A New Tightly-Coupled Transient Electro-Thermal Co-Simulator with Capacitance and Matrix Exponential Method

ABSTRACT

This paper presents a new transient electro-thermal (ET) simulation method for fast 3D chip-level analysis of power electronics with field solver accuracy. The metallization stacks are meshed and solved with 3D field solver using nonlinear temperature-dependent parameters, and the active devices are modeled with nonlinear tabular compact models to avoid time-consuming TCAD simulation. The main contributions include: 1) A tightly-coupled formulation that solves the electrical and thermal responses simultaneously for better convergence property; 2) Explicit account of capacitive effects, including interconnect parasitic capacitance and gate capacitance of power devices, to improve modeling accuracy in high-frequency applications; 3) A specialized transient solver based on the matrix exponential method (MEXP) to address the multi-scale problem caused by the considerably different time scales in electrical and thermal dynamics. Numerical experiments have demonstrated the advantages of the proposed co-simulation framework.

1. INTRODUCTION

Bipolar-CMOS-DMOS (BCD) integration is a key technology for power integrated circuits (ICs) offering many advantages by integrating three distinct types of devices on a single die. New challenges, however, have also been triggered to the thermal managements in the BCD technology due to the closer proximity of high-power DMOS transistors to other temperature-sensitive components and the more complicated geometry and material configurations. Accurate prediction in temperature profile is needed to guide heat removal design and avoid potential reliability issues such as electromigration and negative bias temperature instability (NBTI). To this end, the strong coupling between electrical and thermal dynamics, e.g., the nonlinear temperature dependencies of electrical parameters and device characteristics [10], must be appropriately accounted. In addition, DMOS is often used as switches in BCD and operates under pulse inputs. The peak temperature of devices can go up and down in one pulse period and exceed the maximum threshold momentarily causing permanent damage, which cannot be detected by steady-state analysis [2]. Therefore, an accurate transient electro-thermal (ET) co-simulation is highly desired to detect and avoid thermal failures in the early stage of BCD designs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

Most existing transient ET simulation methods fall into the loose coupling category. Specialized modeling methods, such as field solution or equivalent network analysis, are employed separately to simulate the electrical and thermal responses, while the ET coupling is achieved by communication between the two solvers via an appropriate interface [7, 8]. Despite its convenience to implement, the loose coupling strategy requires a careful bookkeeping of the interaction between electrical and thermal variables, and a complicated software communication scheme to ensure efficient information exchange [12]. The deliberate split of electrical and thermal dynamics may also lead to slow convergence when the ET coupling is strong. Moreover, the iteration between the two solvers remains of a serial nature limiting the parallelizability of the simulation.

Alternatively, a tightly-coupled ET simulation assembles the electrical and thermal dynamics in one numerical system and solves the two sets of variables simultaneously. The concurrent treatment to multiple physics generally leads to a more natural and consistent characterization of the physical interactions, and subsequently to a more automated analysis, a faster convergence for strongly coupled cases and more effective parallelization. A recent tightly-coupled ET co-simulation method was proposed in [12] for transient analysis of power MOSFETs. The method employs a field-based solution to on-chip metallization to address the quest to model metal layers with high spatial resolution to predict accurate voltage drop in the BCD technology and a nonlinear temperature-dependent table model for device currents to avoid detailed time-consuming the TCAD solution. The electrical solver is then coupled with a whole-domain thermal field solver in a tightly-coupled fashion. Being one step closer to the "ab initio" simulation, the ET solver in [12] can determine the voltage drop in the metal structures and the device temperatures with high accuracy and without limiting the applicability to special cases.

One difficulty usually facing tightly-coupled ET methods is the multi-scale nature of the resulting numerical system, due to the large difference in the electrical and thermal time constants. The fast transients in the electrical system require time step size of typically ns scale while the temperature varies generally on μs to ms scale. The tightly-coupled method in [12] avoids this difficulty by assuming the response of the electrical system, including metallic wires and power devices, to be instantaneous. Therefore no temporal differential equation needs to be solved for the electrical part, and the whole ET simulation can be carried out on the thermal scale only. This simplification, however, is not always valid. When the devices contain a large summed gate capacitance, e.g., with many fingers, and operate at relatively high switching frequencies, the electrical time scale becomes relevant and the capacitive effects in the structures should be taken into account for the sake of simula-

tion accuracy. The multi-rate integration [11] is a common strategy to mitigate the high time scale contrast by integrating the subset of fast-varying variables with a smaller step size than the slow-varying subset. Nevertheless, the fastest transients in the ET problem are induced by the small parasitics in on-chip metallization. These transients are generally not important for heat generation, but may dictate unnecessarily small step size even in the multi-rate integration when methods of low accuracy order are applied.

This work proposes a new tightly-coupled transient ET simulation technique for high-speed power devices. The proposed method adopts the "electrical field solver for metallization + compact models for devices + thermal field solver" strategy as in [12]. The capacitive effects in the back-end structure and the front-end devices are explicitly accounted to enhance modeling accuracy. A specialized matrix exponential (MEXP) method with Rosenbrock-type formulation, adaptive time stepping and Jacobian bypass is developed to expedite the time integration involving largely different time scales. Numerical experiments are conducted to confirm the advantages of the proposed ET solver.

2. TIGHTLY-COUPLED ET FORMULATION

2.1 Basic Electrical and Thermal Models

The electrical field solver is applied to the back-end structures, which solves the current-continuity equation under the electro quasistatic assumption because of the relatively weak magnetic effects in power electronics [12]

$$-\nabla\left(\varepsilon\nabla\frac{\partial V}{\partial t}\right)+\nabla\cdot J=0, \tag{1}$$

with

$$\begin{cases} J=\sigma(T)E \\ E=-\nabla V \end{cases}$$

where the first term in (1) describes the displacement current density, J is the conduction current density, E the electrical field, V the electrical potential, and ε the permittivity. The conductivity $\sigma(T)$ for metals has a temperature dependency modeled by the Wiedemann-Franz law

$$\sigma(T)=\sigma_0\left(\frac{T}{T_0}\right)^{\alpha_\sigma}, \tag{2}$$

where the exponent α_σ is a material-dependent parameter.

The drain-source current of semiconductor devices is modeled by the table model taking into account the temperature dependence

$$I_{ds}=f\left(V_{gs},V_{ds},T_{ds}\right), \tag{3}$$

where V_{gs} and V_{ds} are the gate-source and the drain-source voltages, and T_{ds} is the device temperature taken as the average of the temperatures at the source and the drain terminals. To link the table models to the electrical field solvers, the channel layers of the MOS devices, which are geometrically negligible compared to the size of die, are effectively "removed" from the 3D mesh of the substrate. Relevant terminal voltages are measured at proper metal ends and passed to the table models, which in turn generates currents as external current sources at the corresponding terminals in the field solving process. The table model can be extracted from TCAD device simulation or measurement data.

The thermal field solver solves the heat conduction equation in the entire domain:

$$C_T\frac{\partial T}{\partial t}-\nabla\left(\kappa(T)\nabla T\right)-Q=0, \tag{4}$$

Where T is the temperature, C_T the thermal capacitance and Q the heat sources or sinks. For the temperature dependent thermal conductivity $\kappa(T)$, we adopt the widely accepted model [9]:

$$\kappa(T)=\kappa_0\left(\frac{T}{T_0}\right)^{\alpha_\kappa}, \tag{5}$$

in which the exponent α_κ is again material-dependent. In this work we assume the thermal capacitance is temperature-independent.

Solving (5) requires knowledge to the heat generation term Q, which has several contributors: 1) the Joule self-heating of the metal structures and the substrates; 2) the self-heating of the active devices and 3) heat injected or extracted at the boundary of the simulation domain. Similar to [12], the first two contributions are calculated by

$$Q=\begin{cases} E\cdot J=\sigma(T)|\nabla V|^2, & \text{conductors} \\ I_{ds}\cdot V_{ds}, & \text{active devices} \end{cases} \tag{6}$$

from the information provided by the electrical solver. The third one is addressed by thermal contacts placed on the domain boundary enforcing fixed temperatures or fixed heat flux.

2.2 Involvement of Capacitance

One main assumption in [12] is that $\partial V/\partial t=0$, reducing (1) to an algebraic Laplacian equation with the solution solely determined by the instantaneous boundary condition. This numerical convenience, however, comes at the cost of neglecting the capacitive effects of the metallization structures and the devices, which may affect the modeling accuracy for scenarios with nontrivial capacitances, such as devices with a large number of transistor fingers and in microwave applications [13]. Therefore, one contribution of this work is to include these capacitive effects to make the electrical system not respond instantaneously to external excitations.

The parasitic capacitances of the back-end structures is readily accounted for by including the displacement current in (1). To model the gate capacitance of DMOS, each polysilicon finger is divided up into equal sections, and for each section one gate-source capacitance C_{gs} and one gate-drain capacitance C_{gd} are attached, which are modeled as external capacitors in the field solving process of the back-end structures. To account for the charging and discharging currents of the capacitors, extra terms are added to (1) when solved at relevant nodes. For instance, the equation at a gate node reads

$$-\nabla\left(\varepsilon\nabla\frac{\partial V}{\partial t}\right)+\nabla\cdot J+C_{gs}\frac{\partial(V_g-V_s)}{\partial t}+C_{gd}\frac{\partial(V_g-V_d)}{\partial t}=0, \tag{7}$$

where capacitive current paths are introduced between the gate and the source (drain) terminals. Note that the gate capacitance is generally dominant over the parasitic capacitance of metal wires.

The values of the gate capacitors are calculated from a specific capacitance per unit width, a user-input parameter or determined from the material parameter and geometry of the transistor structure. In the current work all the device capacitances are considered independent of terminal voltages and temperatures, though these dependencies will be incorporated by using temperature-aware compact models of DMOS such as [6].

2.3 Tightly-Coupled Formulation

The tightly-coupled system is formed by combining the electrical system (1) and the thermal system (4), as shown in (8). The term C_d collects the gate capacitors of devices, $I(V,T)$ denotes the nonlinear device currents determined by (3) and $b^{(E)},b^{(T)}$ contain respectively the electrical and thermal boundary conditions. The nodal potential vector V of N_V length and the nodal temperature vector T of N_T length are the primary unknowns to be determined.

$$\begin{bmatrix} -\nabla\cdot(\varepsilon\nabla)+C_d & 0 \\ 0 & C_T I \end{bmatrix} \begin{bmatrix} \dot{V} \\ \dot{T} \end{bmatrix} = -\begin{bmatrix} -\nabla\cdot[\sigma(T)\nabla] & 0 \\ 0 & -\nabla\cdot[\kappa(T)\nabla] \end{bmatrix} \begin{bmatrix} V \\ T \end{bmatrix} - \begin{bmatrix} I(V,T) \\ Q(V,T) \end{bmatrix} - \begin{bmatrix} b^{(E)} \\ b^{(T)} \end{bmatrix} \tag{8}$$

After discretizing by the finite volume method (FVM), (8) can be cast into a matrix equation form

$$\mathbf{C}\frac{dx}{dt} = -\mathbf{G}(x)x - F(x) - b, \tag{9}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_E & \\ & \mathbf{C}_T \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{G}_E & \\ & \mathbf{G}_T \end{bmatrix}, F = \begin{bmatrix} I \\ Q \end{bmatrix}, x = \begin{bmatrix} V \\ T \end{bmatrix}. \tag{10}$$

The nonlinear differential equation (9) is commonly solved by the linear multi-step methods (LMM), which reduces (9) to an algebraic equation using a polynomial approximation to the time derivative. For instance, the backward Euler (BE) method results in

$$\frac{\mathbf{C}}{h}x_{n+1} + \mathbf{G}(x_{n+1})x_{n+1} + F(x_{n+1}) - \frac{\mathbf{C}}{h}x_n + b_{n+1} = 0, \tag{11}$$

in which h is the time step size between the nth step and the $(n+1)$th step. Then the Newton's method is applied to solve (11) iteratively

$$\left(\mathbf{J}(x_{n+1}^k) + \frac{\mathbf{C}}{h}\right)\Delta x_{n+1}^{k+1} = -\left(F(x_{n+1}^k) + \mathbf{G}(x_{n+1}^k)x_{n+1}^k + \frac{\mathbf{C}}{h}\left(x_{n+1}^k - x_n\right) + b_{n+1}\right), \tag{12}$$

where Δx_{n+1}^{k+1} denotes the update of x at the $(k+1)$th Newton iteration in the $(n+1)$th step. The Jacobian of the nonlinear function $\mathbf{G}(x)x + F(x)$ is given by

$$\mathbf{J} = \begin{bmatrix} \mathbf{G}_E + \frac{\partial I}{\partial V} & \frac{\partial(\mathbf{G}_E V)}{\partial T} + \frac{\partial I}{\partial T} \\ \frac{\partial Q}{\partial V} & \frac{\partial(\mathbf{G}_T T)}{\partial T} + \frac{\partial Q}{\partial T} \end{bmatrix}. \tag{13}$$

3. TIME INTEGRATION VIA MATRIX EXPONENTIAL METHOD

Solving the tightly-coupled system (8) by LMM suffers from the multi-scale difficulty. The minimum step size has to be sufficiently small to resolve the fastest transients induced by the metal interconnects, which are several orders slower than the thermal dynamics. Nevertheless, an accurate capture of the fine-scale details of electrical waveform is not necessary for temperature prediction. The extremely short duration of these transients limits the amount of heat generation. More importantly, the power dissipation originates mainly from power devices operating with high currents and voltages, and to a lesser degree from interconnect metallization. The pace of power dissipation in power devices, however, can be orders slower than in metal structures due to the larger device capacitance. Put it differently, the temperature variation may still be predicted to a reasonable extent, provided that the relatively slow-varying heat generation from the active devices is properly captured even on a coarser time scale. The accurate characterization of on-chip metallization, in this regard, is more relevant in providing correct voltage drops to determine the operation status of power devices and the power dissipation thereby generated, than in computing the Joule heating of the structures themselves [12].

Increasing time step size to bypass the fast transients, however, is risky for low-order LMM, such as BE and the trapezoidal methods. Failure in resolving the rapidly varying components may result in excessive accumulation and propagation of the local truncation error (LTE), leading to inaccurate and even unstable simulation. Therefore, it is desired to develop high-order transient simulation techniques that can skip the fast and less important transients in metallization in a safe manner, and allow the system to be simulated at the same pace as the operation of the power devices. To this end, we extend the matrix exponential method (MEXP), originally developed for circuit simulation [14–16] and later for EM-TCAD problems [3], to the coupled ET simulation.

3.1 Formulation of MEXP

The MEXP method starts by transforming (9) to a nonlinear ODE [15]:

$$\frac{dx}{dt} = \mathbf{C}^{-1}\mathbf{G}(x)x - \mathbf{C}^{-1}F(x) - \mathbf{C}^{-1}b. \tag{14}$$

Note that the potential singularity in $\mathbf{C}$ can be readily removed by differentiating the Gauss's law [3] and the input b is assumed piecewise linear (PWL) with constant derivative in each step.

To handle the strong nonlinearity, we use a Rosenbrock-type formulation, which employs a linearization of (14) at each time step

$$\begin{aligned} \frac{dx}{dt} &= -\mathbf{C}^{-1}\mathbf{J}_n x - \mathbf{C}^{-1}\left(\mathbf{G}(x)x + F(x) - \mathbf{J}_n x\right) - \mathbf{C}^{-1}b \\ &= -\mathbf{A}_n x - \mathbf{C}^{-1}D_n(x) - \mathbf{C}^{-1}b, \end{aligned} \tag{15}$$

where $\mathbf{J}_n$ is the Jacobian (13) evaluated at x_n and $D_n(x)$ is a nonlinear difference function. The analytical solution of (15) is

$$x_{n+1} = e^{\mathbf{A}h}x_n + \int_0^h e^{\mathbf{A}(h-\tau)}\left[-\mathbf{C}^{-1}D_n(t_n+\tau) - \mathbf{C}^{-1}b(t_n+\tau)\right]d\tau. \tag{16}$$

Applying the first-order approximation to D_n results in the second-order MEXP-Rosenbrock method [1]

$$x_{n+1} = e^{\mathbf{A}h}x_n + \frac{e^{\mathbf{A}h} - I}{\mathbf{A}}\left(-\mathbf{C}^{-1}\left(D_n(x_n) + b_{n+1}\right)\right). \tag{17}$$

MEXP is attractive in this scenario as it solves the linearized subpart of (15), i.e., $\frac{dx}{dt} = \mathbf{A}\mathbf{x}$ exactly and thus avoids in the first place the error from the finite-difference approximation of time derivative underlying most existing methods [4]. The main source of error in MEXP is only from the numerical computation of the matrix exponential (more precisely the product of matrix exponential with a vector), and thus the order of accuracy of MEXP can be made fairly high to allow large step size while maintaining sufficient accuracy [15].

3.2 Computation of $e^{\mathbf{A}h}v$ via Rational Krylov Subspace Method

As shown in (17), the primary computation in MEXP is the product of matrix exponential times a vector $e^{A}v$. Approximation based on Krylov subspace projection is attracting increasing attention in recent years for its capability to handle problems with millions unknowns [3, 15]. Nevertheless, our target step size is rather aggressive ($\sim \mu s$) compared to those reported in [3,16], and MEXP based

on the ordinary Krylov subspace is not efficient for approximating large-magnitude eigenvalues (corresponding to the slow transients) [4]. Therefore, we choose to use the shift-and-invert (SAI) Krylov subspace method, which is essentially an "inverse" version of the standard Krylov subspace methods and can provide a better approximation to slow manifold of the waveform to allow larger step sizes [17, 18]. The main step of SAI-Krylov is an m-step Arnoldi process applied to $(I-\gamma\mathbf{A})^{-1}$

$$(I-\gamma\mathbf{A})^{-1}V_m = V_m H_m + H(m+1,m)v_{m+1}e_m^T, \quad (18)$$

in which V_m is an orthonormal basis of the m-dimensional Krylov subspace $\mathcal{K}_m\left((I-\gamma\mathbf{A})^{-1}, v\right)$ with γ a shift parameter. H is the upper Hessenberg coefficient matrix and H_m the leading $m\times m$ submatrix. Then the MEXP-vector product is approximated as its orthogonal projection onto the SAI-Krylov subspace

$$e^{\mathbf{A}h}v \approx V_m V_m^T e^{\mathbf{A}h}v = \beta V_m e^{\widetilde{H}_m h/\gamma}e_1, \quad (19)$$

with

$$\widetilde{H}_m = \left(I - H_m^{-1}\right), \ \ \beta = \|v\|_2$$

A posteriori error estimate is given by

$$err_{Krylov} = \frac{\beta}{\gamma}H(m+1,m)\|(I-\gamma\mathbf{A})v_{m+1}e_m^T H_m^{-1}e^{\widetilde{H}_m h/\gamma}e_1\|. \quad (20)$$

3.3 Adaptive Time Step and Jacobian Bypass

Adaptive time step size usually offers better accuracy and performance than fixed step size during transient simulation. Such adaptivity is in particular desirable in the ET co-simulation since the conductivities and the electrical time constants may vary substantially with the temperature, rendering an *a priori* step size selection more difficult and less efficient. Like the standard Krylov subspace, the SAI-Krylov subspace also processes the scaling invariant property that allows convenient re-computation of the solution without generating a new subspace. When the step size is changed from h to h_1, the new solution can be updated via

$$e^{\mathbf{A}h_1}v \approx \beta V_m e^{\widetilde{H}_m h_1/\gamma}e_1, \quad (21)$$

where V_m and H_m in (18) are re-used, and only a small-sized matrix exponential of $\widetilde{H}_m h_1/\gamma$ needs to be re-evaluated. This is a marked advantage. LMM methods, in contrast, require a new matrix factorization whenever step size is changed [3]. The adaption of step size h is based on an embedded local error estimation

$$err = x_{n+1} - \tilde{x}_{n+1}, \quad (22)$$

where $\tilde{x}_{n+1}$ is obtained by substituting x_{n+1} into (17)

$$\tilde{x}_{n+1} = e^{\mathbf{A}h}x_n + \frac{e^{\mathbf{A}h}-I}{\mathbf{A}}\left(-\mathbf{C}^{-1}\left(D_n(x_{n+1}) + b_{n+1}\right)\right). \quad (23)$$

One computation-intensive step in the above MEXP method is the factorization of the Jacobian matrix in each step. To improve the computational efficiency, we adopt the Jacobian bypass technique in [5]. For a given tolerance tol, if the solutions of two adjacent steps are close enough, i.e.,

$$\|x_{n+1} - x_n\| < tol\|x_{n+1}\|, \quad (24)$$

the Jacobian (13) can be passed from the nth step to the $(n+1)$th step and the LU factors can be re-used. The Jacobian bypass technique can be interpreted as an adaptive multi-rate approach [5].

3.4 Complexity Analysis

In this subsection we will briefly analyze the complexity of transient ET analysis using the loosely-coupled scheme, the tightly-coupled scheme with LMM and the tightly-coupled scheme with MEXP, denoted as LC, TC+LMM and TC+MEXP hereafter. For simplicity we assume fixed step sizes are applied in all time integration. Since the factorization of the Jacobian matrices is the most expensive step in all the methods, we use the number and the size of matrix factorization as the metric to measure the complexity of different co-simulation schemes.

For the loosely-coupled scheme, suppose n_T thermal steps are used, each of which contains n_E electrical steps and requires an average n_{loop} ET iterations to converge, and each electrical step needs an average of $n_{Newton1}$ Newton iterations, the whole calculation requires $n_{fac1}^{LC} = n_T \times n_E \times n_{loop} \times n_{Newton1}^{LC}$ factorizations of the electrical Jacobian of the size $N_V \times N_V$. If each thermal solving needs $n_{Newton2}$ Newton iterations, the thermal part needs $n_{fac2}^{LC} = n_T \times n_{loop} \times n_{Newton2}$ factorizations of the thermal Jacobian of the size $N_T \times N_T$.

In the tightly-coupled scheme, both the electrical and thermal variables evolve with the same step size and no ET loop is needed. Suppose the LMM methods need n_{ET}^{LMM} time steps; each requires N_{Newton}^{LMM} Newton iterations, then in total $n_{fac}^{TC+LMM} = n_{ET}^{LMM} \times n_{Newton}^{LMM}$ matrix factorizations are involved in the simulation. The size of the Jacobian is $(N_V + N_T) \times (N_V + N_T)$. For the MEXP method, only one Jacobian factorization is required in each step, and thus for n_{ET}^{MEXP} steps only $n_{fac}^{TC+MEXP} = n_{ET}^{MEXP}$ factorizations need to be performed.

In general, the loosely-coupled scheme needs more factorizations of smaller matrices, whereas the tightly-coupled schemes require fewer factorizations of larger matrices. The best choice is problem-dependent and must be made based on a combined consideration of all the relevant factors. For instance, a typical relation $N_T \approx 1.5N_V$ is observed in the examples we have tested. Assume further that the sparse matrix factorization is of a complexity of $O(N^2)$, the cost of the loosely-coupled calculation can be estimated by $\left(n_{fac1}^{LC} + 2.25n_{fac2}^{LC}\right)O(N_V^2)$, while the cost of the tightly-coupled calculation with LMM and MEXP are respectively $6.25n_{fac}^{TC+LMM}O(N_V^2)$ and $6.25n_{fac}^{TC+MEXP}O(N_V^2)$. As will be shown in the next section, the TC+MEXP combination generally offers the smallest pre-factor among the three schemes

4. NUMERICAL RESULTS

The proposed tightly-coupled transient ET simulator is implemented in Matlab. For comparison the loosely-coupled version is also implemented. A practical LDMOS device is used to verify the proposed simulator. The 3D view and the specifications of the test structure are shown in Fig. 1. The 3D FVM discretization results in a mesh with $N_V = 217,202$ potential unknowns and $N_T = 337,869$ temperature unknowns. Since the MEXP method uses a second-order formulation, we apply the second-order backward differentiation formula (BDF2) as the LMM solver to ensure a fair comparison. A 1MHz trapezoidal pulse with 50% duty cycle and 5V amplitude is used as gate driving signal throughout the testing.

We start with a comparison between the three ET co-simulation schemes discussed in Sec. 3.4, i.e., LC, TC+LMM and TC+MEXP, for different degrees of ET coupling. First a small $V_{ds} = 5V$ is used to limit the current and heat generation, leading to a weak ET coupling. Then a large $V_{ds} = 50V$ is applied to enable a strong ET coupling, wherein the current is higher and the temperature variation is more dramatic. All the transient simulations are performed

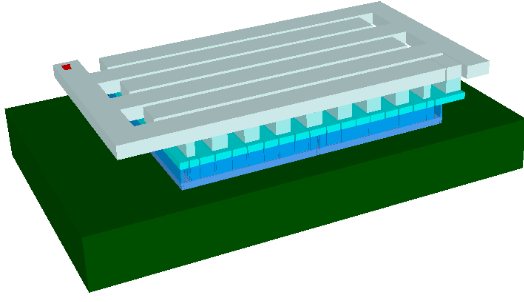

Figure 1: 3D view of the power LDMOS structure (the die part is truncated for better visualization). The back-end structure consists of 6 metal layers. The die has an area of $750 \times 450 \mu m^2$ and a thickness of $400 \mu m$. The LDMOS has 94 transistor fingers with an $18.45mm$ total gate length. The summed C_{gs} and C_{gd} are $420pF$ and $105pF$, respectively.

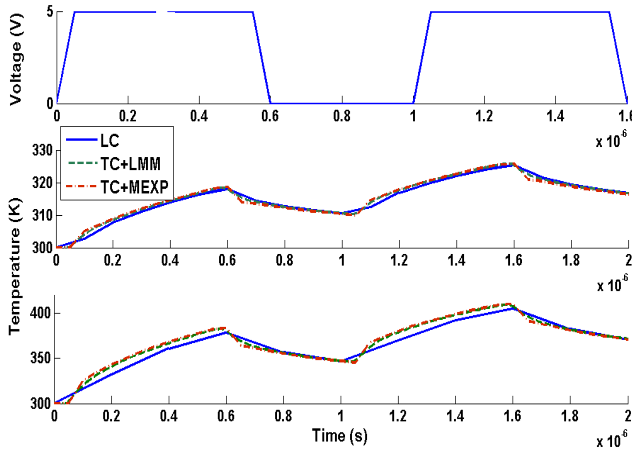

Figure 2: Top: gate driving waveform. Middle: Device 94 temperature with low ET coupling ($V_{ds} = 5V$). Bottom: Device 94 temperature with high ET coupling ($V_{ds} = 50V$).

with fixed step sizes, which are chosen carefully to be the minimal values that can produce accurate results for both the weak and strong coupling cases. The transient temperature waveforms of one device are shown in Fig. 2, in which three schemes agree well with each other. The temperature starts to increase when the MOSFETs are turned on and decrease after the devices are turned off. In the second period of the high ET coupling case, the temperature difference between the peak and the end point can reach $39K$, which will be omitted if one only simulates the temperature at the end point of each period and may cause damage to the devices. This indicates the necessity of a sufficient resolution to the transient temperature variation for high-power applications.

The performance of the three schemes is summarized in Table 1 for one period. The LC scheme uses 5 thermal steps and each contains 20 electrical steps, rendering 100 electrical steps in total. TC+LMM requires the same number of steps to guarantee the same resolution to the fastest electrical transients. On the other hand, TC+MEXP can take a $5X$ larger step size as it can safely bypass the fast transients induced by the metallization, and thus needs only 20 steps. Regarding the number of Jacobian factorizations, LC needs $n_{fac1} = 390$ factorization of electrical Jacobian and $n_{fac2} = 24$ factorizations of thermal Jacobian for the low ET coupling case. When the coupling becomes stronger, the performance of LC deteriorates and n_{fac1} and n_{fac2} increase significantly to 552 and 46, due to more ET loops are needed to achieve the ET convergence. On the other hand, the number of Jacobian factorizations remains nearly constant for the two TC schemes, demonstrating a better convergence property of the TC strategy when applied to handle strong ET interactions. In terms of runtime, in spite of factoring fewer Jacobians than LC, TC-LMM is slower than LC due to the larger matrix size in factorization. The factorization times for the electrical, thermal and combined Jacobian are $4.1s$, $10.3s$ and $39.2s$, respectively. The saving in factorization number in LMM does not suffice to compensate the speed loss in each factorization. In contrast, the TC+MEXP combination delivers the best performance in both cases owing to the more significant reduction in the number of Jacobian factorizations.

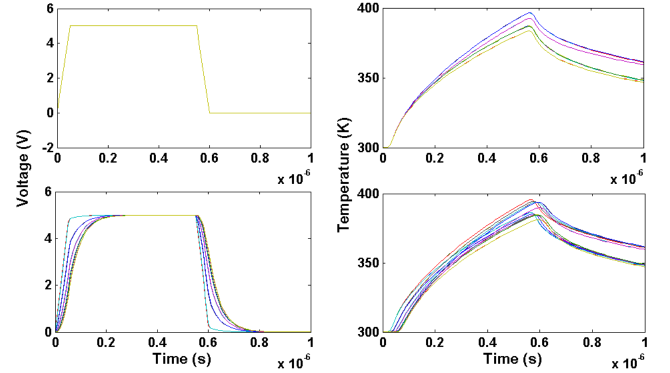

Figure 3: Top left: gate voltages of 20 fingers at different locations without capacitance. Top right: drain temperature of 20 fingers without capacitance. Bottom left: gate voltage with capacitance. Bottom right: drain temperature with capacitance.

Next, we demonstrate the importance of an explicit inclusion of capacitive effects in the ET co-analysis. We simulate the LDMOS with and without including the capacitance in TC+MEXP, and measure the gate voltage and the drain temperature of 20 transistor fingers at different locations of the chip. As shown in Fig. 3, when the capacitive effects are not included (no RC delay), the gate voltages at different locations are identical to the instantaneous applied voltage. When the capacitance is taken into account, the RC delay in the gate voltage becomes obvious and the temperature profiles of the devices at different locations are also modified. Therefore capacitance inclusion is necessary to guarantee accurate prediction of the voltage drop over the gate net and the temperature distribution.

Table 2 shows the performance improvements from the adaptive time step (AD) and the Jacobian bypass (BP) described in Sec. 3.3. The step size adaption is based on the LTE estimate (22) and uses a set of rules similar to [3] with a relative tolerance of $tol = 10^{-3}$. The adaptive stepping nearly halves the number of time steps needed in both LMM and MEXP. However, LMM does not benefit much in terms of computation, as the step size change in LMM requires a new matrix factorization. In contrast, the adaptive stepping is highly efficient for MEXP, owing to the convenient rescaling scheme (21) to update solution without extra matrix factorization. The Jacobian bypass uses the same tolerance to determine when to skip the Jacobian factorization, and is proven to be an effective technique to accelerate the MEXP solver for both constant and adaptive step size.

5. CONCLUSION

Table 1: **Performance comparison of different ET coupling schemes**

Method	# of steps	low ET coupling		high ET coupling	
	$n_T \times n_E$ or n_{ET}	n_{fac}	time (s)	n_{fac}	time (s)
LC	5×20	390 + 24	1,854	552 + 46	2,747
TC+LMM	100	130	5,252	134	5,413
TC+MEXP	20	20	810	20	810

Table 2: Performance of TC-LMM and TC-MEXP with adaptive step size (AD) and Jacobian bypass (BP)

method	# of steps	n_{fac}	runtime (s)
LMM	100	134	5,413
LMM+AD	51	111	4,484
MEXP	20	20	810
MEXP+BP	20	11	451
MEXP+AD	12	12	493
MEXP+AD+BP	12	8	329

We have proposed a new transient ET simulation framework for accurate chip-level ET analysis for BCD integration technology. The framework features a tightly-coupled formulation to provide better handling to the strong ET interactions as a consequence of the ever-increasing functionality integration density. The capacitive effects of on-chip metallization and power devices are explicitly accounted for to improve the modeling accuracy of voltage and temperature distribution. To address the computational bottleneck arising from the different electrical and thermal time constants, a specialized nonlinear MEXP method, augmented by adaptive time stepping and Jacobian bypass, is developed. The numerical results have demonstrated the advantages of the proposed framework.

6. REFERENCES

[1] M. Caliari and A. Ostermann. Implementation of exponential rosenbrock-type integrators. *Applied Numerical Mathematics*, 59(3-4):568–581, 2009.

[2] R. Chandra. Transient temperature analysis. In *EDSSERC*, Sep 2006.

[3] Q. Chen, W. Schoenmaker, S.-H. Weng, C.-K. Cheng, G.-H. Chen, L.-J. Jiang, and N. Wong. A fast time-domain EM-TCAD coupled simulation framework via matrix exponential. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 422–428, Nov 2012.

[4] Q. Chen, W. Zhao, and N. Wong. Efficient matrix exponential method based on extended Krylov subspace for transient simulation of large-scale linear circuits. In *Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific*, pages 262–266, Jan 2014.

[5] M. Culpo. *Numerical algorithms for system level electro-thermal simulation*. PhD thesis, Bergischen Universitat Wuppertal, 2009.

[6] Freescale. Freescale Semiconductor's MET LDMOS model, 2000.

[7] R. Gillon, P. Joris, H. Oprins, B. Vandevelde, A. Srinivasan, and R. Chandra. Practical chip-centric electro-thermal simulations. In *14th International Workshop on Thermal Inveatigation of ICs and Systems (THERMINIC)*, pages 220–223, Sep 2008.

[8] B. Krabbenborg, A. Bosma, H. de Graaff, and A. Mouthaan. Layout to circuit extraction for three-dimensional thermal-electrical circuit simulation of device structures. 15(7):765–774, Jul 1996.

[9] M. Pfost, C. Boianceanu, H. Lohmeyer, and M. Stecher. Electrothermal simulation of self-heating in DMOS transistors up to thermal runaway. 60(2):699–707, Feb 2013.

[10] M. Pfost, D. Costachescu, A. Mayerhofer, M. Stecher, S. Bychikhin, D. Pogany, and E. Gornik. Accurate temperature measurements of DMOS power transistors up to thermal runaway by small embedded sensors. *IEEE Transactions on Semiconductor Manufacturing*, 25(3):294–302, Aug 2012.

[11] V. Savcenco, W. Hundsdorfer, and J. Verwer. A multirate time stepping strategy for stiff ordinary differential equations. *BIT Numerical Mathematics*, 47(1):137–155, 2007.

[12] W. Schoenmaker, O. Dupuis, B. De Smedt, P. Meuns, J. Ocenasek, W. Verhaegen, D. Dumlugol, and M. Pfost. Fully-coupled 3D electro-thermal field simulator for chip-level analysis of power devices. In *19th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC)*, pages 210–215, Sep 2013.

[13] S. Theeuwen and H. Mollee. LDMOS transistors in power microwave applications. *dec. white paper at Micr. Journ. web*, 2008.

[14] S.-H. Weng, Q. Chen, and C.-K. Cheng. Circuit simulation using matrix exponential method. In *ASIC (ASICON), 2011 IEEE 9th International Conference on*, pages 369–372, Oct 2011.

[15] S.-H. Weng, Q. Chen, and C.-K. Cheng. Time-domain analysis of large-scale circuits by matrix exponential method with adaptive control. 31(8):1180–1193, 2012.

[16] S.-H. Weng, Q. Chen, N. Wong, and C.-K. Cheng. Circuit simulation via matrix exponential method for stiffness handling and parallel processing. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 407–414, Nov 2012.

[17] H. Zhuang, S.-H. Weng, and C.-K. Cheng. Power grid simulation using matrix exponential method with rational Krylov subspaces. In *IEEE 9th International Conference on ASIC (ASICON)*, pages 369–372, Oct 2013.

[18] H. Zhuang, S.-H. Weng, J.-H. Lin, and C.-K. Cheng. MATEX: A distributed framework for transient simulation of power distribution networks. In *IEEE/ACM Design Automation Conference (DAC)*, pages 81:1–81:6, 2014.