Enabling Full-System Transient Power Integrity Verification via Model Order Reduction and Waveform Relaxation

Enabling full-system transient power integrity verification via Model Order Reduction and Waveform Relaxation

Antonio Carlucci (1), Alessandro Moglia (1), Stefano Grivet-Talocia (1), Siddharth Kulasekaran (2), Kaladhar Radhakrishnan (2)

(1) Dept. Electronics and Telecommunications, Politecnico di Torino, C. Duca degli Abruzzi 24, 10129 Torino, Italy (email: antonio.carlucci@polito.it, alessandro.moglia@polito.it, stefano.grivet@polito.it)

(2) Intel Corporation, Chandler, AZ, USA (email: siddharth.kulasekaran@intel.com, kaladhar.radhakrishnan@intel.com).

Abstract. We present a comprehensive modeling and simulation framework for transient analysis of multicore power delivery networks equipped with integrated voltage regulator banks. Numerical simulation at the system level of such structures remains a challenging task due to the large-scale nature of the equations to be solved, combined with the nonlinearities of the regulator switches and the feedback loops of the corresponding controls. We propose a solution based on a combination of structure-preserving Model Order Reduction algorithms with parallel time-domain solvers based on system partitioning and Waveform Relaxation. Application to commercial mobile and enterprise server benchmarks demonstrates a speedup as much as 1000X with respect to HSPICE, with negligible loss of accuracy.

I. Technology overview and motivations

Datacenter microprocessors have seen a sharp uptick in power consumption driven by the explosive growth in demand for large AI models (Figure 1). Microprocessors will need to rely on Integrated Voltage Regulators (IVR) to support these increased power levels [1]. IVRs deliver power to the microprocessor at elevated voltages which reduces the current requirements to deliver the same power. As power levels rise, microprocessor architects rely on fine grain power management to optimize power consumption. This leads to an increase in number of power rails as shown in Figure 2. The increase in power rail count is motivating factor in the use of IVRs which can generate a large number of power rails from a single platform level regulator.

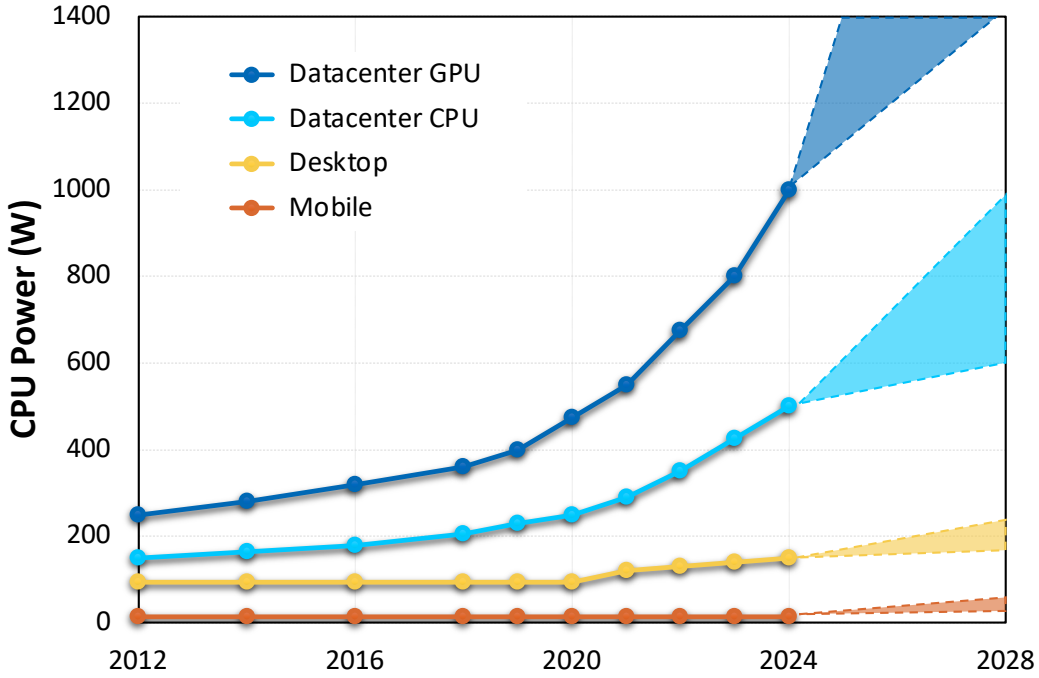


Figure 1: Power scaling of microprocessors for different applications. This figure intended to be one column (half-page) width.

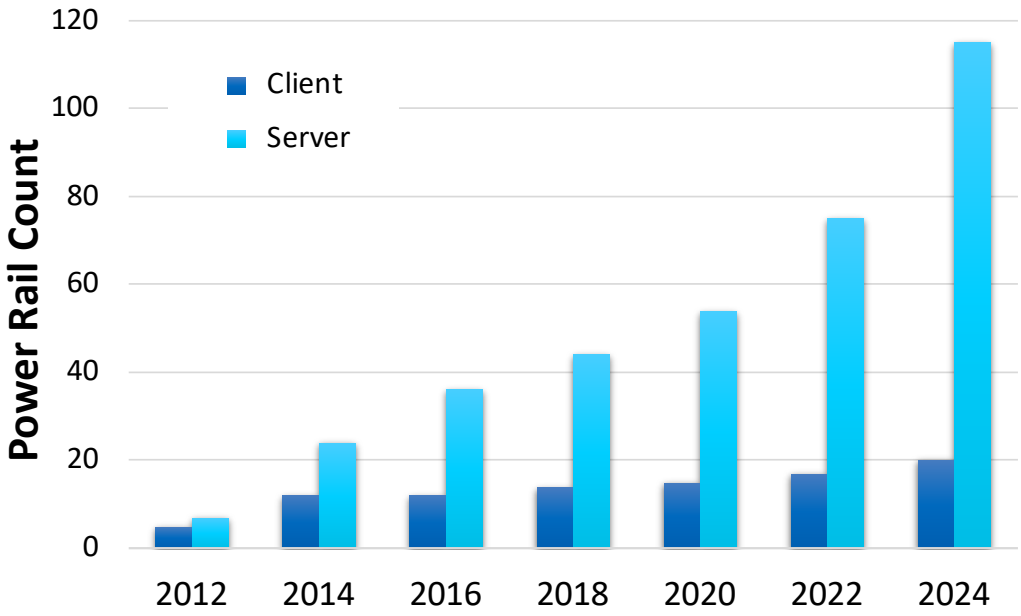


Figure 2: Evolution of the number of Power Rails for Client and Server architectures. This figure intended to be one column (half-page) width.

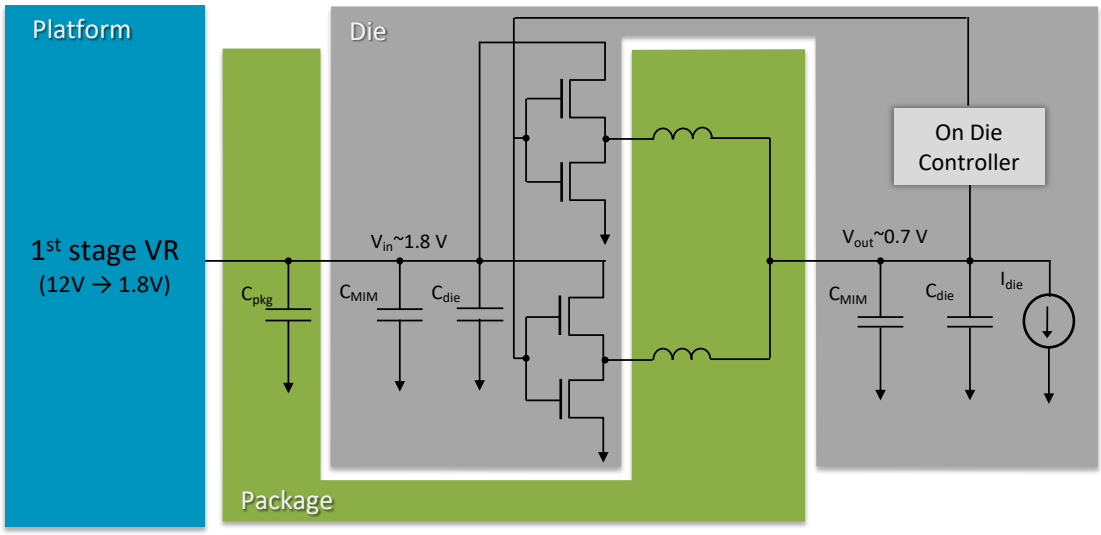


Figure 3: High-level topology of a Fully Integrated Voltage Regulator (FIVR). This figure intended to be full-page width.

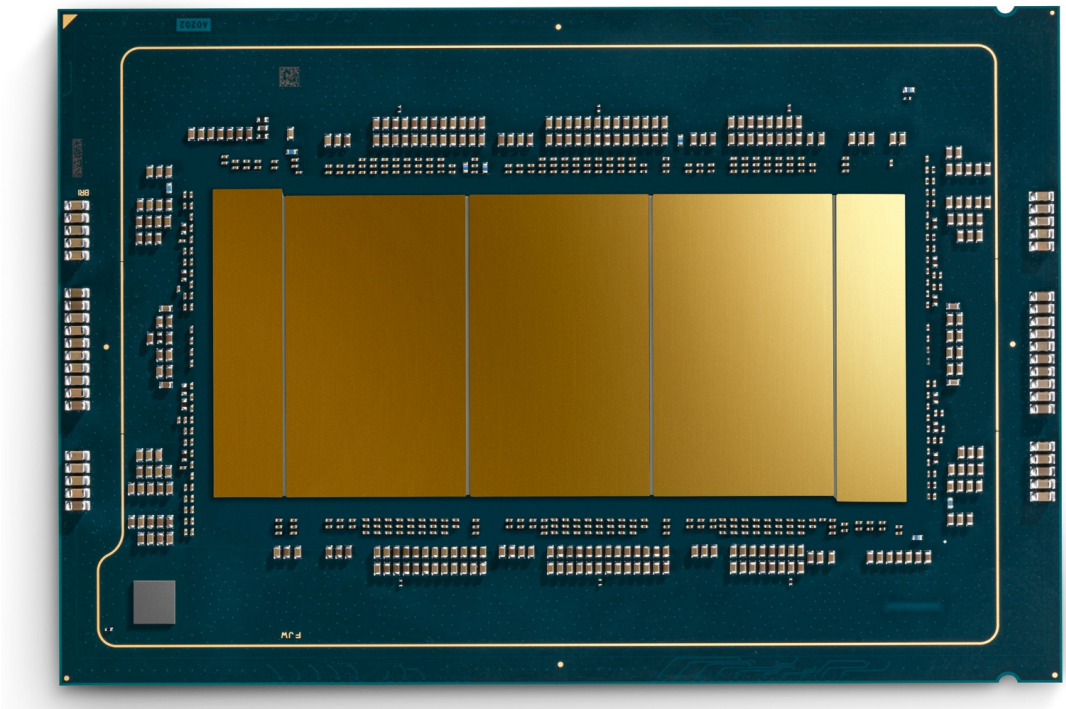


Figure 4: 6th Generation Intel Xeon Processor. This figure intended to be one column (half-page) width.

Intel uses Fully Integrated Voltage Regulators (FIVR) [2] for their datacenter Xeon family of products. FIVR is a multi-phase, synchronous switching, high frequency buck

converter. The switches, control circuits and the output filter capacitor for FIVR are implemented on silicon while the inductors are integrated into the core of the package (Figure 3). The 6th generation Xeon product shown in Figure 4 has about 800 FIVR phases that power over 100 different domains.

While IVRs provide a clear advantage in simplifying the system level power delivery challenges, they add to the Electrical Design and Analysis (EDA) complexity for microprocessors. Commercially available EDA tools and methods are not adequate to accurately analyze the power integrity of IVR based designs. Today's analysis techniques of IVR based systems makes simplifications such as ignoring noise coupling between adjacent voltage domains due to the shared input power supply. Even with these assumptions, numerical run times for SPICE based solvers in IVR systems can be very high. As microprocessors move to larger packages and wafer-scale systems, fully coupled system simulations will become numerically impossible unless numerical methods are developed to provide significant speed-ups over traditional SPICE based solvers.

II. Problem statement

This paper provides a high-level description of a modeling and simulation flow that enables transient analysis of full-scale multicore microprocessor systems equipped with FIVR banks. A detailed description of the theoretical framework and related algorithmic implementation can be found in a series of recent papers [3-6]. The description in this paper is semi-quantitative: here the objective is to illustrate the main flow while motivating why the described methods perform so well and fast. A deep analysis of the technical details is outside the scope of this paper, which is intended for a broad audience of non-specialists in EDA and related algorithms.

We refer to the general topology described in Figure 5, where an ideal voltage source provides the input to the Power Delivery Network (PDN) that is loaded by current sources representing the transient chip workload. The target systems embed high-performance microprocessors with high core count, for which a post-layout, fully coupled transient analysis is required to characterize the voltage droops induced by the interaction between chip activity and PDN parasitics. The analysis is to be conducted while accounting for fine-grained voltage regulation by FIVRs and including inter-core coupling through the common board/package portion of the PDN.

The PDN under investigation includes the following structures

- The Input network collects all interconnects and components at the board and package level up to the input switches of the FIVRs. An appropriate model for the input network must represent all parasitics and resonances of the board and the package, including suitable electrical models of decoupling capacitors and a linear model of the platform Voltage Regulator Module (VRM). The input network feeds a possibly large number N_C of independent cores.
- The FIVR switches include the power transistors that, through a suitable Pulse Width Modulation (PWM) control signal combined with an output LC filter,

provide fine-grained voltage regulation on a per-core level. Each FIVR is a multiphase buck converter with N_P phases.

- The Output network for each core subsystem includes models of the output inductor and capacitor of each FIVR, as well as the required portion of the on-chip RC power grid. For each core, a number N_O of output ports are considered, where independent transient current stimuli are applied to load the PDN. The output networks of individual cores may be identical or may even be different, in case of heterogeneous integrated systems where power needs to be delivered to loads of different nature.
- For each core, one sensing point is considered where the voltage is monitored, compared with a reference, and fed to a controller network $\mathcal{K}$ that filters the error signal and provides on output the instantaneous duty cycle signal that controls the PWM of the FIVRs.

The above-described system is a large-scale, dynamic and nonlinear network with N_C feedback loops. The transient simulation of such a system with general-purpose circuit simulators of the SPICE class is particularly challenging due to the interaction of the nonlinearities (mainly represented by the FIVR switches and by the corresponding control circuitry), with the size and complexity of the PDN interconnect models. This main fact motivates a modeling and simulation flow that exploits the general structure while trying to optimize and reduce the complexity of the various steps.

II-A. Characterizing PDN subsystems

II-A-1. Linear Interconnects

Let us consider the blocks highlighted in green color in Figure 5, namely the Input and the Output networks. All components and submodels that are part of these blocks are Linear and Time-Invariant (LTI) systems, whose behavior is governed by circuit-level and electromagnetic-level descriptions which can be cast in time-domain or equivalently in frequency-domain. In particular,

- Board and package interconnects (multilayer planes, grids, vias) are complex 3D structures that require an accurate ElectroMagnetic (EM) characterization. Commercial full-wave solvers are typically used to derive multiport scattering parameter descriptions, which are usually converted to state-space models using off-the-shelf rational macromodeling software [7,8].
- Decoupling capacitors load some of the board and package ports. Capacitor models may be either circuit-oriented (series RLC or multiple parallel-connected RLC branches), or S-parameter models from the vendors. In the latter case, a preprocessing stage through rational fitting is typically used, so that also capacitors may be represented as state-space models.
- The platform VRM is usually represented with a simple RL model, either a single series branch or multiple RL components connected in various configurations [9]. From a general perspective, this is a simple lumped circuit including few standard linear elements.
- Various other interconnects may also be modeled as lumped circuits (e.g. balls at the package-board interface, bumps at the package-chip interface, the on-

chip RC grid segments). So, the general PDN description may include an arbitrary number of RCLM components.

- The integrated (on-package) coupled inductors providing per-phase FIVR filtering together with the integrated on-chip MIM capacitors also require electromagnetic models to account for all field interactions. The resulting S-parameters from field solvers are converted to state-space macromodels.

We see that the LTI part of the PDN is available as an interconnection of submodels. The EM-based macromodels are available in state-space form

$$\begin{cases} \dot{x}_n = A_n x_n + B_n u_n \\ y_n = C_n x_n + D_n u_n \end{cases} \tag{1}$$

where for the n-th macromodel vectors u_n and y_n collect the input and output signals at all interface ports (e.g. incident or reflected scattering waves), vector x_n collects the macromodel states (loosely speaking, corresponding to the macromodel poles, see [7] for a comprehensive analysis), and matrices A_n, B_n, C_n, D_n are constants from the adopted rational fitting (and passivity enforcement) tool. Conversely, the set of lumped circuit elements and components can be assembled as a multiport system, which is characterized through a standard Modified Nodal Analysis (MNA) formulation [10]

$$G_c\, x_c + C_c\, \dot{x}_c = B_c\, u_c,\;\; y_c = L_c^T\, x_c \tag{2}$$

where x_c collects all MNA variables (nodal voltages and inductor currents), u_c and y_c collect the multiport interface inputs and outputs (port voltages and currents), and matrices G_c, C_c, B_c, L_c are constant. The interconnection of subsystems in MNA and state-space form leads to the general descriptor form

$$\begin{cases} E\, \dot{x} = A\, x + B\, u \\ \quad y = C\, x + D\, u \end{cases} \tag{3}$$

which represents as a single descriptor system all the LTI green blocks of Figure 5. Here, vectors u and y collect all voltages and currents at the input and output of the switches, at the input voltage source, and at all output ports. Matrices E, A, B, C, D are possibly very large in size ($10^4 - 10^6$ or even more) but highly sparse.

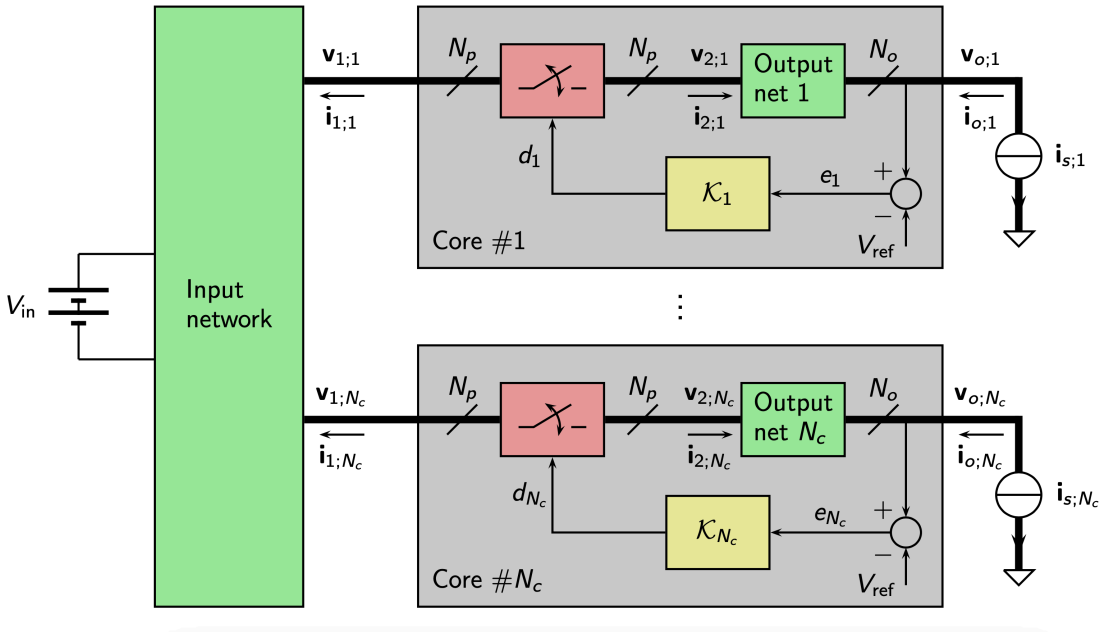


Figure 5: general structure of a Power Distribution Network (PDN) of a microprocessor architecture with N_C cores, each equipped with an N_p-phase Fully Integrated Voltage Regulator. This figure intended to be full-page width.

II-A-2. FIVR controllers

The FIVR controllers are compensators that sense the output voltages and provide on output the duty cycle signals d_k that will drive the corresponding switches. An appropriate model for the k-th core compensator can be cast as

$$\begin{cases} \dot{x}_k = A_k x_k + B_k(N_k v_k - V_{\text{ref}}) \\ d_k = Clip(C_k x_k) \end{cases} \tag{4}$$

where matrix N_k samples the desired feedback voltage from the complete set of output voltages v_k of the k-th core [which in turn are part of the full set of output variables y in (3)]. The clipping operator ensures that the duty cycle signal, for any time, does not escape the allowed range $[d_{min}, d_{max}] \subseteq [0,1]$. For typical controller topologies the number of states (size of A_k) is very small (few units) and does not contribute significantly to the overall complexity.

II-A-3. FIVR switches

The FIVR switches are a key component in the overall PDN system. Depending on the type of analysis, different models may be required. For a system-level transient PDN analysis, averaged models are more than sufficient, since the voltage ripple due to switching is not the main concern as opposed to the global dynamics of the entire PDN system with its resonances and anti-resonances. Therefore, each phase of each FIVR switch is here modeled through basic transformer equations

$$v_{2;k} = d_k\, v_{1;k},\ \ i_{1;k} = -d_k\, i_{2;k} \tag{5}$$

where $(v_{1;k}, i_{1;k})$ and $(v_{2;k}, i_{2;k})$ collect the (differential) voltages and currents on the two sides of the switches, see Figure 5. We see that, although coming from an averaged model, these two equations induce a quadratic (bilinear) nonlinearity due to the multiplication of the duty cycle signal with voltage and current signals. This makes the entire PDN system nonlinear, although the vast majority of its states are governed by linear equations (3). Fortunately, such nonlinearities are low-order and well localized, so that a structured approach is possible to compress the model size without compromising accuracy.

III. Reducing Complexity via Model Order Reduction

In this section, we describe a very effective approach to reduce the complexity of the PDN model in a preprocessing stage, before performing transient analysis. This approach is a Model Order Reduction (MOR) method tailored for the PDN structure under investigation. Based on well-established MOR theory [11,12], a novel algorithm has been developed that allows to reduce the PDN using a structured projection method. In particular, the nonlinear nature of the feedback-regulated PDN system is addressed by exploiting a linear/nonlinear decomposition that allows to reduce the linear part while still preserving the accuracy of the overall system. In addition, the particular system structure is exploited to preserve system passivity in a computationally efficient manner, a task that would not be feasible by directly applying existing methods.

III-A. Linear Model Order Reduction

The basic idea of (linear) MOR of state-space (descriptor) equations is depicted in Figure 6. The large-scale vector of states x is projected as $x = V\,\hat{x}$ onto a lower-dimensional subspace through a "tall and thin" compression matrix V, so that the size q of the reduced state vector $\hat{x}$ is orders of magnitude smaller than the original state-space size. Correspondingly, the number of state-space equations is also reduced by another projection matrix W with the same size as V. The procedure depicted in Figure 6 is denoted as Petrov-Galerkin projection, since $W \neq V$. Applying this projection to the descriptor system in (3) leads to the following reduced-size equations

$$\begin{cases} \hat{E}\,\dot{\hat{x}} = \hat{A}\,\hat{x} + \hat{B}\,u \\ \quad y = \hat{C}\,\hat{x} + D\,u \end{cases} \tag{6}$$

where the input and output vectors are left unchanged and only the matrices that involve operations with the state variables are modified (compressed, labeled with a hat accent).

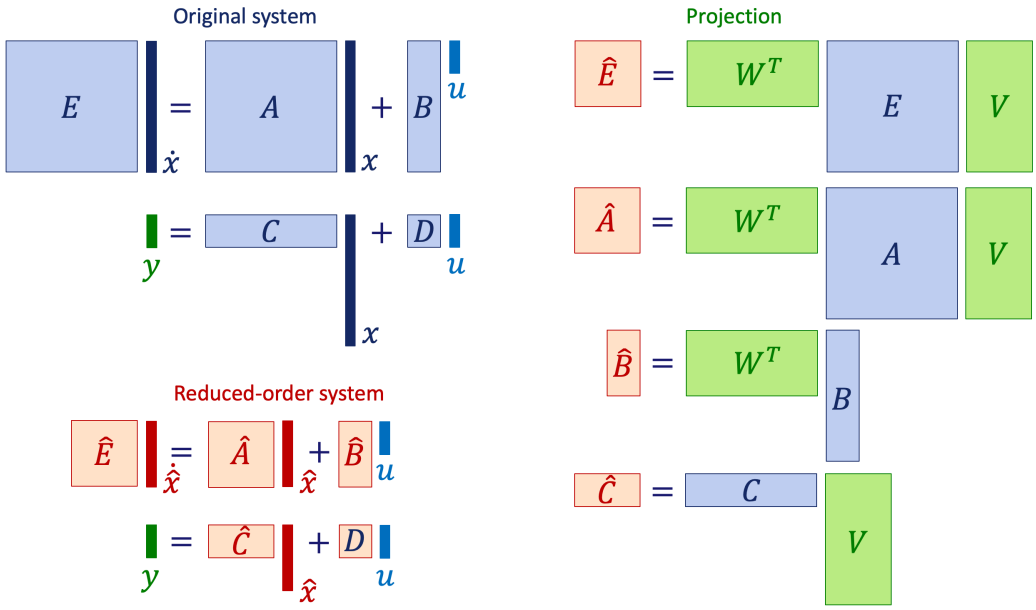


Figure 6: schematic illustration of Petrov-Galerkin projection of a descriptor system. The order of the original system (top-left) is reduced to obtain a compact system (bottom-left). The system matrices are obtained via oblique projection through a pair of matrices V, W. This figure intended to be full-page width.

The MOR procedure forces the system dynamics to evolve in a lower-dimensional space (with less variables) through projected and reduced-size equations: some nonvanishing approximation error is therefore unavoidable. The crucial point is an appropriate selection of the right and left projection matrices V, W such that this error is minimized. An extremely vast literature on MOR exists to address this problem, and the theory is very well developed. Two main approaches are available, depending on how the error $\hat{H}(s) - H(s)$ is controlled between the reduced and the original transfer function, respectively defined as

$$\hat{H}(s) = D + \hat{C}\left(s\hat{E} - \hat{A}\right)^{-1}\hat{B},\ H(s) = D + C(sE - A)^{-1}B \tag{7}$$

These two strategies are discussed below.

III-A-1. Interpolation and Moment Matching

This first approach determines V, W such that $\hat{H}(s)$ matches exactly $H(s)$ at a finite set of frequency points $\{s_i\}$. A key result [13] shows that when V includes as block-columns the matrix $X_i = (s_i E - A)^{-1}B$, then $\hat{H}(s_i) = H(s_i)$, meaning that the reduced model matches exactly the original model at each s_i. The same result is attained when W^T includes as block-rows the matrix $Z_i = C(s_i E - A)^{-1}$. If both conditions are verified, then both the value and the derivative (first-order moment) of $H(s)$ at s_i is preserved in the reduced model. Higher-order derivatives are also preserved when V (resp. W) include as columns some bases of the Krylov subspaces $Kr\{(s_i E - A)^{-1}, B\}$ and $Kr\{(s_i E^T - A^T)^{-1}, C^T\}$, respectively. Details are technical,

for an overview the reader is referred to [12,13]. In practice, a single interpolation condition leading to V was verified to be sufficient in the present application setting. Matrix W is left free to parameterize the reduced model and enforce its passivity, as discussed below.

III-A-2. Balanced Truncation

Balanced Truncation (BT) [14] is a more sophisticated approach that controls the approximation error in terms of the H^∞ norm of the transfer function, instead of its values at discrete and frequency points. The main idea of BT is to convert the state-space or descriptor system (3) by changing the coordinates in the state space through $\tilde{x} = Tx$, where T is square and invertible (Figure 7). Dedicated and numerically reliable strategies exist [15,16] to determine T such that, in the new coordinate system (denoted as *balanced*),

- The last elements of $\tilde{x}$ are poorly controllable, i.e., driving the system with any arbitrary input signals, the effect on such components is negligible.
- The same last elements of $\tilde{x}$ are poorly observable, i.e., these states have a negligible influence on the outputs.

As a result, the components of $\tilde{x}$ that are poorly controllable and observable can be safely eliminated, thus reducing the size of the system. The theory shows that such truncated system is optimal in terms of transfer function norm among all other reduced-order systems having the same size. See [14-17] and references therein for additional details and Figure 7 for a graphical illustration.

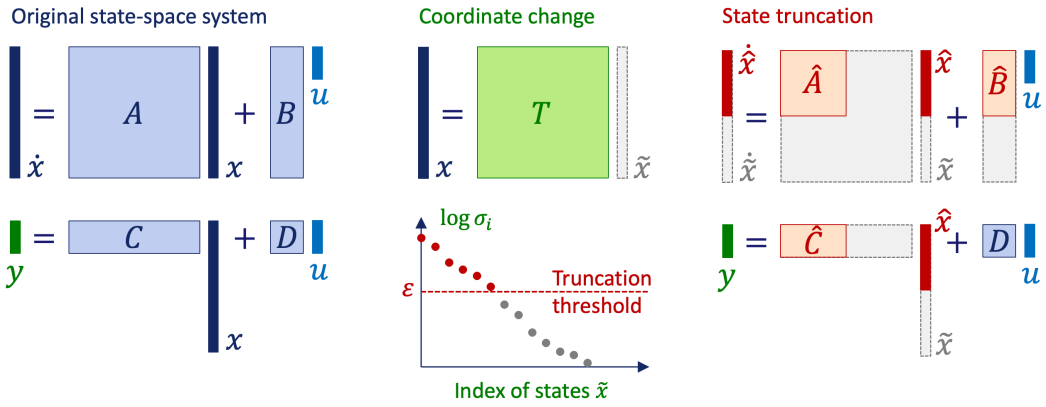


Figure 7: schematic illustration of linear MOR through Balanced truncation. The original system is subjected to a particular coordinate transformation T which reveals the so-called Hankel singular values σ_i. The latter pinpoint the importance of each state, so that truncation with a suitable threshold ε leads to the reduced model. This figure intended to be full-page width.

III-A-3. Preserving Passivity

Both adopting a moment matching (interpolation) or a balanced truncation approach, it is possible to constrain the projection matrices such that the reduced model is

guaranteed passive. This is of course possible and meaningful only if the original large-scale model is passive, which is the case for any combination of electrical interconnect models and lumped RLCM elements). Passivity preservation is of paramount importance to guarantee that the reduced system will not trigger numerical instabilities during transient simulation. Several approaches are possible to achieve this goal [18-20]. Here we advise to use the procedure in [18], which first determines the right projection matrix V using the desired approach, and then defines the left projection matrix as $W = PV$, where P is any solution of the Kalman-Yakubovich-Popov (KYP) equations associated to the original system. Details on this technical aspect are omitted here, for a complete derivation see [18] and references therein.
Compared to [18], we can avoid solving an Algebraic Riccati Equation to find a solution P of the KYP lemma. In fact, this would be an extremely expensive operation to perform on the full-order PDN system matrices. In this case, a solution P can be constructed cheaply by observing that the PDN is always the interconnection of passive macromodels and RLCM networks. It is computationally much easier to obtain solutions P_i for each of these subsystems, which are then combined to obtain the full P, as detailed in [3] (Appendix).

III-B. Handling nonlinearity

Since the PDN model under analysis is nonlinear, a direct application of the linear MOR approaches described in previous section is not feasible. It is of course possible to perform a reduction of the LTI part only expressed by the green block in Figure 5. This approach would however neglect the relation (5) that the switches induce between some of the LTI system port signals, by constraining the voltages and currents at the switches input and output ports. Embedding such switch equations in the system to be reduced would lead to a parameterization of the corresponding equations in terms of the duty cycle signals d_k, making a direct reduction unfeasible since such signals are unknown. A direct parameterized MOR was attempted in [5], but this approach is not scalable to PDN structures for HPC and AI, with hundreds of cores and independent duty cycle signals.
In [3], we propose to exploit the weak nonlinearity to setup a structured MOR that

- still reduces only the LTI subsystem, which in fact is the responsible for the large-scale nature of the overall PDN, but
- constrains the reduction by embedding the constraints induced by the FIVR switches in the evaluation of the projection matrices.

In particular:

1. we solve the for the steady state of all PDN states by choosing a set of independent (constant) loading current excitations within the allowed range $[0, I_{\max}]$, see Figure 8, left.
2. Each of these loading patterns induces an operating point and a corresponding nominal duty cycle signal. At each operating point we evaluate the small-signal LTI subsystem equations, and we evaluate the corresponding small-signal state snapshots X_i at prescribed frequency points (Figure 8, middle).

3. We combine all such state snapshots as matrix columns (Figure 8, right), followed by a column orthogonalization process. The result is the desired right projection matrix V.

This choice of the projection matrix V guarantees that, if the state snapshot X_i of the LTI subsystem is in the range of V, the small-signal transfer function of the overall PDN around the considered operating points will match the original one at the considered frequency points. More precise details are presented in [3], where the idea of structure-preserving reduction (already developed in methods such as SPRIM [21,22,23]) has been adapted to the topology under investigation, so that the reduced part coincides with the linear subsystem of a nonlinear PDN.

Additional tricks and technicalities are used to handle the possibly large number of inputs/outputs, which would lead to a projection matrix with too many columns (hence reduced model size). A randomized Singular Value Decomposition (SVD) [24] is used instead of a pure orthogonalization process, so that only the most important contributions are preserved in the projection matrix. This will induce an approximation error, so that interpolation will not be exact at s_i. This error is easily controlled by the SVD truncation threshold.

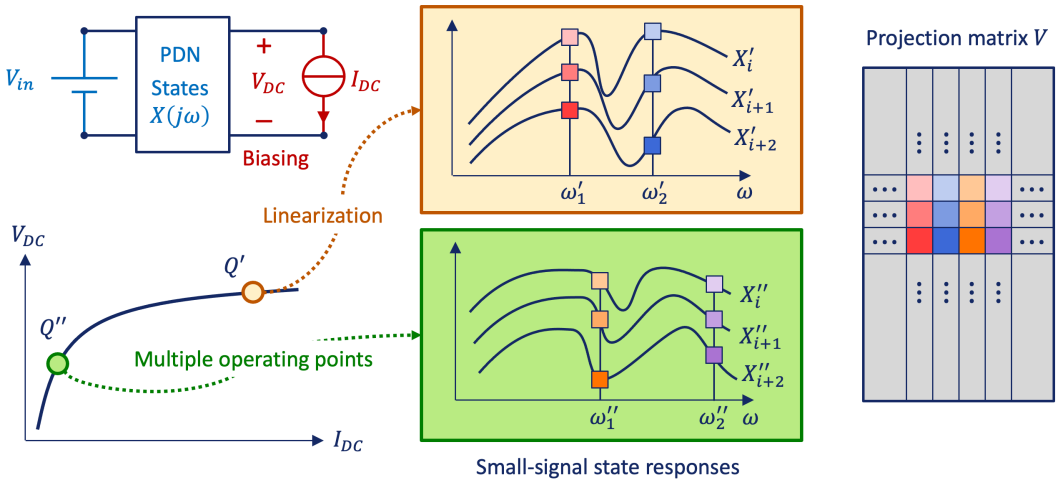


Figure 8: schematic illustration of the process leading to the MOR projection matrix V for the nonlinear moment-matching (interpolation) framework. The PDN is biased to multiple operating points, where the input-state relation is linearized. Frequency samples of the corresponding state responses are collected to assemble the projection matrix, which is finally orthogonalized. This figure intended to be full-page width.

III-C. Summary

We provide here a practical summary in form of high-level algorithm the main steps required by the proposed MOR framework for voltage-regulated PDN models. The Algorithm 1 is general and requires as main prerequisites a SPICE description of the entire PDN model and a general-purpose numerical matrix analysis tool (such as MATLAB or Python with associated numerical libraries) to construct and apply the projection matrices to perform model reduction. Our implementation performs all

tasks in a MATLAB-based script, since we were able to parse and read the native SPICE netlist of the PDN benchmark examples into the MATLAB environment. A parsing utility to assemble the PDN equations from a SPICE description is also a prerequisite.

Algorithm 1: Projection-based MOR
Inputs:

- SPICE netlist of entire PDN system
- Accuracy threshold ϵ
- Typical load current bias points $\{I_\ell\}$

Steps:

1. Read SPICE netlist to assemble a state-space PDN model (3), (4)
2. For each bias point I_ℓ
    a. Compute small-signal linearized model of the feedback-controlled PDN system around this bias point, including (3)-(5).
    b. Run small-signal (AC) analysis on the PDN model to find matrix X^ℓ collecting state snapshots of the linear subsystem (3) at log-spaced frequencies in the bandwidth of interest.
3. Assemble all X^ℓ from different operating points in a global snapshot matrix X
4. Compute principal components (SVD) of X according to truncation threshold ϵ to define an orthogonal matrix V
5. Define $W = PV$, after assembling the passivity matrix P by the analytical procedure of [3, Appendix].
6. Project the linear subsystem (3) to find a reduced model $\hat{G}$ as in (6)
7. Return reduced PDN model made up of $\hat{G}$ coupled with original controllers (4) and switches (5).

IV. Transient simulation

IV-A. Basic time-stepping

The enormous success of SPICE solvers is due to their generality and robustness. SPICE solves practically any circuit for all types of analysis of interest for electrical and electronic designers. To achieve this goal, the numerical algorithms that are embedded in SPICE have been formulated to be as general-purpose as possible. An example is the adaptive time-stepping scheme that drives transient analysis and refines the time resolution while solving the circuit equations, depending on the actual circuit behavior in real time.

The transient analysis of fully coupled PDNs has very particular features that make a general-purpose solver such as SPICE potentially inefficient, namely the large-scale size of the system which interacts with time-varying nonlinearities embedded in the various feedback voltage regulation loops. It turns out that a much simpler time-stepping scheme can be adopted, without losing efficiency or accuracy. In fact, a simple implicit Euler scheme with constant time step is more than sufficient. Time step is mainly dictated by the current excitation stimuli, which for state-of-the-art

microprocessor systems exhibit gradients of several Amperes in few nanoseconds. On the other hand, the relation between excitation currents and transient voltages at the load is typically low-pass due to the on-chip capacitance and substrate behavior. The PDN resonances that are to be assessed, which are responsible for output voltage droops, are located at lower frequencies, typically well below 100 MHz. In fact, PDN models from EM solvers are rarely computed beyond few GHz's. This implies that fixed sub-nanosecond time steps are expected to capture all variations of interest. Adopting such a fixed time step implies that several quantities that are needed to advance in time during transient simulation may be precomputed, leading to major speedup with respect to a general SPICE solver, that manages nonlinearities through Newton-Raphson iterations at each time step. Details on the proposed time-stepping scheme used in the following are available in [3], in short:

- All signals are discretized with fixed timestep Δt, with $h = 0,1,2, \ldots$

$$x^{(h)} \approx x(h\Delta t), \quad u^{(h)} \approx u(h\Delta t), \quad y^{(h)} = y(h\Delta t)$$

- Time-stepping for the LTI part in (3) reads

$$\begin{aligned}(E - A\Delta t)x^{(h+1)} &= Ex^{(h)} + \Delta t B u^{(h+1)} \\ y^{(h+1)} &= Cx^{(h+1)} + Du^{(h+1)}\end{aligned}$$

- Switch voltages/currents in (5) are updated semi-explicitly

$$v_{2;k}^{(h+1)} = d_k^{(h)}\, v_{1;k}^{(h+1)}, \qquad i_{1;k}^{(h+1)} = -d_k^{(h)}\, i_{2;k}^{(h+1)}$$

- Similarly, the compensators in (4) are discretized as

$$\begin{cases}(E_k - A_k\Delta t)x_k^{(h+1)} = E_k x_k^{(h)} + \Delta t B_k\left(N_k v_k^{(h+1)} - V_{\text{ref}}\right) \\ d_k^{(h+1)} = Clip(C_k x_k^{(h+1)})\end{cases}$$

IV-B. Speeding up transient solution via Parallel Waveform Relaxation

The particular PDN structure of massive multicore platforms lends itself to further optimization and speedup in transient analysis, through parallel circuit simulation. In fact, the N_C core subsystems depicted in Figure 5 are almost decoupled and interact only through their connection to the common Input network through the FIVR switches. Supposing for the moment that we know exactly the transient voltage signals $v_{1;k}(t)$ at the input of the k-th core switches, this core subsystem can be decoupled from the rest of the PDN and solved without any approximation error, by exciting its inputs by ideal voltage sources equal to $v_{1;k}(t)$. This is a straightforward application of the well-known circuit Substitution Theorem. Since such decoupled core subsystems are independent, they can be solved concurrently in a parallel (multithreaded) simulation environment, with an ideal reduction in runtime by N_C.

Unfortunately, the interface voltages at the core subsystems are not known a priori, since they arise as the solution of the globally coupled PDN. So, instead of a one-shot exact solution, we can think of refining estimates of such interface signals through an iterative process. This approach is in fact well known and falls under the common denomination of Waveform Relaxation (WR), a technique that was introduced several decades ago to enable parallel circuit simulation [25,26]. Although WR for general circuit simulation has been superseded by more advanced (commercial) solvers, given the difficulties in finding an appropriate circuit partitioning, research is still active on WR for those systems like transmission line networks or PDNs, for which an appropriate decoupling strategy is evident from the system topology [27-29].

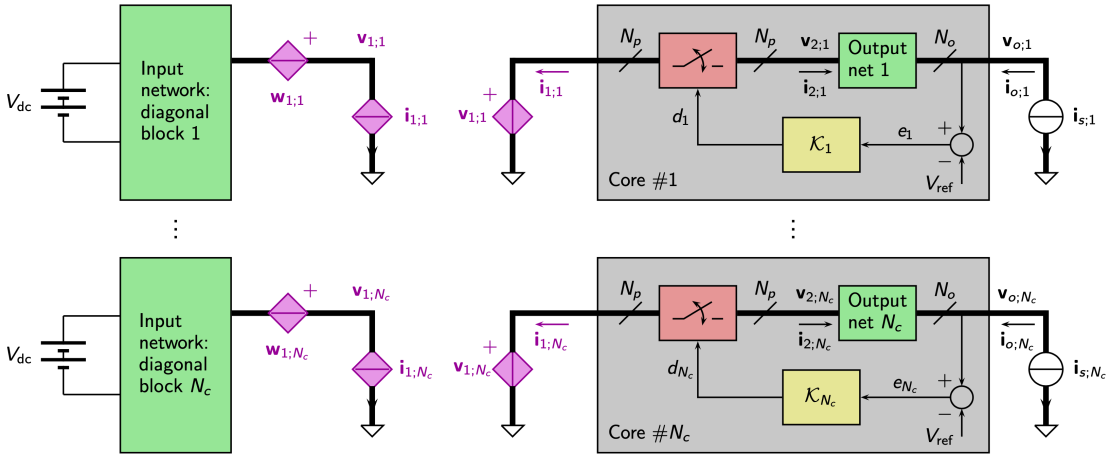


Figure 9: partitioning of the PDN structure depicted in Figure 5, by splitting Input network from the core subsystems (Longitudinal Partitioning) and individual core subsystems with associated Input network (block-diagonal Transverse Partitioning). The purple elements reintroduce the couplings neglected by the partitions in terms of relaxation sources. This figure intended to be full-page width.

The first step in WR is system partitioning. For the PDN structures of interest, a possible partitioning is depicted in Figure 9, where both a Longitudinal Partitioning (LP) that separates Input network and core subsystems at the FIVR switches interface and a Transverse Partitioning (TP) that splits the subblocks pertaining to individual cores and related Input network portions are applied. Such partitions neglect some couplings, which are reintroduced by equivalent sources, usually denoted as *relaxation sources* and highlighted in Figure 9 in purple. Longitudinal and Transverse Partitioning schemes can be applied separately or concurrently, leading to different corresponding WR schemes, see [6] and references therein for a complete description and analysis.

Simulation starts with a given initial estimate of the relaxation sources (e.g., the constant signals corresponding to the DC operating point from which the transient analysis starts). Then, each individual subsystem is solved independently. Next, relaxation source signals are updated based on the just-computed solution, and the process is iterated until convergence. Figure 10 provides a graphical illustration of a

block-partitioning of a template coupled linear system that is solved via WR. The diagonal blocks (green) correspond to the equations of the decoupled system, whereas the off-diagonal blocks (purple) are collected in relaxation sources, which are updated through iterations.

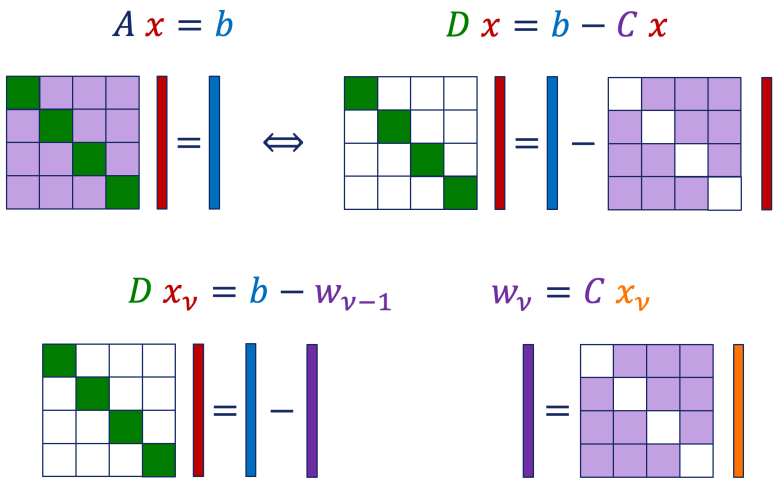


Figure 10: schematic illustration of a basic Waveform Relaxation scheme applied to the solution of a linear system of equations $Ax = b$. The system matrix is decomposed into a block-diagonal term plus off-diagonal block couplings. WR proceeds by solving the block-diagonal system at each iteration $\nu = 1,2, \ldots$ using the estimate of the coupling terms from previous iteration (bottom left); couplings are then updated (bottom right) to prepare next iteration. This figure intended to be one column (half-page) width.

One fundamental issue that may affect WR iterations is convergence, which is only conditional and may be slow. This is in fact the main reason why WR for general circuit simulation was abandoned soon after the general theory had been developed. In short, WR converges well if the couplings that are neglected in the system partitioning are “small” with respect to the contributions from individual decoupled blocks. In the PDN splitting depicted in Figure 9, we see that the main neglected couplings between individual cores are provided by the inter-core couplings provided by the path going from one core to another through FIVRs (twice) and the Input network. These couplings can be roughly quantified by looking at the (small signal) mutual impedances seen from the excitation currents. An example is provided in Figure 11, showing that these mutual impedances are very small compared to the self-impedances. One therefore expects fast convergence of WR for PDN transient analysis. The results in next section will confirm this expectation.

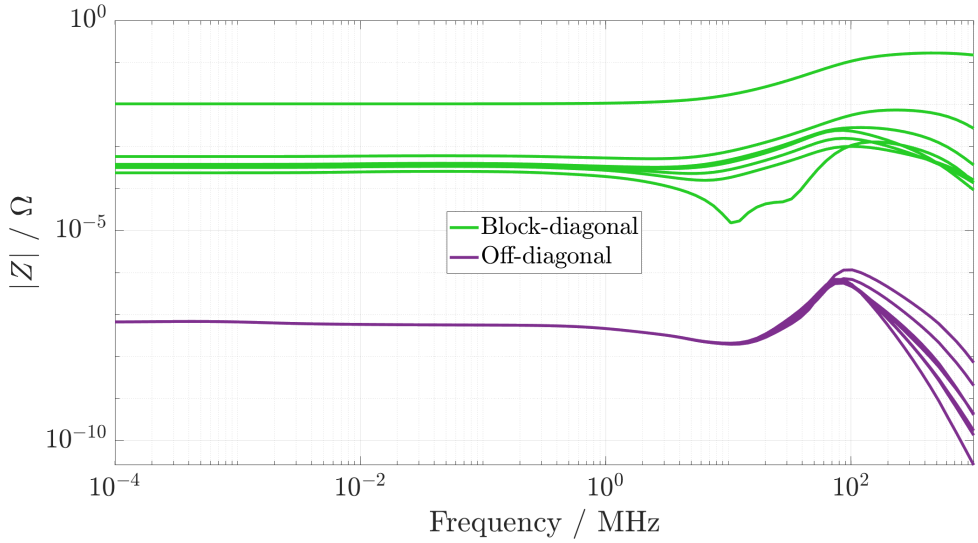


Figure 11: Magnitude of PDN impedance of a mobile platform. The inter-core couplings (off-diagonal blocks, purple color) are at least two orders of magnitude smaller than impedances pertaining to the same core (diagonal blocks, green color). This figure intended to be one column (half-page) width.

V. Benchmark examples

The proposed modeling and simulation flow has been applied to several benchmarks to assess its accuracy and efficiency. In particular, we report numerical results for a 4-core mobile platform and a large-scale 60-core server platform.

V-A. A 4-core mobile platform

A first benchmark that we use as proof of concept is a PDN model from a mobile device with four cores. This example is already relatively large as the original PDN model corresponds to a dynamical system with 2600 internal states. Each core has $N_o = 36$ load ports, and the FIVRs operate each with $N_p = 4$ distinct phases. A reference transient simulation (maximum timestep is 0.1 ns, 50000 timesteps) takes 972 s to be performed with the state-of-the-art solver HSPICE. The proposed MOR algorithm was applied to this PDN, considering four linearization points for the collection of snapshots. The resulting model is more compact, with around 500 states, so that it can be solved through the proposed time-stepping scheme (coded in C language) in only 1.3 seconds. Hence, the speedup provided by the proposed reduction and simulation framework is practically significant even in this simple 4-cores example. The loss in accuracy is negligible, as can be seen by comparing the responses of the original PDN and the reduced model in Figure 12. Here, the load current stimuli are all different and consist in low-pass filtered white noise signals with a 10 MHz bandwidth and a peak amplitude of about 10 A per core. Quantitatively, the error between the time-domain responses is always less than 0.7 mV (peak error), and its RMS value is 0.16 mV. This is much smaller than the load voltage response that spans 100 mV peak to peak.

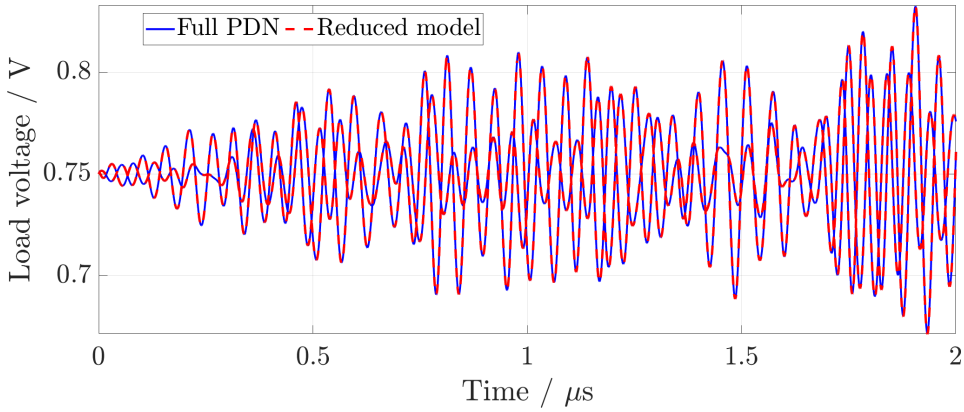


Figure 12: Load voltage response at two selected output ports (located in different cores), in the transient simulation described in Section V-A (4-core mobile example). The visualized time span is here restricted to up to 2 μs to enhance readability. The load current excitations are independent and randomized filtered white noise signals applied to each load port. This figure intended to be one column (half-page) width.

V-B. A 60-core server platform

This test case is the PDN of an Intel-based enterprise server platform with $N_c = 60$ cores. Each core has $N_o = 57$ load ports and is fed by a multi-phase FIVR having $N_p = 3$ phases. Therefore, the entire PDN system is massively MIMO with more than 3000 load ports. Looking at the MNA models of the Input and Output Networks for this PDN, the size of the state vector (number of internal variables) is more than 5×10^4, implying that transient simulation involves solving a very large (nonlinear) system of equations at each timestep. The complexity of this benchmark problem is also confirmed by the fact that a state-of-the-art commercial solver such as HSPICE fails to converge, unless the number of core models included in the simulation is reduced to less than 32.

We applied the structured MOR algorithm described in the previous sections. The construction of the right projection matrix V was carried out by computing snapshots of the full-order linearized system to approximate a balanced truncation basis empirically. In this case the overall PDN behavior is close to ideal in terms of linearity, so that only one linearization point was considered for snapshot collection. Using a relative truncation tolerance of 5×10^{-4}, the resulting reduced PDN system includes a reduced input network with as few as 150 states (much less than the initial 2×10^4) and output network models with only 6 states each.

The accuracy of the reduced model can be checked by considering the results of a typical power-up analysis where individual cores are loaded with step-like excitation currents. In this scenario, the load current changes suddenly from zero to 20 A per core (equally distributed among the 57 load ports), as shown in Figure 13. Note that the very fast switching time (3 ns) required by a full load current swing on this hardware makes this test quite aggressive in terms of induced voltage fluctuation. In

the following simulation setting, groups of eight cores are turned on and off synchronously. A detail of the simulation result is reported in Figure 14, which shows the responses of the first two groups, showing that when a group of cores is turned on, there are significant load voltage oscillations on the load ports of the same core, and small oscillations on the other ones due to inter-core coupling. The results over the full simulation interval are shown in Figure 15. The reduced model response closely replicates the full PDN response and is sufficient to evaluate the transient PDN behavior in terms of peak droop and settling time.

Compared to a reference HSPICE simulation, the reduced model is much faster to evaluate, taking only 31 s using the proposed special-purpose transient solver, with a negligible loss in accuracy (the peak error over all load voltages is 1.3 mV, the RMS value is 0.32 mV). A thorough analysis of the simulation runtime with different number of cores included in the simulation is reported in Figure 16. Starting from the runtime of HSPICE (which fails to converge with more than 32 cores), a speedup of about ten times is obtained by running the same full-order PDN model in the special-purpose transient solver (orange line, without model reduction). Given the trend of the HSPICE runtime, we can assume that a 60 cores simulation would require more than an hour if it were convergent. Using the reduced model, the simulation speed is improved by an additional factor of ten (blue line). This is observed consistently for several number of modeled cores from eight to sixty.

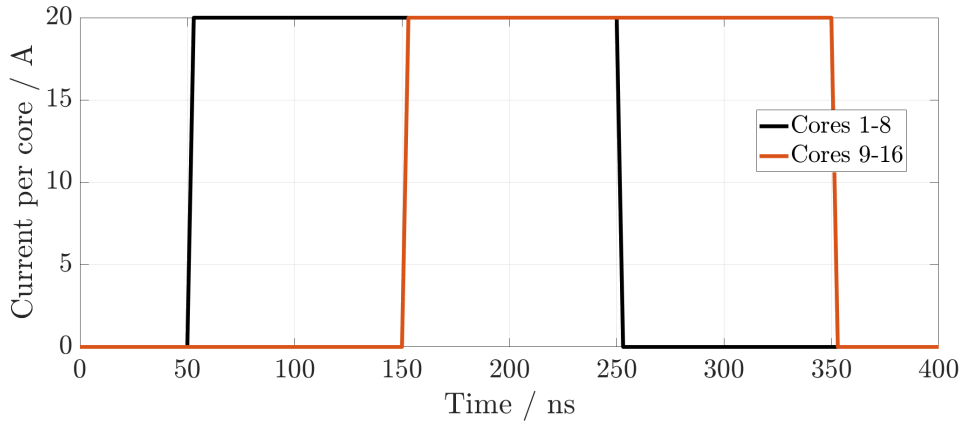


Figure 13: test input signal for the enterprise server example in Section V-B, up to core 16. The load current switches from 0 to 20 A/core in 3 ns. This figure intended to be one column (half-page) width.

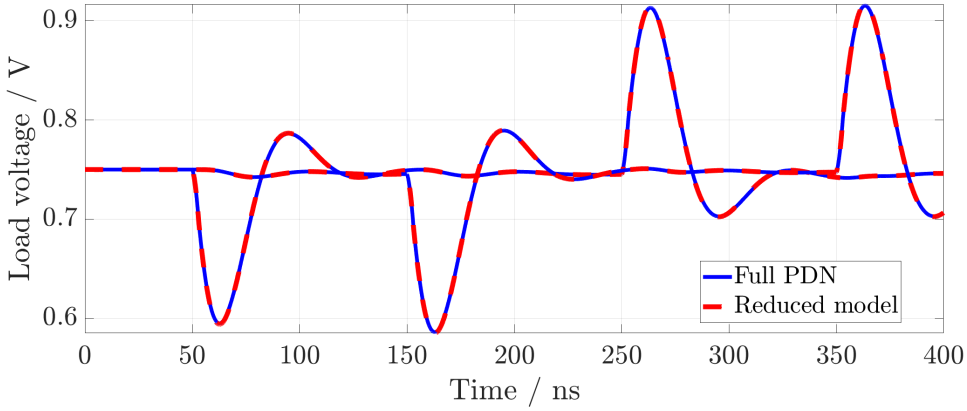


Figure 14: load voltage in time domain for the example in Section V-B. Large voltage droops correspond to rising edges of the current profiles shown in Fig. 13. This figure intended to be one column (half-page) width.

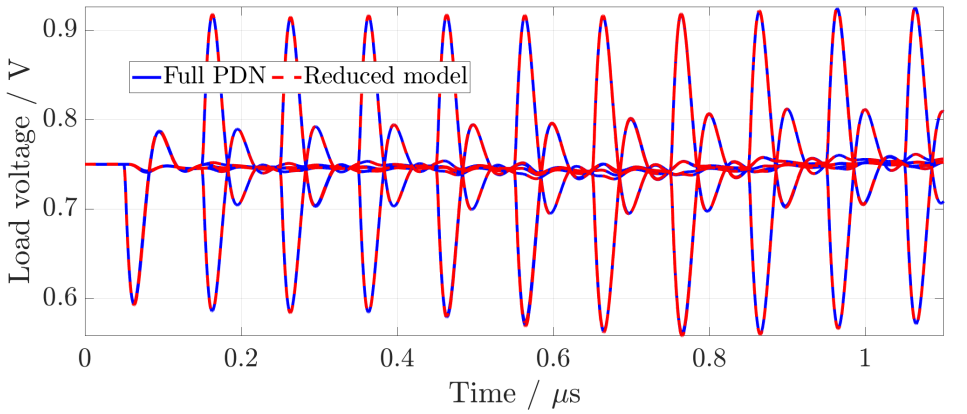


Figure 15: load voltage in time domain over the entire simulation interval (example in Section V-B). The load currents correspond to a ramp-up analysis, all 60 cores are synchronously switched on in groups of eight (as in Figure 13). This figure intended to be one column (half-page) width.

An additional speedup is provided by the parallelized solver based on the WR algorithm. As reported in Figure 17, the parallelized simulation is faster as the number of computing threads is increased. Depending on the partitioning scheme (LP or TP), the actual time is slightly different. However, we can observe in both cases that it scales down linearly with the number of threads until it saturates when the overhead due to thread synchronization becomes dominant. Nonetheless, the final parallel runtime is still close to or lower than one second, meaning that the combination of MOR and WR provide a combined speedup of more than 1000X.

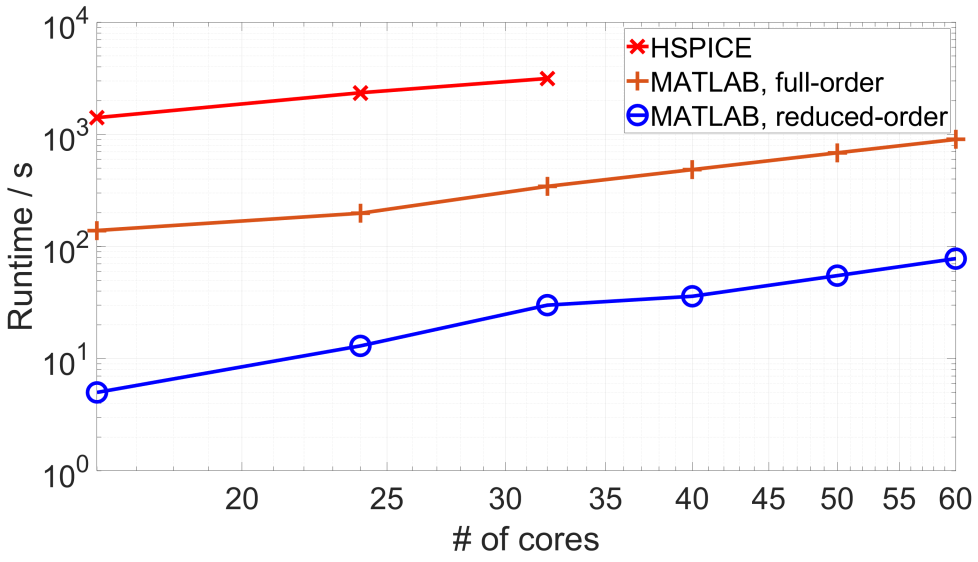


Figure 16: Simulation runtime versus number of modeled cores for the enterprise server example in Section V-B. HSPICE runtime is compared with the full and reduced models solved in MATLAB using the special-purpose transient solver. This figure intended to be one column (half-page) width.

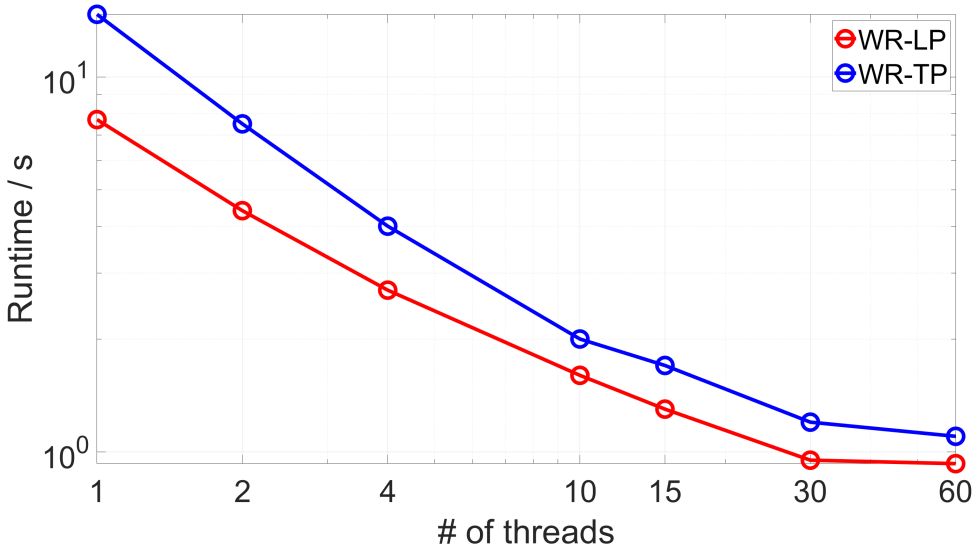


Figure 17: Simulation runtime (in seconds) versus number of parallel threads used in the WR simulation scheme (enterprise server example in Section V-B). The parallel efficiencies of the longitudinal (LP) and transverse (TP) partitioning schemes are compared. This figure intended to be one column (half-page) width.

VI. Conclusions

This paper addressed the problem of accelerating power integrity analyses of power delivery networks of multi-core microprocessors equipped with Integrated Voltage Regulators. This advanced power delivery architecture is widely used in several application domains, yet efficient transient simulation tools at the system levels are still lacking, especially for PDNs of massive many-core systems. This work provided a comprehensive solution that combines model reduction with a special-purpose transient solver. The presented reduction method gives a compressed model that preserves the original PDN structure and guarantees model passivity in a computationally efficient manner. As a result, transient simulation based on the reduced PDN model provides results with a small computational effort, even in those cases where commercial solvers fail to converge, with a runtime that scales very favorably with the number of modeled cores. Furthermore, using a particular form of structure-preserving reduction, the compressed PDN models inherit the modular topology of the original PDN, implying that the transient solution of different cores can be distributed across multiple parallel threads. This can be accomplished through a Waveform Relaxation approach, reaching close-to-ideal parallel scaling. In this regard, three system partitionings have been discussed and compared in terms of parallel efficiency. The numerical results obtained by solving PDN models of commercial mobile and enterprise server platforms confirm that the combination of MOR with parallel transient solver provide a combined speedup of more than 1000X compared to HSPICE.

VII. References

[1] K. Radhakrishnan, M. Swaminathan and B. K. Bhattacharyya, "Power delivery for high-performance microprocessors—Challenges solutions and future trends", *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 4, pp. 655-671, Apr. 2021.

[2] E. A. Burton *et al*., "FIVR — Fully integrated voltage regulators on 4th generation Intel® Core™ SoCs," *2014 IEEE Applied Power Electronics Conference and Exposition - APEC 2014*, Fort Worth, TX, USA, 2014, pp. 432-439.

[3] A. Carlucci, S. Grivet-Talocia, S. Kulasekaran and K. Radhakrishnan, "Structured Model Order Reduction of System-Level Power Delivery Networks," in *IEEE Access*, vol. 12, pp. 18198-18214, 2024

[4] A. Carlucci, S. Grivet-Talocia, S. Mongrain, S. Kulasekaran and K. Radhakrishnan, "A Structured Krylov Subspace Projection Framework for Fast Power Integrity Verification," *2023 IEEE 27th Workshop on Signal and Power Integrity (SPI)*, 7-10 May 2023, Aveiro, Portugal, pp. 1-4

[5] A. Carlucci, T. Bradde, S. Grivet-Talocia, S. Mongrain, S. Kulasekaran and K. Radhakrishnan, "A Compressed Multivariate Macromodeling Framework for Fast Transient Verification of System-Level Power Delivery Networks," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 13, no. 10, pp. 1553-1566, Oct. 2023.

[6] A. Moglia, A. Carlucci, S. Grivet-Talocia, S. Kulasekaran and K. Radhakrishnan, "Fast Transient Simulation of System-Level Power Delivery Networks via Parallel

Waveform Relaxation," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, doi: 10.1109/TCPMT.2024.3410146 (Early Access)

[7] S. Grivet-Talocia and B. Gustavsen. Passive Macromodeling: Theory and Applications. Wiley, 2016.

[8] IdEM R2018, Dassault Systems. [Online]. Available: www.3ds.com/products-services/simulia/products/idem/

[9] M. Swaminathan and A. E. Engin, *Power Integrity Modeling and Design for Semiconductors and Systems*. Upper Saddle River, NJ, USA: Prentice-Hall, 2007.

[10] Chung-Wen Ho, A. Ruehli and P. Brennan, "The modified nodal approach to network analysis," in *IEEE Transactions on Circuits and Systems*, vol. 22, no. 6, pp. 504-509, June 1975.

[11] R. Achar and M. S. Nakhla. Simulation of high-speed interconnects. Proceedings of the IEEE, 89(5):693–728, May 2001.

[12] P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. Schilders, and L. M. Silveira, eds., Model Order Reduction, vol. 1: System- and Data-Driven Methods and Algorithms. Berlin, Boston: De Gruyter, 2021

[13] E. J. Grimme, "Krylov projection methods for model reduction", Ph.D. Thesis, Graduate College of the University of Illinois at Urbana-Champaign, Urbana, Illinois, 1997.

[14] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," in *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 17-32, February 1981.

[15] M. G. Safonov and R. Y. Chiang, "A Schur method for balanced-truncation model reduction," in IEEE Transactions on Automatic Control, vol. 34, no. 7, pp. 729-733, July 1989.

[16] J. R. Phillips and L. M. Silveira, "Poor man's TBR: a simple model reduction scheme," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 1, pp. 43-55, Jan. 2005.

[17] K. Willcox and J. Peraire, "Balanced model reduction via the proper orthogonal decomposition", *AIAA J.*, vol. 40, pp. 2323-2330, Jan. 2002.

[18] Z. Zhang, Q. Wang, N. Wong and L. Daniel, "A moment-matching scheme for the passivity-preserving model order reduction of indefinite descriptor systems with possible polynomial parts," *16th Asia and South Pacific Design Automation Conference (ASP-DAC 2011)*, Yokohama, Japan, 2011, pp. 49-54

[19] A. Odabasioglu, M. Celik and L. T. Pileggi, "PRIMA: passive reduced-order interconnect macromodeling algorithm," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 8, pp. 645-654, Aug. 1998.

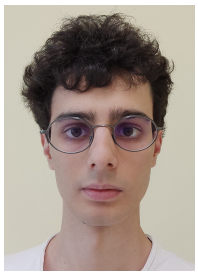
[20] T. Reis and T. Stykel, "PABTEC: Passivity-Preserving Balanced Truncation for Electrical Circuits," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 9, pp. 1354-1367, Sept. 2010.

[21] R. W. Freund, "SPRIM: structure-preserving reduced-order interconnect macromodeling," IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004., San Jose, CA, USA, 2004, pp. 80-87.

[22] H. Sandberg and R.M. Murray, “Model reduction of interconnected linear systems”, in *Optimal Control Applications and Methods*, vol. 30, no. 3, pp. 225-245, 2009.

[23] A. Vandendorpe and P. Van Dooren, "Model reduction of interconnected systems," in *Model Order Reduction: Theory, Research Aspects and Applications*, W. H. A. Schilders, H. A. van der Vorst, and J. Rommes, vol. 13, Mathematics in Industry, Berlin, Heidelberg: Springer, 2008, pp. 305-321.
[24] N. Halko, P. G. Martinsson and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions", *SIAM Rev.*, vol. 53, no. 2, pp. 217-288, Jan. 2011.
[25] E. Lelarasmee, A. Ruehli, and A. Sangiovanni-Vincentelli, "The waveform relaxation method for time-domain analysis of large scale integrated circuits," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 1, no. 3, pp. 131–145, 1982.
[26] J. K. White and A. L. Sangiovanni-Vincentelli, Relaxation techniques for the simulation of VLSI circuits. Springer New York, NY, 1987.
[27] M. J. Gander, M. Al-Khaleel, and A. E. Ruchli, "Optimized waveform relaxation methods for longitudinal partitioning of transmission lines," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 56, no. 8, pp. 1732–1743, 2009.
[28] N.Nakhla, A.Ruehli, M.Nakhla, and R.Achar,"Simulation of coupled interconnects using waveform relaxation and transverse partitioning," IEEE Transactions on Advanced Packaging, vol. 29, no. 1, pp. 78–87, 2006.
[29] T. Menkad and A. Dounavis, "Using strictly dissipative impedance coupling in the waveform relaxation method for the analysis of interconnect circuits," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 3, pp. 1283–1296, 2021.

Biographies



Antonio Carlucci (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in Electronic Engineering respectively in 2019 and 2021, both from Politecnico di Torino, Turin, Italy, where he is currently pursuing a Ph.D. degree within the EMC group. His research focuses on large-scale simulation of electronic systems using macromodeling methods. He received the Best Student Paper award of SPI 2023, the 27th IEEE Workshop on Signal and Power Integrity.

Alessandro Moglia received the B.Sc. in Electronic Engineering in 2023 from Politecnico di Torino, Turin, Italy, where he is currently pursuing a M.Sc. degree in Quantum Engineering. His research focuses on large-scale simulation of electronic systems using parallel computing and relaxation methods and is currently conducted through a research internship at the Department of Electronics and Telecommunications, Politecnico di Torino.

Stefano Grivet-Talocia (M'98–SM'07–F'18) received the Laurea and Ph.D. degrees in electronic engineering from the Politecnico di Torino, Turin, Italy in 1994 and 1998, respectively. From 1994 to 1996, he was with the NASA/Goddard Space Flight Center, Greenbelt, MD, USA. He is currently a Full Professor of electrical engineering with the Politecnico di Torino. He co-founded the academic spinoff company IdemWorks (Turin, Italy) in 2007, serving as the President until its acquisition by CST in 2016. He has authored about 200 journal and conference papers. His current research interests include passive macromodeling of lumped and distributed interconnect structures, model-order reduction, modeling and simulation of fields, circuits, and their interaction, wavelets, time-frequency transforms, and their applications. Dr. Grivet-Talocia was a co-recipient of the 2007 Best Paper Award of the IEEE TRANSACTIONS ON ADVANCED PACKAGING. He received the IBM Shared University Research Award in 2007, 2008, and 2009 and an Intel Strategic Research Segment Grant in 2022, 2023 and 2024. He was an Associate Editor of the IEEE TRANS- ACTIONS ON ELECTROMAGNETIC COMPATIBILITY from 1999 to 2001 and He is currently serving as Associate Editor for the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING AND MANUFACTURING TECHNOLOGY. He was the General Chair of the 20th and 21st IEEE Workshops on Signal and Power Integrity (SPI2016 and SPI2017), the co-Chair of SPI2023 and the Program co-Chair of SPI2024.

Siddharth Kulasekaran (Member, IEEE) received the B.Tech degree in Electrical Engineering from NIT, Trichy, India in 2010 and M.Sc. degree and the Ph.D. degree in electrical, electronic and communications engineering from Arizona State University, Tempe, USA in 2012 and 2017 respectively. He is currently working at Intel, Arizona as a Senior Analog Engineer at the Power Delivery Core Competency team since 2017. His areas of expertise are in integrated voltage regulators, advanced packaging and passives technologies. At Intel, he focuses on developing modelling methodologies and measurement metrologies for characterizing power delivery networks. His work involves closing the gap between modeling and measurements through accurate 3D EM extraction and loss estimation.

Kaladhar Radhakrishnan (Senior Member, IEEE) is an Intel Fellow and a Power Delivery Architect with the Technology Development group at Intel. He has played a significant role in shaping and driving power delivery technologies for Intel micro-processors. His areas of expertise are in integrated voltage regulators, advanced packaging and passives technologies. Kaladhar is a two-time recipient of the Intel Achievement Award. He has authored four book chapters, over 50 technical papers in peer reviewed journals, and has been awarded 40 US patents. Kaladhar joined Intel in

2000 after he received his Ph.D. in Electrical Engineering from the University of Illinois at Urbana-Champaign.