence and stability of the method on multiple windows it is crucial to control the error propagation from the previous window to the current one, hence we need to inspect

$$\mathbf{e}_{\mathbf{y},n} = \Delta_{\mathbf{y},n}^{k_n}(\Phi_n\mathbf{x}|_{(t_{n-1},t_n]}, \Phi_n\tilde{\mathbf{x}}|_{(t_{n-1},t_n]})$$

and the similar expression for $\mathbf{e}_{\mathbf{z},n}$ (here $\mathbf{x}$ denotes the analytic solution and $\tilde{\mathbf{x}}$ an approximation). The following result is again a consequence of Proposition 3.1 (cf. [1]):

Proposition 3.3 (Propagation Error) *Let an continuous extrapolation (3.11) be given, that is of accuracy $\mathscr{O}(H)$ and satisfies a uniform Lipschitz condition (with*

constant L_Φ) and a dynamic iteration (3.13), which fulfill the assumptions of Proposition 3.1 with $\mu_n < 1$, then there is a constant $C_{\mathbf{e}^\star} > 0$, such that the propagation error is bounded by

$$\begin{pmatrix} ||\mathbf{e}_{\mathbf{y},n}|| \\ ||\mathbf{e}_{\mathbf{z},n}|| \end{pmatrix} \leq \begin{pmatrix} 1 + C_{\mathbf{e}^\star} & C_{\mathbf{e}^\star} H \\ C_{\mathbf{e}^\star} & \alpha_{n^\star} \end{pmatrix} \cdot \begin{pmatrix} ||\boldsymbol{\epsilon}_{\mathbf{y},n-1}|| \\ ||\boldsymbol{\epsilon}_{\mathbf{z},n-1}|| \end{pmatrix} \tag{3.47}$$

with $\alpha_n^\star$ depending on the Lipschitz constant L_Φ of the extrapolation operator

$$\alpha_n^\star := L_\Phi(\mu_n^{k_n} + (\mu_n - \alpha_n)^{k_n}) \tag{3.48}$$

Proof When applying Proposition 3.1 to the extrapolation of exact and erroneous functions of the previous time window

$$\mathbf{X} := \Phi_n \mathbf{x}|_{(t_{n-1},t_n]} \quad \textit{and} \quad \tilde{\mathbf{X}} := \Phi_n \tilde{\mathbf{x}}|_{(t_{n-1},t_n]}\,,$$

we will have an offset in the initial values (at t_n), which is bounded by the total error on the interval

$$||\Delta^0_{\mathbf{y},n}(\Phi_n \mathbf{x}|_{(t_{n-1},t_n]}, \Phi_n \tilde{\mathbf{x}}|_{(t_{n-1},t_n]})(t_n)|| \leq ||\mathbf{y} - \tilde{\mathbf{y}}||_{(t_{n-1},t_n]}.$$

Furthermore the extrapolation operator is a uniformly Lipschitz continuous mapping with Lipschitz constant L_Φ, hence we have

$$\begin{aligned} \delta^0_n(\Phi_n \mathbf{x}|_{(t_{n-1},t_n]}, \Phi_n \tilde{\mathbf{x}}|_{(t_{n-1},t_n]}) &\leq L_\Phi \begin{pmatrix} ||\mathbf{y} - \tilde{\mathbf{y}}||_{(t_{n-1},t_n]} \\ ||\mathbf{z} - \tilde{\mathbf{z}}||_{(t_{n-1},t_n]} \end{pmatrix} \\ &= L_\Phi \begin{pmatrix} ||\mathbf{e}_{\mathbf{y},n-1}|| \\ ||\mathbf{e}_{\mathbf{z},n-1}|| \end{pmatrix}, \end{aligned}$$

that completes together with Eq. (3.40) of Proposition 3.1 the proof. □

Now bringing all pieces together, we obtain the following result on stability and convergence

Theorem 3.2 (Stability) *Let a continuous extrapolation Φ (3.11) be given, that is of accuracy order $\mathcal{O}(H)$ and satisfies a uniform Lipschitz condition (L_Φ), further a dynamic iteration (3.13), where the splitting functions $\mathbf{F}$, $\mathbf{G}$ are consistent and for the current time window $[t_n, t_{n+1}]$ let Assumption 3.1 hold true, furthermore the contractivity constant is bounded*

$$\alpha_n \leq \bar{\alpha} < 1 \qquad \textit{and} \qquad L_\Phi \alpha_n^{k_n} \leq \bar{\alpha}$$

and the numerical solution remains close to the exact solution

$$||\boldsymbol{\epsilon}_{\mathbf{y},m}|| + ||\boldsymbol{\epsilon}_{\mathbf{z},m}|| \leq d \quad \textit{for } 0 \leq m < n,$$

then there is a constant $C^\star > 0$, independent of n and H, such that the total error on the time window $[t_n, t_{n+1}]$ is bounded by

$$||\epsilon_{\mathbf{y},n}|| + ||\epsilon_{\mathbf{z},n}|| \leq C^\star \max_{0 \leq m < n} \delta_m^0 \leq d \tag{3.49}$$

all for a sufficiently small step size $0 < H < H_0$.

Proof According to Eq. (3.41) we have $\mu_n = \alpha_n + \mathscr{O}(H)$ and by assumption $L_\Phi \alpha_n^{k_n} \leq \bar{\alpha}$, hence

$$\alpha_n^\star = L_\Phi \left((\mu_n^{k_n})^{k_n} + (\mu_n - \alpha_n)^{k_n} \right) = L_\Phi \left((\alpha_n + \mathscr{O}(H))^{k_n} + \mathscr{O}(H)^{k_n} \right) < 1,$$

and therefore the maximum is bounded as well

$$\alpha^\star := \max_{0 \leq m \leq n} \alpha_m^\star < 1.$$

Now combining the results from Propositions 3.2 and 3.3 yields

$$\begin{pmatrix} ||\epsilon_{\mathbf{y},n}|| \\ ||\epsilon_{\mathbf{z},n}|| \end{pmatrix} \leq \begin{pmatrix} 1 + C_{\mathbf{e}^\star} & C_{\mathbf{e}^\star} H \\ C_{\mathbf{e}^\star} & \alpha^\star \end{pmatrix} \cdot \begin{pmatrix} ||\epsilon_{\mathbf{y},n-1}|| \\ ||\epsilon_{\mathbf{z},n-1}|| \end{pmatrix} + \begin{pmatrix} C_{\mathbf{d}^\star} H \delta_n^0 \\ C_{\mathbf{d}^\star} \delta_n^0 \end{pmatrix}$$

and this proves the left half of (3.49), the right bound is enforceable since the extrapolation error $\delta_m^0 = \mathscr{O}(H)$ decreases with the step size. □

One can use Theorem 3.2 to prove by induction that the numerical solution remains close to the exact solution, analogously to the application in [1], then the overall convergence and stability follows by

Corollary 3.2 (Global Convergence and Stability) *Let the assumptions of Theorem 3.2 be fulfilled, then there is a constant $C^\star$, such that the estimate holds*

$$||\tilde{y} - y||_{[0,t_e]} + ||\tilde{z} - z||_{[0,t_e]} \leq C^\star \cdot \max_{0 \leq n < N} \delta_m^0,$$

where δ_m^0 is the extrapolation error on the m-th window.

This result shows convergence and stability, since the global error can by controlled in terms of the step size H, which determines the extrapolation error.

3.3 Applications in Electrical Engineering

In this section we show how the dynamic iteration theory can be used to study the convergence of iteration schemes for the main coupled models introduced in the previous chapter. These problems basically stem from chip design.

3.3.1 Refined Network Models

We consider an electric network with semiconductor devices, modeled by drift-diffusion equations. The electric network is described by the MNA equations, which can be written in the form:

$$\begin{aligned} \mathbf{A}_C \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}_C(\mathbf{A}_C^T\mathbf{u}) + \mathbf{A}_R\mathbf{r}(\mathbf{A}_R^T\mathbf{u}) + \mathbf{A}_L\mathbf{i}_L + \mathbf{A}_V\mathbf{i}_V + \mathbf{A}_I\mathbf{i}_I + \boldsymbol{\lambda} &= 0, \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}_L(\mathbf{i}_L) - \mathbf{A}_L^T\mathbf{u} &= 0, \\ \mathbf{A}_V^T\mathbf{u} - \mathbf{v}_V &= 0. \end{aligned} \tag{3.50}$$

This system is supplemented with initial data for the differential part,

$$\mathbf{P}_C\mathbf{u}(t_0) = \mathbf{P}_C\mathbf{u}_0, \quad \mathbf{i}_L(t_0) = \mathbf{i}_{L,0}. \tag{3.51}$$

Here, we have $\mathbf{P}_C = \mathbf{I} - \mathbf{Q}_C$, where $\mathbf{Q}_C$ is a projector onto the null-space of $\mathbf{A}_C^T$, and we are assuming index-1 conditions for the uncoupled MNA system.

The above equations are coupled, through the current term $\boldsymbol{\lambda}$, to the drift-diffusion equations which describe the devices contained in the circuit. Here, as an exemplification, we use the space-discretization derived in the previous Chapter, by means of the Box Integration method. Then, assuming for simplicity that the circuit contains a single device, this device will be described by the time-dependent vectors $\boldsymbol{\phi}$, $\boldsymbol{n}$, $\boldsymbol{p}$, comprising the values of the electric potential ϕ, the electron concentration n and the hole concentration p, evaluated on the inner grid points, and by the time-dependent vectors $\boldsymbol{\phi}^\partial$, $\boldsymbol{n}^\partial$, $\boldsymbol{p}^\partial$, comprising the values of ϕ, n and p on the boundary grid points. As we have seen in the previous Chapter, these vector functions satisfy the following equations:

$$\begin{aligned} \mathbf{A}_\phi\boldsymbol{\phi} + \mathbf{A}_\phi^\partial\boldsymbol{\phi}^\partial &= \mathbf{b}_\phi(\boldsymbol{n}, \boldsymbol{p}), \\ \mathbf{A}^\partial\boldsymbol{\phi} + \boldsymbol{\phi}^\partial &= \mathbf{b}_\phi^\partial(\mathbf{u}_D), \\ \mathbf{A}_0\frac{\mathrm{d}\boldsymbol{n}}{\mathrm{d}t} + \mathbf{A}_n(\boldsymbol{\phi})\boldsymbol{n} + \mathbf{A}_n^\partial(\boldsymbol{\phi})\boldsymbol{n}^\partial &= \mathbf{b}_n(\boldsymbol{n}, \boldsymbol{p}), \\ \mathbf{A}^\partial\boldsymbol{n} + \boldsymbol{n}^\partial &= \mathbf{b}_n^\partial, \\ \mathbf{A}_0\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} + \mathbf{A}_p(\boldsymbol{\phi})\boldsymbol{p} + \mathbf{A}_p^\partial(\boldsymbol{\phi})\boldsymbol{p}^\partial &= \mathbf{b}_p(\boldsymbol{n}, \boldsymbol{p}), \\ \mathbf{A}^\partial\boldsymbol{p} + \boldsymbol{p}^\partial &= \mathbf{b}_p^\partial. \end{aligned} \tag{3.52}$$

These equations must be supplemented with initial data for the differential variables,

$$\boldsymbol{n}(t_0) = \boldsymbol{n}_0, \quad \boldsymbol{p}(t_0) = \boldsymbol{p}_0. \tag{3.53}$$

The MNA equations (3.50) and the device equations (3.52) are coupled by means of appropriate relations which express, on the one hand, the boundary electric potential $\mathbf{u}_D$ in (3.52) in terms of the network variables (network-to-device coupling), and on the other hand, the device current source term $\boldsymbol{\lambda}$ in (3.50) in terms of the device variables (device-to-network coupling). The network-to-device coupling is given by:

$$\mathbf{u}_D = \mathbf{S}_D^T \mathbf{u}. \tag{3.54}$$

The device-to-network coupling is more involved. In Chap. 1 we have introduced two alternative formulations. In the first formulation, the device-to-network coupling relation is given by:

$$\boldsymbol{\lambda} = \mathbf{A}_D \mathbf{i}_D, \quad \mathbf{i}_D = \mathbf{A}^c \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} + \mathbf{A}_n^c(\boldsymbol{\phi})\mathbf{n} + \mathbf{A}_p^c(\boldsymbol{\phi})\mathbf{p}, \tag{3.55}$$

with $\mathbf{A}_D = \mathbf{S}_D \hat{\mathbf{A}}_D$. This formulation is problematic, because of the appearance of the time derivative of $\boldsymbol{\phi}$, which is an algebraic variable for the uncoupled device system. Thus, after the coupling, the set of differential variables generally differs from the union of the differential variables for the network and the device system, considered as uncoupled.

For this reason, we consider the alternative formulation,

$$\boldsymbol{\lambda} = \mathbf{A}_D \frac{\mathrm{d}}{\mathrm{d}t}(\tilde{\mathbf{C}}_D \mathbf{A}_D^\top \mathbf{u}) + \mathbf{A}_D \tilde{\mathscr{I}}_D, \quad \tilde{\mathscr{I}}_D = \mathbf{A}_n^c(\boldsymbol{\phi})\mathbf{n} + \mathbf{A}_p^c(\boldsymbol{\phi})\mathbf{p}. \tag{3.56}$$

In this formulation, it is simpler to see that the differential variables for the coupled system are $\mathbf{P}_C\mathbf{u}$, $\mathbf{i}_L$, $\boldsymbol{n}$, $\boldsymbol{p}$, provided the additional condition

$$\mathbf{A}_D^T \mathbf{Q}_C = 0. \tag{3.57}$$

Under this condition, we can identify the differential and algebraic components, and we set

$$\mathbf{y}_c = \begin{pmatrix} \mathbf{P}_C\mathbf{u} \\ \mathbf{i}_L \end{pmatrix}, \qquad \mathbf{z}_c = \begin{pmatrix} \mathbf{Q}_C\mathbf{u} \\ \mathbf{i}_V \end{pmatrix},$$

and

$$\mathbf{y}_d = \begin{pmatrix} \boldsymbol{n} \\ \boldsymbol{p} \end{pmatrix}, \qquad \mathbf{z}_d = \begin{pmatrix} \boldsymbol{\phi} \\ \boldsymbol{n}^\partial \\ \boldsymbol{p}^\partial \\ \boldsymbol{\phi}^\partial \end{pmatrix}.$$

Then, using the standard reduction to differential and algebraic equations, by means of appropriate projectors, the two systems of equations can be written in the following form:

$$\begin{aligned} \dot{\mathbf{y}}_c &= \mathbf{f}_c(\mathbf{y}_c, \mathbf{z}_c, \tilde{\boldsymbol{\lambda}}), \\ 0 &= \mathbf{g}_c(\mathbf{y}_c, \mathbf{z}_c), \\ \dot{\mathbf{y}}_d &= \mathbf{f}_d(\mathbf{y}_d, \mathbf{z}_d), \\ 0 &= \mathbf{g}_d(\mathbf{y}_d, \mathbf{z}_d, \mathbf{u}_D), \end{aligned} \tag{3.58}$$

with

$$\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}(\mathbf{y}_d, \mathbf{z}_d) := \mathbf{A}_D[\mathbf{A}_n^c(\boldsymbol{\phi})\boldsymbol{n} + \mathbf{A}_p^c(\boldsymbol{\phi})\boldsymbol{p}].$$

Also, we have

$$\mathbf{u}_D = \mathbf{S}_D^T(\mathbf{P}_C\mathbf{u} + \mathbf{Q}_C\mathbf{u}) = \mathbf{S}_D^T(\mathbf{y}_c + \mathbf{z}_c),$$

so the above system becomes

$$\begin{aligned} \dot{\mathbf{y}}_c &= \mathbf{f}_c^*(\mathbf{y}_c, \mathbf{z}_c, \mathbf{y}_d, \mathbf{z}_d), \\ 0 &= \mathbf{g}_c(\mathbf{y}_c, \mathbf{z}_c), \\ \dot{\mathbf{y}}_d &= \mathbf{f}_d(\mathbf{y}_d, \mathbf{z}_d), \\ 0 &= \mathbf{g}_d^*(\mathbf{y}_d, \mathbf{z}_d, \mathbf{y}_c, \mathbf{z}_c). \end{aligned} \tag{3.59}$$

Next, we apply the dynamic iteration theory, expounded in this Chapter, to the coupled system (3.59), by using the Gauss-Seidel method. We can use to different strategies: circuit-device iteration, and device-circuit iteration. For the circuit-device coupling, we have[1]:

$$\begin{aligned} \dot{\tilde{\mathbf{y}}}_c^{(k)} &= \mathbf{f}_c^*(\tilde{\mathbf{y}}_c^{(k)}, \tilde{\mathbf{z}}_c^{(k)}, \tilde{\mathbf{y}}_d^{(k-1)}, \tilde{\mathbf{z}}_d^{(k-1)}), \\ 0 &= \mathbf{g}_c(\tilde{\mathbf{y}}_c^{(k)}, \tilde{\mathbf{z}}_c^{(k)}), \\ \dot{\tilde{\mathbf{y}}}_d^{(k)} &= \mathbf{f}_d(\tilde{\mathbf{y}}_d^{(k)}, \tilde{\mathbf{z}}_d^{(k)}), \\ 0 &= \mathbf{g}_d^*(\tilde{\mathbf{y}}_c^{(k)}, \tilde{\mathbf{z}}_c^{(k)}, \tilde{\mathbf{y}}_d^{(k)}, \tilde{\mathbf{z}}_d^{(k)}). \end{aligned} \tag{3.60}$$

We can observe that in this case the matrix $\mathbf{G}_{\mathbf{z}^{(k-1)}}$ is identically zero. Thus, by Corollary 3.1, this scheme leads to an unconditionally, strongly contractive map.

[1] For simplicity we omit the subscript n.

By contrast, if we consider the device-circuit iteration scheme, we have

$$\begin{aligned}
\dot{\tilde{\mathbf{y}}}_d^{(k)} &= \mathbf{f}_d(\tilde{\mathbf{y}}_d^{(k)}, \tilde{\mathbf{z}}_d^{(k)}),\\
0 &= \mathbf{g}_d^*(\tilde{\mathbf{y}}_c^{(k-1)}, \tilde{\mathbf{z}}_c^{(k-1)}, \tilde{\mathbf{y}}_d^{(k)}, \tilde{\mathbf{z}}_d^{(k)}),\\
\dot{\tilde{\mathbf{y}}}_c^{(k)} &= \mathbf{f}_c^*(\tilde{\mathbf{y}}_c^{(k)}\tilde{\mathbf{z}}_c^{(k)}, \tilde{\mathbf{y}}_d^{(k)}, \tilde{\mathbf{z}}_d^{(k)}),\\
0 &= \mathbf{g}_c(\tilde{\mathbf{y}}_c^{(k)}, \tilde{\mathbf{z}}_c^{(k)}).
\end{aligned}$$

and the condition (3.43), in Theorem 3.1, which ensure the contractivity of the dynamic iteration map, is verified if and only if

$$\left\| \left(\frac{\partial \mathbf{g}_d^*}{\partial \mathbf{z}_d^{(k)}} \right)^{-1} \frac{\partial \mathbf{g}_d^*}{\partial \mathbf{z}_c^{(k-1)}} \right\| < 1.$$

Explicitly, this condition is equivalent to

$$\left\| (\mathbf{A}_\phi - \mathbf{A}_\phi^\partial \mathbf{A}^\partial)^{-1} \frac{\partial \mathbf{b}_\phi^\partial}{\partial \mathbf{u}_D} \mathbf{S}_D^T \mathbf{Q}_C \right\| < 1, \tag{3.61}$$

where, by definition, we have

$$\frac{\partial b_{\phi,i}^\partial}{\partial u_{D,j}} = \begin{cases} 1, \text{ if } \mathbf{x}_i \in \Gamma_{D,j}, \\ 0, \text{ otherwise.} \end{cases}$$

The matrix $\mathbf{A}_\phi - \mathbf{A}_\phi^\partial \mathbf{A}^\partial$ depends on the space-discretization, so the above condition implies a smallness assumption on the spacing of the grid, unless $\mathbf{S}_D^T \mathbf{Q}_C = 0$. This condition is stronger than the additional topological condition (3.57), and in general is not satisfied.

In conclusion, the circuit-device iteration scheme is preferable to the device-circuit scheme.

3.3.2 *Electro-Thermal Coupling*

Similarly, the coupling of heat effects with electric systems plays an important role in electric circuit simulation, see Sect. 2.2.2 and, e.g., [3, 14]. Spatial discretization of certain thermal models (e.g. for heat conduction) can yield a DAE-ODE coupling. This type of coupling is less problematic, since no coupling via old algebraic variables will occur. Therefore no contraction is needed in this case, see, e.g., [3]. In other models, e.g. with patches, the situation is a bit more complicated—for details we refer to [14].

3.3.3 Coupled System of Electric Networks and Maxwell's Magnetoquasistatic Equations and Their Properties

3.3.3.1 Introduction

Let us apply the dynamic iteration theory to the field/circuit coupling as introduced in Sect. 2.2.3.

There are two subproblems, on one hand the electric circuit and on the other hand the magnetoquasistatic field problem ("eddy current problem"). The circuit equations can abstractly by described by the semi-explicit initial value problem

$$\begin{aligned} \dot{\mathbf{y}}_c &= \mathbf{f}_c(\mathbf{y}_c, \mathbf{z}_c, \mathbf{i}_\mathrm{m}), \quad \text{with} \quad \mathbf{y}_c(0) = \mathbf{y}_{c,0} \\ \mathbf{0} &= \mathbf{g}_c(\mathbf{y}_c, \mathbf{z}_c, \mathbf{i}_\mathrm{m}), \end{aligned} \tag{3.62}$$

similar to the derivation in Sect. 3.3.1. We assume an index-1 circuit, i.e., the topological conditions as given in [17] to be fulfilled, such that

$$\frac{\partial \mathbf{g}_c}{\partial \mathbf{z}_c} \text{ is nonsingular.} \tag{3.63}$$

The unknowns are given by

$$\mathbf{y}_c := (\mathbf{q}, \boldsymbol{\phi})^\top, \quad \mathbf{z}_c := (\mathbf{u}\, \mathbf{i}_\mathrm{L}, \mathbf{i}_\mathrm{V})^\top, \quad \mathbf{i}_\mathrm{m} := (\mathbf{i}_\mathrm{str}, \mathbf{i}_\mathrm{sol})^\top.$$

where $\mathbf{u}$ denotes node potentials, $\mathbf{q}$ charges, $\boldsymbol{\phi}$ fluxes and $\mathbf{i}_\mathrm{L}$, $\mathbf{i}_\mathrm{V}$ currents through inductances and voltage sources. The additional variables $\mathbf{i}_\mathrm{str}$ and $\mathbf{i}_\mathrm{sol}$ define currents through stranded and solid conductors and are treated separately since they are determined by the field model. This field model describes a relation between those currents and the voltage drops

$$\mathbf{v}_\mathrm{str} := \mathbf{A}_\mathrm{str}^\top \mathbf{u}, \qquad \mathbf{v}_\mathrm{sol} := \mathbf{A}_\mathrm{sol}^\top \mathbf{u}$$

by one common PDE for the whole domain Ω and an additional differential equation for the coupling of each stranded ($k = 1, \ldots, N_\mathrm{sol}$) and solid conductor ($l = 1, \ldots, N_\mathrm{sol}$) in the corresponding subdomains $\Omega_{\mathrm{str},k}$ and $\Omega_{\mathrm{sol},l}$ to the circuit

$$\sigma \frac{\partial \boldsymbol{A}}{\partial t} + \nabla \times (\nu \nabla \times \boldsymbol{A}) = \sum_k \boldsymbol{\chi}_{\mathrm{str},k} \, (\mathbf{i}_\mathrm{str})_k + \sum_l \sigma \boldsymbol{\chi}_{\mathrm{sol},l} \, (\mathbf{v}_\mathrm{sol})_l \tag{3.64a}$$

$$\int_\Omega \boldsymbol{\chi}_{\mathrm{str},k} \cdot \frac{\partial \boldsymbol{A}}{\partial t} \, \mathrm{d}\Omega = (\mathbf{v}_\mathrm{str})_k - (\mathbf{R}_\mathrm{str})_{k,k} \cdot (\mathbf{i}_\mathrm{str})_k \,, \tag{3.64b}$$

$$\int_\Omega \sigma \boldsymbol{\chi}_{\mathrm{sol},l} \cdot \frac{\partial \boldsymbol{A}}{\partial t} \, \mathrm{d}\Omega = (\mathbf{G}_\mathrm{sol})_{l,l} \cdot (\mathbf{v}_\mathrm{sol})_l - (\mathbf{i}_\mathrm{sol})_l \,, \tag{3.64c}$$

with Coulomb gauging, flux wall boundary and initial conditions

$$\nabla \cdot \boldsymbol{A} = 0, \qquad \boldsymbol{A} \times \mathbf{n}_{\perp} = 0 \text{ on } \partial\Omega, \qquad \boldsymbol{A} = \boldsymbol{A}_0 \text{ at } t = t_0, \tag{3.64d}$$

where $\boldsymbol{A}$ denotes the magnetic vector potential, $\mathbf{n}_{\perp}$ is the vector normal to the boundary, $\nu = \nu(\boldsymbol{A})$ the reluctivity tensor and σ the conductivity tensor vanishing on stranded conductor domains, i.e.

$$\sigma \frac{\partial \boldsymbol{A}}{\partial t}\Big|_{\Omega_{\mathrm{str},k}} = \mathbf{0} \tag{3.65}$$

since it is assumed that the diameter of the individual strands in the those conductors is thinner that the skin depth. Each *distribution function* $\boldsymbol{\chi}_{\mathrm{str},k}$ and $\boldsymbol{\chi}_{\mathrm{sol},k}$ distributes the current in the corresponding conductor domains $\Omega_{\mathrm{str},k}$ and $\Omega_{\mathrm{sol},k}$. The diagonal matrices

$$(\mathbf{R}_{\mathrm{str}})_{k,k} = \int_{\Omega} \frac{1}{f_{\mathrm{str}}} \sigma^{-1} \boldsymbol{\chi}_{\mathrm{str},k} \cdot \boldsymbol{\chi}_{\mathrm{str},k} \mathrm{d}\Omega \quad \text{and} \quad (\mathbf{G}_{\mathrm{sol}})_{l,l} = \int_{\Omega} \sigma \boldsymbol{\chi}_{\mathrm{sol},l} \cdot \boldsymbol{\chi}_{\mathrm{sol},l} \mathrm{d}\Omega$$

describe lumped DC resistances $\mathbf{R}_{\mathrm{str}}$ for stranded conductors using the fill factor $f_{\mathrm{str}} \in (0, 1]$ and DC conductivities $\mathbf{G}_{\mathrm{sol}}$ for the solid conductors.

According to Sect. 2.2.3.3, the spatial discretization of the field PDE yields a DAE, describing a unique vector potential in time. The discrete field problem reads in the FIT notation, [12]

$$\mathbf{M}_{\sigma} \frac{\mathrm{d}}{\mathrm{d}t} \widehat{\mathbf{a}} + \mathbf{K}_{\nu}(\widehat{\widehat{\mathbf{b}}}) \widehat{\mathbf{a}} = \mathbf{Q}_{\mathrm{str}} \mathbf{i}_{\mathrm{str}} + \mathbf{M}_{\sigma} \mathbf{Q}_{\mathrm{sol}} \mathbf{v}_{\mathrm{sol}} \tag{3.66a}$$

$$\mathbf{Q}_{\mathrm{str}}^{\top} \frac{\mathrm{d}}{\mathrm{d}t} \widehat{\mathbf{a}} = \mathbf{v}_{\mathrm{str}} - \mathbf{R}_{\mathrm{str}} \mathbf{i}_{\mathrm{str}} \tag{3.66b}$$

$$\mathbf{Q}_{\mathrm{sol}}^{\top} \mathbf{M}_{\sigma} \frac{\mathrm{d}}{\mathrm{d}t} \widehat{\mathbf{a}} = \mathbf{G}_{\mathrm{sol}} \mathbf{v}_{\mathrm{sol}} - \mathbf{i}_{\mathrm{sol}}, \tag{3.66c}$$

where $\widehat{\mathbf{a}}$ denotes the discrete magnetic vector potential with consistent initial value $\widehat{\mathbf{a}}(0) = \widehat{\mathbf{a}}_0$, the mass matrix $\mathbf{M}_{\sigma}$ is symmetric positive semi-definite describing the conductivities, $\mathbf{K}_{\nu}$ is a symmetric curl-curl matrix composed of the discrete curl-operators and the reluctivities. We assume a regularization on $\mathbf{K}_{\nu}$, e.g. by the Coulomb gauging, such that

$$\widehat{\mathbf{e}}^{\top} \Big(\alpha \mathbf{M}_{\sigma} + \frac{\partial}{\partial \widehat{\mathbf{a}}} \big(\mathbf{K}_{\nu}(\widehat{\widehat{\mathbf{b}}}) \widehat{\mathbf{a}} \big) \Big) \widehat{\mathbf{e}} > 0 \quad \text{for all } \widehat{\mathbf{e}} \neq \mathbf{0} \text{ and } \alpha \neq 0. \tag{3.67}$$

which ensures a (symmetric) positive definite matrix pencil and hence allows for the application of iterative solvers, e.g. Krylov subspace methods, [13]. The matrix $\mathbf{Q} = [\mathbf{Q}_{\mathrm{sol}}, \mathbf{Q}_{\mathrm{str}}]$ is the discrete counterpart to the characteristic functions $\boldsymbol{\chi}$ in the

continuous model, it imposes currents and voltages onto edges in the computational grid.

The matrices of lumped resistances and conductivities can be extracted from the discrete field model by

$$\mathbf{R}_{\mathrm{str}} := \mathbf{Q}_{\mathrm{str}}^{\top}\mathbf{M}_{\sigma,\mathrm{str}}^{+}\mathbf{Q}_{\mathrm{str}} \qquad \text{and} \qquad \mathbf{G}_{\mathrm{sol}} := \mathbf{Q}_{\mathrm{sol}}^{\top}\mathbf{M}_{\sigma}\mathbf{Q}_{\mathrm{sol}}, \tag{3.68}$$

where $\mathbf{M}_{\sigma,\mathrm{str}}^{+}$ is the pseudo inverse of a conductivity matrix with positive conductivities in the stranded conductor domains.

3.3.3.2 Coupling Analysis

To apply the schemes of Sect. 3.2 to the field/circuit coupled problem we need to verify, that the DAE index of the field problem is one, see Eq. (3.4), and that the contractivity condition (3.30) is fulfilled. Here comes the decomposition of the field system into differential and algebraic parts into play: according to the Lemma the field system (3.66) can be interpreted as the semi-explicit initial value problem

$$\begin{aligned} \dot{\mathbf{y}}_m &= \mathbf{f}_m(\mathbf{y}_m, \mathbf{z}_{ma}, \mathbf{v}_{\mathrm{c}}), \quad \text{with} \quad \mathbf{y}_m(0) = \mathbf{y}_{m,0}, \\ \mathbf{0} &= \mathbf{g}_{ma}(\mathbf{y}_m, \mathbf{z}_{ma}), \\ \mathbf{0} &= \mathbf{g}_{mb}(\mathbf{y}_m, \mathbf{z}_{ma}, \mathbf{z}_{mb}), \end{aligned} \tag{3.69}$$

where $\mathbf{y}_m := \mathscr{P}_\sigma \widehat{\mathbf{a}}$, $\mathbf{z}_{ma} := \mathscr{Q}_\sigma \widehat{\mathbf{a}}$, $\mathbf{z}_{mb} := (\mathbf{i}_{\mathrm{str}}, \mathbf{i}_{\mathrm{sol}})^{\top}$ and $\mathbf{v}_{\mathrm{c}} := (\mathbf{v}_{\mathrm{str}}, \mathbf{v}_{\mathrm{sol}})^{\top}$. Now using this semi-explicit problem formulation we obtain the following result

Lemma 3.2 *The field System* (3.66) *is an index-1 DAEs, i.e.,*

$$\frac{\partial \mathbf{g}_m}{\partial \mathbf{z}_m} \text{ is nonsingular,}$$

for given voltages $\mathbf{v}_{\mathrm{str}}$ *and* $\mathbf{v}_{\mathrm{sol}}$ *and the matrix pencil of the curl-curl equation* (3.67) *is positive definite.*

Proof The DAE-indices of Systems (3.66) and (3.69) are equal, since the second system was obtained only by merely algebraic operations, proof of Lemma 2.1, hence it is sufficient to consider the more abstract system only; with the definitions

$$\mathbf{g}_m := (\mathbf{g}_{ma}, \mathbf{g}_{mb})^{\top} \qquad \text{and} \qquad \mathbf{z}_m := (\mathbf{z}_{ma}, \mathbf{z}_{mb})^{\top}$$

the index-1 requirement corresponds to the non-singularity of the Jacobian

$$\frac{\partial \mathbf{g}_m}{\partial \mathbf{z}_m} = \begin{pmatrix} \frac{\partial \mathbf{g}_{ma}}{\partial \mathbf{z}_{ma}} & \frac{\partial \mathbf{g}_{ma}}{\partial \mathbf{z}_{mb}} \\ \frac{\partial \mathbf{g}_{mb}}{\partial \mathbf{z}_{ma}} & \frac{\partial \mathbf{g}_{mb}}{\partial \mathbf{z}_{mb}} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathbf{g}_{ma}}{\partial \mathbf{z}_{ma}} & \mathbf{0} \\ \frac{\partial \mathbf{g}_{mb}}{\partial \mathbf{z}_{ma}} & \mathbf{I} \end{pmatrix},$$

where the upper left vanishes, since there is no coupling in the algebraic part of the curl-curl equation. The lower right block $\partial \mathbf{g}_{mb}/\partial \mathbf{z}_{mb} = \mathbf{I}$ is the identity, since the function $\mathbf{g}_{mb}$ is just an assignment of the currents through solid and stranded conductors and hence trivially regular. On the other hand the upper left block coming from Eq. (2.163) reads

$$\begin{aligned} \frac{\partial \mathbf{g}_{ma}}{\partial \mathbf{z}_{ma}} &= \frac{\partial}{\partial \mathbf{z}_{ma}} \left(\mathscr{Q}_\sigma \mathbf{K}_\nu \mathscr{P}_\sigma^\top \mathbf{y}_2 + \mathscr{Q}_\sigma \mathbf{K}_\nu \mathscr{Q}_\sigma^\top \mathbf{z}_{ma} \right) \\ &= \frac{\partial}{\partial \widehat{\mathbf{a}}} \left(\mathscr{Q}_\sigma \mathbf{K}_\nu(\widehat{\mathbf{b}})\widehat{\mathbf{a}} \right) \frac{\partial \widehat{\mathbf{a}}}{\partial \mathbf{z}_{ma}} = \mathscr{Q}_\sigma \frac{\partial}{\partial \widehat{\mathbf{a}}} \left(\mathbf{K}_\nu(\widehat{\mathbf{b}})\widehat{\mathbf{a}} \right) \mathscr{Q}_\sigma^\top \end{aligned}$$

which is surely regular since the matrix pencil was assumed to be positive definite and thus the transformation

$$\mathscr{Q}_\sigma \left(\lambda \left(\mathbf{M}_\sigma + \mathbf{Q}_{\mathrm{str}} \mathbf{R}_{\mathrm{str}}^{-1} \mathbf{Q}_{\mathrm{str}}^\top \right) + \frac{\partial}{\partial \widehat{\mathbf{a}}} \left(\mathbf{K}_\nu(\widehat{\mathbf{b}})\widehat{\mathbf{a}} \right) \right) \mathscr{Q}_\sigma^\top$$

is still positive definite because $\mathscr{Q}_\sigma^\top$ has full rank and the mass matrix does not contribute by construction to this submatrix

$$\mathscr{Q}_\sigma \left(\mathbf{M}_\sigma + \mathbf{Q}_{\mathrm{str}} \mathbf{R}_{\mathrm{str}}^{-1} \mathbf{Q}_{\mathrm{str}}^\top \right) \mathscr{Q}_\sigma^\top = \mathbf{0} \,.$$

Hence we obtain the positive definiteness of $\frac{\partial}{\partial \widehat{\mathbf{a}}} \left(\mathbf{K}_\nu(\widehat{\mathbf{b}})\widehat{\mathbf{a}} \right)$, furthermore this shows the regularity of the minor $\partial \mathbf{g}_{ma}/\partial \mathbf{z}_{ma}$ and thus we have finally proven System (3.66) being an index-1 DAE. □

Theorem 3.3 *The field/circuit coupled system* (3.62)+(3.69)*, i.e.,*

$$\begin{aligned} \dot{\mathbf{y}}_c &= \mathbf{f}_c(\mathbf{y}_c, \mathbf{z}_c, \boxed{\mathbf{z}_m}), \qquad \text{and} \qquad \dot{\mathbf{y}}_m = \mathbf{f}_m(\mathbf{y}_m, \mathbf{z}_m, \boxed{\mathbf{z}_c}), \\ \mathbf{0} &= \mathbf{g}_c(\mathbf{y}_c, \mathbf{z}_c, \boxed{\mathbf{z}_m}), \qquad\qquad\quad \mathbf{0} = \mathbf{g}_m(\mathbf{y}_m, \mathbf{z}_m), \end{aligned}$$

is index-1, if the circuit fulfills the index-1 assumption (3.63)*, and the matrix pencil of the underlying curl-curl equation* (3.67) *is positive definite.*

Proof We proceed similarly to the proof of Lemma 3.2, where we inspected the algebraic constraints for the field DAE. For the algebraic constraints and variables

of the whole coupled system

$$\mathbf{g} := (\mathbf{g}_c, \mathbf{g}_m)^\top = (\mathbf{g}_c, \mathbf{g}_{ma}, \mathbf{g}_{mb})^\top \quad \text{and} \quad \mathbf{z} := (\mathbf{z}_c, \mathbf{z}_m)^\top = (\mathbf{z}_c, \mathbf{z}_{ma}, \mathbf{z}_{mb})^\top,$$

follows analogously the Jacobian

$$\frac{\partial \mathbf{g}}{\partial \mathbf{z}} = \begin{pmatrix} \frac{\partial \mathbf{g}_c}{\partial \mathbf{z}_c} & \frac{\partial \mathbf{g}_c}{\partial \mathbf{z}_m} \\ \frac{\partial \mathbf{g}_m}{\partial \mathbf{z}_c} & \frac{\partial \mathbf{g}_m}{\partial \mathbf{z}_m} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathbf{g}_c}{\partial \mathbf{z}_c} & \frac{\partial \mathbf{g}_c}{\partial \mathbf{z}_m} \\ \mathbf{0} & \frac{\partial \mathbf{g}_m}{\partial \mathbf{z}_m} \end{pmatrix},$$

which is nonsingular, because the index-1 assumption for the circuit guarantees the regularity of $\partial \mathbf{g}_c / \partial \mathbf{z}_c$ and finally Lemma 3.2 gives the regularity of $\partial \mathbf{g}_m / \partial \mathbf{z}_m$. □

To allow the coupling of already existing simulator packages, the coupled system (3.62)+(3.64) is split such both that sub-problems can be computed independently. The dynamic iteration method will call each simulator to integrate the sub-problem on a time window and then exchange the obtained voltages and currents at the synchronization points. During the computation of a sub-problem on a window the data of the other system is frozen and represented by a source. Since each system describes for current/voltage relations, we have to decide which quantities are considered as known for each branch and conductor. This question is crucial for the field/circuit coupling since the DAE-index of the field system and hence the applicability of the dynamic iteration method depends on this decision:

Corollary 3.3 *The field system* (3.66) *is index-1, if all voltages* ($\mathbf{v}_{\text{sol}}$, $\mathbf{v}_{\text{str}}$) *are given, and in all other cases, i.e., given* ($\mathbf{i}_{\text{sol}}$, $\mathbf{v}_{\text{str}}$), ($\mathbf{v}_{\text{sol}}$, $\mathbf{i}_{\text{str}}$) *or* ($\mathbf{i}_{\text{sol}}$, $\mathbf{i}_{\text{str}}$), *it is at least index-2.*

Proof The first part for given voltages is proven by Lemma 3.2, since the currents $\mathbf{z}_{mb}$ in (3.69) can be obtained by just evaluating the algebraic equation

$$\mathbf{0} = \mathbf{g}_{mb}(\mathbf{z}_m, \mathbf{z}_{ma}, \mathbf{z}_{mb})$$

but if instead a current is prescribed, then the function $\mathbf{f}_2$ depends on an unknown voltage ($\mathbf{v}_{\text{str}}$ or $\mathbf{v}_{\text{sol}}$) and hence the coupling equation $\mathbf{g}_{mb}$ must be differentiated once with respect to time to obtain a hidden algebraic constraint for the missing voltage, such that the overall system is at least index-2. □

That it is just index-2 has been shown for the case of given currents in solid conductors in [37] and more generally in [34] was proven, that (3.66) is in fact an index-2 Hessenberg system, [8], with some additional algebraic (index-1) equations due to the singularity of the mass matrix.

3.3.3.3 Field-Circuit Scheme

Now having obtained semi-explicit index-1 formulations of both sub-systems, (3.62) and (3.69), we give a more abstract description of the coupling that fits into the framework of dynamic iteration methods in Sect. 3.2.2, i.e., System (3.4).

On hand we have the circuit DAE-IVP

$$\begin{aligned} \dot{\mathbf{y}}_c &= \mathbf{f}_c(\mathbf{y}_c, \mathbf{z}_c, \boxed{\mathbf{z}_{mb}}), & \mathbf{y}_c(0) &= \mathbf{y}_{c,0}, & \mathbf{y}_c &:= \left(\mathbf{q}^\top, \boldsymbol{\phi}^\top\right)^\top, \\ \mathbf{0} &= \mathbf{g}_c(\mathbf{y}_c, \mathbf{z}_c, \boxed{\mathbf{z}_{mb}}), & & & \mathbf{z}_c &:= \left(\mathbf{u}^\top, \mathbf{i}_{\mathrm{L}}^\top, \mathbf{i}_{\mathrm{V}}^\top\right)^\top, \end{aligned}$$

and on the other hand the field DAE-IVP

$$\begin{aligned} \dot{\mathbf{y}}_m &= \mathbf{f}_m(\mathbf{y}_m, \mathbf{z}_{ma}, \boxed{\mathbf{z}_c}), & \mathbf{y}_m(0) &= \mathbf{y}_{m,0}, & \mathbf{y}_m &= \mathscr{P}_\sigma \widehat{\mathbf{a}}, \\ \mathbf{0} &= \mathbf{g}_{ma}(\mathbf{y}_m, \mathbf{z}_{ma}), & & & \mathbf{z}_{ma} &= \mathscr{Q}_\sigma \widehat{\mathbf{a}}, \\ \mathbf{0} &= \mathbf{g}_{mb}(\mathbf{y}_m, \mathbf{z}_{ma}, \mathbf{z}_{mb}), & & & \mathbf{z}_{mb} &= \left(\mathbf{i}_{\mathrm{str}}^\top, \mathbf{i}_{\mathrm{sol}}^\top\right)^\top \end{aligned}$$

where a slight abuse of notation is introduced when inserting all algebraic circuit unknowns $\mathbf{z}_c$ into $\mathbf{f}_{ma}$ instead of only the actually needed voltage drops $\mathbf{v}$.

Let us discuss a dynamic iteration of Gauss-Seidel type on the time interval $[0, t_e]$, with $1 \leq n \leq N$ windows $[t_n, t_{n+1}] \subset [0, t_e]$ and adequate initial values for each window

$$\begin{pmatrix} \mathbf{y}_{c,n} \\ \mathbf{y}_{m,n} \end{pmatrix} := \begin{pmatrix} \mathbf{y}_c(t_n) \\ \mathbf{y}_m(t_n) \end{pmatrix} =: \mathbf{y}(t_n).$$

We start each iteration with the integration of the field DAE-IVP. It depends on data from the circuit DAE-IVP (denoted by $\mathbf{y}_c$, $\mathbf{z}_c$). These missing data, i.e., the voltage drops $\mathbf{v}$ at the conductors, are unknown at start time. Hence we extrapolate the initial value to the current time. We choose the following constant extrapolation of the differential variables

$$\begin{pmatrix} \mathbf{y}_{c,n}^{(0)} \\ \mathbf{y}_{m,n}^{(0)} \end{pmatrix} = \Phi_{\mathbf{y},n}(\mathbf{y}|_{(t_{n-1},t_n)}) := \mathbf{y}(t_n), \tag{3.70}$$

from which a consistent supplement $\mathbf{z}_{1,n}^{(0)}$ and $\mathbf{z}_{2,n}^{(0)}$ for the algebraic variables is obtained. Providing this data the field system can be solved for the first time ($k = 1$)

on the time window $[t_n, t_{n+1}]$

$$\begin{aligned}\dot{\mathbf{y}}_m^{(k)} &= \mathbf{f}_m(\mathbf{y}_m^{(k)}, \mathbf{z}_{ma}^{(k)}, \boxed{\mathbf{z}_c^{(k-1)}}), \quad \mathbf{y}_m^{(k)}(0) = \mathbf{y}_{m,n},\\ \mathbf{0} &= \mathbf{g}_{ma}(\mathbf{y}_m^{(k)}, \mathbf{z}_{ma}^{(k)}),\\ \mathbf{0} &= \mathbf{g}_{mb}(\mathbf{y}_m^{(k)}, \mathbf{z}_{ma}^{(k)}, \mathbf{z}_{mb}^{(k)}).\end{aligned} \tag{3.71a}$$

Having obtained a first algebraic iterate $\mathbf{z}_{mb}^{(k)}$ (at $k = 1$) for the currents $\mathbf{i}_{\text{str}}$ and $\mathbf{i}_{\text{sol}}$ we may continue to solve the circuit subsystem

$$\begin{aligned}\dot{\mathbf{y}}_c^{(k)} &= \mathbf{f}_c(\mathbf{y}_c^{(k)}, \mathbf{z}_c^{(k)}, \boxed{\mathbf{z}_{mb}^{(k)}}), \quad \mathbf{y}_c^{(k)}(0) = \mathbf{y}_{c,n},\\ \mathbf{0} &= \mathbf{g}_c(\mathbf{y}_c^{(k)}, \mathbf{z}_c^{(k)}, \boxed{\mathbf{z}_{mb}^{(k)}}).\end{aligned} \tag{3.71b}$$

After the first iteration the functions $\mathbf{y}^{(k)}$ and $\mathbf{z}^{(k)}$ ($k = 1$) are obtained and we may restart the scheme for $k + 1$ until k_n sweeps of the n-th time window are completed. After that we proceed to the next time window ($n + 1$) and start again with the constant extrapolation (3.11) and the following k_{n+1} Gauss-Seidel iterations, until the end of the integration interval $[0, t_e]$ is reached.

In this application of the Gauss-Seidel-Scheme the splitting functions, as introduced in Eqs. (3.13), are defined as the mappings

$$\mathbf{F}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}, \mathbf{z}^{(k-1)}) := \begin{pmatrix} \mathbf{f}_c(\mathbf{y}_c^{(k)}, \mathbf{z}_c^{(k)}, \mathbf{z}_{mb}^{(k)}) \\ \mathbf{f}_m(\mathbf{y}_m^{(k)}, \mathbf{z}_{ma}^{(k)}, \mathbf{z}_c^{(k-1)}) \end{pmatrix}$$

and

$$\mathbf{G}(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}) := \begin{pmatrix} \mathbf{g}_c(\mathbf{y}_c^{(k)}, \mathbf{z}_c^{(k)}, \mathbf{z}_{mb}^{(k)}) \\ \mathbf{g}_{ma}(\mathbf{y}_m^{(k)}, \mathbf{z}_{ma}^{(k)}) \\ \mathbf{g}_{mb}(\mathbf{y}_m^{(k)}, \mathbf{z}_{ma}^{(k)}, \mathbf{z}_{mb}^{(k)}) \end{pmatrix}$$

where $\mathbf{G}$ does not depend on an old algebraic variable $\mathbf{z}^{(k-1)}$. Therefore Corollary 3.1 applies here

Corollary 3.4 *The dynamic iteration scheme* (3.71) *is is unconditionally stable on the time interval* $[0, t_e]$.

By contrast, if we consider the circuit-field Gauss-Seidel scheme or a Jacobi-Scheme, we have to deal with the partial derivatives as in the case of the device-circuit scheme in Sect. 3.3.

3.3.3.4 Multimethod and Multirate Benefits

Besides advantages in software engineering, there are other benefits for the coupling of simulation packages: the most important are benefits due the use of problem-specific methods for time integration (multimethod) and the possibility of different step sizes (multirate) for each subproblem. Thus adaptive time stepping schemes will apply automatically the time step sizes, that are inherently given by the subproblem and not the minimum of all those step size as in the monolithic approach. This will yield a computational more efficient integration.

The benefit of the multimethod approach is obviously present since the packages for field simulation are commonly applying the implicit Euler scheme or implicit Runge-Kutta schemes for time integration, [28], while circuit simulators are typically based on schemes from the BDF family, [21].

The advantage due multirate behavior depend highly on the specific configuration of the problem considered, since different time scales do not occur in the field/circuit coupling as natural as in the thermal coupling (Sect. 3.3.2), where the effects are clearly from multiphysics. In contrast to this, the described phenomena of the field/circuit coupling originate all from Maxwell's equations and hence there is no guarantee of multirate effects. Even if present, e.g. due to switches or filters, the partition of the subsystems according to the network DAE and field PDE model does not necessarily correspond to time constants of different magnitude. Moreover a partition into fast and slow switching components would require to split the circuit at arbitrary nodes and could hence destroy the advantages of the simulator coupling approach.

Anyhow if the circuit contains only a small number of devices that are active at a time, while others remain latent and the field model belongs to such a latent part, then the computational expensive solution of the PDE can be obtained using less time steps than the circuit solution requires. This *weak* coupling will be naturally exploited by the dynamic iteration method, if the step sizes for the time integration of the sub-problems are chosen accordingly (or are automatically determined by an adaptive time integrator) and increase its efficiency when compared to a single-rate integration method.

For such configurations an efficient special case of the dynamic iteration method is the *multirate co-simulation*, where only one sweep ($k_n = 1$) is made, but obviously smaller synchronization steps have to be chosen.

3.3.3.5 Numerical Example

Let us discuss a classical example from engineering: a transformer is excited at its primary coil by an alternating voltage source with $\mathbf{v}_{\text{eff}} = 250\,\text{V}$ at $f = 50\,\text{Hz}$ and is connected to a rectifier circuit at its secondary coil with a load resistance of $R_{\text{load}} = 100\,\Omega$, Fig. 3.1a. The diodes are described by Shockley's model with $I_s = 10\mu\text{A}$. The transformer is represented by a PDE in 3D, discretized by EM

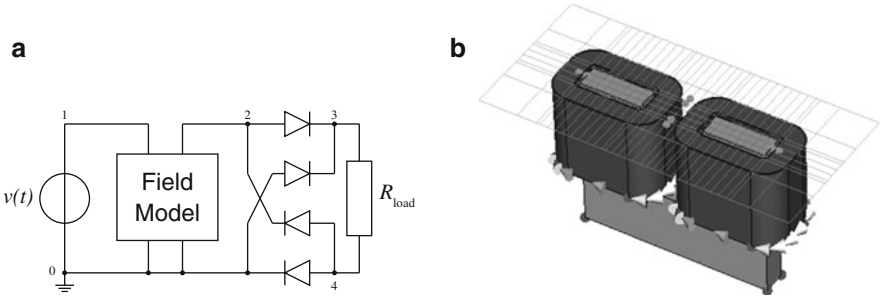

Fig. 3.1 (**a**) Example circuit: rectifier. (**b**) Field model: transformer

`Studio` from CST Software,[2] where each coil is connected to the circuit using the a stranded conductor model.

The simulation software was implemented within the `COMSON DP` and applies either the classic monolithic strategy or the dynamic iteration method by using Gauss-Seidel's scheme. Simulation results are presented in Fig. 3.2.

3.3.3.6 Summary

In this section we shown how nonlinear index-analysis of DAEs can be used to prove the convergence of dynamic iteration methods applied coupled problems. In the case of Maxwell's magnetoquasistatic equations coupled to electric circuits we find that there is no dependence of the algebraic equations on previous algebraic iterates. This guarantees an index-1 problem in the case of monolithic coupling and furthermore proofs convergence and stability of the proposed dynamic iteration scheme. To obtain this result it is not even necessary to validating the contractivity condition given in Sect. 3.2.3.

3.4 Coupled Numerical Simulations of the Thermal Effects in Silicon Devices

In this section we analyze the discretization of the model presented in Sect. 2.2.4, describing the coupling between the transport of electrons and the heating of the crystal lattice. Results of the simulation of a MOSFET with a nanoscale channel are presented and the influence of the thermal effects on the electrical performance is analyzed. This section is based on reference [30, 31] where the interested reader can find more details.

[2] see http://www.cst.com/

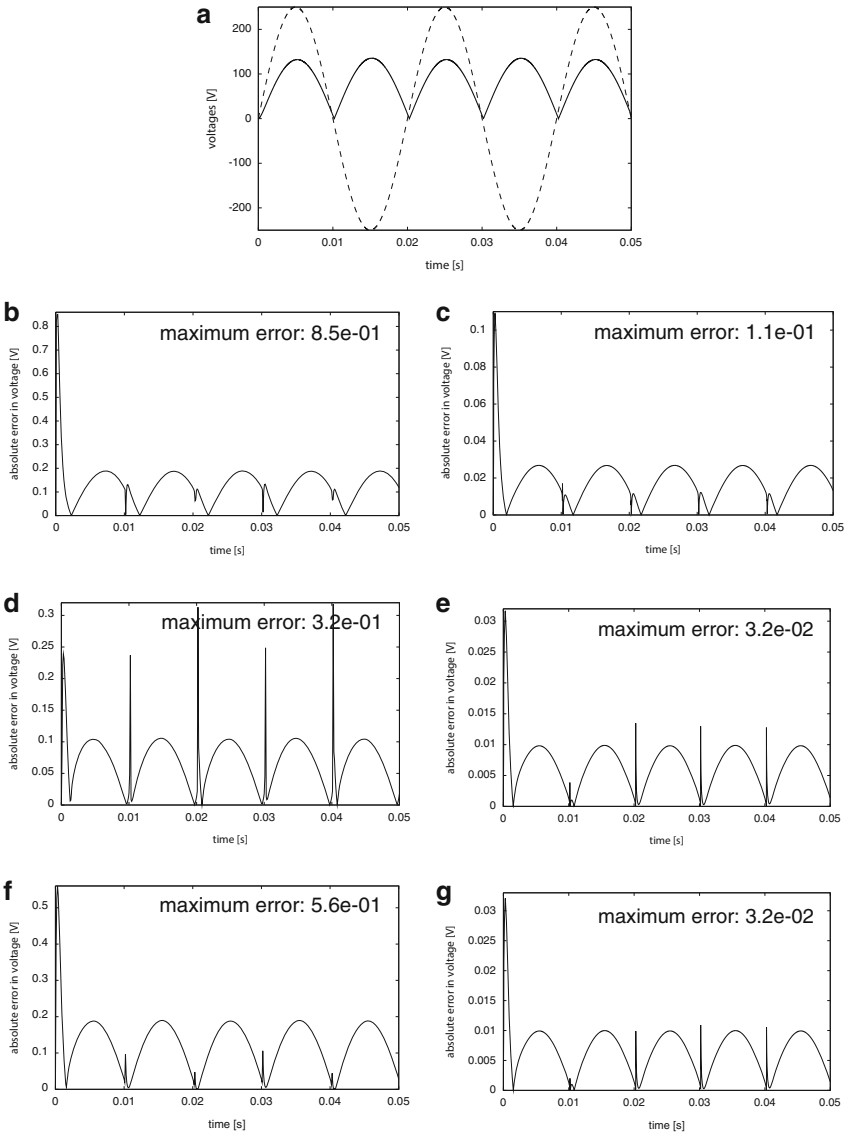

Fig. 3.2 *Numerical Example.* Error plots with respect to the reference solution. (**a**) Input (*dashed*) and output (*solid*) voltages of the reference solution, obtained by monolithic simulation with step size $H = 1e-5$. (**b**) Error in output voltage in multirate co-simulation with window size $H = 1e-4$. (**c**) Error in output voltage in multirate co-simulation with window size $H = 1e-5$. (**d**) Error in output voltage in dynamic iteration with 3 sweeps and window size $H = 1e-4$. (**e**) Error in output voltage in dynamic iteration with 3 sweeps and window size $H = 1e-5$. (**f**) Error in output voltage in monolithic simulation with step size $H = 1e-4$. (**g**) Error in output voltage in monolithic simulation with step size $H = 1e-5$

In addition to the model presented in Chap. 2, also the holes will be included with a simple drift-diffusion equation.

The complete mathematical model is given by the equations

$$\frac{\partial n}{\partial t} + \operatorname{div}(n\,\mathbf{V}) = -R, \tag{3.72}$$

$$\frac{\partial p}{\partial t} + \operatorname{div}\left(p\,\mathbf{V_p}\right) = -R, \tag{3.73}$$

$$\frac{\partial (nW)}{\partial t} + \operatorname{div}(n\,\mathbf{S}) + nq\mathbf{V}\cdot\nabla\phi = nC_W, \tag{3.74}$$

$$\rho c_V \frac{\partial T_L}{\partial t} - \operatorname{div}\left[K(T_L)\nabla T_L\right] = H, \tag{3.75}$$

$$\mathbf{E} = -\nabla\phi, \quad \epsilon\Delta\phi = -q(N_D - N_A - n + p), \tag{3.76}$$

with n and p the electron and holes density respectively, W the electron energy, T_L the lattice temperature, ϕ the electrostatic potential and $\mathbf{E} = -\nabla\phi$ the electric field. N_D and N_A are the density of donors and acceptors respectively (assumed as known function of the position). q is the elementary charge, ρ the silicon density, c_V the specific heat, C_W the energy production term, which can be written in a relaxation form as

$$C_W = -\frac{W - W_0}{\tau_W}, \tag{3.77}$$

with $W_0 = 3/2k_BT_L$ and $\tau_W(W)$ the energy relaxation time. k_B is the Boltzmann constant and ϵ is the dielectric constant.

The closure relations for the electron velocity $\mathbf{V}$, the energy flux $\mathbf{S}$, the thermal conductivity $K(T_L)$ and the crystal energy production term H have been obtained in [32, 33] by employing MEP and are reported in Chap. 2. The holes are described by a standard drift-diffusion model with constant mobility. $\mathbf{V_p}$ is the velocity of holes.

Since the electron production terms are slowly changing with respect to k_BT_L, we adopt the simplification that they are evaluated at $T_L = 300\,\mathrm{K}$.

The phonon energy production is given by

$$H = -(1 + P_S)\,n\,C_W + P_S\;\mathbf{J}\cdot\mathbf{E}, \tag{3.78}$$

where $P_S = -c^2\,\tau_R\,c_{12}^{(p)}$ plays the role of a thermopower coefficient and τ_R is the phonon relaxation time in resistive processes.

R is the generation-recombination term (see [35] for a complete review) which splits into the Shockley-Read-Hall (SRH) and the Auger contribution (AU) $R = R^{SRH} + R^{AU}$ where

$$R^{SRH} = \frac{np - n_i^2}{\tau_p(n + n_1) + \tau_n(p + p_1)}, \quad R^{AU} = \left(C_{cn}n + C_{cp}\,p\right)(np - n_i^2), \tag{3.79}$$

We will take the values $C_{cn} = 2.8\times10^{-31}$ cm^6 s^{-1} and $C_{cp} = 9.9\times10^{-32}$ cm^6 s^{-1}. In our numerical experiments we set $n_1 = p_1 = n_i$, n_i being the intrinsic concentration. The expressions of τ_p and τ_n we will use are [35]

$$\tau_n = \frac{\tau_{n0}}{1 + \frac{N_D(x)+N_A(x)}{N_n^{ref}}}, \quad \tau_p = \frac{\tau_{p0}}{1 + \frac{N_D(x)+N_A(x)}{N_p^{ref}}}, \tag{3.80}$$

where $\tau_{n0} = 3.95 \times 10^{-4}$, $\tau_{p0} = 3.25 \times 10^{-5}s$, $N_n^{ref} = N_p^{ref} = 7.1 \times 10^{15}$cm^{-3}.

At the source and drain contacts the Robin boundary condition

$$-k_L \frac{\partial T_L}{\partial n} = R_{th}^{-1}(T_L - T_{env}), \tag{3.81}$$

is assumed, R_{th} being the thermal resistivity of the contact and T_{env} the environment temperature. We use no-flux condition for the temperature on the lateral boundary and oxide silicon interface and Dirichlet condition at the bulk contact. The electron energy at the source, drain and bulk contacts is set equal to the lattice energy. The other boundary conditions needed for integrating the Mosfet model are described in [29].

3.4.1 The Numerical Method

The crystal lattice temperature T_L changes much slower than other variables. For instance the typical relaxation time for the temperature in our simulations is in the order of thousand picoseconds, while relaxation time of the other fields is in the order of picoseconds. We exploit this double-scale behavior by applying a variant of the multirate integration scheme [18, 20] which is a popular choice in coupled electro-thermal circuit simulation [5]. For the simulation of the transient response of the model we solve the balance equations by adopting the following multirate integration scheme:

- Step 1. We first integrate the balance equations for electrons and holes with the crystal lattice energy and the electric field frozen at the time step $k-1$. This gives the density of the electrons and holes and the electron energy at the time step k and schematically can be written as

$$\frac{\partial \mathscr{U}^k}{\partial t} + F(\mathscr{U}^k, \phi^{k-1}, T_L^{k-1}) = 0, \tag{3.82}$$

  with $\mathscr{U} = (n, p, W)$. Here $k = 1, \ldots, N$ is the index of the integration interval $[t_{k-1}, t_k]$, with $t_k = t_{k-1} + \Delta t$, Δt being the time size of the synchronization window.

- Step 2. We integrate the lattice energy balance equation with n and W given by the step 1

$$\rho c_V \frac{\partial T_L^k}{\partial t} - \operatorname{div}\left[K(T_L^k)\nabla T_L^k\right] = H(\mathscr{U}^k, T_L^k) \tag{3.83}$$

  along with the Poisson equation with $n = n^k$ and $p = p^k$.

For steps 1 and 2 different time steps for the numerical integration over the interval $[t_{k-1}, t_k]$ are used. Typically the time step for integration of (3.83) we can use is 100 times larger than the time step for (3.82).

This sequence can be considered as steps of a splitting technique [26] and we expect that such a numerical scheme is a stable first-order approximation with respect to time, as confirmed by the numerical experiments presented in the next section.

3.4.1.1 Step 1

The numerical scheme is based on an exponential fitting like that employed in the Scharfetter-Gummel scheme for the drift-diffusion model of semiconductors. The basic idea is to split the particle and energy density currents as the difference of two terms. Each of them is written by introducing suitable mean mobilities in order to get expressions of the currents similar to those arising in other energy-transport models known in literature [6, 7, 11, 25, 36]. A simple explicit discretization in time with constant time step proves satisfactorily efficient and avoids the problem related to the high nonlinear coupling of the discretized equations of [27]. The equations are spatially discretized on a regular grid. The details of the numerical scheme can be found in [29]. Here a brief account is given.

For the sake of simplicity, the numerical method is presented only for the electron part, putting equal to zero the generation-recombination term. The inclusion of holes and the coupling with electrons is performed straightforwardly in an explicit way.

First the current density $\mathbf{J} = n\mathbf{V}$ and the energy-flux density $\mathbf{Z} = n\mathbf{S}$ are rewritten as

$$\mathbf{J} = \mathbf{J}^{(1)} - \mathbf{J}^{(2)}, \quad \mathbf{Z} = \mathbf{Z}^{(1)} - \mathbf{Z}^{(2)} \tag{3.84}$$

and then each term is put into a drift-diffusion form

$$\mathbf{J}^{(1)} = \frac{c_{22}}{D}\left[\nabla(nU) - qn\nabla\phi\right], \quad \mathbf{J}^{(2)} = \frac{c_{12}}{D}\left[\nabla(nF) - qn\frac{F}{U}\nabla\phi\right], \tag{3.85}$$

$$\mathbf{Z}^{(1)} = \frac{c_{11}}{D}\left[\nabla(nF) - qn\frac{F}{U}\nabla\phi\right], \quad \mathbf{Z}^{(2)} = \frac{c_{12}}{D}\left[\nabla(nU) - qn\nabla\phi\right], \tag{3.86}$$

with $D = c_{11}c_{22} - c_{12}c_{21}$.

We introduce the grid points (x_i, y_j) with $x_{i+1} - x_i = h$ = constant and $y_{j+1} - y_j = k$ = constant, and the middle points $(x_i, y_{j\pm1/2}) = (x_i, y_j \pm k/2)$ and $(x_{i\pm1/2}, y_j) = (x_i \pm h/2, y_j)$. A uniform time step Δt is used and we set $u^l_{i,j} = u(x_i, y_j, l\,\Delta t)$.

By indicating with J_x and J_y the x and y component of the current density $\mathbf{J}$ and by Z_x and Z_y the x and y component of $\mathbf{Z}$, we discretize the balance equations (3.72) and (3.74) up to terms of order $O(h^2, k^2, \Delta t)$ in the bidimensional case as

$$\frac{n_i^{l+1} - n_i^l}{\Delta t} + \frac{(J_x)_{i+1/2,j} - (J_x)_{i-1/2,j}}{h} + \frac{(J_y)_{i,j+1/2} - (J_y)_{i,j-1/2}}{k} = 0, \tag{3.87}$$

$$\begin{aligned}&\frac{(n\,W)_i^{l+1} - (n\,W)_i^l}{\Delta t} + \frac{(Z_x)_{i+1/2,j} - (Z_x)_{i-1/2,j}}{h} + \frac{(Z_y)_{i,j+1/2} - (Z_y)_{i,j-1/2}}{k} + \\ &- q\frac{(J_x)_{i+1/2,j} + (J_x)_{i-1/2,j}}{2}\frac{\phi_{i+1,j} - \phi_{i-1,j}}{2h} \\ &- q\frac{(J_y)_{i,j+1/2} + (J_y)_{i,j-1/2}}{2}\frac{\phi_{i,j+1} - \phi_{i,j-1}}{2k} + n_{i,j}\frac{W_{i,j} - W_0}{(\tau_W)_{i,j}} = 0.\end{aligned} \tag{3.88}$$

The variables without temporal index must be considered evaluated at time level l.

In order to evaluate the components of the currents in the middle points, let us consider the sets

$$I_{i+1/2,j} = [x_i, x_{i+1}] \times [y_{j-1/2}, y_{j+1/2}], \quad I_{i,j+1/2} = [x_{i-1/2}, x_{i+1/2}] \times [y_j, y_{j+1}]$$

and expand $J_x^{(r)}$, r= 1, 2, in Taylor's series in $I_{i+1/2,j}$

$$J_x^{(r)}(x, y) \approx (J_x^{(r)})_{i+1/2,j} + (x - x_{i+1/2})\left(\frac{\partial J_x^{(r)}}{\partial x}\right)_{i+1/2,j} + (y - y_j)\left(\frac{\partial J_x^{(r)}}{\partial y}\right)_{i+1/2,j}$$

and $J_y^{(r)}$, r =1, 2, in Taylor's series in $I_{i,j+1/2}$

$$J_y^{(r)}(x, y) \approx (J_y^{(r)})_{i,j+1/2} + (x - x_i)\left(\frac{\partial J_y^{(r)}}{\partial x}\right)_{i,j+1/2} + (y - y_{j+1/2})\left(\frac{\partial J_y^{(r)}}{\partial y}\right)_{i,j+1/2}.$$

First, we introduce $U_T = U(W)/q$, which plays the role of a thermal potential (see [29] for more details) and indicate by $\overline{U}_T$ its piecewise constant approximation, which is given by $\overline{U}_T = \frac{U(W_{i,j}) + U(W_{i+1,j})}{2q}$ in the cell $I_{i+1/2,j}$ and by

$\overline{U}_T = \frac{U(W_{i,j+1})+U(W_{i,j})}{2q}$ in the cell $I_{i,j+1/2}$. Then we introduce the *local* mobilities

$$g_{11} = -\frac{\overline{c}_{22}}{\overline{D}}nU, \quad g_{12} = -\frac{\overline{c}_{12}}{\overline{D}}nF, \quad g_{21} = -\frac{\overline{c}_{11}}{\overline{D}}nF, \quad g_{22} = -\frac{\overline{c}_{12}}{\overline{D}}nU, \tag{3.89}$$

where $\overline{c}_{pq}$ is a piecewise constant approximation of $c_{pq}\, p, q = 1, 2$, given by $\overline{c}_{pq} = c_{pq}\left(\frac{W_{i,j}+W_{i+1,j}}{2}\right)$ in the cell $I_{i+1/2,j}$ and by $\overline{c}_{pq} = c_{pq}\left(\frac{W_{i,j}+W_{i,j+1}}{2}\right)$ in the cell $I_{i,j+1/2}$, and, as in [15], the *local* Slotboom variables

$$s_{kr} = \exp\left(-\phi/\overline{U}_T\right) g_{kr} \quad k, r = 1, 2$$

that satisfy

$$\nabla s_{1r} \simeq -\exp\left(-\phi/\overline{U}_T\right)\mathbf{J}^{(r)}, \quad \nabla s_{2r} \simeq -\exp\left(-\phi/\overline{U}_T\right)\mathbf{H}^{(r)} \quad r = 1, 2. \tag{3.90}$$

From the x component of $(3.90)_1$, one has

$$\frac{\partial s_{1r}(x, y_j)}{\partial x} \simeq -\exp\left(-\phi/\overline{U}_T\right) J_x^{(r)}(x, y_j) =$$

$$-\exp\left(-\phi/\overline{U}_T\right)\left\{(J_x^{(r)})_{i+1/2,j} + (x - x_{i+1/2})\left(\frac{\partial J_x^{(r)}}{\partial x}\right)_{i+1/2,j} + o(\Delta x, \Delta y)\right\},$$

which, after integration over $[x_i, x_{i+1}]$ and some algebra, gives

$$(J_x^{(r)})_{i+1/2,j} = -z_{i+1/2,j}\coth z_{i+1/2,j}\ \frac{(g_{1r})_{i+1,j} - (g_{1r})_{i,j}}{h}$$
$$+z_{i+1/2,j}\frac{(g_{1r})_{i+1,j} + (g_{1r})_{i,j}}{h}, \quad r = 1, 2 \tag{3.91}$$

where $z_{i+1/2,j} = \frac{\phi_{i+1,j}-\phi_{i,j}}{2\overline{U}_T}$.

Likewise by evaluating the y component of $(3.90)_2$ and integrating over $[y_j, y_{j+1}]$ we find

$$(J_y^{(r)})_{i,j+1/2} = -z_{i,j+1/2}\coth z_{i,j+1/2}\ \frac{(g_{1r})_{i,j+1} - (g_{1r})_{i,j}}{k}$$
$$+z_{i,j+1/2}\frac{(g_{1r})_{i,j+1} + (g_{1r})_{i,j}}{k}, \quad r = 1, 2 \tag{3.92}$$

where $z_{i,j+1/2} = \frac{\phi_{i,j+1}-\phi_{i,j}}{2\overline{U}_T}$. With the same procedure the following discrete expression for the components of the energy flux are obtained

$$(H_x^{(r)})_{i+1/2,j} = -z_{i+1/2,j} \coth z_{i+1/2,j} \frac{(g_{2r})_{i+1,j} - (g_{2r})_{i,j}}{h} + z_{i+1/2,j} \frac{(g_{2r})_{i+1,j} + (g_{2r})_{i,j}}{h}, \tag{3.93}$$

$$(H_y^{(r)})_{i,j+1/2} = -z_{i,j+1/2} \coth z_{i,j+1/2} \frac{(g_{2r})_{i,j+1} - (g_{2r})_{i,j}}{k} + z_{i,j+1/2} \frac{(g_{2r})_{i,j+1} + (g_{2r})_{i,j}}{k}, \quad r = 1, 2. \tag{3.94}$$

The error in formulas (3.91)–(3.94) is $O(h, k)$.

The Poisson equation is solved by replacing it with

$$\phi_t - \operatorname{div}(\epsilon \nabla \phi) = q(N_D - N_A - n). \tag{3.95}$$

The solution of (3.95) as $t \mapsto +\infty$ is the same as that of the original Poisson equation, at least in the smooth case.

If we introduce a time step $\Delta\hat{t}$ and set $\phi_{ij}^r = \phi(x_i, y_j, r\Delta\hat{t})$, (3.95) can be discretized in an explicit way as

$$\phi_{ij}^{r+1} = \phi_{ij}^r + \epsilon\Delta\hat{t}\left[\frac{1}{h^2}\left(\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j}\right) + \frac{1}{k^2}\left(\phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1}\right) + q(C_{i,j} - n_{i,j})\right] \tag{3.96}$$

with the notable advantage to take easily into account the different types of boundary conditions, that will be considered in more detail in the next sections. The price to pay is that at each time step, we need to reach the stationary state of (3.95) by using a time step satisfying the CFL condition, usual for parabolic equations,

$$\Delta\hat{t} \le \frac{1}{2}\frac{1}{\frac{1}{h^2} + \frac{1}{k^2}}.$$

However the computational effort is comparable with that required by direct methods.

3.4.1.2 Step 2

A coordinate splitting technique [26] is used for the solution of the lattice energy equation for the variable $u = k_B T$ with time step Δt_T. The splitting technique allows an efficient usage of stable implicit time schemes. The procedure contains

two steps with the two sub operators

$$\rho c_V \frac{u^{n+1/2} - u^n}{\Delta t_T} = \frac{\partial}{\partial x}\left[K(T_L^n)\frac{\partial u^{n+1/2}}{\partial x}\right] + \frac{k_B}{2} H^{n+1/2}, \tag{3.97}$$

$$\rho c_V \frac{u^{n+1} - u^{n+1/2}}{\Delta t_T} = \frac{\partial}{\partial y}\left[K(T_L^n)\frac{\partial u^{n+1}}{\partial y}\right] + \frac{k_B}{2} H^{n+1}. \tag{3.98}$$

This scheme is absolutely stable and approximates the equation of the lattice energy with first order accuracy in time. For the approximation of the spatial derivatives, the standard stencil with three points has been chosen. For instance, the approximation of (3.98) is the following

$$\begin{aligned}\rho c_V u_{i,j}^{n+1} = \rho c_V u_{i,j}^{n+1/2} + \frac{\Delta t_T}{k^2}\Bigg[&\frac{\tilde{K}_{i,j} + \tilde{K}_{i,j+1}}{2}(u_{i,j+1}^{n+1} - u_{i,j}^{n+1}) - \\ &\frac{\tilde{K}_{i,j} + \tilde{K}_{i,j-1}}{2}(u_{i,j}^{n+1} - u_{i,j-1}^{n+1})\Bigg] \\ &+ \frac{k_B}{2}\frac{\Delta t_T}{\tau_W}(1 + P_S) n_{i,j}^{n+1}\left(W_{ij}^{n+1} - \frac{3}{2}u_{i,j}^{n+1}\right) + \frac{k_B}{2}\Delta t_T\, J_{i,j}^{n+1}\, E_{i,j}^{n+1},\end{aligned}$$

where $\tilde{K}_{i,j} = K(T_{Li,j})$. Of course such a discretization is valid in the interior points of the mesh.

The Robin boundary condition (3.81) is approximated as

$$-k_L \frac{u_{i,1}^{n+1} - u_{i,0}^{n+1}}{k} = R_{th}^{-1}(u_{i,0}^{n+1} - k_B T_{env}). \tag{3.99}$$

Here we have assumed that at the portion of boundary where the Robin condition holds, one has $j = 0$ and the closest interior points have $j = 1$.

The obtained linear system can be solved efficiently with the tridiagonal matrix factorization procedure.

3.4.2 *Numerical Simulation of the Crystal Lattice Heating in MOSFETs*

We apply the above numerical method for the simulation of the heating of the crystal lattice in a MOSFET described by the MEP model.

We have modeled the thermal conductivity with the fitting formula $K(T_L) = 1.5486\,(T_L/300\,\mathrm{K})^{-4/3}$ V A/cm K and have set $c_V = 703\,\mathrm{m^2/s^2\,K}$ (see [35]). The

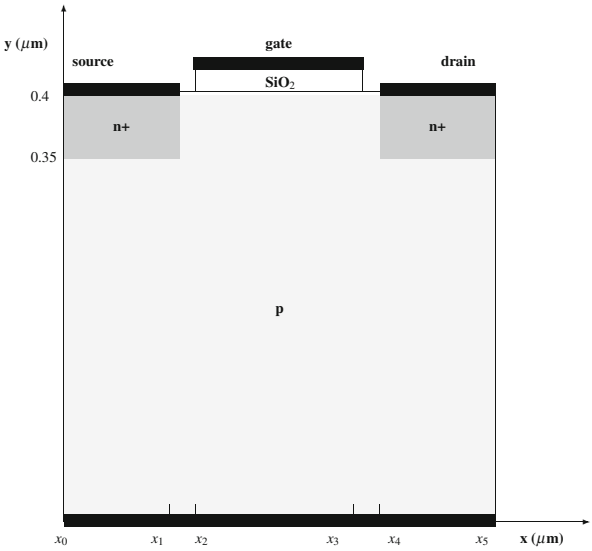

Fig. 3.3 Schematic representation of a bidimensional MOSFET

mobility of holes has been considered as constant and equal to $500\,\mathrm{cm}^2/\mathrm{V\,s}$. More details about the values of the physical parameters can be found in [30].

The shape of the device is shown in Fig. 3.3. The length of the channel ($x_4 - x_1$ in the figure) is L_c, the length of source and drain ($x_1 - x_0$) is $L = L_c/2$. We will consider $L_c = 50\,\mathrm{nm}$ and $L_c = 200\,\mathrm{nm}$. The source and drain depths are $0.1\,\mu\mathrm{m}$. The gate oxide is 20 nm thick. The substrate thickness is $0.4\,\mu\mathrm{m}$. An environment temperature $T_{env} = 300\,\mathrm{K}$ has been considered. In most of our numerical experiments we will take $R_{th} = 10^{-8}\,\mathrm{K\,m}^2/\mathrm{W}$ as in [9].

The doping concentration is

$$N_D(x) - N_A(x) = \begin{cases} n_+ \text{ in the } n^+ \text{ regions} \\ -p_- = -10^{14}\mathrm{cm}^{-3} \text{in the } p \text{ region} \end{cases} \tag{3.100}$$

with abrupt junctions. We will consider different values of n_+ in the simulations.

First a MOSFET with 200 nm channel length has been simulated. The stationary solution is shown in the Figs. 3.4–3.8. The distance between gate and source ($x_2 - x_1$) and between drain and gate ($x_4 - x_3$) is 25 nm. The thermal resistivity of the contact is set equal to $R_{th} = 10^{-8}\ \mathrm{K\ m}^2/\mathrm{W}$. The donor concentration is $n_+ = 10^{17}\mathrm{cm}^{-3}$. In Fig. 3.4 one can see a relatively small heating of the crystal, just a maximum of about 7° above the environment temperature. The maximum of the crystal temperature is attained near the drain contact where also the maximum of the electron energy is observed (see Fig. 3.5). It is worth remarking that there is almost no influence of the device self-heating on the current through the device as shown in Fig. 3.8, where the characteristic curves with the lattice temperature fixed at 300 K are compared with those obtained with varying T_L.

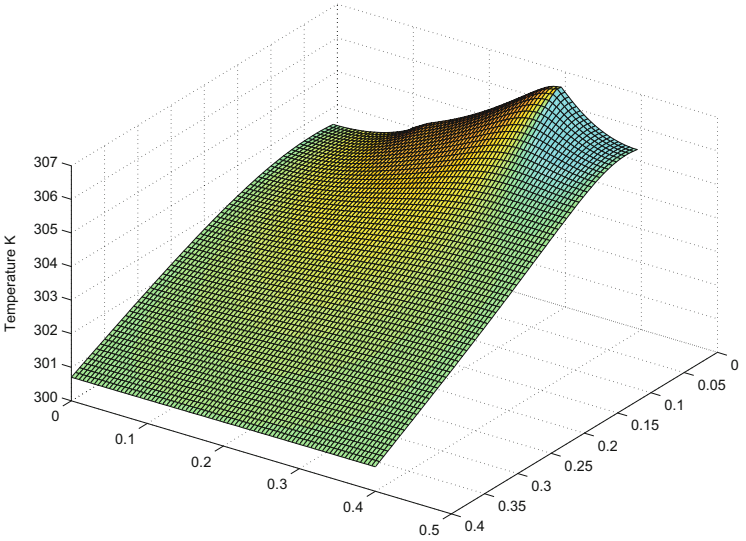

Fig. 3.4 Stationary solution of the lattice temperature in the MOSFET with channel of 200 nm by setting $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

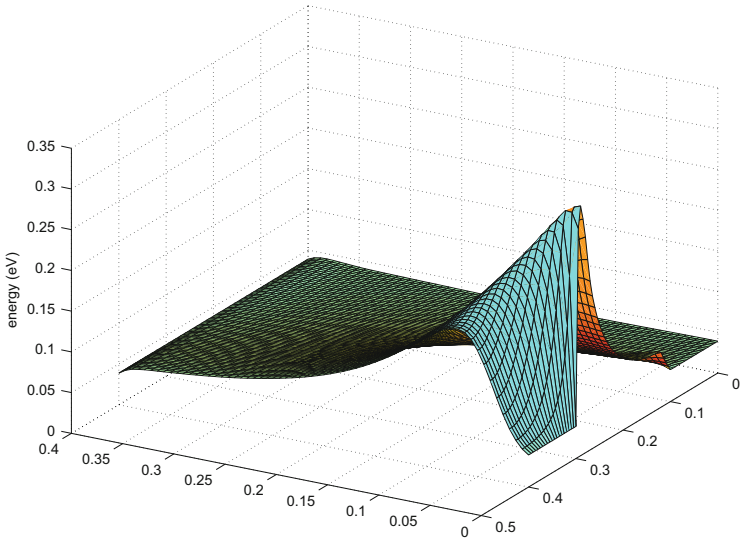

Fig. 3.5 Stationary solution of the electron energy in the MOSFET with channel of 200 nm by setting $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

As second example we have simulated a nanoscale MOSFET device with a channel of length 50 nm . The gate length is 45 nm and the gate voltage $V_{\mathrm{DG}} = 0.8$ V. The donor concentration is $n_+ = 10^{17}\mathrm{cm}^{-3}$. In the Figs. 3.9 and 3.10 is plotted the stationary solution of the lattice temperature and the electron energy. In contrast to

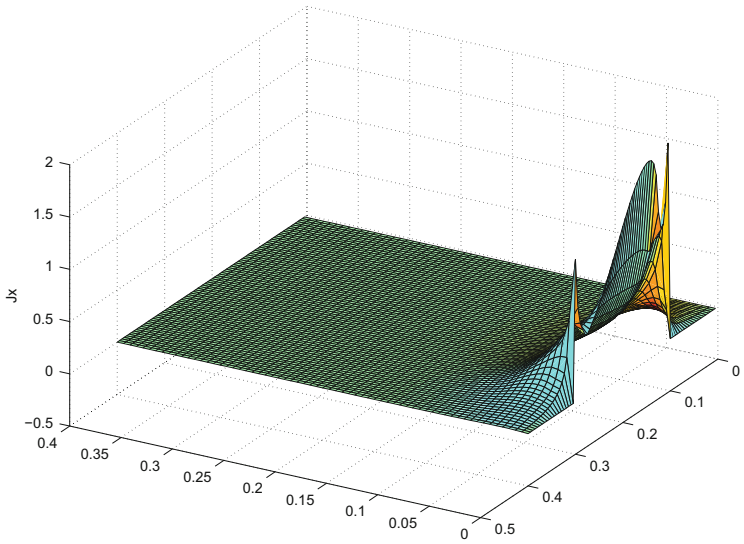

Fig. 3.6 Stationary solution of the x component of current in the MOSFET with channel of 200 nm by setting $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

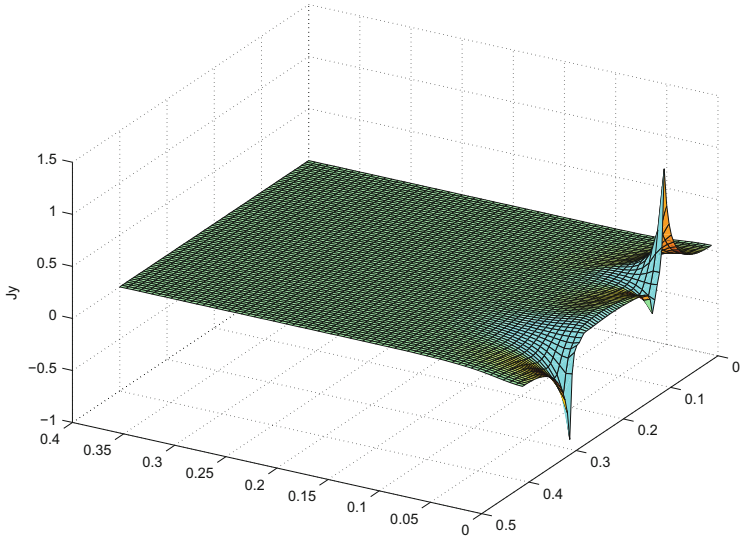

Fig. 3.7 Stationary solution of the y component of current in the MOSFET with channel of 200 nm by setting $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

the previous case, now the lattice temperature raises up to 380 K in the area near the gate. We argue that this temperature raise should depend, beside the strength of the electric field, on the density of the hot electrons and might be higher for higher

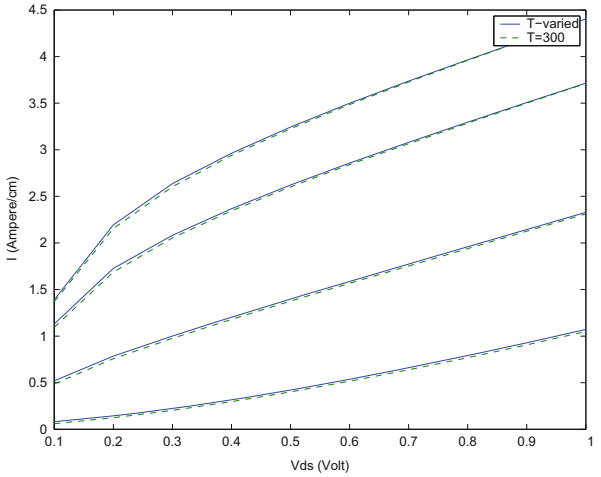

Fig. 3.8 Drain current for $V_{DG} = 0.4, 0.6, 0.8, 0.9$ V. The current increases by increasing V_{DG}

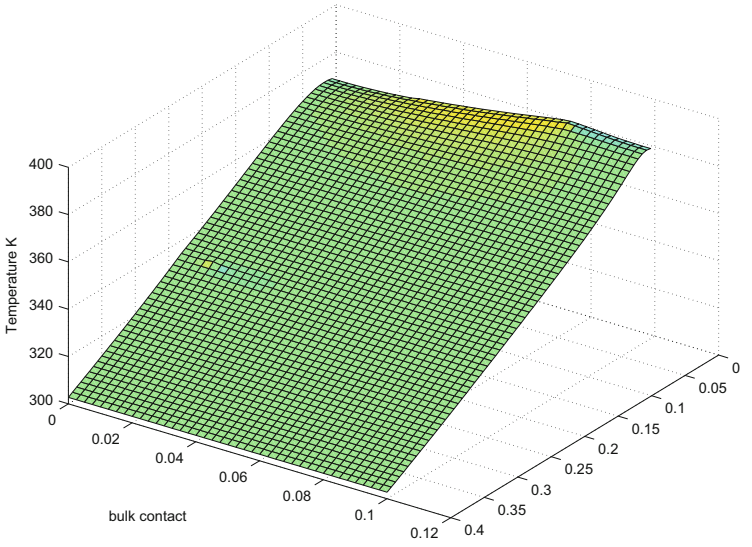

Fig. 3.9 Stationary solution of the lattice temperature in the MOSFET with channel of 50 nm by setting $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

doping concentration. In order to investigate this assumption, a simulation with $n_+ = 5 \times 10^{17}\mathrm{cm}^{-3}$ has been performed too. As expected one can see in Fig. 3.11 that the maximum lattice temperature attains about 550 K. In Fig. 3.12 the result of the lattice temperature for the even higher donor concentration $n_+ = 10^{18}\,\mathrm{cm}^{-3}$ is reported. The maximum lattice temperature achieves about 700 K, confirming the dependence of it on the density of the electron current.

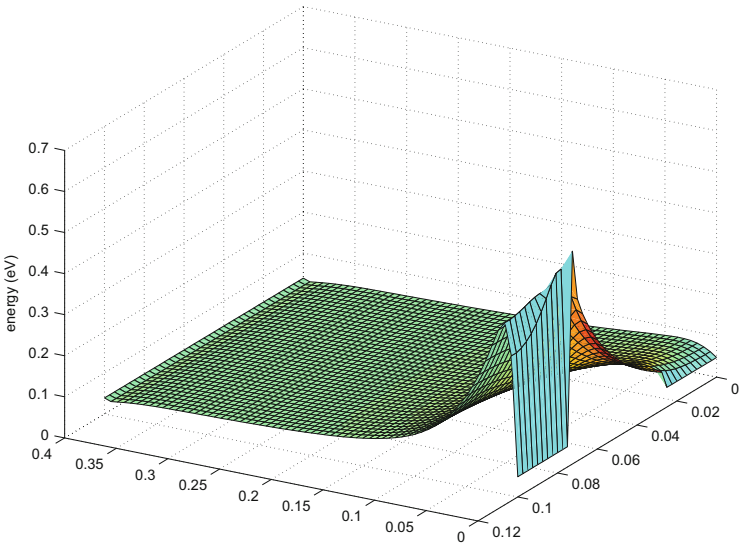

Fig. 3.10 Stationary solution of the electron energy in the MOSFET with channel of 50 nm by setting $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

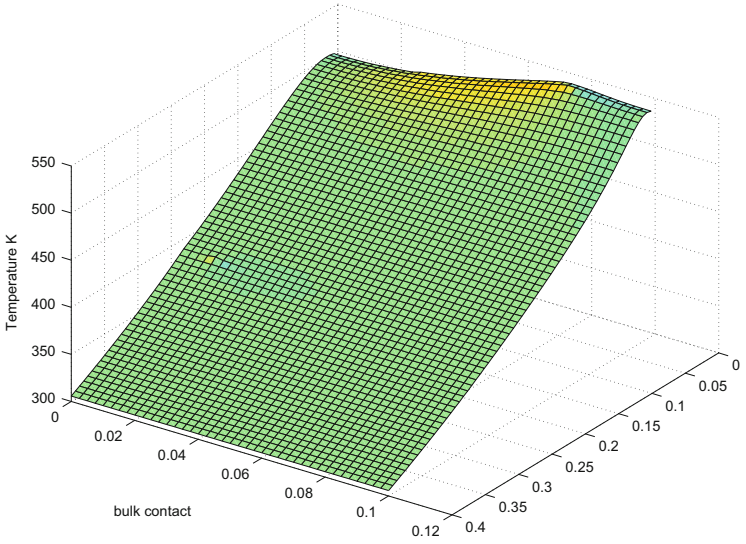

Fig. 3.11 Stationary solution of the lattice temperature in the MOSFET with channel of 50 nm by setting $n_+ = 5 \times 10^{17}\,\mathrm{cm}^{-3}$ and $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

By shrinking the dimension of the device the thermal effects have also a non negligible influence on the current through the device. In Fig. 3.13 current Ãś voltage characteristics for the device with $n_+ = 10^{17}\,\mathrm{cm}^{-3}$ are shown. With

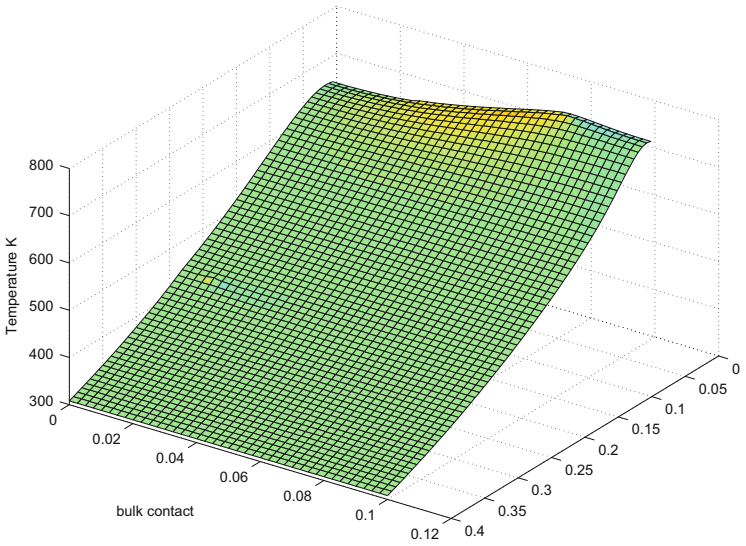

Fig. 3.12 Stationary solution of the lattice temperature in the MOSFET with channel of 50 nm by setting $n_+ = 10^{18}\,\mathrm{cm}^{-3}$ and $R_{th} = 10^{-8}\,\mathrm{Km}^2/\mathrm{W}$

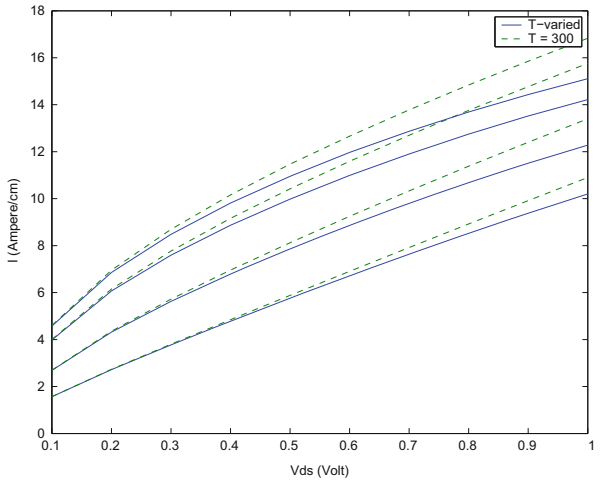

Fig. 3.13 Drain current with constant and varying lattice temperature in the MOSFET with channel of 50 nm by setting $n_+ = 10^{18}\mathrm{cm}^{-3}$ and $R_{th} = 10^{-8}$ for $V_{DG} = 0.4, 0.6, 0.8, 0.9\,\mathrm{V}$. The current increases by increasing V_{DG}

increasing electric field strength, we observe a rising deviation of the characteristic curves corresponding to a constant lattice temperature from those with varying T_L.

The lattice temperature in the device is also strongly influenced by the thermal resistivity of the contact R_{th}. This value depends on the manufacturing process. In Figs. 3.14 and 3.15 the lattice temperature is shown for $R_{th} = 10^{-10}\,\mathrm{Km}^2/\mathrm{W}$ and $R_{th} = 10^{-9}\,\mathrm{Km}^2/\mathrm{W}$ with $n_+ = 10^{17}\,\mathrm{cm}^{-3}$.

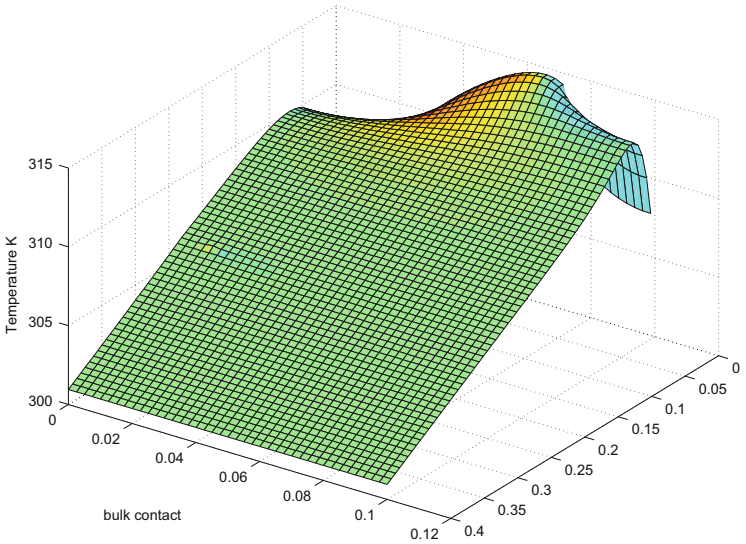

Fig. 3.14 Stationary solution of the lattice temperature in the MOSFET with channel of 50 nm by setting $R_{th} = 10^{-10}\,\mathrm{Km}^2/\mathrm{W}$ and $n_+ = 10^{17}\,\mathrm{cm}^{-3}$

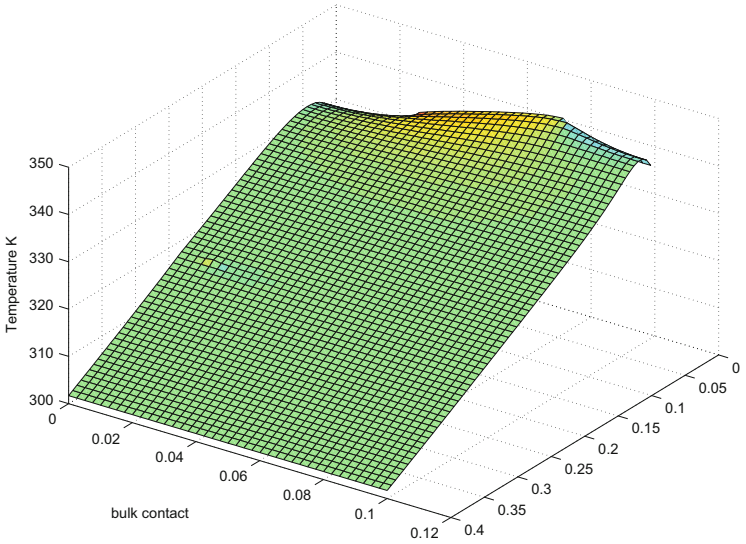

Fig. 3.15 Stationary solution of the lattice temperature in the MOSFET with channel of 50 nm by setting $R_{th} = 10^{-9}\,\mathrm{Km}^2/\mathrm{W}$ and $n_+ = 10^{17}\,\mathrm{cm}^{-3}$

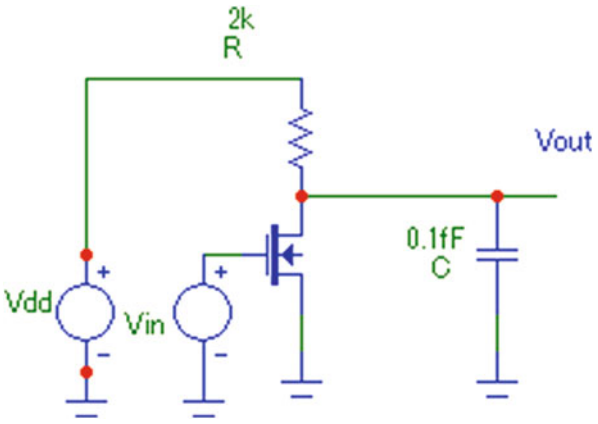

Fig. 3.16 Simulated inverter circuit

3.4.3 Coupled Circuit-Device Simulation

At last a case of coupling between a Mosfet and a circuit is present. We simulate the heating of a transistor in the electrical circuit representing an inverter. The inverter circuit is plot in Fig. 3.16. Input voltage on the gate contact is (in Volt) $V_{in} = 0.3\cos(\omega t) + 0.5$, with frequency $\omega = 2\pi\ 10^9$ rad/s and power voltage $V_{dd} = 1V$. The width of the transistor (length in the orthogonal direction with respect to the considered 2D cross section) is set equal to 200 nm. Modified nodal analysis gives us for the output voltage V_{out}:

$$C\frac{dV_{out}}{dt} + \frac{V_{out} - V_{dd}}{R} + j(V_{in}, V_{out}, t) = 0, \tag{3.101}$$

where current through the transistor $j(V_{in}, V_{out}, t)$ is computed by the energy-transport model. We refer for instance to [9] for details of device-circuits coupled modeling algorithm.

The output voltage simulated with and without transistor self heating and maximum temperature in the transistor are plot in the Fig. 3.17. One can see that lattice temperature in the transistor does not achieve 400 K as we have observed in the single transistor simulation. It can be explained with smaller average voltage at the gate and consequently smaller average electrical field. However there is still a shift in the minimum values of the output voltage and a clear indication of the crystal heating.

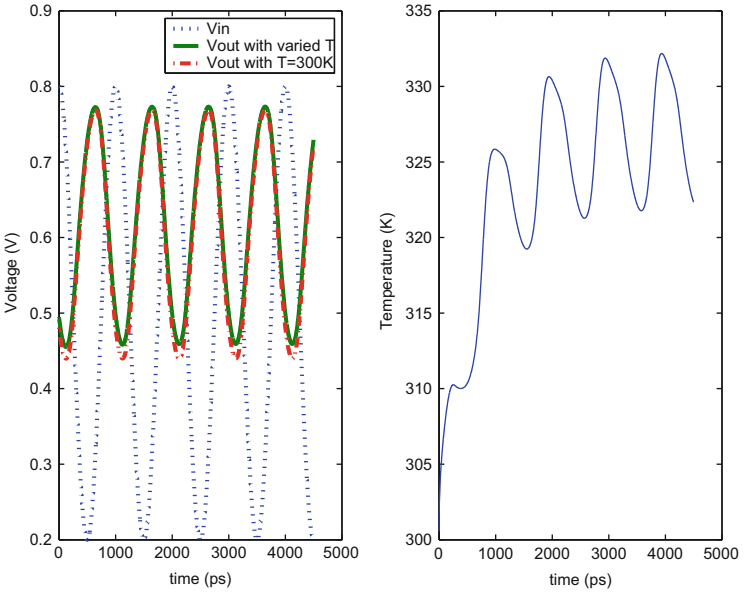

Fig. 3.17 On the *left* input and output voltages versus time. On the *right* maximum value of the lattice temperature in the MOSFET versus time

References

1. Arnold, M., Günther, M.: Preconditioned dynamic iteration for coupled differential-algebraic systems. BIT **41**(1), 1–25 (2001)
2. Arnold, M., Heckmann, A.: From multibody dynamics to multidisciplinary applications. In: García Orden, J., Goicolea, J., Cuadrado, J. (eds.) Multibody Dynamics. Computational Methods and Applications, pp. 273–294. Springer, Dordrecht (2007)
3. Bartel, A.: Partial Differential-Algebraic Models in Chip Design – Thermal and Semiconductor Problems. Forschrittsberichte. VDI-Verlag, Düsseldorf (2004)
4. Bartel, A., Brunk, M., Günther, M., Schöps, S.: Dynamic iteration for coupled problems of electric circuits and distributed devices. SIAM J. Sci. Comput. **35**(2), B315–B335 (2012)
5. Bartel, A., Günther, M.: Multirate co-simulation of first order thermal models in electric circuit design. In: Schilders, W., ter Maten, E., Houben, S. (eds.) Scientific Computing in Electrical Engineering SCEE 2002, Mathematics in Industry, pp. 23–28. Springer, Berlin (2004)
6. Ben Abdallah, N., Degond, P.: On a hierarchy of macroscopic models for semiconductors. J. Math. Phys. **37**, 3306–3333 (1996)
7. Ben Abdallah, N., Degond, P., Genieys, S.: An energy-transport model for semiconductors derived from the boltzmann equation. J. Stat. Phys. **84**, 205–231 (1996)
8. Brenan, K.E., Campbell, S.L.V., Petzold, L.R.: Numerical solution of initial-value problems in differential-algebraic equations. SIAM, Philadelphia (1995)
9. Brunk, M., Jüngel, A.: Numerical coupling of electric circuit equations and energy-transport models for semiconducotrs. SIAM J. Sci. Comput. **30**, 873–894 (2008)
10. Burrage, K.: Parallel and Sequential Methods for Ordinary Differential Equations. Clarendon, Oxford (1995)

11. Chen, D., Kan, E., Ravaioli, U., Shu, C.W., Dutton, R.: An improved energy-transport model including nonparabolicity and non-maxwellian distribution effects. IEEE Electron Device Lett. **13**, 26–28 (1992)
12. Clemens, M.: Large systems of equations in a discrete electromagnetism: formulations and numerical algorithms. IEE Proceedings - Science, Measurement and Technology **152**(2), 50–72 (2005). doi:10.1049/ip-smt:20050849
13. Clemens, M., Schuhmann, R., van Rienen, U., Weiland, T.: Modern Krylov subspace methods in electromagnetic field computation using the finite integration theory. Appl. Comput. Electromagn. Soc. J. **11**(1), 70–84 (1996)
14. Culpo, M.: Numerical Algorithms for System Level Electro-Thermal Simulation. Ph.D. thesis, Bergische Universität Wuppertal (2009)
15. Degond, P., Jüngel, A., Pietra, P.: Numerical discretization of energy-transport models for semiconductors with nonparabolic band structure. SIAM J. Sci. Comput. **22**, 986–1007 (2000)
16. Deuflhard, P., Hairer, E., Zugck, J.: One-step and extrapolation methods for differential-algebraic systems. Numer. Math. **51**, 501–516 (1987)
17. Estévez Schwarz, D., Tischendorf, C.: Structural analysis of electric circuits and consequences for mna. Int. J. Circuit Theory Appl. **28**(2), 131–162 (2000)
18. Gear, C., Wells, R.: Multirate linear multistep methods. BIT **24**, 484–502 (1984)
19. Günther, M.: Preconditioned splitting in dynamic iteration schemes for coupled dae systems in rc network design. In: Buikis, A., Ciegis, R., Fitt, A. (eds.) Progress in Industrial Mathematics at ECMI 2002, Mathematics in Industry, pp. 173–177. Springer, Berlin (2004)
20. Günther, M., Rentrop, P.: Multirate row methods and latency of electric circuits. Appl. Numer. Math. **13**, 83–102 (1993)
21. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. Springer Series in Computational Mathematics. Springer, Berlin (1996)
22. Jackiewicz, Z., Kwapisz, M.: Convergence of waveform relaxation methods for differential-algebraic systems. SIAM J. Numer. Anal. **33**, 2303–2317 (1996)
23. Kübler, R., Schielen, W.: Two methods for simulator coupling. Math. Comput. Model. Dyn. Syst. **6**, 93–113 (2000)
24. Lelarasmee, E., Ruehli, A., Sangiovanni-Vincentelli, A.: The waveform relaxation method for time domain analysis of large scale integrated circuits. IEEE Trans. on CAD of IC and Syst. **1**, 131–145 (1982)
25. Lyumkis, E., Polsky, B., Shir, A., Visocky, P.: Transient semiconductor device simulation including energy balance equation. Compel **11**, 311–325 (1992)
26. Marchuk, G.: Splitting and alternating direction method. In: Ciarlet, P., Lions, J. (eds.) Handbook of Numerical Analysis. Vol. 1: Finite Difference Methods (Part 1) and Solution of Equations in $\mathbb{R}^n$ (Part 1), pp. 197–462. North-Holland, Amsterdam (1990)
27. Marrocco, A., Anile, A., Romano, V., Sellier, J.M.: 2d numerical simulation of the mep energy-transport model with a mixed finite elements scheme. J. Comput. Electron. **4**, 231–259 (2005)
28. Nicolet, A., Delincé, F.: Implicit Runge-Kutta methods for transient magnetic field computation. IEEE Trans. Magn. **32**(3), 1405–1408 (1996)
29. Romano, V.: 2d numerical simulation of the mep energy-transport model with a finite difference scheme. J. Comput. Phys. **221**, 439–468 (2007)
30. Romano, V., Rusakov, A.: 2d numerical simulations of an electron-phonon hydrodynamical model based on the maximum entropy principle. Comput. Meth. Appl. Mech. Eng **199**(41–44), 2741–2751 (2010)
31. Romano, V., Rusakov, A.: Numerical simulation of coupled electron devices and circuits by the mep hydrodynamical model for semiconductors with crystal heating. Il Nuovo Cimento C (2010). doi:10.1393/ncc/i2010-10573-5
32. Romano, V., Scordia, C.: Simulations of an electron-phonon hydrodynamical model based on the maximum entropy principle. In: Roos, J., Costa, L.R.J. (eds.) Scientific Computing in Electrical Engineering SCEE 2008, Mathematics in Industry. Springer, Berlin/Heidelberg (2010)

33. Romano, V., Zwierz, M.: Electron-phonon hydrodynamical model for semiconductors. 2741–2751 (2008) doi:10.1016/j.cma.2010.06.005
34. Schöps, S., Bartel, A., De Gersem, H., Günther, M.: DAE-index and convergence analysis of lumped electric circuits refined by 3-d magnetoquasistatic conductor models. Preprint 08/06, Bergische Universität Wuppertal, Wuppertal (2008)
35. Selberherr, S.: Analysis and simulation of semiconductor devices. Springer, Wien/New York (1984)
36. Stratton, R.: Diffusion of hot and cold electrons in semiconductor barriers. Phys. Rev. **126**, 2002–2014 (1962)
37. Tsukerman, I.: Finite element differential-algebraic systems for eddy current problems. Numer. Algorithms **31**(1), 319–335 (2002)