Ivory: Early-Stage Design Space Exploration Tool for Integrated Voltage Regulators

An Zou[1], Jingwen Leng[2], Yazhou Zu[3], Tao Tong[4],
Vijay Janapa Reddi[3], David Brooks[5], Gu-Yeon Wei[5], Xuan Zhang[1]
[1]Washington University in St. Louis, St. Louis, U.S. [2]Shanghai Jiao Tong University, Shanghai, China.
[3]University of Texas at Austin, Austin, U.S. [4]Kolmostar, Inc., Fremont, U.S. [5]Harvard University, Cambridge, U.S.

ABSTRACT

Despite being employed in burgeoning efforts to improve power delivery efficiency, integrated voltage regulators (IVRs) have yet to be evaluated in a rigorous, systematic, or quantitative manner. To fulfill this need, we present *Ivory*, a high-level design space exploration tool capable of providing accurate conversion efficiency, static performance characteristics, and dynamic transient responses of an IVR-enabled power delivery subsystem (PDS), enabling rapid trade-off exploration at early design stage, approximately 1000x faster than SPICE simulation. We demonstrate and validate *Ivory* with a wide spectrum of IVR topologies. In addition, we present a case study using *Ivory* to reveal the optimal PDS configurations, with underlying power break-downs and area overheads for the GPU manycore architecture, which has yet to embrace IVRs.

1 INTRODUCTION

With the demise of Dennard scaling, power and energy efficiency restrict single thread performance [1] and designers are looking for new ways to deliver power more efficiently to microprocessors. Integrated voltage regulators (IVRs) can enhance supply integrity and enable flexible voltage scaling by moving power conversion closer to the point-of-load. Distributed IVRs (Fig. 1) can deliver per-core, fine-grain, fast dynamic voltage and frequency scaling (DVFS) [2] at a level unattainable with traditional off-chip regulators, and also suppress voltage noise more effectively[3]. Leveraging these benefits improves both performance and efficiency. Also, 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 [4]. As IVR becomes a viable solution for power delivery in modern microprocessors, it is 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 tend to focus on circuit-level implementation to improve conversion efficiency [5]. Real implementation benefits in IVR-enabled 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 resorted to either over-simplified assumptions of IVR efficiency [6–8], overlooking important design considerations such as voltage ripples, or a fixed IVR design covering only a fraction of the entire design space [2].

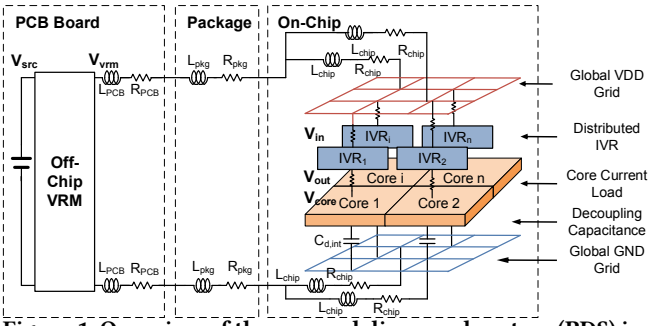


Figure 1: Overview of the power delivery subsystem (PDS) in modern microprocessors with distributed IVRs.

To address these shortcomings, we propose *Ivory* (Fig. 2), a high-level IVR modeling tool for early-stage design space exploration in combination with architecture-level performance and power simulators. *Ivory* captures the complex yet subtle design trade-offs among different IVR typologies to evaluate the performance benefits and implementation costs in full-system settings. It abstracts away the complexities of low-level IVR circuit details to facilitate architects, system engineers, and other experts at the upper levels of the system stack to effectively explore new design spaces that are enabled by such embedded fine-grain voltage regulation capability, similar to what Cacti [9] did for memory systems and ORION [10] for network-on-chip designs. *Ivory* seamlessly incorporates several advanced features that were previously lacking and makes the following key contributions:

- A fast, accurate, and validated (using both SPICE simulations and measured silicon data) parameterized IVR model to estimate conversion efficiency, voltage ripple/droop, and die/board area of multiple IVR topologies in different technology nodes or processes.
- A novel method to derive IVR's dynamic feedback response to fast DVFS and load current changes by combining a *cycle-by-cycle* model together with an *in-cycle* model. This combination facilitates the complete capture of an IVR's dynamic voltage waveform and noise characteristics, given power traces from real-world workloads.
- Comprehensive design explorations covering 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 that are made available with compatible architecture simulator interfaces.
- A case study investigating the optimal power delivery architecture in the manycore GPU architecture, which reveals that a distributed IVR configuration outperforms the conventional off-chip VRM's output delivery efficiency by 9.5% in the 130nm technology node.

DAC '17 June 18-22, 2017, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/3061639.3062268

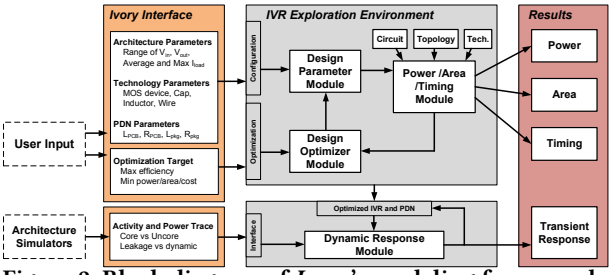


Figure 2: Block diagram of *Ivory*'s modeling framework.

2 BACKGROUND

The benefits of integrated fine-grain voltage regulation [2] have precipitated recent advancements in device fabrication [5, 11], circuit implementation [4, 12], and system integration of IVR [6, 8]. In this section, we review the current state of IVR design and implementation, especially in the context of the entire power delivery system of modern processors.

2.1 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. Linear and switching regulators are two main types, and they differ most notably in the efficiency ranges. The linear regulator's efficiency is determined by the input/output voltage ratio, whereas the switching regulator yields higher efficiency even with a higher conversion ratio.

Switching regulators usually require large discrete passive components such as capacitors and inductors due to lower switching frequencies ($< 10MHz$). Recent technology advancements [5, 11] make it possible for switching regulators to operate at much higher frequencies and to be integrated on the same die as the processors. Buck converters [13] and switched-capacitor converters [4, 5, 14] are two types of topology commonly adopted for such IVRs, in addition to low dropout linear regulators (LDO). While buck converter requires both an inductor and a capacitor, it can sustain a relatively constant conversion efficiency over a wider 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 efficiency of both the switch-capacitor and the buck converter is sensitive to device parameters which 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/process, topology, input/output voltage ratios, and load current levels [2, 3]. Therefore, the findings cannot easily be extended to different use cases. While analytical models of the buck [15] and switched-capacitor converters [16] exist, they primarily focus on modeling individual IVRs as stand-alone blocks, and thus are unable to handle integration with the entire power delivery subsystem.

2.2 Power Delivery Subsystem

A typical PDS can be broken into on-chip and off-chip components, as shown in Fig. 1. The off-chip portion consists of an voltage regulator module (VRM), cascaded power delivery networks (PDNs) at the PCB board level, and the package level, consisting of discrete RLC components. C4 bumps interface the off-chip PDN with the on-chip power grid, consisting of a distributed PDN, IVRs, and processors as the current load. The IVR not only decouples the on-chip power grid and the off-chip network but also provides extra voltage regulation and noise isolation to the digital loads.

In this paper, the power source (V_{src}) supplied to the input of the off-chip VRM is assumed to be ideal. We consider the conversion loss of the off-chip VRM and assume its output voltage (V_{vrm}) is stable and does not experience transient voltage ripples or droops. Given the ample amount of decoupling capacitance near the VRM and local feedback control, these assumptions are accepted for well-designed power delivery systems [2]. We assume V_{in} and V_{out} to be the input and output voltages of the on-chip IVR and V_{core} the voltage delivered to the core. In a typical computing system, the power delivery efficiency, depends on not only the conversion efficiency of the VRM and/or the IVR, but also the extra voltage guardbands inserted between V_{vrm} and V_{in} and between V_{out} and V_{core} for reliable operation. These voltage margins have to be accurately estimated to account for the supply noise caused by the combined effects of load current transients and PDN impedance.

To summarize, given the complexity of IVR and its associated power delivery subsystem, significant low-level understanding is required to navigate the different IVR design options, PDS architectures, and control schemes, making it difficult for system engineers and computer architects without such expertise to effectively explore the hidden opportunities in power delivery subsystem co-design with the microprocessor. We believe that *Ivory*'s ability to accurately abstract all the circuit-level implementation details in an IVR-enabled PDS will provide system architects with an accessible tool to adeptly reap such co-design benefits.

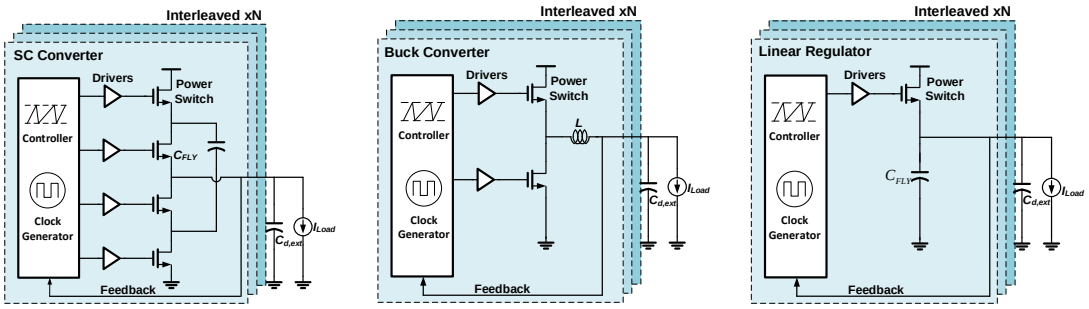
3 MODELING METHODOLOGY

Ivory enables rapid evaluation of an IVR's impact on power delivery efficiency for design exploration in computing systems. Towards this end, it is crucial to capture the two main factors that critically determine the overall power delivery efficiency: 1) the power consumption (loss) of each component in the PDS under static load conditions, and 2) the voltage margins required for the worst-case load transients. Here, we present a detailed description of the modeling framework and methodology employed in *Ivory* to obtain accurate estimates of these parameters.

3.1 Ivory Framework

An overview of *Ivory*'s modeling framework is shown in Fig. 2. Users input high-level parameters, such as the input/output voltage range and maximum load current. 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 130nm down to 10nm, based on ITRS and PTM models [17] as well as surface-mounted-inductor and integrated-inductor data recently published [11, 13]. By default, *Ivory* optimizes for maximum conversion efficiency (to reduce power delivery overhead); *Ivory* also allows users to specify a different optimization target, such as area or supply noise. The internal structure of *Ivory* consists of the followings:

- **System 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.
- **Static design trade-off module**: calculates power consumption, static voltage ripple, timing delay, and die/board area for various building blocks in an IVR, based on design parameters.
- **Dynamic feedback response module**: rapidly models the dynamic voltage waveforms in response to load current transients and/or external DVFS scheduling.
- **Design optimization module**: calculates the design constraints based on the specified technology, architecture configurations



(a) Switched-Capacitor Converter (b) Buck Converter (c) Linear Regulator

Figure 3: Three types of converter topologies.

and basic circuit design guidelines. *Ivory* then employs optimization algorithms to achieve the desired design targets.

Advanced users familiar with IVR design trade-offs can leverage built-in interfaces to specify design parameters directly. Our model considers both the static performance characteristics and the dynamic behaviors of the IVR and applies distinctive modeling strategies, which we elaborate in the remaining sections.

3.2 IVR Static Modeling

By static modeling, we refer to the calculation of the IVR conversion efficiency and voltage ripples based on static assumption of average load conditions and statistics. In contrast, the dynamic modeling described in Section 3.3 deals with an IVR's output voltage feedback response to load current transients from dynamic power traces. In *Ivory*, the static model applies to switched-capacitor converters, buck converters, and linear regulators, which are the most commonly used IVR topologies.

Switched-capacitor converters: Fig. 3(a) illustrates a basic switched-capacitor circuit. *Ivory* adopts the analytical methodology introduced by Seeman [16]. 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 switching (R_{FSL}) limit output impedance. R_{SSL} and R_{FSL} can be expressed as:

$$R_{SSL} = \frac{(\sum_i |a_{c,i}|)^2}{C_{tot} f_{sw}} \qquad R_{FSL} = \frac{(\sum_i |a_{r,i}|)^2}{G_{tot} D_{cyc}} \tag{1}$$

C_{tot} is the total amount of fly capacitance, G_{tot} is the total amount of switch resistance, f_{sw} is the switching frequency, and D_{cyc} is the duty cycle of the switching phase signals in a switched-capacitor IVR. The power loss due to the series of output impedance is $I_{load}^2 \sqrt{R_{SSL}^2 + R_{FSL}^2}$. The loss due to the switch parasitic capacitance, bottom plate parasitic, and gate leakage current from the fly capacitors are calculated to model the total power loss from the switching cells. *Ivory* models 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 [16]. *Ivory's* built-in, analytical formula calculates the charge multiplier vectors for *any* conversion ratio of these two topologies, automating the tedious derivation. Advanced users can plug-in their own switch topology by providing the charge multiplier vectors explicitly.

Buck converters: A typical buck converter is shown in Fig. 3(b). An accepted and validated analytical model that calculates the power loss of buck converters can be found in previous work on off-chip voltage regulators [15]. This model is based on high-side and low-side switch resistance/capacitance, inductor size, parasitic

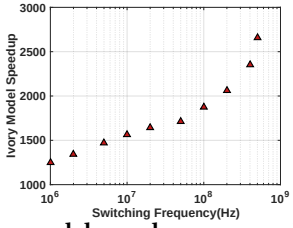


Figure 4: Ivory model speedup compared with SPICE.

resistance, capacitance, switching frequency, and PWM signal duty cycle. *Ivory* extends 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 modeled in *Ivory* by a polynomial-fitted frequency-dependent coefficient of the inductance.

Linear regulators: Analog G_m amplifiers are traditionally used in linear regulators. Recent design trends [18] have increasingly adopted digital comparators and controllers to achieve faster transient responses. Therefore, *Ivory* models linear regulators with a digital feedback path, as illustrated in Fig. 3(c). Since current efficiency close to 99% can usually be achieved by state-of-the-art linear regulator design for moderate load current, the conversion efficiency of a linear regulator 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, digital controller, and clock generator – not to mention the basic capacitor and inductor devices. By commensurately modeling these shared building blocks across all topologies, *Ivory* 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 the transient response analysis and the scalability study of IVR designs, and therefore is taken into account in *Ivory*. We also embed the dynamic and leakage current model of a typical digital logic load to handle DVFS natively—once the maximal load current is specified, the tool will automatically calculate the load current at different voltage and activity levels.

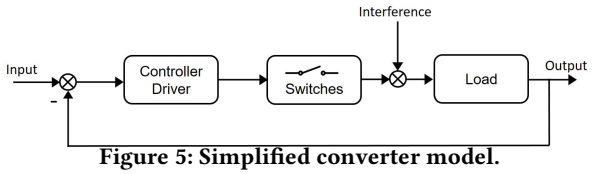


Figure 5: Simplified converter model.

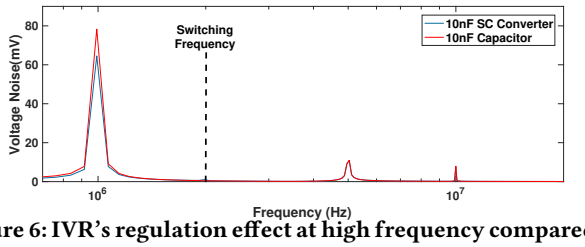


Figure 6: IVR's regulation effect at high frequency compared to a decoupling capacitor.

3.3 IVR Dynamic Modeling

Ivory models the feedback control and voltage dynamic response of the three main types of IVRs to reflect the effects due to fast DVFS and/or load current transients. In such circuit-level analytical modeling of transient response, there is always a conflict between accuracy and speed. In order to balance these two considerations, we propose a method combining a *cycle-by-cycle* model with an *in-cycle* model, whose speed is 1000x faster than SPICE, as shown in Fig. 4. The basic switched-capacitor circuit can be simplified as a simplified model[16], of which the discrete time model can be expressed as

$$V_{out}[k+1] = V_{out}[k] + \frac{1}{C_O}(I_{out}[k]T[k] + (nV_{in} - V_{out}[k])C_{eq}(1 - e^{(-T/2R_{eq}C_{eq})}) \quad (2)$$

The sampling time T is equal to $1/f_{sw}$. Using (2), we can accurately model the dynamic response *cycle-by-cycle*. However, the dynamic response caused by voltage noise from load current variation is usually at a higher frequency than the converter switching frequency f_{sw}. This part of the dynamic response cannot be effectively modeled by the *cycle-by-cycle* model. Thus we present an *in-cycle* model, accounting for high frequency dynamic response. In the *in-cycle* model, only the flying capacitor connected directly to the load is taken into consideration, since it is the only component that has regulation effect at high frequency noise. To demonstrate, we constructed a synthetic voltage noise waveform with representative noise components at 1MHz, 5M and 10MHz and simulated the regulation effect from a 2MHz SC converter with 10nF fly capacitor, in comparison with a single 10nF capacitor. Analyzing their FFT spectrum (Fig. 6), we find that when the voltage noise frequency is equal to or higher than the converter switching frequency, the converter and the capacitor have the same regulation effect, which further proves that the *in-cycle* model can accurately model high frequency voltage noise. Below the switching frequency, the regulating effect of the converter is adequately captured by the *cycle-by-cycle* model. Similar findings have been reported in earlier work on power delivery subsystems with on-chip linear regulators [19].

The generalized model of a typical converter consists of a controller, driver, switches, current load, and feedback, as shown in Fig. 5. The voltage noise can be regarded as the interference to the load. The frequency response of the interference at the output port is

$$V_{out}(j\omega) = \frac{F_L(j\omega) \cdot V_{Noise}(j\omega)}{1 + F_L(j\omega) \cdot F_{Ctl,Dri}(j\omega) \cdot F_{sw}(j\omega)} \quad (3)$$

in which the switches are modeled as Zero-Order Holder.

$$F_{sw}(j\omega) = \frac{1}{j\omega}\left(1 - e^{\frac{-j\omega}{j\omega_{sw}}}\right) \quad (4)$$

When the frequency of the voltage noise ω is higher than the converter switching frequency f_{sw}, the switches frequency response $F_{sw} \approx 0$. Also, the interference frequency response in (3) will be like (5), which demonstrates that the converter does not have the ability to regulate such high frequency noise.

$$V_{out}(j\omega) \approx F_L(j\omega) \cdot V_{Noise}(j\omega) \quad (5)$$

To summarize, the *cycle-by-cycle* model accurately captures the regulation effect from the converter below the switching frequency; meanwhile, the *in-cycle* model, with decoupling effect mainly from the fly capacitor, dictates the dynamic response above the switching frequency. In this way, our *cycle-by-cycle* + *in-cycle* model effectively yields the dynamic response of an IVR's output voltage to fast DVFS and load current transients for the full frequency range.

The dynamic response model for the buck converter and the linear regulator also adopt the *cycle-by-cycle* + *in-cycle* method. For buck converters, our *cycle-by-cycle* model, derived from the Continuous Conduction Mode (CCM) topology, uses discrete transfer function with a feedback controller. Previously, another challenge of modeling a buck converter's dynamic response lies in the treatment of an interleaved circuit architecture. Our *cycle-by-cycle* model takes advantage of the averaging effect in the N-interleaved buck converter and transforms it equivalently to N parallel-connected buck converters for dynamic response derivation.

4 MODEL VALIDATION

This section validates *Ivory*'s modeling accuracy against both SPICE simulation results and measurement data from recent publications, spanning different technology nodes, input/output voltage ranges, and power levels. The *Ivory* dynamic response model is validated under various line regulation, reference regulation, and load regulation scenarios. All these results demonstrate that *Ivory* can faithfully model the design space of realistic voltage regulator configurations. Validation data for the switched-capacitor IVR model is presented in Fig. 7. On the left, *Ivory* is compared against silicon measurements taken from a reconfigurable switched-capacitor implemented in 32nm SOI process [14]. It is clear that *Ivory* adequately models the measured data for the 3:2 and the 2:1 configurations until an efficiency drop occurs past 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 *Ivory* 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 [4]. Regular CMOS capacitors are used for the low-power density design, whereas embedded trench capacitors [5] are used for the high-power density design. The data validates *Ivory's* 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 [13]. On the right data is from our buck design simulated in a 40nm CMOS process. *Ivory* again proves capable of modeling voltage regulator efficiency, validating its internal buck converter modeling framework. Additionally, the analytical buck model used in *Ivory* has previously been validated

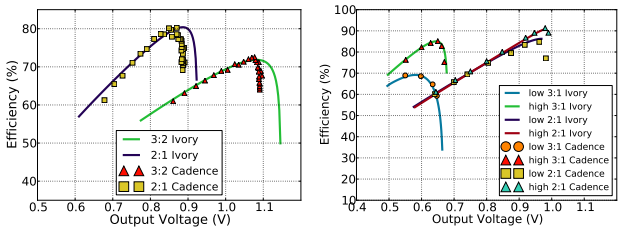


Figure 7: Efficiency validation for SC converters.

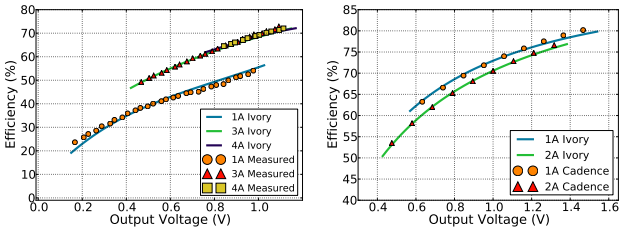
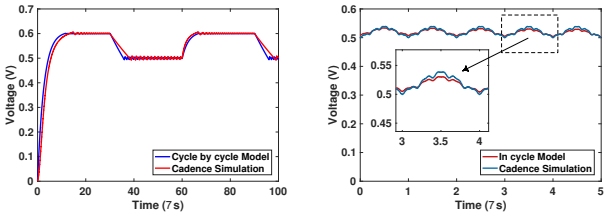


Figure 8: Efficiency validation for buck converters.



(a) Cycle by Cycle (b) in Cycle

Figure 9: Transient voltage response validation between *Ivory* and SPICE simulation.

Table 1: Summary of *Ivory* Input Parameters

Max. Area(mm^2)	200
Total Average Power(W)	20
Input Voltage(V)/Output Voltage(V)	3.3/1
Max Number of Distributed IVRS	4
$R_{sw}(\Omega{\cdot}\mu m^2)$/L($nH/mm^2$)/C($nF/mm^2$)	40/1/10
Off/On-Chip PDN parameter	$R_{off,on}/L_{off,on}$

against off-chip VRMs [15]. For the dynamic model, the comparison of the *cycle-by-cycle* model with SPICE transient waveforms is shown in Fig. 9(a) and the comparison of the *in-cycle* model with SPICE transient waveforms is shown in Fig. 9(b).

5 CASE STUDY ON MANYCORE 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 optimum power delivery subsystem configuration in the context of a GPU-style manycore processor. Our goal is *not* to champion any one particular configuration, rather it is to demonstrate how *Ivory* can be used for design exploration.

5.1 System Configuration

We focus on the comparison between the IVR and conventional off-chip VRM based power delivery system (PDS). We assume an embedded GPU system with four cores (i.e. Streaming Multiprocessors, SMs), although *Ivory* allows an arbitrary number of cores. Each SM adopts the Fermi architecture and has an average power of 5 W. This system uses the same off-chip PDN equivalent circuit as in GPUVolt[20], with a 3.3V supply at the board and a 0.85V SM nominal voltage + 0.15V voltage guardband. The maximum area budget for IVR is 200 mm^2 scaled to be similar to the IVR area in a 4-core Intel CPU with 45 nm technology [21]. The other input parameters to *Ivory* is summarized in Table 1.

5.2 IVR Design Space Exploration

In this study, we set the max efficiency as the optimization target, and use *Ivory* to find the optimal IVR converter design (Fig. 12). We find the buck has higher efficiency than the SC converter with more stringent area budget, 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 concludes the optimal IVR solution shown in Table 2.

Table 2: Summary of Design Space Exploration

Topology	3:1 SC	Buck	LR
Distribute No.	1/2/4	1/2/4	1/2/4
Efficiency(%)	81.3/81.2/81/	80.4/80.2/80	33.2/30.1/30
Ripple(mv)	1.68/1.63/1.56	1.64/1.56/1.25	5.11/4.75/4.16
f_{sw}(MHz)	141/139/137	59/57/56	300/300/300

5.3 Workload-Aware Dynamic Optimization

We find that a 32 interleaved 3:1 switched-capacitor converter has the highest efficiency for this GPU system. We use this converter for dynamic response and power delivery subsystem optimization.

We perform the dynamic response optimization to explore the design space of centralized and distributed IVR design and we compare the results from previous optimization with the conventional off-chip VRM design. The dynamic response analysis optimizes the IVR design through a workload dependent analysis. We integrate *Ivory* with the GPU performance and power simulation infrastructure [22] and use large programs from the CUDA SDK and Rodinia suite. The dynamic analysis in *Ivory* uses the optimal converter design from the static analysis to calculate the voltage noise. Since distributed IVRs can suppress voltage noise more effectively and the max number of distributed IVRs for this on-chip System is 4, *Ivory* allow us to compare the dynamic response of all centralized and distributed IVR configurations.

The voltage statistics of the GPU system running different workloads are shown by box plot in Fig. 10. *Ivory* shows that the design with four distributed IVRs is the optimal solution. Fig. 11 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.

5.4 Putting It Together: Power Eff. Analysis

Ivory lets designers rapidly evaluate the final PDS efficiency through the combined static and dynamic analysis. The static converter design analysis finds the optimal converter with high converter efficiency and low IR-drop loss. *Ivory* optimizes the voltage margin by identifying the IVR design with the minimal voltage noise that accounts for the most of the voltage margin [23]. Fig. 13 shows the breakdown of different overheads for different PDS designs. The power efficiency is the percentage of power consumed by cores that perform the actual computation over total power. The optimal PDS solution by *Ivory* achieves a 9.5% power efficiency improvement over the previous off-chip VRM-based PDS, without any performance loss. A Fast DVFS could yield further improvement and can also be explored using *Ivory*, but detailed evaluation is left for future work.

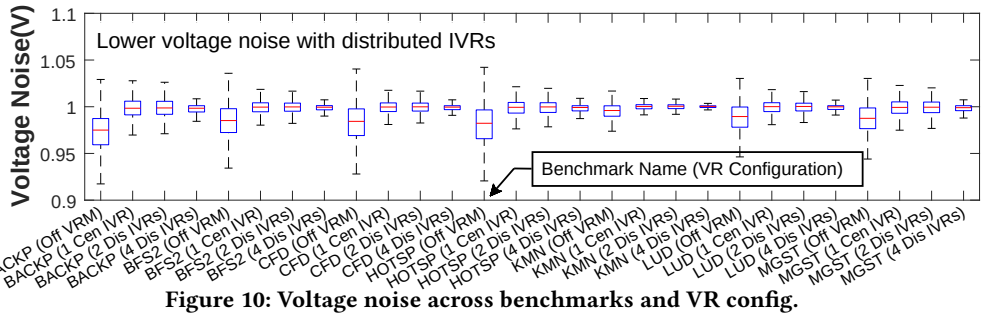


Figure 10: Voltage noise across benchmarks and VR config.

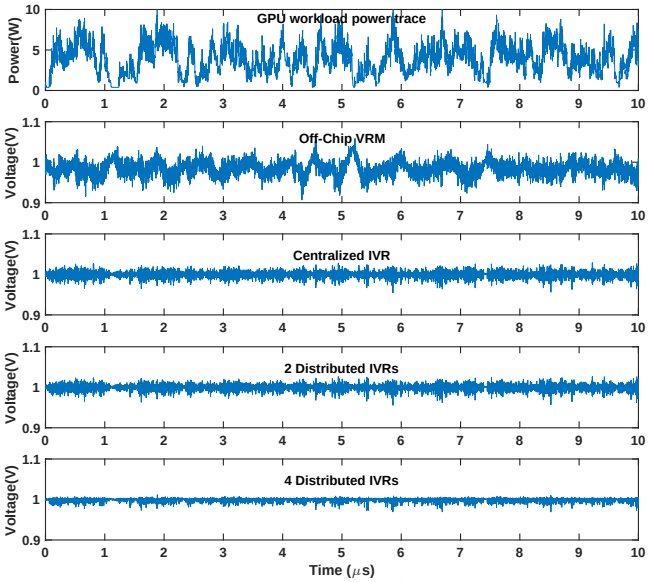


Figure 11: Voltage noise waveforms (CFD) with varying VR configurations.

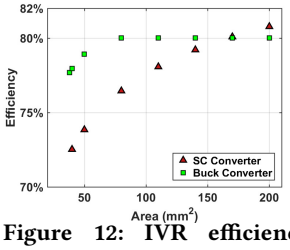


Figure 12: IVR efficiency trade-off with area.

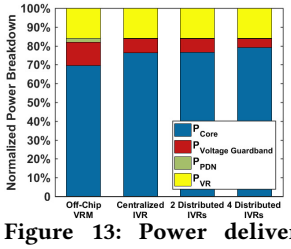


Figure 13: Power delivery system optimization.

6 CONCLUSIONS

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

7 ACKNOWLEDGEMENT

This work was supported in part by NSF 1657562, NSF 1528045 and DARPA HR0011-13-C-0022.

REFERENCES

[1] Hadi Esmaeilzadeh et al. Dark silicon and the end of multicore scaling. In *ACM SIGARCH Computer Architecture News*, volume 39, pages 365–376. ACM, 2011.
[2] Wonyoung Kim et al. System level analysis of fast, per-core dvfs using on-chip switching regulators. In *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*, pages 123–134. IEEE, 2008.
[3] Pingqiang Zhou et al. 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*, pages 1–4. IEEE, 2011.
[4] Tao Tong et al. A fully integrated battery-connected switched-capacitor 4: 1 voltage regulator with 70% peak efficiency using bottom-plate charge recycling. In *Custom Integrated Circuits Conference (CICC), 2013 IEEE*, pages 1–4. IEEE, 2013.
[5] Leland Chang et al. 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*, pages 55–56. IEEE, 2010.
[6] Hamid Reza Ghasemi et al. Cost-effective power delivery to support per-core voltage domains for power-constrained processors. In *Proceedings of the 49th Annual Design Automation Conference*, pages 56–61. ACM, 2012.
[7] Ulya R Karpuzcu et al. Energysmart: Toward energy-efficient manycores for near-threshold computing. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, pages 542–553. IEEE, 2013.
[8] Guihai Yan et al. 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*, pages 1–12. IEEE, 2012.
[9] Steven JE Wilton et al. Cacti: An enhanced cache access and cycle time model. *IEEE Journal of Solid-State Circuits*, 31(5):677–688, 1996.
[10] Hang-Sheng Wang et al. Orion: a power-performance simulator for interconnection networks. In *Microarchitecture, 2002.(MICRO-35). Proceedings. 35th Annual IEEE/ACM International Symposium on*, pages 294–305. IEEE, 2002.
[11] Donald S Gardner et al. Review of on-chip inductor structures with magnetic films. *IEEE Transactions on Magnetics*, 45(10):4760–4766, 2009.
[12] Wonyoung Kim et al. A fully-integrated 3-level dc-dc converter for nanosecond-scale dvfs. *IEEE Journal of Solid-State Circuits*, 47(1):206–219, 2012.
[13] Noah Sturcken et al. A 2.5 d integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer. *IEEE Journal of Solid-State Circuits*, 48(1):244–254, 2013.
[14] Hanh-Phuc Le et al. Design techniques for fully integrated switched-capacitor dc-dc converters. *IEEE Journal of Solid-State Circuits*, 46(9):2120–2131, 2011.
[15] Yongseok Choi et al. Dc–dc converter-aware power management for low-power embedded systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(8):1367–1381, 2007.
[16] Michael D Seeman. Analytical et al. Technical report, DTIC Document, 2006.
[17] Saurabh Sinha et al. Exploring sub-20nm finfet design with predictive technology models. In *Proceedings of the 49th Annual Design Automation Conference*, pages 283–288. ACM, 2012.
[18] Mohammad Al-Shyoukh et al. A transient-enhanced low-quiescent current low-dropout regulator with buffer impedance attenuation. *IEEE journal of solid-state circuits*, 42(8):1732–1742, 2007.
[19] Zhiyu Zeng et al. Tradeoff analysis and optimization of power delivery networks with on-chip voltage regulation. In *Proceedings of the 47th Design Automation Conference*, pages 831–836. ACM, 2010.
[20] Jingwen Leng et al. Gpuvolt: Modeling and characterizing voltage noise in gpu architectures. In *Proceedings of the 2014 international symposium on Low power electronics and design*, pages 141–146. ACM, 2014.
[21] Edward A Burton et al. Fivr-fully integrated voltage regulators on 4th generation intel® core^{TM} socs. In *Applied Power Electronics Conference and Exposition (APEC), 2014 Twenty-Ninth Annual IEEE*, pages 432–439. IEEE, 2014.
[22] Jingwen Leng et al. Gpuwattch: enabling energy optimizations in gpgpus. volume 41, pages 487–498. ACM, 2013.
[23] Jingwen Leng et al. Safe limits on voltage reduction efficiency in gpus: a direct measurement approach. In *Proceedings of the 48th International Symposium on Microarchitecture*, pages 294–307. ACM, 2015.