System-Level Early-Stage Modeling and Evaluation of IVR-assisted Processor Power Delivery System

AN ZOU*, Shanghai Jiao Tong University, China, Washington University in St. Louis, USA,
HUIFENG ZHU, Washington University in St. Louis, USA
JINGWEN LENG, Shanghai Jiao Tong University, China
XIN HE, University of Michigan, USA
VIJAY JANAPA REDDI, Harvard University, USA
CHRISTOPHER D. GILL, XUAN ZHANG, Washington University in St. Louis, USA

Despite being employed in numerous efforts to improve power delivery efficiency, the integrated voltage regulator (IVR) approach has yet to be evaluated rigorously and quantitatively in a full power delivery system (PDS) setting. To fulfill this need, we present a system-level modeling and design space exploration framework called *Ivory* for IVR-assisted power delivery systems. Using a novel modeling methodology, it can accurately estimate power delivery efficiency, static performance characteristics, and dynamic transient responses under different load variations and external voltage/frequency scaling conditions. We validate the model over a wide range of IVR topologies with silicon measurement and SPICE simulation. Finally, we present two case studies using architecture-level performance and power simulators. The first case study focuses on optimal PDS design for multi-core systems, which achieves 8.6% power efficiency improvement over conventional off-chip voltage regulator module (VRM)-based PDS. The second case study explores the design trade-offs for IVR-assisted PDSs in CPU and GPU systems with fast per-core dynamic voltage and frequency scaling (DVFS). We find 2 μs to be the optimal DVFS time scale, which not only reaps energy benefits (12.5% improvement in CPU and 50.0% improvement in GPU), but also avoids costly IVR overheads.

CCS Concepts: • **Hardware → Chip-level power issues** .

Additional Key Words and Phrases: integrated voltage regulators, power delivery systems, processor power efficiency, voltage noise, fast dynamic voltage and frequency scaling, CPU, GPU

ACM Reference Format:

An Zou, Huifeng Zhu, Jingwen Leng, Xin He, Vijay Janapa Reddi, and Christopher D. Gill, Xuan Zhang. 2021. System-Level Early-Stage Modeling and Evaluation of IVR-assisted Processor Power Delivery System. *ACM Trans. Arch. Code Optim.* 1, 1, Article 1 (January 2021), 25 pages. https://doi.org/10.1145/3468145

1 INTRODUCTION

With the decline of Dennard scaling, thermal design power and energy efficiency restrict single thread performance [14], and designers are looking for more efficient ways to deliver power to microprocessors. Integrated voltage regulators

*An Zou was with Washington University in St. Louis, USA. He is now with Shanghai Jiao Tong University. The initial work was done at Washington University in St. Louis and the major revision was done at Shanghai Jiao Tong University.

Authors' addresses: An Zou, Shanghai Jiao Tong University, China, Washington University in St. Louis, USA, ; Huifeng Zhu, Washington University in St. Louis, USA; Jingwen Leng, Shanghai Jiao Tong University, China; Xin He, University of Michigan, USA; Vijay Janapa Reddi, Harvard University, USA; Christopher D. Gill, Xuan Zhang, Washington University in St. Louis, USA.

Extension of Conference Paper: Ivory: Early-stage design space exploration tool for integrated voltage regulators (IVRs) [89].

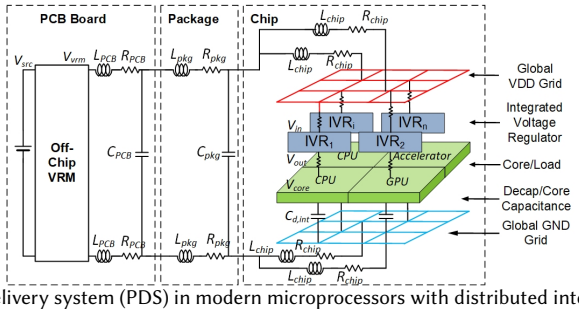


Fig. 1. Overview of the power delivery system (PDS) in modern microprocessors with distributed integrated voltage regulators (IVRs).

(IVRs) can enhance supply integrity and enable flexible voltage scaling by moving power conversion closer to the point-of-load; *distributed* IVRs (shown in Fig. 1) can further provide per-core, fine-grain, and fast dynamic voltage and frequency scaling (DVFS) [29] and effective supply noise suppression [82] at a level unattainable with traditional off-chip regulators. These benefits improve both performance and efficiency, and IVR solutions save precious board/package area compared to bulky off-chip regulators with large discrete passive components, making them especially attractive for mobile SoCs [66]. As IVRs become viable solutions for power delivery in modern microprocessors, it is thus important to explore various design alternatives and thoroughly evaluate their impacts on performance and efficiency at the system level.

Despite the recent proliferation of IVR research, prior studies often focus on circuit-level implementation to improve conversion efficiency [10]. Real implementation benefits in IVR-assisted power delivery subsystems remain elusive due to the lack of modeling tools and evaluation frameworks to explore the design space and investigate the performance and efficiency implications of IVRs in a full system setting. Given the absence of high-level user-friendly IVR models, previous studies resort to either over-simplified assumptions of IVR efficiency [17, 24, 76] that overlook important design considerations such as dynamic response, or a fixed IVR design covering only a fraction of the entire design space [29].

To address these shortcomings, we propose an analytical modeling framework for early-stage design space exploration that is compatible with architecture-level performance and power simulators. Our system-level model captures the complex yet subtle design trade-offs among different IVR typologies to evaluate the performance benefits and implementation costs in full power delivery subsystem settings. It abstracts away the details of low-level IVR circuit implementation to enable architects, system engineers, and other experts working with the upper levels of the system stack to effectively explore new design spaces enabled by IVR's fine-grain voltage regulation capability, similar to what Cacti [73] did for memory systems and ORION [71] did for network-on-chip designs. Our modeling framework incorporates several advanced features that were previously lacking, and makes the following key contributions:

- A fast, accurate, parameterized IVR static model is introduced, and validated using both SPICE simulations and measured silicon data, to estimate static characteristics such as conversion efficiency, static voltage ripple/droop, and die/board area of multiple IVR topologies in different technology nodes or processes.
- A novel method to derive an IVR's dynamic model as a two-port network is described, which allows direct drop-in of IVR modules into the power delivery system. This model facilitates the complete capture of an IVR-assisted PDS's dynamic voltage/current waveform, noise characteristics, and power efficiency, given power traces from real-world workloads or voltage scaling.

- A comprehensive design exploration framework is presented - it covers a wide spectrum of IVR topologies and a variety of IVR metrics for hierarchical composition of multi-stage on-chip and off-chip power delivery networks and provides compatible interfaces with architecture simulators.

Two case studies of system-level design exploration are presented:

- Case study I investigates an optimal power delivery system in a manycore GPU architecture, and reveals that a distributed IVR configuration can outperform a conventional off-chip VRM's output efficiency by 8.6%.
- Case study II explores IVR-assisted hierarchical power delivery with a microsecond level DVFS for CPU and GPU systems. This DVFS can achieve 12.5% and 50.0% net energy improvement for CPUs and GPUs respectively.

2 BACKGROUND

The benefits of integrated fine-grain voltage regulation [29] have driven recent advances in device fabrication [10, 16], circuit implementation [28, 66], and system integration of integrated voltage regulators (IVRs) [17, 76]. In this section, we review the current state of IVR designs and implementations, especially in the context of the entire PDS of modern processors.

2.1 Conventional Power Delivery Systems and Efficiency

The underlying physical mechanisms to convert and transfer electron charges from the higher supply voltage on the motherboard to the much lower supply voltage on the microprocessor chip invariably causes energy loss. The energy loss in power delivery can be broken down into three parts:

First, energy is lost in voltage conversion to step down the supply voltage [23]. We define the conversion efficiency of a voltage regulator (η_{VR}) as the ratio between the power it delivers at the voltage regulator output over the power it consumes at the input. η_{VR} is usually a function of the step-down conversion ratio α. A high performance off-chip switching VRM can deliver over 90% conversion efficiency, but the efficiency is degraded at a lower output voltage with a higher step-down ratio [64].

The second part of the energy loss occurs in the power delivery networks mostly because of heat dissipation when current runs through the parasitic resistance that exists along the path of the power delivery network. This loss is related to the IR-drop component of the supply voltage noise [7, 20, 51]:

$$\eta_{PDN} = \frac{I_{core}V_{core}}{R_{PDN}I_{PDN}^2 + I_{core}V_{core}}, \tag{1}$$

where R_{PDN} and I_{PDN} represent the total parasitic resistance contributed by the power delivery network and the current that goes through the power delivery network. I_{core} and V_{core} represent the current and supply voltage of the computational load. In the off-chip VRM-based power delivery network, the current that goes through the power delivery network I_{PDN} is equal to the current of the computational load I_{core}.

The third and often overlooked part is the energy overhead incurred by raising the supply by a non-negligible voltage margin, $\Delta V = V_{core} - V_{min}$, to accommodate the supply voltage noise and sustain fault-free operation [18, 30, 39, 52, 55, 90]. V_{core} is the actual voltage applied on the core and V_{min} is the minimal ideal supply voltage needed by the core without any supply voltage noise or process variations. We can express this component as $\eta_{\Delta V}$:

$$\eta_{\Delta V} = \frac{P_{core}(V_{min})}{P_{core}(V_{core})} = \frac{V_{min}I_{core}(V_{min})}{V_{core}I_{core}(V_{core})}, \tag{2}$$

where P_{core} and I_{core} represent the power consumption and the current load of the processor core as a function of the core supply voltage (V_{core} and V_{min}).

Based on the analysis above, the full power delivery efficiency can be expressed as

$$\eta_{PDS} = \frac{P_{core}(V_{min})}{P_{src}} = \eta_{VR} \cdot \eta_{PDN} \cdot \eta_{\Delta V}, \tag{3}$$

where P_{src} is the total power drawn from the source.

2.2 Integrated Voltage Regulator

A voltage regulator converts an input voltage to an output voltage at a different level that serves as the supply to load circuits. Digital low-dropout regulators (digital LDOs) and switching regulators are the two main types, and they differ most notably in their efficiency ranges. The digital LDO's efficiency is determined by the input/output voltage ratio, whereas the switching regulator yields higher efficiency even with a higher conversion ratio.

Due to their lower switching frequencies ($< 10MHz$), switching regulators usually require large discrete passive components such as capacitors and inductors to mitigate static ripples. Recent technology advances make it possible for switching regulators to operate at much higher frequencies and to be integrated on the same die as processors [10, 16]. Buck converters [61] and switched-capacitor converters [10, 36, 66] are two types of topologies commonly adopted for such IVRs, in addition to digital LDOs. While a buck converter requires both an inductor and a capacitor, it can sustain a relatively constant conversion efficiency over a wide output range. In contrast, the inductor-free switched-capacitor topology benefits from higher capacitor density with technology scaling, but incurs a linear drop in efficiency when its output voltage deviates from its peak efficiency points. The efficiencies of both the switched-capacitor and the buck converter are sensitive to device parameters that depend on technology and process options.

Prior work on the system-level impact of IVR provides fragmented evaluations on a few fixed configurations of technology/ processes, topologies, input/output voltage ratios, and load current levels [29, 82]. Therefore, the findings cannot easily be extended to different use cases. While analytical models of buck converters [13] and switched-capacitor converters [36, 57] exist, they primarily focus on modeling individual IVRs as stand-alone blocks, and thus are unable to handle integration with the entire PDS.

2.3 IVR-assisted Power Delivery System and Efficiency

As shown in Fig. 1, in an IVR-assisted PDS, voltage conversion is moved from off-chip to on-chip. Because the on-chip die space is limited, IVRs adopt high frequency switches to compensate for the reduced size of passive components like capacitors and inductors. As the high frequency switches may cause more power loss, IVRs usually suffer from lower conversion efficiencies than the conventional off-chip voltage regulator modules (VRMs).

As described in Section 2.1 and E.q. (1) , in conventional off-chip VRM based power delivery systems, the power lost in a power delivery network is described by $R_{PDN}I_{PDN}^2$. In the off-chip VRM-based power delivery network, the current that goes through the power delivery network I_{PDN} is equal to the current of the computational load I_{core}. In the IVR-assisted power delivery system, the voltage regulator (voltage conversion) is moved from off-chip to on-chip. The power P=VI is almost equal before and after the voltage conversion by the voltage regulator. Before the voltage conversion by the voltage regulator, the voltage is at the high value (same as the motherboard voltage) and therefore the current is smaller. After the voltage conversion by the voltage regulator, the voltage is stepped down to a low value (same as the voltage of the microprocessor chip) and therefore the current is large. Therefore, after the voltage conversion is moved from off-chip to on-chip by IVR, the voltage at the off-chip and the package power delivery network

(which is before the voltage regulator now) is still high, therefore the current that goes through the off-chip and the package power delivery network is smaller. As the voltage conversion ratio is α, the smaller current that goes through the off-chip and the package power delivery network in IVR-based power delivery system is $1/\alpha$ of the current in the off-chip voltage regulator-based power delivery system going through the off-chip and the package power delivery network. The power loss on parasitic resistance is $I_{PDN}^2 R_{PDN}$. Therefore, given the same parasitic resistance, the power loss on parasitic resistance is reduced by $1/\alpha^2$ [45].

As voltage regulation is now located closer to the load, an IVR-assisted PDS enjoys multiple intrinsic benefits. In a conventional PDS with an off-chip VRM, the voltage margin, $\Delta V = V_{core} - V_{min}$, causes a non-negligible power loss. In an IVR-assisted PDS, we can potentially reduce the voltage margin to mitigate the energy overhead. Besides, IVRs open up the opportunity for faster power management at the microsecond level. In this paper, we present two case studies to reveal the benefits of IVR-assisted PDS.

2.4 Related Work

Proof-of-concept circuits [3, 21, 35, 47, 62, 63] and silicon prototypes [27, 32, 50, 53, 60, 70] have been presented previously to explore the designs and benefits of integrated voltage regulators (IVRs) and IVR-assisted PDSs. Burton et al. [8] presented a fully integrated voltage regulator design (FIVR) on commercial 4th generation Intel® Core SoCs with improved power delivery efficiency. Fluhr et al. [15] presented the design of a POWER8 Processor powered by integrated voltage regulation. Zimmer et al. [85] designed an integrated switched capacitor voltage regulator that can support a sub-microsecond scale fast DVFS power management.

On the system side, Zhuo et al. [84] and Zhou et al. [82] proposed cross-layer infrastructures for the co-exploration of power delivery and system architecture, especially focusing on the power delivery network supply noises from parasitic components. Kim et al. [29] evaluated the system-level benefits from fast DVFS supported by a fixed IVR-assisted PDS. Zeng et al. [79] studied the system dynamic stability of integrating a large number of LDO on-chip voltage regulators, and found the design offers strong local load regulation and facilitates system-level power management. Wang et al. [72] developed PowerSoc which is a modeling, analysis, and optimization platform for buck converter based PDS. Based on analytical models, PowerSoc provides an accurate and fast evaluation of static characteristics, such as power efficiency, transient response, and cost. Zeng [78], Gjanci [19] and Vaisband [69] conducted systematic design analyses on power delivery networks that incorporate off-chip buck voltage regulator and on-chip LDOs for the entire chip power supply. Kose [31] proposed an unified design methodology to determine the optimal on-chip location of the power supplies and decoupling capacitors. Zhan et al. [80] proposed a heterogeneous voltage regulation (HVR) architecture, exploring the rich heterogeneity and tunability of HVR. They developed systematic workload-aware power management policies to adapt heterogeneous VRs with respect to workload change at multiple temporal scales.

Besides the supply voltage conversion, many hidden benefits are also brought by IVRs. Voltage stacked power delivery systems with IVRs [22, 41, 58, 74, 86–88] are proposed to improve the power delivery efficiency by reducing power loss in voltage regulators and power delivery networks. After moving voltage regulation from on-chip to on-chip, a more secure power delivery is available compared with previous off-chip voltage regulator based power delivery systems [34, 68, 83].

These policies significantly improved the system's power efficiency while providing a guarantee for power integrity. However, none of these previous works provide a comprehensive study and fair comparison across different IVR topologies and IVR-assisted PDSs in either static or dynamic characteristics at the system level. To fill this need, we

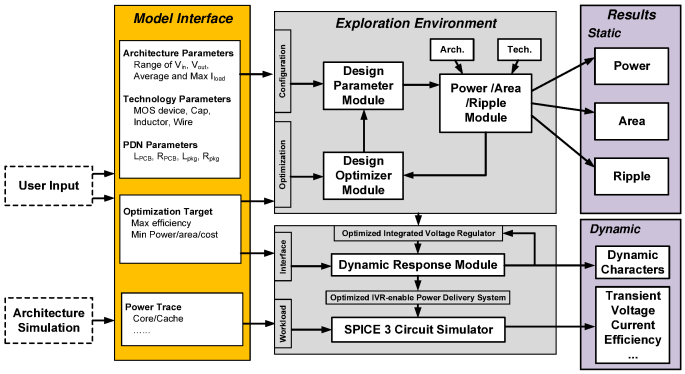
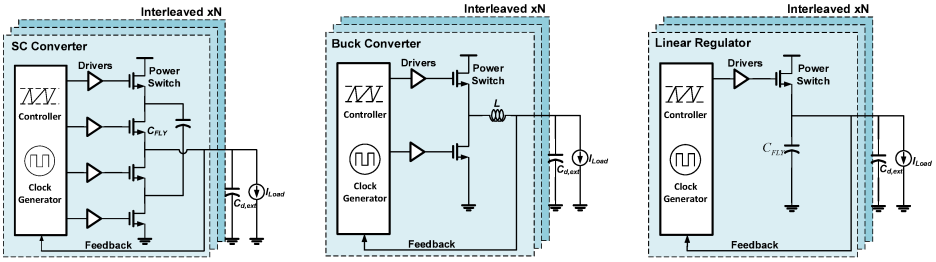


Fig. 2. Block diagram of the IVR and IVR-assisted PDS system-level modeling framework.



(a) Switched-Capacitor Converter (b) Buck Converter (c) Digital LDO

Fig. 3. Three types of converter topologies.

present a system-level early-stage modeling framework, which can accurately estimate both the static and dynamic behaviors of IVRs and IVR-assisted PDSs.

3 MODELING METHODOLOGY

The system-level model enables rapid design exploration of IVR-assisted PDSs for computing systems with diverse configurations. Towards this end, it is crucial to capture the main parameters that critically determine the overall PDS characteristics such as the power consumption (loss) of each component in the PDS under static load conditions, and the dynamic transient voltage, current and power variations and the system's responses under different scenarios. Here, we present a detailed description of the modeling framework and methodology to obtain accurate estimates of these characteristics.

3.1 System-Level Modeling Framework

An overview of the IVR and the IVR-assisted system-level modeling framework is shown in Fig. 2. Users input high-level parameters, such as the input/output voltage range, maximum load current, and power delivery network (PDN) parameters. The input PDN parameters are obtained from PDN design tools such as the Intel Power Distribution Network and the Altera PDN Tool for initial estimation. After this early-stage study where the IVR designs with key parameters are determined, several iterations can be further performed to optimize a detailed PDN with signal integrity, as the PDN parasitics can be partly impacted by IVR designs. Technology parameters that characterize CMOS switches, capacitors, and inductors in the IVR are built-in and extensible when necessary, with a comprehensively-compiled database containing MOSFET and capacitor data from 130 nm down to 10 nm, based on ITRS and PTM models [59] as

well as recently published surface-mounted-inductor and integrated-inductor data [16, 61]. By default, the static module optimizes for maximum conversion efficiency (to reduce power delivery overhead); it also allows users to specify a different optimization target, such as area. The dynamic module considers the dynamic responses in IVR-assisted PDSs. The internal structure consists of the following key modules:

- The **design parameter module** reads in user input and technology information, such as input/output voltage, load power, power switch width, capacitor/inductor density and so on.
- The **power/area/ripple static module** calculates power consumption, static voltage ripple, and die/board area for various building blocks across different IVR topologies, based on design parameters.
- The **design optimizer module** calculates the optimal IVR designs based on the specified technology, architecture configurations and basic circuit design guidelines. The system-level modeling can further support run-time optimization to achieve the desired power delivering performance considering the PDS dynamic responses.
- The **dynamic response module** rapidly models the dynamic responses of IVRs and IVR-assisted full PDSs under load current transients and/or external commands with the help of the SPICE 3 circuit simulator.

Advanced users familiar with IVR design trade-offs can leverage built-in interfaces to specify design parameters directly. Our model not only considers the static performance characteristics of the IVR-assisted PDSs, but also applies distinctive modeling strategies to accurately capture dynamic system behaviors, which we will elaborate in the remaining sections.

3.2 Power/Area/Ripple Static Module

By power/area/ripple static modeling, we refer to the calculation of the IVR conversion efficiency, area, and voltage ripples based on static assumptions of average load conditions and statistics. In contrast, the dynamic module described in Section 3.3 deals with IVRs' dynamic responses to load current transients from dynamic power traces. The static model applies to switched-capacitor converters, buck converters, and digital LDOs, which are the most commonly used IVR topologies in processor's PDSs.

Switched-capacitor converters: Fig. 3(a) illustrates a basic switched-capacitor circuit. System-level modeling adopts the analytical model introduced by Seeman [57] and Le[36]. The model derives the charge multiplier vectors ($a_{c,i}$ and $a_{r,i}$) based on the switch topology, and uses these vectors to calculate both the slow (R_{SSL}) and fast (R_{FSL}) switching limit output impedances. R_{SSL} and R_{FSL} can be expressed as:

$$R_{SSL} = \sum_i \frac{(a_{c,i})^2}{C_i f_{sw}} \qquad R_{FSL} = \sum_i \frac{R_i (a_{r,i})^2}{D_i} \tag{4}$$

where C_i is the capacitance of the i-th capacitor assuming linear capacitors, R_i is the on-state resistance of switch i, f_{sw} is the switching frequency, and D_i is the duty cycle of the i-th switch in a switched-capacitor IVR. The power loss due to the series of output impedances is $I_{load}^2 \sqrt{R_{SSL}^2 + R_{FSL}^2}$. The losses due to the switch parasitic capacitance, bottom plate parasitic capacitance, and the gate leakage current from the fly capacitors are calculated to model the total power loss from the switching cells. Our model considers the commonly used Series-Parallel and Symmetric Ladder switched-capacitor topologies because both require capacitors with the same voltage rating and thus are suitable for on-chip implementation [57]. Researchers can plug in their own switched-capacitor topology by providing the charge multiplier vectors explicitly. Meanwhile, the high switching frequencies in the integrated voltage regulator move SC IVR into operating areas in which stray inductance becomes more obvious [48, 49, 77]. The power loss in this stray inductance can also be modeled and its equivalent resistance. The equivalent resistance from stray inductances can be

calculated by [48]:

$$R_{EQ}(f_{sw}) = R_{ISL}(f_{sw})[1 + [\frac{R_{FSL}}{R_{ISL}(f_{sw})}]^{1.224}]^{1/1.224} \tag{5}$$

$$R_{ISL} = \frac{4L}{D^2} f_{sw} \tag{6}$$

where $R_{EQ}(f_{sw})$ is the equivalent resistance from stray inductances. R_{ISL} is the asymptote of the equivalent resistance that is caused by stray inductance. D is the duty cycle of the swichted capacitor IVR, L is the loop inductance.

Buck converters: A typical buck converter is shown in Fig. 3(b). We adopt an existing validated analytical model that calculates the power loss of buck converters, as can be found in previous work on off-chip voltage regulators [13]. This model is based on the high-side and low-side switch resistance/capacitance, inductor size, parasitic resistance, capacitance, switching frequency, and PWM signal duty cycle. We extend this model to on-chip regulators by deriving the required parameters from the technology characteristics of switches and inductors, using parameters stored in its internal device database. Compared to an off-chip voltage regulator with a low switching frequency, the change of inductor characteristics with frequency is more pronounced in buck IVRs, and this effect is considered in the proposed system-level model by a polynomial-fitted frequency-dependent coefficient of the inductance.

Digital LDOs: Analog G_m amplifiers have been traditionally used in digital LDOs. Recent design trends [2] have increasingly adopted digital comparators and controllers to achieve faster transient responses. Therefore, our model evaluates digital LDOs with a digital feedback path, as illustrated in Fig. 3(c). Since a current efficiency close to 99% usually can be achieved by state-of-the-art digital LDO design for moderate load currents, the conversion efficiency of a digital LDO in this load range will closely follow a linear relationship satisfying V_{out}/V_{in}.

Common building blocks: As illustrated in Fig. 3, different IVR topologies share many of the same circuit building blocks, such as power switches, drivers, comparators, a digital controller, and a clock generator – not to mention the basic capacitor and inductor devices. By commensurately modeling these shared building blocks across all topologies, the system-level modeling guarantees fair comparisons between different topologies, given the same technology and design constraints, which is of paramount importance for the efficiency-driven design exploration discussed in Section 5.2. For advanced digital technology, the power consumed and the area occupied by the digital feedback system are minimal compared to the moderate load current (10s of mA) and the on-chip capacitor and inductor needed for IVRs. Despite its insignificant power and area proportion, such peripheral circuitry is still important for transient response analysis and the scalability study of IVR designs, and therefore is taken into account in the dynamic module of this system-level modeling approach.

Optimization and Power Delivery System: For design optimization, we adopt the traditional hyperparameter optimization approach called *grid search* (a.k.a. *parameter sweep*), which is simply an exhaustive search through a manually specified subset of the hyperparameter space of a learning algorithm. The grid search algorithm is guided by performance metrics such as conversion efficiency or die area. Its complexity is determined by $Grid^N$ where $Grid$ is the number of searching points of one design parameter and N is the number of design parameters. Meanwhile, we also tried and compared other optimization algorithms, such as "differential evolution", "dual annealing", and so on. The grid search is relatively slow but can achieve the best results in all the tested experiments.

A full power delivery system includes voltage regulators and power delivery networks. From the input PDN parameters [5, 75], such as the parasitic resistance, the efficiency of a full power delivery system can be calculated based on the losses in the voltage regulators and power delivery networks. By exploring the design parameters of power

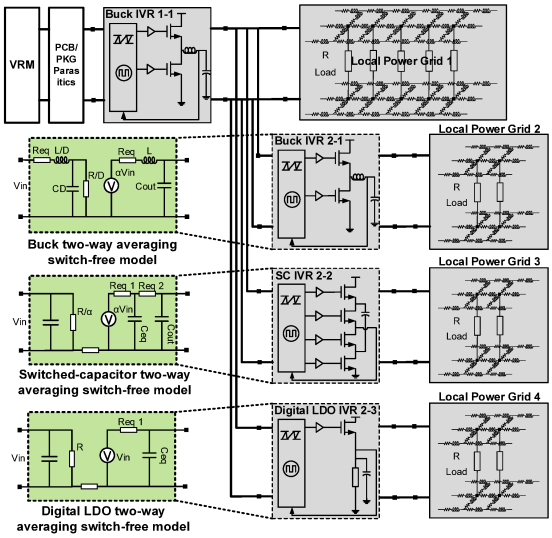


Fig. 4. Hierarchical power delivery system with integrated voltage regulator (IVR) dynamic models.

delivery systems, such as the middle voltages in a multiple stage power delivery system, the optimal power delivery system design for these input PDN parameters can be generated.

3.3 Dynamic Response Module

Besides static characteristics, the dynamic responses of IVRs also determine critical properties of the PDS, such as system reliability, efficiency and power management flexibility. The dynamic module models the dynamic responses of the three main types of IVRs in PDSs. Fig. 4 shows a hierarchical IVR-assisted PDS where the supply voltage is stepped down by multiple off-chip VRMs and on-chip IVRs before reaching the workloads. To effectively model these coexisting “serial and parallel" voltage regulators and the dynamic responses of the full PDSs, we propose a two-way averaging switch-free model which models each voltage regulator as a two port network without periodic switches. This model not only can capture all the critical dynamic responses but also can filter out the static voltage ripples from periodic switches, whose magnitudes are negligible in modern multi-phase IVR designs.

This two-way averaging switch-free model uses a power delivery network side and a load side to model each IVR as shown in the dynamic models of Fig. 4. As it models IVR as a switch-free two-port network, the model can be directly plugged into the power delivery network. In this model, the IVR switch dynamics are considered as average values of currents and voltages within a switching period by employing a weighted combination of the state equations of switching phases in pulse-width modulated (PWM) converters. By avoiding the periodic switches in the dynamic model, this model improves the simulation speed by 1000× the direct SPICE simulation, and also enables the AC analysis of the hierarchical PDS including multiple IVRs. Compared with a real voltage regulator, this averaging approach only neglects the static voltage ripple effects by using the switching state-space averaging (SSA) method [46]. A generalized transfer function (GTF) can be further deployed to evaluate the influence from this periodic switch ripples. Here, we use a classic buck converter to derive and demonstrate how this model captures the IVR dynamic responses in the PDS. That derivation and demonstration are not limited to buck converters, and can also be applied to other switching

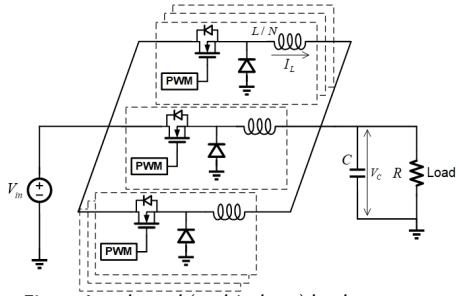


Fig. 5. Interleaved (multi-phase) buck converter.

voltage regulators like switched capacitor voltage regulators and digital LDOs, whose dynamic models are shown in Fig. 4. We will start with the two-way averaging model and then present the GTF analysis for the static voltage ripples.

An integrated buck converter is shown in Fig. 5. Its single-phase state space model can be described as:

$$\begin{aligned} \dot{X} &= AX(t) + B_i u(t), i = 1, 2, \\ y &= CX(t) + Du(t), \end{aligned} \tag{7}$$

where

$$X(t) = \begin{bmatrix} V_C(t) \\ I_L(t) \end{bmatrix}, A = \begin{bmatrix} -\frac{1}{RC} & \frac{1}{C} \\ -\frac{1}{L} & 0 \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ \frac{1}{L} \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$y(t) = V_o(t), C = \begin{bmatrix} 1 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 \end{bmatrix}, u(t) = V_{in}(t).$$

Modeling the switch period with the averaging model, the input matrix B is written as:

$$B = \alpha B_1 + (1 - \alpha) B_2 = \begin{bmatrix} 0 \\ \frac{\alpha}{L} \end{bmatrix},$$

where α is the duty ratio of the periodic switch, which is also the voltage conversion ratio of the integrated buck converter.

Thus, the above system can be modeled as an averaging model system with input $V_{in}^{'}$ and $B^{'}$.

$$V_{in}^{'} = \alpha V_{in}, B = \begin{bmatrix} 0 \\ \frac{1}{L} \end{bmatrix},$$

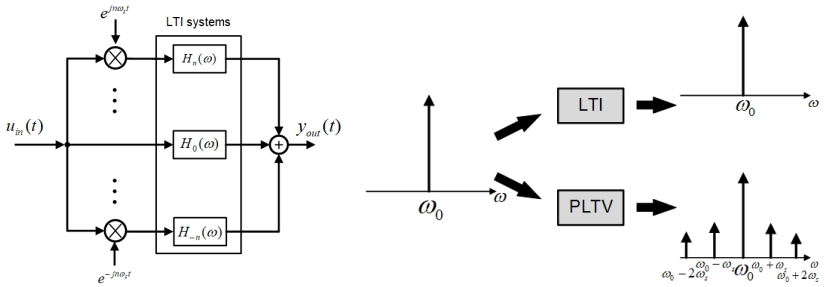
Similarly, the IVR and its loads can be modeled as Eq. (8).

$$\dot{X} = AX(t) + Bu(t), \tag{8}$$

where

$$X(t) = \begin{bmatrix} \frac{V_C(t)}{\alpha} \\ I_L(t) \end{bmatrix}, A = \begin{bmatrix} -\frac{1}{\frac{R}{\alpha} C\alpha} & \frac{1}{C\alpha} \\ -\frac{1}{\frac{L}{\alpha}} & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ \frac{1}{\frac{L}{\alpha}} \end{bmatrix}, u(t) = V_{in}(t).$$

This two-way averaging model supports the analysis and simulation of hierarchical power delivery networks by bridging the lower level and higher level power delivery network through IVRs. Similarly, the dynamic model of switched capacitor IVRs can be derived from the two-way averaging model [57].



(a) PLTV system. (b) Output of PLTV system.

Fig. 6. Periodical linear time-varying (PLTV) systems.

The two-way averaging switch-free model discussed above ignores the static voltage ripples from periodic switches. The GTF model gives out the system transfer function step by step from the switch between on and off, including the disturbances of periodic switches filtered out in the two-way averaging switch-free model. Continued from Eq. (7), the time domain solution of the integrated buck converter is

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}B_i u(\tau)\,d\tau, t \geq t_0. \tag{9}$$

Phase 1 (switch on): When $t \in [kT, kT + DT)$,

$$x(kT+DT)=e^{ADT}x(kT)+\int_{kT}^{kT+DT} e^{A(kT+DT-\tau)}B_1 u(\tau)\,d\tau. \tag{10}$$

Phase 2 (switch off): When $t \in [kT + DT, (k + 1)T)$,

$$x(t) = e^{A(t-kT-DT)}x(kT + DT). \tag{11}$$

At the end of period $t = (k + 1)T$,

$$x((k+1)T)=e^{AT}x(kT)+e^{A(k+1)T}\int_{kT}^{kT+DT} e^{-A\tau}B_1 u(\tau)\,d\tau \tag{12}$$

Because the buck converter is a non-linear system, small signal analysis is used in analyzing its dynamic response. The input can be expressed as the combination of a DC value and an AC component of frequency ω. $u(t) = u_0 + \widetilde{u}e^{j\omega t}$. The output at steady-state contains a DC component and an AC component of the same frequency. The GTF for the system above is given by

$$H_{GTF}(j\Omega) = C(e^{j\Omega T}I - e^{AT})e^{AT}\int_0^{DT} e^{-A\tau}B_1 e^{j\Omega\tau}\,d\tau, \tag{13}$$

where I is the identity matrix. The impulse response and transfer function can be extended to time-varying systems. The output is

$$y(t) = \int_{-\infty}^{\infty} h(t,\tau)u(\tau)d\tau, h(t,\tau) = R[\delta(t-\tau)], \tag{14}$$

where $h(t, \tau)$ is the generalized impulse response, R is an operator describing the system behavior, t is the observation time, and τ is the excitation time.

The bi-frequency transfer function is

$$H(\omega,\Omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t,\tau)e^{-j(\omega t-\Omega\tau)}dtd\tau, \tag{15}$$

where Ω and ω are the input and output frequencies. The time-varying transfer function can be written as

$$H(t,\Omega) = \int_{-\infty}^{\infty} h(t,\tau) e^{-j\Omega(t-\tau)} d\tau. \tag{16}$$

Here, $H(t,\Omega)$ is a periodic function of t, w.r.t. $\omega_s = \frac{2\pi}{T}$, and the system is a periodic linear time-varying (PLTV) system [67] as shown in Fig. 6. The LTI relationship can be recovered for $n = 0$, which is exactly modelled by the two-way averaging switch-free model.

$$H(t,\Omega) = \sum_{n=-\infty}^{n=\infty} H_n(\Omega) e^{jn\omega_s t}. \tag{17}$$

The frequency-dependent Fourier coefficients $H_n(\Omega)$ are called aliasing transfer functions:

$$H_n(\Omega) = \frac{1}{T} \int_0^T H(t,\Omega) e^{jn\omega_s t} dt. \tag{18}$$

The switches in the buck converter make the system non-linear by introducing new harmonics at multiples of ω_s, which can be evaluated by GTF analysis.

Based on the model of single-phase buck converter, the dynamic model of the modern interleaved buck converter (also called a multi-phase buck converter) can be derived as follows. For a N-phase buck converter, its state space model is

$$\begin{aligned} \dot{X} &= AX(t) + B_i u(t), i = 1, 2, ..., 2N, \\ y &= CX(t) + Du(t), \end{aligned} \tag{19}$$

where

$$X(t) = \begin{bmatrix} V_C(t) \\ I_{L_1}(t) \\ ... \\ I_{L_N}(t) \end{bmatrix}, A = \begin{bmatrix} -\frac{1}{RC} & \frac{1}{C} & \cdots & \frac{1}{C} \\ -\frac{1}{NL} & 0 & \cdots & 0 \\ & & \cdots & \\ -\frac{1}{NL} & 0 & \cdots & 0 \end{bmatrix},$$

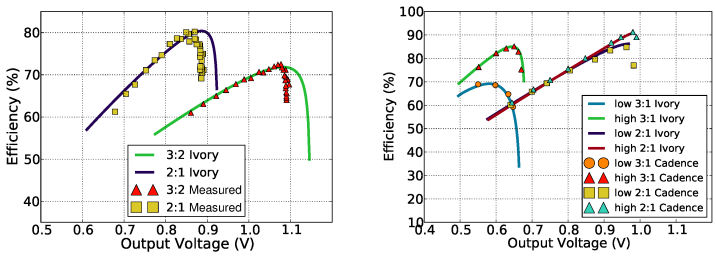
$$B_1 = \begin{bmatrix} \frac{1}{NL} \\ 0 \\ \cdots \\ 0 \end{bmatrix}, B_2 = \begin{bmatrix} \frac{1}{NL} \\ \frac{1}{NL} \\ \cdots \\ 0 \end{bmatrix}, \cdots, B_N = \begin{bmatrix} \frac{1}{NL} \\ \frac{1}{NL} \\ \cdots \\ \frac{1}{NL} \end{bmatrix},$$

$$B_{N+1} = \begin{bmatrix} 0 \\ \frac{1}{NL} \\ \cdots \\ \frac{1}{NL} \end{bmatrix}, B_{N+2} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ \frac{1}{NL} \end{bmatrix}, \cdots, B_{2N} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 \cdots 0 \end{bmatrix}, D = \begin{bmatrix} 0 \end{bmatrix},$$

$$y(t) = V_o(t), u(t) = V_{in}(t).$$

According to the state-space description of the multi-phase interleaved buck converter in Eq. (19), it has the same two-way averaging model as the conventional single phase buck converter. For an N phase interleaved buck converter, the GTF model can be derived with 2N phases in Eq. (9) - (12). In modern multi-phase interleaved voltage regulator



(a) Silicon measurements. (b) SPICE simulations.

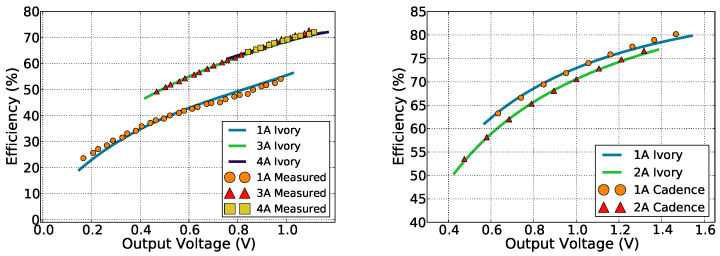
Fig. 7. Efficiency validation for SC converters.

designs, the static voltage ripple effects from periodic switches are sufficiently mitigated [4, 54]. When the number of interleaved phases $N \rightarrow \infty$, which is the ideal voltage regulator, there will not be any ripple effects from periodic switches and the two-way averaging switch-free model will reflect all the dynamic behaviors. Therefore, the two-way averaging switch-free model captures the essential dynamic responses of IVRs in power delivery systems. Now, we have presented a general IVR dynamic model, and any customized feedback laws for IVRs could also be easily reflected in this model by adjusting the duty ratio α, which is a part of the IVR dynamic module. For example in the voltage regulator with PID control, the duty ratio is controlled by the PID controller so that the duty ratio in a dynamic model will be expressed by $\alpha + k_{PID}(V_{ref} - V_{out})$.

4 MODEL VALIDATION

We validate the analytical model against both SPICE simulation results and measurement data from recent publications, spanning different technology nodes, input/output voltage ranges, and power levels. All these results demonstrate that the system-level model can faithfully represent and explore the design space of voltage regulator configurations in realistic PDS settings.

For the static model, validation data for the switched-capacitor IVR model is presented in Fig. 7. On the left, the model is compared against silicon measurements taken from a reconfigurable switched-capacitor implemented in 32nm SOI process [36]. It is clear that the model adequately matches the measured data for the 3:2 and the 2:1 configurations until an efficiency drop occurs past the peak efficiency. Normal switched-capacitors do not function past the efficiency *cliff region*. Given that these points are non-functional and are mostly likely caused by aggravated leakage current when the power switch exceeds its intended operating range, we conclude that the model is sufficiently accurate over the realistic, functional range of operation. Data points on the right plot were generated by SPICE simulations of two sets of 2:1 and 3:1 switched-capacitor converter designs in 40nm CMOS process [66]. Regular CMOS capacitors are used for the low-power density design, whereas embedded trench capacitors [10] are used for the high-power density design. The data validate the ability to model the conversion efficiency across all four designs. The buck converter IVR topologies are validated in Fig. 8. The measured data on the left is obtained from a 2.5D buck converter using an integrated inductor-on-silicon interposer, a 45nm SOI process and an embedded trench capacitor. The buck converter operates at different load current levels [61]. On the right is data from our buck design simulated in a 40nm CMOS process. it again proves capable of modeling voltage regulator efficiency, validating its internal buck converter modeling framework. Additionally, the analytical buck model used in the framework has previously been validated against off-chip VRMs [13].



(a) Silicon measurements. (b) SPICE simulations.

Fig. 8. Efficiency validation for buck converters.

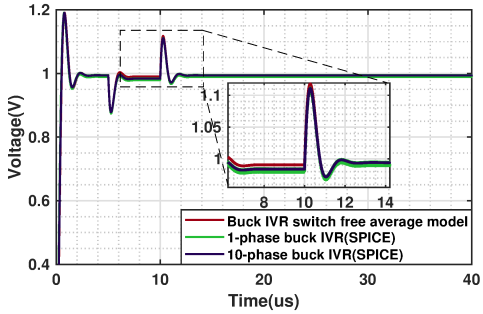


Fig. 9. Interleaved (multi-phase) buck converter dynamic responses in time domains.

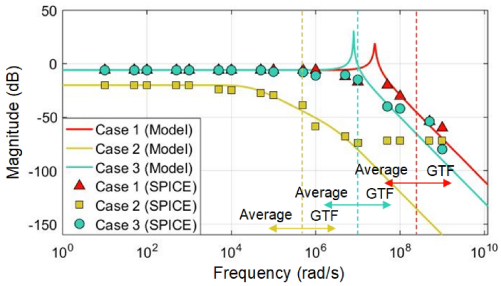


Fig. 10. Interleaved (multi-phase) buck converter frequency responses in frequency domains.

For the dynamic model, we validate the IVR two-way averaging switch-free model with SPICE simulations of recent IVR designs in both the time and frequency domains. Fig. 9 shows the comparison of the step responses starting from 0 μs, change of load from low (2 W) to high (2.5 W) at 5 μs, and change of load from high (2.5 W) to low (2 W) at 10 μs from the proposed two-way averaging switch-free model and the measurements of integrated buck voltage regulators [42] SPICE simulation with $L = 0.1uH$, $C = 0.5uF$, $f_{sw} = 20MHz$, $D = 0.2$, load $R = 0.5\Omega$ and different phases. To stress-test our method by evaluating the worst-case deviation of the proposed model and the measurements of IVRs, we specifically generate the voltage fluctuations (both overshoot and droop) with large magnitudes using a small output capacitor and without decoupling capacitors in our test. Although the 1-phase buck converter has voltage ripples from the periodic switches, the voltage ripples are effectively mitigated in the interleaved multi-phase (10-phase) designs. These kinds of static voltage ripples become trivial in real designs as interleaved multi-phase designs are widely used in modern IVRs [64]. The two-way averaging switch-free model naturally filters out the ripples from periodic switches but accurately captures all critical dynamic responses. Therefore, the two-way averaging switch-free models effectively and efficiently capture the dynamic response of modern IVRs. Fig. 10 shows the output and input frequency responses of recent integrated buck voltage regulator designs (case 1 [1], case 2 [42], and case 3 [56]). The two-way averaging switch-free models (plotted with curves) match the SPICE simulations (plotted by points) of the full integrated buck converters below half of the switching frequency. Above half of the switching frequency, the magnitude of frequency responses is the magnitude of static ripples, which is a constant value that is lower than -50 dB. As this magnitude is small, constant and mitigated in modern multi-phase IVR designs, it is ignorable compared to other dynamic responses. If this small constant magnitude from static voltage ripples needed to be analyzed in specific scenarios, the generalized transfer function (GTF) can be further used as described in Section 3.3.

Table 1. Summary of *Ivory* input parameters.

Parameter	Value
Max. Area(mm^2)	200
Total Average Power(W)	20
Total Peak Power(W)	56
Input Voltage(V)/Output Voltage(V)	3.3/1
Max Number of Distributed IVRS	4
R_{sw}(Ω·μm)/L(nH/mm^2)/C(nF/mm^2)	40/1/10
Off/On-Chip PDN parameter	$R_{off,on}$/$L_{off,on}$

5 CASE STUDY I: MANY-CORE GPU PDS

To demonstrate how *Ivory* enables early-stage design exploration at upper levels of the system stack, we present a case study on finding the optimal PDS configuration in the context of a GPU style many-core processor. Our goal is *not* to champion any one particular configuration, but rather to demonstrate how *Ivory* can be used for the early-stage design exploration of the PDS.

5.1 System Configuration

In this case study, we focus on the comparison between the IVR-assisted and a conventional off-chip VRM-based PDS. We assume an embedded GPU system with four cores (i.e., Streaming Multiprocessors, SMs) that form a 2×2 grid. The Fermi architecture based SM has an average power of 5 W and a peak power of 14 W. In the early-stage design space, this system uses the same off-chip and on-chip PDN equivalent circuit with previous IVR-assisted power delivery system [81, 82] and the corresponding GPU PDN parameters in GPUVolt[40], with a 3.3 V supply at the board and a 0.85 V SM nominal voltage + 0.15 V voltage guardband. The four SMs are modelled with 12×12 on-chip power grids, where each SM is modelled with 3×3 grid points. For a fair comparison, we assume that the on-chip metal resources are the same for each power delivery system. The parasitic resistance of the power grid is inversely proportional to the metal resource allocated to that power grid. Therefore, the metal resource allocations are formulated as an optimization problem. The objective function is the total power loss in the power delivery network: $min : I_1^2 \frac{k}{S_1} + I_2^2 \frac{k}{S_2} + I_3^2 \frac{k}{S_3} + ... + I_N^2 \frac{k}{S_N}$. The constraints are the total metal resource: $S_1 + S_2 + S_3 + ... + S_N \leq S$. S_i is the metal resource allocated to the ith voltage domain. S is the total metal resource. $\frac{k}{S_i}$ is the parasitic resistance in the ith voltage domain. I_i is the current of the ith voltage domain. In this case study with homogeneous cores as the loads, the metal resources for on-chip power grid are evenly allocated to each core. The area budget assigned to IVRs is swept from $30mm^2$ to $200mm^2$ and the budget includes the area for both the active circuitry and on-die passive components. This budget is set according to the inevitable physical constraint of IVR designs - the power density, and the load current of the GPU. The typical value of the former is down to $0.1W/mm^2$ [33, 36, 72]. More constraints (e.g., the available on-die area) can be included when available. Note that in this experiment we only consider on-die IVRs, but we can extend the design space exploration by including the package-based IVRs where extra parasitic resistance will be considered in the power grid. The other input parameters of *Ivory* are summarized in Table 1.

5.2 IVR Design Space Exploration

In this study, we set the maximum efficiency as the optimization target, and use *Ivory* to find the optimal IVR design (Fig. 14). We find the buck converter has higher efficiency than the SC converter with a more stringent area budget,

Table 2. Summary of design space exploration.

Topology	3:1 SC	Buck	LR
Distri. No.	1/2/4	1/2/4	1/2/4
Eff.(%)	80.1/80.3/80.1/	75.5/75.3/70.5	33.2/30.1/30
Ripple(mV)	$\leq$ 10 mV	$\leq$ 10 mV	$\leq$ 10 mV
f_{sw}(MHz)	106/109/106	83/94/179	300/300/300

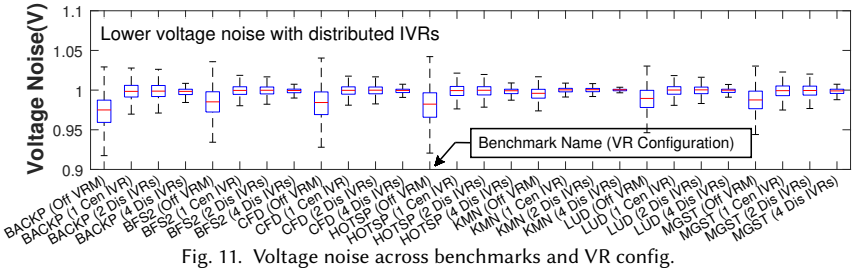


Fig. 11. Voltage noise across benchmarks and VR config.

although a high capacitor density process can be used to alleviate such hurdles. With the design constraints shown in Table 1, *Ivory* performs the design space exploration and gives the optimal IVR solution shown in Table 2.

5.3 Power Delivery System Dynamic Behaviors

We find that a 32-phase interleaved 3:1 switched-capacitor converter has the highest efficiency for this GPU system, and use it to optimize the dynamic response and for PDS optimization. We use the dynamic module to explore the centralized and distributed IVR designs, and we compare the results from previous default setting with the conventional off-chip VRM design which adopts a 6-phase buck converter [12]. The dynamic response analysis compares the IVR designs through a workload-dependent analysis. We feed *Ivory* GPU SM power traces from performance and power simulation infrastructures (GPGPUSim 3.2.0 [6] and GPUWattch [38]) in running large programs from the *Rodinia* suite [11] and NVIDIA CUDA SDK.

Ivory allows us to compare the run-time voltage noise of all centralized and distributed IVR configurations. In the centralized IVR configuration, the IVR is located in the middle of the 12x12 on-chip power grids, while in the distributed IVR configurations, the four IVRs are evenly distributed in the 12x12 on-chip power grids. The voltage statistics of the GPU system running different workloads are shown in box plots in Fig. 11. As indicated by the tight boxes with short whiskers, the design with four distributed IVRs is the optimal solution in supply voltage noise mitigation. Fig. 12 shows the supply voltage trace of the workload "CFD" with different VR designs. The voltage noise range in the off-chip VRM, the centralized IVR, the two distributed IVRs, and the four distributed IVRs scenarios are 125 mV, 59 mV, 55 mV, and 25 mV, respectively.

Besides the exhaustive time domain simulation of supply voltage noise with a run-time workload power trace, *Ivory* also supports the AC analysis of the full PDS including customized IVR feedback controls. The impedance plot is from power delivery network AC analysis. In the impedance plot the x-axis is the frequency and the y-axis is the impedance Z. The supply voltage noise can be calculated by the impedance and current through PDN: V=Z(f)*I(f). Previously, the AC analysis cannot take voltage regulator and its feedback control into consideration due to the switching nature of the voltage regulator. The improved two-way averaging switch-free model in Ivory can support the AC analysis of the full power delivery system including power delivery network and IVR. Fig. 13(a) (from *Ivory*) presents the effective impedance (normalized to core current) to load variations of centralized IVR-assisted, and distributed IVRs-assisted

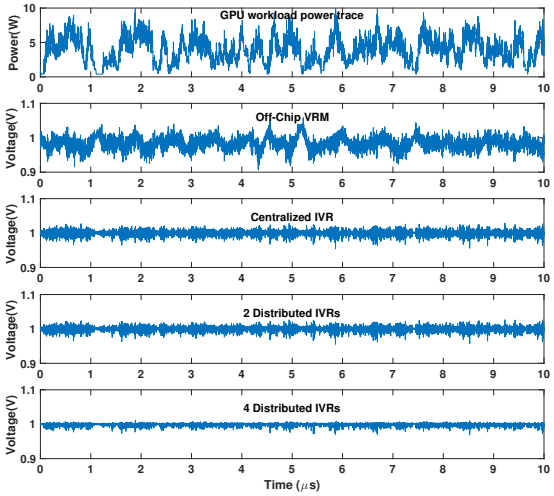
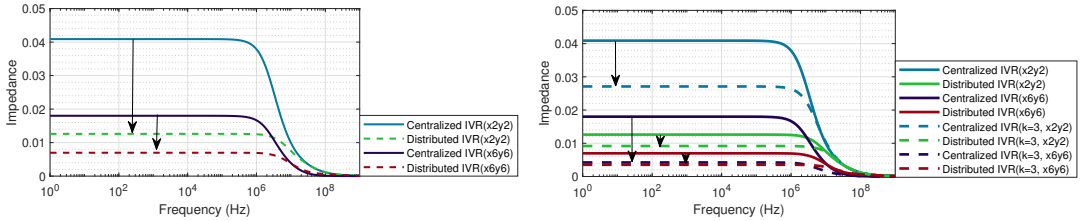


Fig. 12. Voltage noise waveforms (CFD) with VR config.



(a) Effective impedance in centralized and distributed IVR PDS. (b) Effective impedance in centralized and distributed IVRs with feedback.

Fig. 13. Supply voltage noise effective impedance.

PDSs. For a closer examination of each point, we index the 12×12 on-chip power grid with 12×12 X-Y coordinates where grid point (x6,y6) is the middle grid point of the power delivery network. The effective impedance plot directly demonstrates that the distributed IVR configuration has a lower effective impedance and less supply voltage noise than the centralized IVR configuration, especially for load points located far away from the IVRs, like the grid point (x2,y2).

The constant values in the low frequency range are from the voltage regulator's internal resistance and the power delivery network's parasitic resistance, which were ignored in previous work. In [20], for example, VRMs are directly modeled as a fixed voltage source for simplicity. To account for the voltage regulator's internal resistance and power delivery network's parasitic resistance (which is also called IR drop), an extra voltage margin is added to the fixed voltage source as load line compensation. However, in real voltage regulator designs and PDSs, feedback control plays an important role in mitigating the static voltage drop and the low frequency voltage noise within the regulation frequency. Fig. 13(b) shows the effective impedance after introducing feedback control (for example proportional feedback control k=3). In the high frequency range, the resonant impedance in an off-chip VRM-based PDS exceeds the feedback regulation frequency, but it can be mitigated by IVR, especially by a distributed IVR-assisted PDS. Furthermore, the distributed IVR-assisted PDS has a lower supply noise impedance over the full range, especially at the resonant frequency. In the medium and low frequency ranges, and within the regulation frequency ranges, the feedback control can further help mitigate static IR drop and low frequency voltage noise.

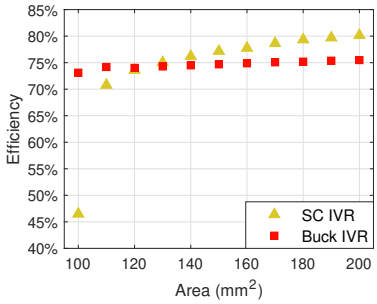


Fig. 14. IVR efficiency trade-off with area.

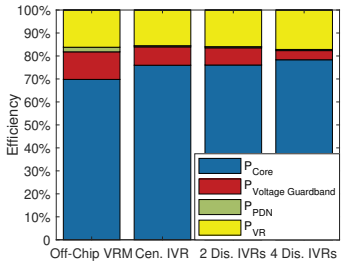


Fig. 15. Power efficiency breakdown.

Table 3. CPU and GPU many-core system.

Configuration	Value	Configuration	Value	Configuration	Value	Configuration	Value
PCB Supply Volt.	5 V	CPU Core Num.	16	Process Tech.	40 nm	GPU SMs Num.	15
CPU PDN Para.	DisPDN	Dispatch Width	4	GPU PDN Para.	GPUvolt	SM L1 $	16 KB
CPU Core Arch.	Nehalem	Replace Policy	LRU	GPU Core Arch.	Fermi	Chip L2 $	768 KB
CPU Core Volt.	0.6-1 V	I-TLB Entries	128	GPU Core Volt.	0.8-1 V	SM L1 $ Asso.	4
CPU Core Power	0-5 W	I-TLB Asso.	4	GPU Core Power	0-14 W	Chip L2 $ Asso.	16
CPU Core L1$.	32 KB	D-TLB Entries	512	Threads per SM	1536	SM L1 $ Block	128B
CPU Core L2$.	512 KB	D-TLB Asso.	4	Registers per SM	128 KB	Chip L2 $ Block	128 B
CPU Core L3$.	8 MB	2nd TLB Entries	512	Warp Threads	32	Shared Memory	48 KB
Execution Order	OoOE	2nd TLB Asso.	4	Warp Scheduler	GTO	Mem Bandwidth	179.2GB/s

5.4 Putting It Together: Power Efficiency Analysis

Ivory lets designers rapidly evaluate the final PDS efficiency through combined static and dynamic analysis. The static converter design analysis finds the optimal converter with high converter efficiency and low IR-drop loss. *Ivory* further optimizes the voltage margin by identifying the IVR design with the minimal voltage noise that accounts for most of the voltage margin [37]. Fig. 15 shows the power delivery efficiency breakdowns of different PDS designs. The power efficiency is the percentage of power consumed by cores that perform the actual computation over total power input to the PDS. The optimal PDS solution obtained by *Ivory* achieves a 8.6% power efficiency improvement over the previous off-chip VRM-based PDS, without any performance loss.

6 CASE STUDY II: PDS WITH FAST PER-CORE DVFS

Another significant benefit of an IVR-assisted PDS is fast power management, such as microsecond level fast DVFS. Computer architects keep pursuing faster power management, because faster voltage scaling means higher power and energy efficiency. Voltage regulator circuit designers usually focus on voltage conversion efficiency under area constraints. In this case study, we demonstrate *Ivory* as the downstream platform, after architecture level performance analysis and power simulation, to analyze power delivery for many-core computing systems and bridge the gap between computer architects and voltage regulator circuit designers.

6.1 System Configuration

In this case study, we apply the fast DVFS supported by IVR-assisted PDS on both CPU cores with an Intel Nehalem (x86) architecture and GPU streaming multi-processors (SMs) with an NVIDIA Fermi architecture. The detailed specifications

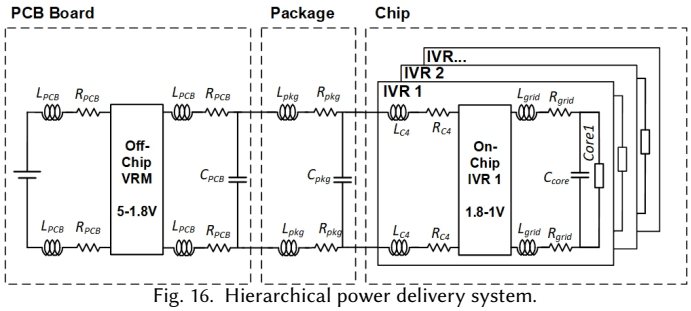


Fig. 16. Hierarchical power delivery system.

Table 4. CPU core DVFS frequency and voltage pairs.

Core Freq. (MHz)	2000	1800	1500	1000	800
Core Voltage (V)	1.2	0.9	0.8	0.7	0.6

Table 5. GPU core DVFS frequency and voltage pairs.

Core Freq. (MHz)	700	650	600	550	300
Core Voltage (V)	1.00	0.95	0.91	0.87	0.46

Table 6. Summary of design space explorations of 16-phase buck IVRs.

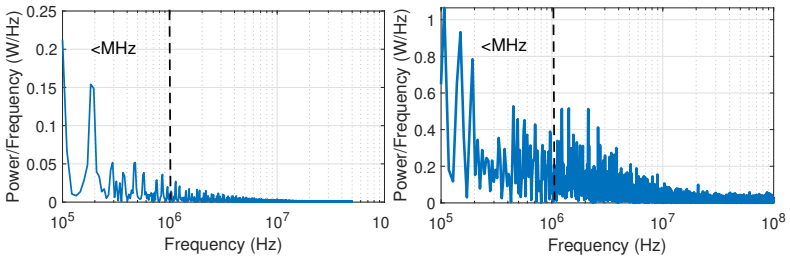
DVFS Speed	**500ns**	**2μs**	**5μs**	**50μs**
Efficiency (%)	77.0	83.3	83.4	84.8
Switch Freq. f_{sw} (MHz)	189	62	53	44
L per-phase (nH)	0.25	0.5	0.75	0.5
C per-phase (μF)	0.125	0.56	0.56	1.125
Area (mm^2)	42.9	184.9	187.0	365.9

of this CPU and GPU system are shown in Table 3. The system is powered with a hierarchical IVR-assisted PDS [29, 72, 80], shown in Fig. 16. The hierarchical IVR-assisted PDS [72, 80] is proposed to adopt an off-chip VRM to step down the voltage from board level to an intermediate level (for example 1.8V) with higher efficiency. Then the per-core IVR further regulates the intermediate voltage to the desired core voltage with more flexibility. The off-chip VRM is modeled based on a commercial product [12]. The CPU on-chip power grid is scaled from the distributed power delivery network[20], and the GPU on-chip power grid is from GPUVolt[40], which are validated with CPU and GPU systems respectively. We use *Ivory* to find the IVR designs that can support the desired microsecond per-core DVFS.

On the architecture side, we use Sniper[9] (with Mcpat) and GPGPUsim[6, 26] (compatible with GPUWattch) to simulate the architecture level performance and power activities of CPU and GPU systems. Sniper (with Mcpat) simulates the CPU part, generating run-time statistics with a granularity of 100 ns, and GPGPUsim 3.1.1 (with GPUWattch) simulates the GPU part at 700 MHz. We use representative benchmarks that cover a wide range of scientific and computational domains from CPU benchmarks *parsec* and *splash2*, and also the GPU benchmarks from the *Rodinia* suite [11] and NVIDIA CUDA SDK. Here, we use Fermi GPU architecture mainly for keeping it consistent and comparable with previous work with widely accepted simulation tools. Fermi architecture is the most classic and representative GPU architecture and widely used GPU in the study of GPU power delivery and power management such as supply voltage noise mitigation and fast power management [37, 43, 44, 65]. The CPU and GPU voltage and frequency scaling pairs are shown in Table 4 and Table 5, respectively.

6.2 IVR Support for Fast DVFS

To guide the IVR designs especially for fast DVFS, we perform a frequency analysis on the power activities of the CPU and GPU cores in running benchmarks. For example, the CPU and GPU core power frequency analyses in executing the *blackscholes* and *backp* system benchmarks are shown in Fig. 17, where both the CPU and GPU power variations



(a) CPU frequency analysis. (b) GPU frequency analysis.

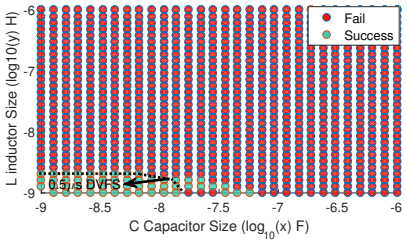
Fig. 17. CPU and GPU power activity frequency analysis in executing blackscholes and backp benchmarks.

can reach MHz. To capture the majority of the power variations, we correspondingly explore the IVR designs that can support fast DVFS, which is up to microsecond timescales.

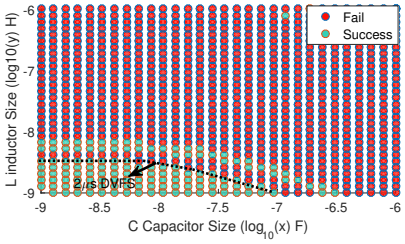
From the perspective of DVFS control policy, the faster (smaller period) control means that the voltage and frequency can act faster and usually can achieve a higher efficiency. Therefore, the DVFS control policy prefers a fast speed (small period). However, in the power delivery system, the voltage transition time of the DVFS is limited by the physical constraints. For example, the passive components like inductors and capacitors in the integrated voltage regulators and power delivery network limit the voltage transition time. Based on the hierarchical IVR-assisted PDS in Fig. 16, we use the *Ivory* dynamic module to explore the design spaces of IVRs with hierarchical PDSs that can support different fast DVFS. Here, we set the voltage scaling rise time to within 1% of DVFS intervals [25, 27, 72, 80] and the voltage overshoot to less than 5%. In the design space of an integrated buck voltage regulator, the passive inductors and capacitors directly and significantly affect the voltage scaling speed. Fig. 18 shows the design space explorations of the inductor and capacitor sizes for desired voltage scaling speeds, where the green points indicate that the inductor and capacitor design parameters can support the desired CPU fast DVFS. These parameters further form new design space boundaries and are passed to the static module to find proper IVR design configurations. Similar approaches are also applied to the GPU SM cores. The key design parameters for the IVRs that support different speeds of DVFS are summarized in Table 6. When supporting fast DVFS, IVR designs keep reducing the size of on-die inductors and capacitors to achieve a faster voltage transition, and one prominent side effect is pushing the switching frequency from tens to hundreds of MHz. The higher frequency switching comes at the cost of degrading the conversion efficiency of the IVRs as the switching loss becomes more significant.

6.3 Power Delivery System and Architecture Co-Design

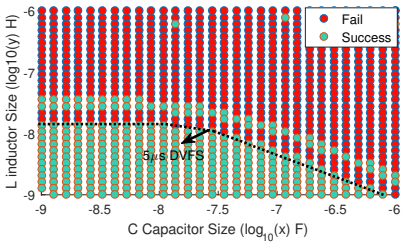
Finally, we evaluate the system's energy efficiency with fast per-core voltage scaling supported by this hierarchical PDS. For a fair comparison of the raw benefit from the fast per-core DVFS given by an IVR, we use a native DVFS mechanism where the instructions-per-cycle (IPC) value is monitored to adjust the frequency and voltage at run-time. The energy benefits for different speeds of fast DVFS supported by IVRs on CPUs and GPUs are shown in Fig. 19 and Fig. 20 respectively. The fast DVFS supported by IVR can reach finer granularity and save more energy for CPUs and GPUs. On the CPU side, the 50 μs, 2 μs, and 500ns DVFS have energy saving of 7.65%, 12.5%, and 15.7% on average, and 20.7%, 33.5%, and 43.2% on specific workloads like *Canneal*. Also, on the GPU side, the 50 μs, 2 μs, and 500 ns DVFS offer energy savings of 18.2%, 50.0%, and 55.4% on average and 55.8%, 66.6% and 66.9% on specific workloads like *Srad* and *LavaMD*. Together with the results from *Ivory*, although the fastest DVFS (0.5μs DVFS) achieves the greatest energy saving, the implementation overheads of IVRs offset the fast DVFS benefits. The 2μs DVFS is the proper candidate for this hetergenous system especially for the GPU SMs, because it not only reaps the energy benefits from fast DVFS and



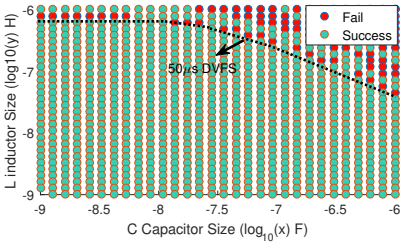
(a) 0.5μs voltage scaling.



(b) 2μs voltage scaling.



(c) 5μs voltage scaling.



(d) 50μs voltage scaling.

Fig. 18. Inductor and capacitor sizes for different voltage scaling speeds.

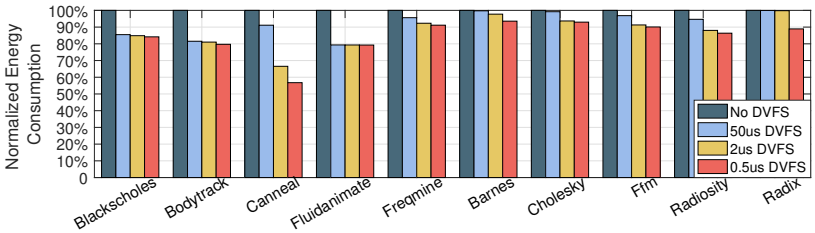


Fig. 19. CPU energy benefit from different speed fast DVFS.

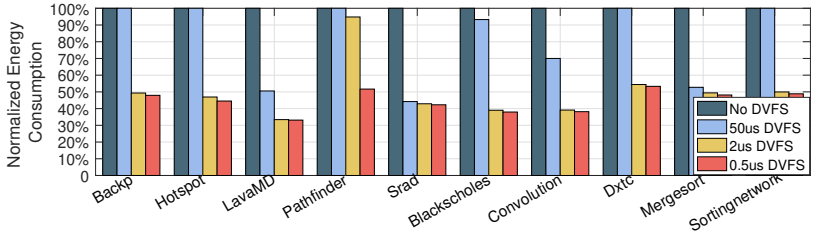


Fig. 20. GPU energy benefit from different speed fast DVFS.

the power delivery efficiency improvement seen in case study I, but also avoids the costly IVR overheads. Note that this is an initial design at the early-stage since we use the core average load current and the nominal supply voltage for estimation to achieve the trade-off between design accuracy and speedup. This early-stage design will work as a starting point for future detailed IVR-assisted power delivery system designs, such as run-time re-configurable IVR designs for various frequencies of fast DVFS, and IP core designs with customized fast power management mechanisms.

7 CONCLUSIONS

Subtle trade-offs and topology choices in IVRs can make efficiency decisions unintuitive, forcing researchers to use inaccurate or incomplete models. As IVRs continue to grow in popularity and become more beneficial, the system-level model exposes design space trade-offs and supports dynamic response optimization without manual effort and without the circuit expertise otherwise required, making the system-level model and tool useful to system architects. Using design space exploration, we show cases where optimizing across technologies, topologies, and dynamic responses can yield area and efficiency savings that would otherwise be missed without such a high-level model.

8 ACKNOWLEDGEMENTS

The research described in this paper was partly supported by NSF Award CNS-1739643, Semiconductor Research Corporation (SRC) Task 2810.003 through the University of Texas at the Dallas Texas Analog Center of Excellence (TxACE), and the National Natural Science Foundation of China (NSFC) 62072297. We are also grateful to the reviewers for their constructive feedback.

REFERENCES

[1] Siamak Abedinpour, Bertan Bakkaloglu, and Sayfe Kiaei. 2007. A multistage interleaved synchronous Buck converter with integrated output filter in 0.18 *mu* m SiGe process. *IEEE Transactions on Power Electronics* 22, 6 (2007), 2164–2175.

[2] Mohammad Al-Shyoukh, Hoi Lee, and Raul Perez. 2007. A transient-enhanced low-quiescent current low-dropout regulator with buffer impedance attenuation. *IEEE journal of solid-state circuits* 42, 8 (2007), 1732–1742.

[3] Toke Meyer Andersen, Florian Krismer, Johann Walter Kolar, Thomas Toifl, Christian Menolfi, Lukas Kull, Thomas Morf, Marcel Kossel, Matthias Brändli, Peter Buchmann, et al. 2014. 4.7 A sub-ns response on-chip switched-capacitor DC-DC voltage regulator delivering 3.7 W/mm 2 at 90% efficiency using deep-trench capacitors in 32nm SOI CMOS. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE, 90–91.

[4] David Baba. 2012. Benefits of a multiphase buck converter. *Texas Instruments Incorporated* 2012 (2012).

[5] Rassul Bairamkulov, Kan Xu, Mikhail Popovich, Juan S Ochoa, Vaishnav Srinivas, and Eby G Friedman. 2019. Power Delivery Exploration Methodology based on Constrained Optimization. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2019).

[6] Ali Bakhoda, George L Yuan, Wilson WL Fung, Henry Wong, and Tor M Aamodt. 2009. Analyzing CUDA workloads using a detailed GPU simulator. In *2009 IEEE International Symposium on Performance Analysis of Systems and Software*. IEEE, 163–174.

[7] Lakshmi Bhamidipati, Bhoopal Gunna, Houman Homayoun, and Avesta Sasan. 2017. A power delivery network and cell placement aware IR-drop mitigation technique: Harvesting unused timing slacks to schedule useful skews. In *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 272–277.

[8] Edward A Burton, Gerhard Schrom, Fabrice Paillet, Jonathan Douglas, William J Lambert, Kaladhar Radhakrishnan, and Michael J Hill. 2014. FIVR—Fully integrated voltage regulators on 4th generation Intel® Core™ SoCs. In *Applied Power Electronics Conference and Exposition (APEC), 2014 Twenty-Ninth Annual IEEE*. IEEE, 432–439.

[9] Trevor E Carlson, Wim Heirmant, and Lieven Eeckhout. 2011. Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation. In *SC'11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–12.

[10] Leland Chang, Robert K Montoye, Brian L Ji, Alan J Weger, Kevin G Stawiasz, and Robert H Dennard. 2010. A fully-integrated switched-capacitor 2 1 voltage converter with regulation capability and 90% efficiency at 2.3 A/mm 2. In *VLSI Circuits (VLSIC), 2010 IEEE Symposium on*. IEEE, 55–56.

[11] Shuai Che, Michael Boyer, Jiayuan Meng, David Tarjan, Jeremy W Sheaffer, Sang-Ha Lee, and Kevin Skadron. 2009. Rodinia: A benchmark suite for heterogeneous computing. In *2009 IEEE international symposium on workload characterization (IISWC)*. Ieee, 44–54.

[12] CHL8266. 2011. *Digital Multi-Phase GPU Buck Controller*. Technical Report. infineon.

[13] Yongseok Choi, Naehyuck Chang, and Taewhan Kim. 2007. DC–DC converter-aware power management for low-power embedded systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 26, 8 (2007), 1367–1381.

[14] Hadi Esmaeilzadeh, Emily Blem, Renee St Amant, Karthikeyan Sankaralingam, and Doug Burger. 2011. Dark silicon and the end of multicore scaling. In *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*. IEEE, 365–376.

[15] Eric J Fluhr, Steve Baumgartner, David Boerstler, John F Bulzacchelli, Timothy Diemoz, Daniel Dreps, George English, Joshua Friedrich, Anne Gattiker, Tilman Gloekler, et al. 2015. The 12-Core POWER8™ Processor With 7.6 Tb/s IO Bandwidth, Integrated Voltage Regulation, and Resonant Clocking. *IEEE Journal of Solid-State Circuits* 50, 1 (2015), 10–23.

[16] Donald S Gardner, Gerhard Schrom, Fabrice Paillet, Brice Jamieson, Tanay Karnik, and Shekhar Borkar. 2009. Review of on-chip inductor structures with magnetic films. *IEEE Transactions on Magnetics* 45, 10 (2009), 4760–4766.

[17] Hamid Reza Ghasemi, Abhishek A Sinkar, Michael J Schulte, and Nam Sung Kim. 2012. Cost-effective power delivery to support per-core voltage domains for power-constrained processors. In *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*. IEEE, 56–61.

[18] Dimitris Gizopoulos, George Papadimitriou, Athanasios Chatzidimitriou, Vijay Janapa Reddi, Behzad Salami, Osman S. Unsal, Adrian Cristal Kestelman, and Jingwen Leng. 2019. Modern Hardware Margins: CPUs, GPUs, FPGAs Recent System-Level Studies. In *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*.

[19] Juliana Gjanci and Masud H Chowdhury. 2010. A hybrid scheme for on-chip voltage regulation in system-on-a-chip (SOC). *IEEE transactions on very large scale integration (VLSI) systems* 19, 11 (2010), 1949–1959.

[20] Meeta S Gupta, Jarod L Oatley, Russ Joseph, Gu-Yeon Wei, and David M Brooks. 2007. Understanding voltage variations in chip multiprocessors using a distributed power-delivery network. In *Design, Automation & Test in Europe Conference & Exhibition*.

[21] Edward NY Ho and Philip KT Mok. 2010. A capacitor-less CMOS active feedback low-dropout regulator with slew-rate enhancement for portable on-chip application. *IEEE Transactions on Circuits and Systems II: Express Briefs* 57, 2 (2010), 80–84.

[22] MD Shazzad Hossain and Ioannis Savidis. 2020. Recycling of unused leakage current for energy efficient multi-voltage systems. *Microelectronics Journal* 101 (2020), 104782.

[23] Rinkle Jain, Bibiche M Geuskens, Stephen T Kim, Muhammad M Khellah, Jaydeep Kulkarni, James W Tschanz, and Vivek De. 2014. A 0.45–1 V fully-integrated distributed switched capacitor DC-DC converter with high density MIM capacitor in 22 nm tri-gate CMOS. *IEEE Journal of Solid-State Circuits* 49, 4 (2014), 917–927.

[24] Ulya R Karpuzcu, Abhishek Sinkar, Nam Sung Kim, and Josep Torrellas. 2013. Energysmart: Toward energy-efficient manycores for near-threshold computing. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*. IEEE, 542–553.

[25] Ben Keller, Martin Cochet, Brian Zimmer, Jaehwa Kwak, Alberto Puggelli, Yunsup Lee, Milovan Blagojević, Stevo Bailey, Pi-Feng Chiu, Palmer Dabbelt, et al. 2017. A RISC-V processor SoC with integrated power management at submicrosecond timescales in 28 nm FD-SOI. *IEEE Journal of Solid-State Circuits* 52, 7 (2017), 1863–1875.

[26] Mahmoud Khairy, Zhesheng Shen, Tor M Aamodt, and Timothy G Rogers. 2020. Accel-Sim: An extensible simulation framework for validated GPU modeling. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 473–486.

[27] Stephen T Kim, Yi-Chun Shih, Kaushik Mazumdar, Rinkle Jain, Joseph F Ryan, Carlos Tokunaga, Charles Augustine, Jaydeep P Kulkarni, Krishnan Ravichandran, James W Tschanz, et al. 2015. Enabling wide autonomous DVFS in a 22 nm graphics execution core using a digitally controlled fully integrated voltage regulator. *IEEE Journal of Solid-State Circuits* 51, 1 (2015), 18–30.

[28] Wonyoung Kim, David Brooks, and Gu-Yeon Wei. 2012. A fully-integrated 3-level DC-DC converter for nanosecond-scale DVFS. *IEEE Journal of Solid-State Circuits* 47, 1 (2012), 206–219.

[29] Wonyoung Kim, Meeta S Gupta, Gu-Yeon Wei, and David Brooks. 2008. System level analysis of fast, per-core DVFS using on-chip switching regulators. In *2008 IEEE 14th International Symposium on High Performance Computer Architecture*. IEEE, 123–134.

[30] Youngtaek Kim and Lizy Kurian John. 2011. Automated di/dt stressmark generation for microprocessor power delivery networks. In *Low Power Electronics and Design (ISLPED) 2011 International Symposium on*. IEEE, 253–258.

[31] Selçuk Kose and Eby G Friedman. 2012. Distributed on-chip power delivery. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 2, 4 (2012), 704–713.

[32] Harish K Krishnamurthy, Vaibhav Vaidya, Pavan Kumar, Rinkle Jain, Sheldon Weng, Stephen T Kim, George E Matthew, Nachiket Desai, Xiaosen Liu, Krishnan Ravichandran, et al. 2017. A digitally controlled fully integrated voltage regulator with on-die solenoid inductor with planar magnetic core in 14-nm tri-gate CMOS. *IEEE Journal of Solid-State Circuits* 53, 1 (2017), 8–19.

[33] Pavan Kumar, Vaibhav A Vaidya, Harish Krishnamurthy, Stephen Kim, George E Matthew, Sheldon Weng, Bharani Thiruvengadam, Wayne Proefrock, Krishnan Ravichandran, and Vivek De. 2015. A 0.4 V 1V 0.2 A/mm 2 70% efficient 500MHz fully integrated digitally controlled 3-level buck voltage regulator with on-die high density MIM capacitor in 22nm tri-gate CMOS. In *2015 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 1–4.

[34] Sebastian Kutzner, Axel Poschmann, and Marc Stöttinger. 2013. TROJANUS: An ultra-lightweight side-channel leakage generator for FPGAs. In *2013 International Conference on Field-Programmable Technology (FPT)*. IEEE, 160–167.

[35] Hanh-Phuc Le, John Crossley, Seth R Sanders, and Elad Alon. 2013. A sub-ns response fully integrated battery-connected switched-capacitor voltage regulator delivering 0.19 W/mm 2 at 73% efficiency. In *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*. IEEE, 372–373.

[36] Hanh-Phuc Le, Seth R Sanders, and Elad Alon. 2011. Design techniques for fully integrated switched-capacitor DC-DC converters. *IEEE Journal of Solid-State Circuits* 46, 9 (2011), 2120–2131.

[37] Jingwen Leng, Alper Buyuktosunoglu, Ramon Bertran, Pradip Bose, and Vijay Janapa Reddi. 2015. Safe limits on voltage reduction efficiency in GPUs: a direct measurement approach. In *Microarchitecture (MICRO), 2015 48th Annual IEEE/ACM International Symposium on*. IEEE, 294–307.

[38] Jingwen Leng, Tayler Hetherington, Ahmed ElTantawy, Syed Gilani, Nam Sung Kim, Tor M Aamodt, and Vijay Janapa Reddi. 2013. GPUWattch: enabling energy optimizations in GPGPUs. In *ACM SIGARCH Computer Architecture News*, Vol. 41. ACM, 487–498.

[39] Jingwen Leng, Yazhou Zu, and Vijay Janapa Reddi. 2014. Energy Efficiency Benefits of Reducing the Voltage Guardband on the Kepler GPU Architecture. In *Workshop on Silicon Errors in Logic - System Effects (SELSE)*.

[40] Jingwen Leng, Yazhou Zu, Minsoo Rhu, Meeta Gupta, and Vijay Janapa Reddi. 2014. GPUVolt: Modeling and characterizing voltage noise in GPU architectures. In *Proceedings of the 2014 international symposium on Low power electronics and design*. ACM, 141–146.

[41] Jules Mace. 2020. *Design and Control of Topologies for Voltage Stacking*. Ph.D. Dissertation.

[42] Ashis Maity, Amit Patra, Norihisa Yamamura, and Jonathan Knight. 2011. Design of a 20 MHz DC-DC buck converter with 84 percent efficiency for portable applications. In *2011 24th Internatioal Conference on VLSI Design*. IEEE, 316–321.

[43] Xinxin Mei, Xiaowen Chu, Hai Liu, Yiu-Wing Leung, and Zongpeng Li. 2017. Energy efficient real-time task scheduling on CPU-GPU hybrid clusters. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[44] Xinxin Mei, Qiang Wang, and Xiaowen Chu. 2017. A survey and measurement study of GPU DVFS on energy conservation. *Digital Communications and Networks* 3, 2 (2017), 89–100.

[45] Pascal Meinerzhagen, Carlos Tokunaga, Andres Malavasi, Vaibhav Vaidya, Ashwin Mendon, Deepak Mathaikutty, Jaydeep Kulkarni, Charles Augustine, Minki Cho, Stephen Kim, et al. 2018. An energy-efficient graphics processor featuring fine-grain DVFS with integrated voltage regulators, execution-unit turbo, and retentive sleep in 14nm tri-gate CMOS. In *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 38–40.

[46] Richard D Middlebrook and Slobodan Cuk. 1976. A general unified approach to modelling switching-converter power stages. In *1976 IEEE Power Electronics Specialists Conference*. IEEE, 18–34.

[47] Robert J Milliken, Jose Silva-Martínez, and Edgar Sánchez-Sinencio. 2007. Full on-chip CMOS low-dropout voltage regulator. *IEEE Transactions on Circuits and Systems I: Regular Papers* 54, 9 (2007), 1879–1890.

[48] Lukas Müller and Jonathan W Kimball. 2014. Effects of stray inductance on hard-switched switched capacitor converters. *IEEE Transactions on Power Electronics* 29, 12 (2014), 6276–6280.

[49] Yerzhan Mustafa, Vivekanandan Subburaj, and Alex Ruderman. 2019. Revisited SCC Equivalent Resistance High Frequency Limit Accounting for Stray Inductance Effect. *IEEE Journal of Emerging and Selected Topics in Power Electronics* (2019).

[50] James Myers, Anand Savanth, Rohan Gaddh, David Howard, Pranay Prabhat, and David Flynn. 2015. A subthreshold ARM cortex-M0+ subsystem in 65 nm CMOS for WSN applications with 14 power domains, 10T SRAM, and integrated voltage regulator. *IEEE Journal of Solid-State Circuits* 51, 1 (2015), 31–44.

[51] Chi-Hsien Pao, An-Yu Su, and Yu-Min Lee. 2020. XGBIR: an xgboost-based IR drop predictor for power delivery network. In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 1307–1310.

[52] George Papadimitriou, Athanasios Chatzidimitriou, Dimitris Gizopoulos, Vijay Janapa Reddi, Jingwen Leng, Behzad Salami, Osman Sabri Unsal, and Adrian Cristal Kestelman. 2020. Exceeding Conservative Limits: A Consolidated Analysis on Modern Hardware Margins. *IEEE Transactions on Device and Materials Reliability* (2020).

[53] George Patounakis, Yee William Li, and Kenneth L Shepard. 2004. A fully integrated on-chip DC-DC conversion and power management system. *IEEE Journal of Solid-State Circuits* 39, 3 (2004), 443–451.

[54] Gerard Villar Piqué. 2012. A 41-phase switched-capacitor power converter with 3.8 mV output ripple and 81% efficiency in baseline 90nm CMOS. In *2012 IEEE International Solid-State Circuits Conference*. IEEE, 98–100.

[55] Reddi, V.J. and Kanev, S. and Campanoni, S. and Smith, M.D. and Wei, G.Y. and Brooks, D. 2010. Voltage Smoothing: Characterizing and Mitigating Voltage Noise in Production Processors Using Software-Guided Thread Scheduling. In *Proc. Annual IEEE/ACM Int. Symp. on Microarchitecture*.

[56] Gerhard Schrom, P Hazucha, Fabrice Paillet, DJ Rennie, ST Moon, DS Gardner, T Kamik, P Sun, TT Nguyen, MJ Hill, et al. 2007. A 100MHz eight-phase buck converter delivering 12A in 25mm2 using air-core inductors. In *APEC 07-Twenty-Second Annual IEEE Applied Power Electronics Conference and Exposition*. IEEE, 727–730.

[57] Michael D Seeman. 2006. *Analytical and practical analysis of switched-capacitor dc-dc converters*. Technical Report. CALIFORNIA UNIV BERKELEY DEPT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE.

[58] Kamlesh Singh, Barry de Bruin, Jos Huisken, Hailong Jiao, and José Pineda de Gyvez. 2019. Voltage stacked design of a microcontroller for near/sub-threshold operation. In *2019 32nd IEEE International System-on-Chip Conference (SOCC)*. IEEE, 370–375.

[59] Saurabh Sinha, Greg Yeric, Vikas Chandra, Brian Cline, and Yu Cao. 2012. Exploring sub-20nm FinFET design with predictive technology models. In *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*. IEEE, 283–288.

[60] Noah Sturcken, Eugene J O'Sullivan, Naigang Wang, Philipp Herget, Bucknell C Webb, Lubomyr T Romankiw, Michele Petracca, Ryan Davies, Robert E Fontana, Gary M Decad, et al. 2012. A 2.5 D integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer. *IEEE Journal of solid-state circuits* 48, 1 (2012), 244–254.

[61] Noah Sturcken, Eugene J O'Sullivan, Naigang Wang, Philipp Herget, Bucknell C Webb, Lubomyr T Romankiw, Michele Petracca, Ryan Davies, Robert E Fontana, Gary M Decad, et al. 2013. A 2.5 D integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer. *IEEE Journal of solid-state circuits* 48, 1 (2013), 244–254.

[62] Noah Sturcken, Michele Petracca, Steven Warren, Paolo Mantovani, Luca P Carloni, Angel V Peterchev, and Kenneth L Shepard. 2012. A switched-inductor integrated voltage regulator with nonlinear feedback and network-on-chip load in 45 nm SOI. *IEEE Journal of Solid-State Circuits* 47, 8 (2012), 1935–1945.

[63] Nghia Tang, Wookpyo Hong, Bai Nguyen, Zhiyuan Zhou, Jong-Hoon Kim, and Deukhyoun Heo. 2020. Fully Integrated Switched-Inductor-Capacitor Voltage Regulator With 0.82-A/mm^2 Peak Current Density and 78% Peak Power Efficiency. *IEEE Journal of Solid-State Circuits* (2020).

[64] Texus Instruments. [n.d.]. LMZ10501 1-A SIMPLE SWITCHER® Nano Module With 5.5-V Maximum Input Voltage. http://www.ti.com/product/LMZ10501.

[65] Renji Thomas, Naser Sedaghati, and Radu Teodorescu. 2016. EmerGPU: Understanding and mitigating resonance-induced voltage noise in GPU architectures. In *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 79–89.

[66] Tao Tong, Xuan Zhang, Wonyoung Kim, David Brooks, and Gu-Yeon Wei. 2013. A fully integrated battery-connected switched-capacitor 4: 1 voltage regulator with 70% peak efficiency using bottom-plate charge recycling. In *Proceedings of the IEEE 2013 Custom Integrated Circuits Conference*. IEEE, 1–4.

[67] Riccardo Trinchero. 2015. EMI Analysis and Modeling of Switching Circuits. In *PhD thesis*. Politecnico di Torino.

[68] Akihiro Tsukioka, Karthik Srinivasan, Shan Wan, Lang Lin, Ying-Shiun Li, Norman Chang, and Makoto Nagata. 2019. A Fast Side-Channel Leakage Simulation Technique Based on IC Chip Power Modeling. *IEEE Letters on Electromagnetic Compatibility Practice and Applications* 1, 4 (2019), 83–87.

[69] Inna Vaisband and Eby G Friedman. 2012. Heterogeneous methodology for energy efficient distribution of on-chip power supplies. *IEEE Transactions on Power Electronics* 28, 9 (2012), 4267–4280.

[70] Tom Van Breussegem and Michiel Steyaert. 2009. A 82% efficiency 0.5% ripple 16-phase fully integrated capacitive voltage doubler. In *2009 Symposium on VLSI Circuits*. IEEE, 198–199.

[71] Hang-Sheng Wang, Xinping Zhu, Li-Shiuan Peh, and Sharad Malik. 2002. Orion: a power-performance simulator for interconnection networks. In *Microarchitecture, 2002.(MICRO-35). Proceedings. 35th Annual IEEE/ACM International Symposium on*. IEEE, 294–305.

[72] Xuan Wang, Jiang Xu, Zhe Wang, Kevin J Chen, Xiaowen Wu, Zhehui Wang, Peng Yang, and Luan HK Duong. 2015. An analytical study of power delivery systems for many-core processors using on-chip and off-chip voltage regulators. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, 9 (2015), 1401–1414.

[73] Steven JE Wilton and Norman P Jouppi. 1996. CACTI: An enhanced cache access and cycle time model. *IEEE Journal of Solid-State Circuits* 31, 5 (1996), 677–688.

[74] Kan Xu and Eby G Friedman. 2020. Challenges in High Current On-Chip Voltage Stacked Systems. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.

[75] Kan Xu, Ravi Patel, Praveen Raghavan, and Eby G Friedman. 2018. Exploratory design of on-chip power delivery for 14, 10, and 7 nm and beyond FinFET ICs. *Integration* 61 (2018), 11–19.

[76] Guihai Yan, Yingmin Li, Yinhe Han, Xiaowei Li, Minyi Guo, and Xiaoyao Liang. 2012. AgileRegulator: A hybrid voltage regulator scheme redeeming dark silicon for power efficiency in a multicore architecture. In *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*. IEEE, 1–12.

[77] Yuanmao Ye and Ka Wai Eric Cheng. 2016. Analysis and optimization of switched capacitor power conversion circuits with parasitic resistances and inductances. *IEEE Transactions on Power Electronics* 32, 3 (2016), 2018–2028.

[78] Zhiyu Zeng, Suming Lai, and Peng Li. 2013. IC power delivery: Voltage regulation and conversion, system-level cooptimization and technology implications. *ACM Transactions on Design Automation of Electronic Systems (TODAES)* 18, 2 (2013), 1–21.

[79] Zhiyu Zeng, Xiaoji Ye, Zhuo Feng, and Peng Li. 2010. Tradeoff analysis and optimization of power delivery networks with on-chip voltage regulation. In *Proceedings of the 47th Design Automation Conference*. ACM, 831–836.

[80] Xin Zhan, Jianhao Chen, Edgar Sánchez-Sinencio, and Peng Li. 2019. Power Management for Multicore Processors via Heterogeneous Voltage Regulation and Machine Learning Enabled Adaptation. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (2019).

[81] Runjie Zhang. 2015. *Pre-RTL On-Chip Power Delivery Modeling and Analysis*. Ph.D. Dissertation. University of Virginia.

[82] Pingqiang Zhou, Dong Jiao, Chris H Kim, and Sachin S Sapatnekar. 2011. Exploration of on-chip switched-capacitor DC-DC converter for multicore processors using a distributed power delivery network. In *Custom Integrated Circuits Conference (CICC), 2011 IEEE*. IEEE, 1–4.

[83] Huifeng Zhu, Xiaolong Guo, Yier Jin, and Xuan Zhang. 2020. PowerScout: A Security-Oriented Power Delivery Network Modeling Framework for Cross-Domain Side-Channel Analysis. In *2020 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 1–6.

[84] Cheng Zhuo, Kassan Unda, Yiyu Shi, and Wei-Kai Shih. 2018. From layout to system: Early stage power delivery and architecture co-exploration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 38, 7 (2018), 1291–1304.

[85] Brian Zimmer, Yunsup Lee, Alberto Puggelli, Jaehwa Kwak, Ruzica Jevtić, Ben Keller, Steven Bailey, Milovan Blagojević, Pi-Feng Chiu, Hanh-Phuc Le, et al. 2016. A RISC-V vector processor with simultaneous-switching switched-capacitor DC–DC converters in 28 nm FDSOI. *IEEE Journal of Solid-State Circuits* 51, 4 (2016), 930–942.

[86] An Zou, Jingwen Leng, Xin He, Yazhou Zu, Christopher D Gill, Vijay Janapa Reddi, and Xuan Zhang. 2018. Voltage-stacked gpus: A control theory driven cross-layer solution for practical voltage stacking in gpus. In *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 390–402.

[87] An Zou, Jingwen Leng, Xin He, Yazhou Zu, Christopher D Gill, Vijay Janapa Reddi, and Xuan Zhang. 2020. Voltage-Stacked Power Delivery Systems: Reliability, Efficiency, and Power Management. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39, 12 (2020), 5142–5155.

[88] An Zou, Jingwen Leng, Xin He, Yazhou Zu, Vijay Janapa Reddi, and Xuan Zhang. 2018. Efficient and reliable power delivery in voltage-stacked manycore system with hybrid charge-recycling regulators. In *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[89] An Zou, Jingwen Leng, Yazhou Zu, Tao Tong, Vijay Janapa Reddi, David Brooks, Gu-Yeon Wei, and Xuan Zhang. 2017. Ivory: Early-Stage Design Space Exploration Tool for Integrated Voltage Regulators. In *Proceedings of the 54th Annual Design Automation Conference 2017*. ACM, 1.

[90] Yazhou Zu, Charles R. Lefurgy, Jingwen Leng, Matthew Halpern, Michael S. Floyd, and Vijay Janapa Reddi. 2015. Adaptive Guardband Scheduling to Improve System-Level Efficiency of the POWER7+. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.