Multi-Story Power Distribution Networks for GPUs

Qixiang Zhang*, Liangzhen Lai, Mark Gottscho, and Puneet Gupta
zhangqixiang@outlook.com.cn, {liangzhen, mgottscho}@ucla.edu, puneet@ee.ucla.edu
Department of Electrical Engineering, University of California, Los Angeles, USA
*College of Electrical Engineering, Zhejiang University, China

***Abstract*—High-performance chips require many power pins to support large currents, which increases fabrication cost, limits scalability, and degrades power efficiency. Multi-story serial power distribution networks (PDNs) are a promising approach to reducing pin counts and power losses. We study the feasibility of 2-story PDNs for graphics processing units (GPUs). These PDNs use either an auxiliary off-chip regulator or integrated on-die supercapacitors to stabilize the virtual rail voltage. Static SIMT thread scheduling (SSTS) and dynamic current compensation (DCC) can reduce transient impedance mismatch when the auxiliary regulator is omitted. Simulation results show that compared to a traditional 1-story design, our 2-story GPU architectures can reduce the required number of core power pins by up to 2X, power losses in the PDN by up to 3.6X, and/or maximum voltage swing by up to 2X without any performance degradation. Our results demonstrate the efficiency and cost advantages of multi-story PDNs for GPUs without any impact on performance.**

I. Introduction

High-performance chips such as graphics processing units (GPUs) suffer from limited availability of signal I/O pins. Many pins are dedicated to power delivery, which stabilize supply (VDD) and ground voltages against quickly-changing and high-magnitude currents. This limits the performance scalability of high-power designs and incurs significant manufacturing and packaging costs to support high pin counts. A primary cause of this problem is the conventional *single-story power distribution network (PDN)*. Logic macros are connected in parallel to low-voltage power domains while sharing a common ground plane; this results in high chip currents.

One solution is to use a *multi-story PDN* [1], where separate voltage domains do not share the same ground plane. Instead, they are stacked in series, forming *virtual supply* and *virtual ground* reference voltages for each domain through voltage division. The *real supply* voltage is a multiple of the required logic level, which reduces the drawn chip current proportionally and lowers pin counts, power losses from IR drop [2], and $L\frac{di}{dt}$ noise. However, to ensure functional correctness, power consumption must be evenly distributed across all of the stories at all times to ensure the voltage divider maintains stable virtual supply and ground references. This is challenging to guarantee in typical systems-on-chip (SoCs) that are comprised of a mix of heterogeneous functional units, making multi-story PDNs impractical for most chip designs. But certain classes of systems such as GPUs could have sufficient design homogeneity to use a multi-story PDN effectively.

The regularity of a GPU would result in evenly-distributed power consumption across its single-instruction, multiple-thread (SIMT) cores, which comprise the majority of chip area. Massively-parallel applications divide their work evenly across individual threads such that SIMT cores behave similarly during runtime. Cores are laid out in using a tiled floorplan, making it relatively straightforward to partition them into separate stories. Fig. 1 shows how a GPU's SIMT cores could be divided into N stories given an input VDD that is approximately N times the required VDD_{core} logic level. Yet even in a GPU, small differences in core-to-core power that arise from minor workload variation could result in unstable virtual rails.

In this paper, we propose to use multi-story PDNs for GPUs, focusing on the simplest case with two stories. We explore two specific and complementary techniques to reduce dynamic imbalances in power demand between the stories. The first approach is to include an low-overhead off-chip *auxiliary regulator* in addition to the *primary regulator*. The auxiliary regulator would have low area, pin, and power overheads, because it only needs to deliver a small differential current to stabilize the virtual rail. The second approach is to add integrated *on-die supercapacitors* enabled by recent advances [3]–[7]. Our contributions include the following.

- We propose 2-story PDN architectures for GPUs and evaluate their feasibility with general-purpose (GPGPU) workloads. Our 2-story PDN with auxiliary regulator can reduce the required number of core power pads by 1.9X, maximum voltage swing (MVS) by 2X, and PDN power losses by 2X without any impact on performance compared to a conventional 1-story PDN.
- We explore technology, hardware, and software techniques to improve our two-story PDN. By eliminating the auxiliary regulator and adding on-chip supercapacitors and *dynamic current compensation (DCC)* circuits, we can reduce core power pin requirements by 2X and MVS by 2X for approximately the same power losses as the 1-story design, also without any impact on performance. Effective use of *Static SIMT thread scheduling (SSTS)* can also improve MVS by up to 2X.

To the best of our knowledge, this is the first work that studies multi-story power architectures for GPUs. The rest of the paper is organized as follows. In Sec. II, we briefly describe related work and state how our work advances the state-of-the-art. In Sec. III, we discuss the details of our proposed two-story PDNs for GPUs. In Secs. IV and V, we outline our experimental setup and present the results. We conclude the paper in Sec. VI with suggestions for future work.

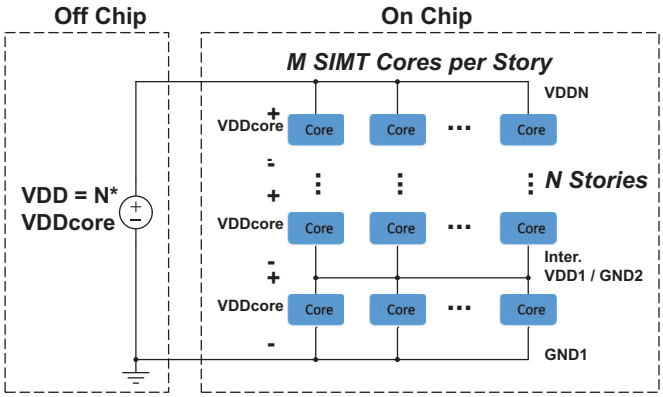


Fig. 1. A multi-story power distribution network reduces the total GPU current by grouping cores into stacked voltage domains. Intermediate virtual rails are formed through balanced voltage division that arises from the inherent architecture and workload regularity of the GPU.

978-3-9815370-6-2/DATE16/

II. Related Work

There are many approaches to managing power consumption of GPUs which use conventional 1-story PDNs. The most popular methods include power and clock gating of idle resources and dynamic voltage/frequency scaling (DVFS) for active resources [8]–[10]. However, none of these approaches can solve the power pin problem that emerges in high-performance designs.

The multi-story power architecture was initially proposed by Gu et al. [1]. This work was followed up by several other researchers [11]–[14]. In [1], the authors studied how to reduce power supply noise in very small circuits (e.g., shift registers), evaluated using a simple benchmark. The energy-saving benefits of the multi-story architecture on some specific systems has also been studied [15]–[18]. These prior works show that multi-story architectures can significantly reduce power losses for different chips, but none have considered their use in GPUs.

We address several shortcomings of prior research. Specifically, no research has focused on multi-story PDNs for GPUs, although they seem especially well-suited for it. Previous work has also not considered the impact on power pin requirements. Reducing pin count can lead to an improvement in fabrication cost. Conventional power saving techniques could be used with our multi-story PDN, as long as each story in the multi-story architecture receives balanced dynamic power-saving treatment.

III. Proposed Multi-Story GPU Power Architecture

GPUs are well-suited to multi-story PDNs because of their architectural regularity. For a SoC with heterogeneous functional units, in order to use a multi-story PDN, the designer must statically partition different functional units among the *stories* such that their time-varying current demands are closely matched at runtime for a variety of workloads. In contrast, at a high level, the functional units in a GPU are largely homogeneous and laid out in a regular fashion. Thus, aside from within-die process variation and aging effects, logic can be more easily partitioned into multiple stories.

There are many ways to partition a GPU into voltage stories using various levels of granularity. A fine-grain approach would grant more flexibility in matching power consumption between different stories, but could incur higher design overhead and circuit complexity. For example, a fine-grained approach might stripe adjacent processing elements (PEs, known as CUDA cores or streaming processors in NVIDIA parlance) within a SIMT core (streaming multiprocessor) into adjacent stories that span all SIMT cores. All PEs within each SIMT core execute the same instruction sequence in lockstep, so this method avoids imbalances between fetch and decode units across SIMT cores. However, this fine-grained approach would require voltage level-converters at the buses for each of the PEs, potentially having a large impact on performance, power, and area. A viable partitioning strategy, in addition to ensuring balanced current demands, should also minimize the implementation overheads like the number of voltage level-converters and power routing complexity.

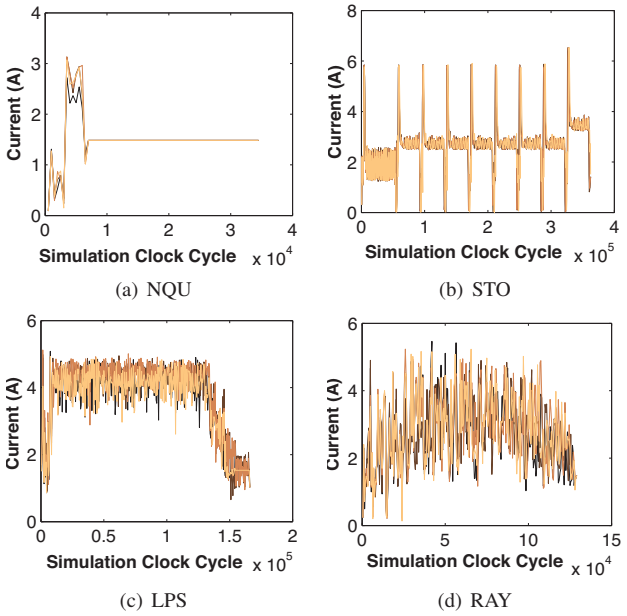
We propose a coarse-grain partitioning approach that assigns groups of SIMT cores to separate stories. This minimizes the level converter cost – as they are required only for the interconnection network between the cores and memory – while still satisfying the balanced-current requirement. Shared uncore components such as the on-chip L2 cache, interconnection networks, and chip I/O are still powered conventionally using separate 1-story power rails. This helps ensure power demand is homogeneous across stories in the 2-story PDN. Because the majority of chip power is consumed by SIMT cores in GPGPU workloads (see Table I), placing uncore components on the 1-story PDN would not have a major effect on required power pin count or losses. Moreover, memory and I/O are often on separate voltage domains anyway. Thus, throughout this paper, when we refer to 2-story PDNs, we only include the GPU SIMT cores.

There are several advantages to our 2-story network architecture. The aggregate chip current demand decreases by approximately one half compared to the traditional 1-story network. While the power consumed by each core is approximately the same as the 1-story network, *resistive losses in the power network* are reduced roughly by 4X because $P = I^2R$. Due to reduced current demand, one half of the number of power pins are required for all SIMT cores, reducing chip, package, and board costs.

To verify that GPUs actually exhibit closely-balanced current demands across SIMT cores at runtime, we collected traces from simulation of four GPGPU workloads on an NVIDIA GTX 480, shown in Fig. 2 (the experimental setup is described later in Sec. IV). For simplicity, it is assumed that the supply voltage of each core is fixed at 1 V. It is clear that for each workload, the cores exhibit similar power profiles over time. We quantify these similarities in Fig. 2(e), which shows that the average current of each core is

TABLE I
Simulated GPGPU benchmarks show that most of chip power is consumed by SIMT cores.

Benchmark	Description	GPU Cycles	Avg. Power (W) Cores	NoC	Total
NQU	N-Queen solv.	34500	24.25	0.06	34.37
STO	StoreGPU	362500	42.65	0.75	52.5
LPS	Laplace disc.	166000	70.47	7.86	107.01
RAY	Ray tracing	129000	46.15	2.58	61.92



(a) NQU (b) STO (c) LPS (d) RAY

Benchmark	Core 1 Avg.	Core 1 MSE	Core 2 Avg.	Core 2 MSE	Core 3 Avg.	Core 3 MSE	Core 4 Avg.	Core 4 MSE
NQU	1.48	0.001	1.51	0.002	1.52	0.002	1.51	0.001
STO	2.65	0.072	2.65	0.116	2.65	0.072	2.65	0.355
LPS	3.61	0.112	3.88	0.089	3.88	0.108	3.63	0.103
RAY	2.72	0.216	2.72	0.483	2.82	0.284	2.82	0.260

(e) Average current and mean squared error (MSE)

Fig. 2. Current profiles of four distinct SIMT cores demonstrate the inherent power symmetry of the GPU architecture for each of the four simulated GPGPU benchmarks.

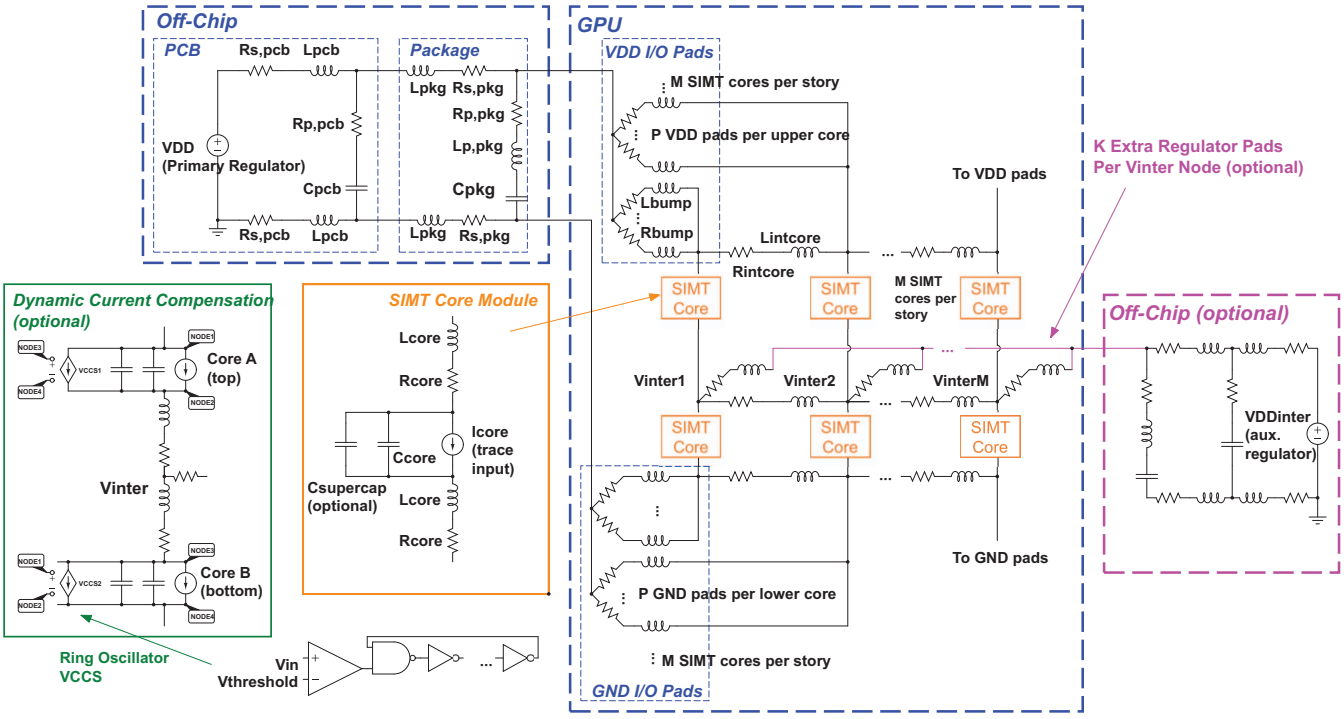


Fig. 3. Circuit model for 2-story GPU PDNs.

nearly the same, and that the average mean squared error (MSE) of their time-sliced currents with respect to the time-sliced average current is small. Nevertheless, as a consequence of our coarse-grain approach, SIMT thread (warp) scheduling could cause differences in the power consumption of distinct cores. To overcome this, some form of voltage regulation must be used on the intermediate virtual rail. We cover the circuit models and methods for stabilizing the rail in the rest of the section.

A. Circuit Models for Power Distribution Networks

We now describe several variants of a circuit model that captures the effects of different PDNs on energy loss and logic-level voltage noise in a GPU. We consider two general types of PDN: (1) the traditional 1-story, 1-regulator design, which we adopt as our baseline, and (2) a 2-story design with several different techniques to stabilize the intermediate virtual rail voltage, V_{inter}. We consider two alternative designs for the 2-story version: (2a) 1-regulator or (2b) 2-regulator. In the latter, an auxiliary off-chip voltage regulator is used to help stabilize V_{inter}. We also explore the application of on-chip integrated supercapacitors [3] in both 2-story designs. Supercapacitors can help stabilize V_{inter} against rapid voltage fluctuations while reducing the need for the auxiliary off-chip voltage regulator.

We base our PDN circuit models on those from [19]. Although the exact circuit model is GPU implementation-dependent, the approach is still representative. Our overall circuit model is shown in Fig. 3, which we will refer to throughout the paper. The off-chip network captures impedances from the PCB and package as a ladder RLC network. These elements lump parasitics and discrete RLC components together. The power pads are represented as parallel RL branches (R_{bump} and L_{bump}) that connect the off-chip network to different locations on the grid. The grid models individual GPU SIMT cores and RL branches for parasitic impedance between adjacent cores ($R_{intcore}$ and $L_{intcore}$). The individual GPU cores are abstracted as ideal current sources (I_{core}) in parallel with parasitic capacitors (C_{core}) and in series with resistive and inductive parasitics (R_{core} and L_{core}). The time-varying current profile I_{core} is controlled by the traces collected from an architectural simulator that executes an application. We discuss this aspect later in Sec. IV.

SIMT cores are evenly partitioned such that there are M cores per story. In this paper, we use a simulation model from an NVIDIA GTX 480, which has 15 SIMT cores. We omit one core so that the remaining 14 cores can be evenly partitioned into two stories. For our 1-story, 1-regulator baseline network, $N = 1$ and $M = 14$. Similarly, for our proposed 2-story networks, $N = 2$ and $M = 7$. Power and ground pads are evenly divided among SIMT cores in all variations of the network topology such that each core in the top story has P real VDD pads, and each core in the bottom story has P real ground pads. The intermediate virtual rail is generally referred to as V_{inter}, although we actually model it as a set of distributed nodes $V_{inter,i}$ where $1 \leq i \leq M$ in order to capture the RL parasitics (as shown in Fig. 3). For the 1-story network, the real supply voltage would be $VDD = VDD_{core} = 1$ V, while in the 2-story networks, the real $VDD = 2VDD_{core} = 2$ V.

The 2-Story, 2-Regulator Network. We now discuss the simplest solution to stabilize the intermediate virtual rail V_{inter} in a 2-story PDN. Thus, one might consider adding an additional auxiliary off-chip low power regulator with voltage $VDD_{inter} = VDD_{core} = 1$ V to control $V_{inter,i}$. This addition is shown on the right side of Fig. 3 in pink. K extra pins per $V_{inter,i}$ node are used to connect to the auxiliary regulator. Although we use identical circuit models and parameters for the two regulators for simplicity, the auxiliary regulator would likely be designed differently; it only needs to supply a small amount of current to stabilize V_{inter}. It may also be possible to integrate the regulator on-die due to the low current requirement. Although this network has additional small pin, package, and board overheads as well as some extra power losses, it can generally stabilize V_{inter} effectively, while still reducing overall overheads and losses compared to a conventional 1-story PDN.

The auxiliary regulator can be augmented with on-chip supercapacitors ($C_{supercap}$) connected in parallel with each SIMT core.

These can more effectively smooth local and/or fast-moving voltage fluctuations while reducing the burden on the auxiliary regulator. Recent work has demonstrated 23 pF/μm^2 [3]. To obtain 20 μF per $V_{inter,i}$ node, this would require at worst approximately 6 mm^2 of chip area ($\approx$ 1% of overall GTX 480 die area). Supercapacitors can be stacked on top of logic and metal layers, so the actual area impact would be far less. On-chip supercapacitors is an active area of research, and density is expected to continue increasing [6]. We explore the regulator/supercapacitor design space later in our evaluation by varying the values of $C_{supercap}$ and K.

The 2-Story, 1-Regulator Network. It may be desirable to omit the auxiliary regulator entirely to save pins, board overheads, and further reduce power losses. However, the burden of stabilizing V_{inter} falls completely on the now-required on-chip supercapacitors. Such a 2-story, 1-regulator network is represented in Fig. 3 if the optional auxiliary off-chip regulator and its K extra per-core power pads (shown in pink) are omitted. However, $C_{supercap}$ would need to be considerably larger than was needed for the 2-story, 2-regulator scenario.

B. Runtime Techniques for Stabilizing the Intermediate Rail

We propose two "runtime" methods to improve the stability of the virtual supply rail in both 2-story network designs: (1) *static SIMT thread scheduling (SSTS)* and (2) *dynamic current compensation (DCC)*. These techniques can temper the voltage fluctuations caused by variability in per-core current demand, reducing the design-time regulator and/or supercapacitor requirements.

Static SIMT Thread Scheduling (SSTS). A compile-time or runtime thread scheduler could assign an application's SIMT threads (warps) to specific cores in the GPU to reduce variability in V_{inter} at runtime. Even with the presence of the auxiliary regulator and/or on-chip supercapacitors, there is still the possibility that certain types of application behaviors might upset the stability of the GPU.

We model the potential benefits of *static SIMT thread scheduling (SSTS)* that would execute at compile time or runtime, depending on the availability of detailed profiling information. This is done with a greedy offline partitioning algorithm that estimates the potential improvement in MVS from a software scheduler. It attempts to best assign warps to SIMT cores with prior knowledge of each warp's power trace, where the objective is to reduce the overall maximum voltage swing (MVS) on V_{inter}. Our offline method is similar to the well-known Fiduccia-Mattheyses (FM) [20] algorithm commonly used for VLSI logic partitioning. The algorithm can be divided into four main steps.

Dynamic Current Compensation (DCC). In addition to a good SIMT thread scheduler, on-chip supercapacitors, and an optional auxiliary regulator, it may still be helpful to have another hardware-based stabilizing method. We propose *dynamic current compensation (DCC)* as a technique to reactively balance top/bottom story current demand at runtime.

DCC are essentially dummy voltage-controlled current sources (VCCSes) accompanying each SIMT core. They can be controlled in hardware at runtime to compensate for small mismatches in power consumption between directly adjacent cores on separate stories. Fig. 3 illustrates the arrangement of DCC circuitry (shown in green). In general, these dummy current sources could be fully-adaptive analog circuits. We model them as digital ring oscillator (RO) circuits, shown in the bottom left of the figure. Each RO can be enabled or disabled according to a control voltage. The output current is a step function of the local logic supply voltage $V_{in} = V_{core,opposite}$ of the adjacent core in the opposite story. When V_{in} falls below $V_{threshold} = 0.95$ V, DCC is turned on and the RO draws a fixed current to compensate for the power mismatch between cores and stabilize V_{inter}.

Voltage oscillation could occur due to the presence of a feedback loop between adjacent VCCSes. This can be avoided with sufficiently low control latency. A tunable delay element models control latency of the RO (not shown in the figure). We evaluate the effect of DCC as a function of current and latency values in Sec. V-C.

C. Practical Considerations

There are some practical considerations that have to be addressed for the proposed 2-story PDN architecture to work. Our PDN architecture requires multiple virtual ground planes on the same silicon die. This could be achieved by a triple-well or moat isolation process [21] between the two stories. At boot time, the voltage at the intermediate rail should be carefully controlled so that the voltage on any of the stories does not exceed the limit for gate oxide break-down. This can be achieved by supplying the intermediate voltage first and/or slowly ramping-up the primary off-chip voltage.

Process variations and aging could cause power mismatch between stories. A software thread scheduler may need to have prior knowledge about the chip variation signature to optimally assign threads to cores. Both of these issues can be mitigated with the auxiliary regulator, on-die supercapacitors, and DCC technique. The life expectancy of supercapacitors may be limited. However, our 2-story PDNs are designed such that the intermediate voltage should be DC-stable. This would keep the amount of charge on supercapacitors relatively stable, increasing their expected lifetime [22].

IV. Experimental Setup

We used GPGPU-Sim [23], [24], which offers a detailed and precise simulation model of a modern GPU's performance and power. The simulator was used to obtain per-core current traces from the NVIDIA GTX 480, running four distinct GPGPU benchmarks: NQU, STO, LPS, and RAY. Basic benchmark information is shown earlier in Table I. We configured GPGPU-Sim to report power at a component-level (cores, uncore, memory) and assumed that the chip had an ideal 1 V supply voltage, allowing us to obtain the current traces shown earlier in Fig. 2. Each core was treated as an independent current source in the circuit model that was simulated offline using SPICE.

Our circuit parameters are based on those from [19]. Table II lists the values for circuit elements in each of our simulated networks. We use the same values for the off-chip part of the PDN (i.e., the board and package). The on-chip PDN parameters were derived by assuming that each SIMT core is equivalent to 3x3 grids in the on-chip PDN in [19]. A lumped model is used for each SIMT core, instead of a distributed one as in [19]: our C_{core} is 9X, while R_{core}, $R_{intcore}$, L_{core}, and $L_{intcore}$ are all 1/9X the respective values from [19].

V. Simulation Results

We analyzed the SPICE simulation results for maximum voltage swing (MVS) of core supply voltages, overall power loss in the PDN, and the required number of power pins for SIMT cores. We discuss the baseline results without supercapacitors, SSTS, or DCC first. All presented results are for the SIMT cores and omit the other power components, which are not expected to be significantly affected by our architecture; they are on separate power domains.

TABLE II
Our parameters for simulated PDNs are based on [19].

Resistance (mΩ)		**Inductance** (pH)		**Capacitance** (μF)	
$R_{\text{s,pcb}}$	0.094	L_{pcb}	21	C_{pcb}	240
$R_{\text{p,pcb}}$	0.166				
$R_{\text{s,pkg}}$	1	L_{pkg}	120	C_{pkg}	26
$R_{\text{p,pkg}}$	0.542	$L_{\text{p,pkg}}$	5.61		
R_{bump}	40	L_{bump}	72		
R_{core}	0.011	L_{core}	0.0111	C_{core}	3
R_{intcore}	0.011	L_{intcore}	0.0111		

TABLE III
BASELINE SIMULATION RESULTS WITHOUT ON-CHIP SUPERCAPACITORS, SSTS, OR DCC. THESE TECHNIQUES ARE ESSENTIAL FOR THE 2-STORY, 1-REGULATOR DESIGN DUE TO HIGH MVS.

Core PDN	Pwr. Pins	MVS (%)				Avg. Power Loss (%)			
		RAY	LPS	NQU	STO	RAY	LPS	NQU	STO
1-story 1-reg	420	11.23	15.38	5.25	10.55	10.16	12.25	11.73	9.14
2-story 2-reg	210+14	5.68	7.85	2.82	4.93	5.27	6.13	5.54	4.82
2-story 1-reg	210	62.29	79.36	25.86	155.10	3.21	3.41	3.20	3.14

A. Baseline 2-Story Networks

Both 2-story networks significantly reduce the required number of power pads. As shown in Table III, the 2-story, 1-regulator network requires half the number of pins of the traditional 1-story, 1-regulator architecture. In the 2-story, 2-regulator network, the addition of some pins for the auxiliary regulator results in a overall core pin reduction of 46% instead of 50%. Thus, both 2-story GPU designs would likely reduce packaging cost significantly, perhaps as much as 25%, as it scales linearly with die area and pin count [25].

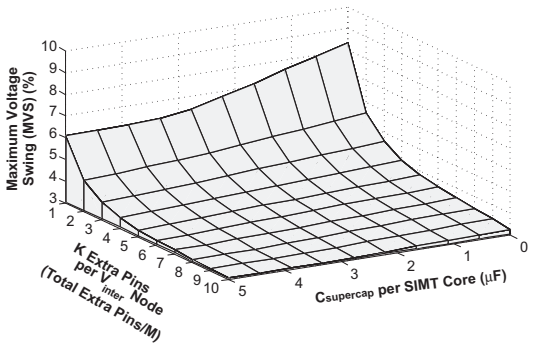
The total resistance of the core PDN is inversely proportional to the number of power pads, which impacts MVS through IR drop and $L\frac{di}{dt}$ noise. For the 420-pad 1-story GPU, MVS is no more than 15.38% across all four benchmarks, which we consider an acceptable value. In the 2-story, 2-regulator network, the auxiliary regulator keeps the intermediate rail stable, with a MVS of no more than 7.85%. However, the MVS of the *baseline* 2-story, 1-regulator design is unacceptably high, ranging from 25.86% on the NQU benchmark to 155.10% for STO. This is because V_{inter} is formed by a noisy voltage divider. Without any stabilization mechanism, it fluctuates wildly according to the transient variations in each story's power consumption.

2-story PDNs bring improvements to energy efficiency. The baseline 2-story 2-regulator network improves power loss (5.27%) compared with the 1-story 1-regulator design (10.16%). This comes at the cost, however, of $K = 2$ extra pins per V_{inter}, i node, as well as an extra voltage regulator on the board. But because the maximum transient current through the auxiliary regulator did not exceed 20% of the average total current drawn by all cores, it can be optimized for small loads, improving its power efficiency and area. The 2-story, 1-regulator design has the best and most consistent power efficiency (as little as 3.14% loss) due to the lack of any auxiliary regulator.

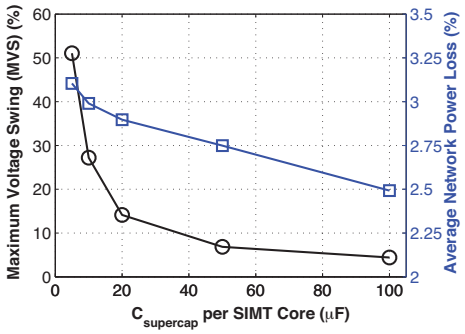
B. 2-Story PDNs with On-Chip Supercapacitors

Integrated on-chip supercapacitors can further reduce MVS and improve power efficiency for 2-story networks. Fig. 4(a) depicts the MVS for the 2-story, 2-regulator network as a function of added per-core capacitance and the number K of extra auxiliary power pins per bottom core. The figure indicates that supercapacitors are most useful when the number of auxiliary regulator pins is low.

Supercapacitors can make the 2-story, 1-regulator scenario viable by addressing its baseline MVS weakness. Fig. 4(b) shows the impact of supercapacitor sizing on MVS and power loss. To achieve 10% MVS, a 38 μF capacitor is required for each core, corresponding to a worst-case of 1.65 mm^2 per core (pessimistically assuming that supercapacitors have a density of 23 pF/μm^2 and that they are not able to be stacked on top of metal and logic). This would result in a 4.6% die area overhead on the GTX 480 (also pessimistically assuming that the 50% reduction in core power pin count afforded by the 2-story, 1-regulator design does not save any area). To approach the $\approx 5\%$ MVS of the 2-story, 2-regulator baseline PDN, each core requires $C_{supercap} = 80\mu$F, which may be too costly. Meanwhile, the average network power loss reduces roughly linearly for $C_{supercap} > 5\mu$F. This is because the supercapacitors buffer energy nearby that



(a) 2-story, 2-regulator with supercapacitors



(b) 2-story, 1-regulator with supercapacitors

Fig. 4. Supercapacitors are most effective when the auxiliary voltage regulator is weakly connected or removed entirely. DCC and SSTS are not used here. The benchmark is RAY.

can be used to mitigate the difference in current demand between stories without passing through too many resistive wires.

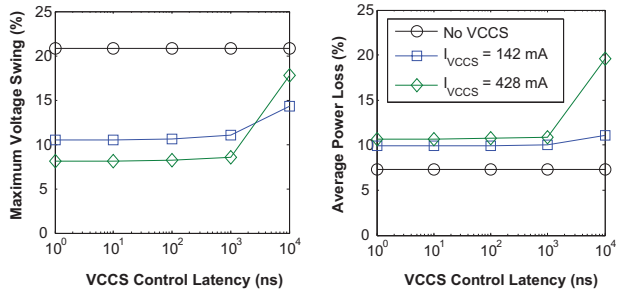
C. SSTS and DCC for 2-Story, 1-Regulator PDNs

Finally, to understand the potential benefits from a good SIMT thread scheduler, we simulated the 2-story, 1-regulator network with $C_{supercap} = 10\mu$F where SIMT threads were statically mapped to cores according to our SSTS algorithm. The results are shown in Table IV; all benchmarks demonstrate a significant drop in MVS, with three out of four matching or bettering the baseline 1-story PDN. The worst-case total area overhead for this supercapacitor configuration is just 6.52 mm^2, or about 1% of total die area, once again pessimistically assuming total chip area is not reduced by the 50% core power pin savings. This result makes the 2-story, 1-regulator PDN better than the baseline results in Table III would suggest. Moreover, an effective thread scheduler can significantly reduce design overheads while ensuring V_{inter} is stable.

We studied DCC separately from SSTS to evaluate its effectiveness in the 2-story, 1-regulator network with supercapacitors. The VCCS control latency is an important parameter that can affect the stability of V_{inter}. The results are shown in Fig. 5 for $C_{supercap} = 8\mu$F. When the VCCS latency is less than 1 μs, both MVS and power loss remain nearly constant compared to the ideal zero-latency case. MVS improves by 50% when $I_{VCCS} = 1/M$ A $= 142$ mA, and by 61% when $I_{VCCS} = 3/M$ A $= 428$ mA. Power loss, however, increases moderately because of the extra current drawn by the VCCSes, but still remains under 4.5% if the VCCS latency is under 100 ns. This is less than half of the power loss in our baseline 1-story, 1-regulator network. For more than 1 μs VCCS latency, both MVS and power loss begin to suffer, with a higher I_{VCCS} exacerbating the problem. This is because the compensation currents from VCCSes lag

TABLE IV
A COMPARISON OF BASELINE AND SSTS THREAD SCHEDULING ON MVS FOR THE 2-STORY, 1-REGULATOR NETWORK WITH $C_{supercap} = 10\mu$F AND NO DCC SHOWS A CONSIDERABLE OPPORTUNITY FOR IMPROVEMENT OF VOLTAGE STABILITY THROUGH SOFTWARE.

	RAY	LPS	NQU	STO
Baseline	27.3%	17.4%	4.2%	20.4%
SSTS	11.6%	8.7%	3.9%	14.5%



(a) MVS (b) Average Power Loss

Fig. 5. VCCS control latency should remain below 1 μs for DCC to be effective on the 2-story, 1-regulator network. Here, $C_{supercap} = 8\mu$F and the benchmark is LPS. No SSTS is used.

behind the time-varying control voltage, and can induce oscillations, decreases V_{inter} stability and wasting power.

VI. CONCLUSION

In this paper, we proposed novel multi-story power delivery networks (PDNs) for GPUs. Our key insight was that GPUs are well-suited to multi-story PDNs because of their architecture and workload regularity. We explored two primary variants of a 2-story PDN. The first included an auxiliary off-chip voltage regulator to stabilize the virtual supply voltage. The second omitted the auxiliary regulator in favor of integrated on-die supercapacitors. To improve the runtime chip stability, we used static SIMT thread scheduling (SSTS) and dynamic current compensation (DCC). Our results showed that 2-story PDNs on GPUs could reduce the total core power pin requirement by up to 2X, power losses by up to 3.6X, and/or maximum voltage swing by up to 2X without affecting application performance.

There are several promising directions for future work. Development of on-chip integrated supercapacitors is rapidly progressing and could have far-reaching benefits beyond those outlined in this paper. The design of practical multi-story PDNs will need to support power gating, clock gating, dynamic voltage/frequency scaling (DVFS); these popular power management techniques would be challenging to implement correctly. Process variability, aging, and poorly-balanced or malicious applications could hamper the reliability of multi-story GPUs. We look forward to future research developments in this exciting area.

ACKNOWLEDGMENT

The authors thank the reviewers for their constructive feedback. Mr. Gottscho thanks Dr. Dana Watson and Maria Yefimova at UCLA for their writing suggestions. This work was supported by the US National Science Foundation (NSF) Grant No. CCF-1029030. Qixiang Zhang contributed to this work while a summer intern at the UCLA NanoCAD Lab. Dr. Liangzhen Lai contributed to this work while a Ph.D. candidate at the UCLA NanoCAD Lab; he is now at ARM Research.

REFERENCES

[1] J. Gu *et al.*, "Multi-story power delivery for supply noise reduction and low voltage operation," in *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2005.
[2] T. Yu and M. D. Wong, "A novel and efficient method for power pad placement optimization," in *IEEE International Symposium on Quality Electronic Design (ISQED)*, 2013.
[3] M. F. El-Kady and R. B. Kaner, "Scalable fabrication of high-power graphene micro-supercapacitors for flexible and on-chip energy storage," *Nature Communications*, vol. 4, 2013.
[4] L. Smith *et al.*, "Scaled carbon-ionogel supercapacitors for electronic circuits," in *IEEE National Aerospace and Electronics Conference (NAECON)*, 2014.
[5] D. Membreno *et al.*, "A high-energy-density quasi-solid-state carbon nanotube electrochemical double-layer capacitor with ionogel electrolyte," *Translational Materials Research*, vol. 2, no. 1, 2015.
[6] G. K.-W. Leung, "Variability and heterogeneous integration of emerging device technologies," Ph.D. Dissertation, UCLA, 2015.
[7] P. Sharma *et al.*, "A review on electrochemical double-layer capacitors," *Elsevier Energy Conversion and Management*, vol. 51, no. 12, 2010.
[8] D. E. Lackey *et al.*, "Managing power and performance for system-on-chip designs using voltage islands," in *IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2002.
[9] J. Hu *et al.*, "Architecting voltage islands in core-based system-on-a-chip designs," in *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2004.
[10] P. Shenoy *et al.*, "System level trade-offs of microprocessor supply voltage reduction," in *IEEE International Conference on Energy Aware Computing (ICEAC)*, 2010.
[11] S. Rajapandian *et al.*, "Implicit dc-dc downconversion through charge-recycling," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 40, no. 4, 2005.
[12] ——, "High-voltage power delivery through charge recycling," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 41, no. 6, 2006.
[13] P. S. Shenoy *et al.*, "Overcoming the power wall: Connecting voltage domains in series," in *IEEE International Conference on Energy Aware Computing (ICEAC)*, 2011.
[14] ——, "Power delivery for series connected voltage domains in digital circuits," in *IEEE International Conference on Energy Aware Computing (ICEAC)*, 2011.
[15] P. Jain *et al.*, "A multi-story power delivery technique for 3d integrated circuits," in *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2008.
[16] J. McClurg *et al.*, "Re-thinking data center power delivery: Regulating series-connected voltage domains in software," in *IEEE Power and Energy Conference at Illinois (PECI)*, 2013.
[17] P. S. Shenoy *et al.*, "Differential power processing for increased energy production and reliability of photovoltaic systems," *IEEE Transactions on Power Electronics*, vol. 28, no. 6, 2013.
[18] J. McClurg *et al.*, "A series-stacked architecture for high-efficiency data center power delivery," in *IEEE Energy Conversion Congress and Exposition (ECCE)*, 2014.
[19] M. S. Gupta *et al.*, "Understanding voltage variations in chip multiprocessors using a distributed power-delivery network," in *Design, Automation and Test in Europe (DATE)*. IEEE, 2007, pp. 1–6.
[20] C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," in *IEEE Design Automation Conference (DAC)*, 1982.
[21] C. Pei *et al.*, "0.026 m2 high performance embedded dram in 22nm technology for server and soc applications," in *IEEE International Electron Devices Meeting (IEDM)*, 2014.
[22] D. Linzen, S. Buller, E. Karden, and R. W. De Doncker, "Analysis and evaluation of charge-balancing circuits on performance, reliability, and lifetime of supercapacitor systems," *IEEE Transactions on Industry Applications*, vol. 41, no. 5, 2005.
[23] A. Bakhoda *et al.*, "Analyzing cuda workloads using a detailed gpu simulator," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2009.
[24] "GPGPU-Sim," http://gpgpu-sim.org.
[25] X. Dong *et al.*, "Fabrication cost analysis and cost-aware design space exploration for 3-d ics," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2010.