Voltage-Stacked Power Delivery Systems: Reliability, Efficiency, and Power Management

An Zou, Student Member, IEEE, Jingwen Leng, Xin He, Member, IEEE, Yazhou Zu, Christopher D. Gill, Senior Member, IEEE, Vijay Janapa Reddi, Xuan Zhang, Member, IEEE

Abstract—In today’s manycore processors, energy loss of more than 20% may result from inherent inefficiencies of conventional power delivery system (PDS) design. By stacking multiple voltage domains in series to lower the step-down conversion ratio of the off-chip voltage regulator module (VRM) and reduce energy loss along the path of the power delivery network (PDN), voltage stacking (VS) offers a novel alternative power delivery technique to fundamentally improve power delivery efficiency (PDE). However, voltage stacking suffers from aggravated supply voltage noise from current imbalance, which hinders its adoption. In this paper, we investigate practical voltage stacking implementation in manycore processors to improve power delivery efficiency (PDE) and achieve reliable performance, while maintaining compatibility with advanced power management techniques. We first present the system configuration of a voltage-stacked manycore processor. We then systematically characterize supply voltage noise in voltage stacking, identify global and residual differential currents as its dominant contributors, and calculate the possible worst supply voltage noise. We next propose a hybrid voltage regulation solution, based on a charge-recycling off-chip voltage regulator and distributed integrated voltage regulators, to mitigate supply voltage noise effectively. We also study the compatibility of voltage stacking with higher level power management techniques. Finally, the performance of a voltage-stacked GPU system is comprehensively evaluated. Simulation results show that our approach can achieve 93.5% power delivery efficiency, reducing the power loss by 13.6% compared to conventional single-layer power delivery system.

Index Terms—Power Delivery System, Manycore Architecture, Voltage Stacking, Supply Noise, Integrated Voltage Regulator

I. INTRODUCTION

Computers consume a non-trivial proportion of the total electricity energy both globally and in the U.S. [1], [2]. For example, it is predicted that the power consumption of world data centers alone will approach 1,000 TWh within a decade, which is more than the amount now consumed for all purposes by Japan and Germany combined [3], [4]. A closer examination of the power delivery path in modern computing systems reveals a provocative finding: transmitting...
down voltage conversion and reduces the currents flowing through the PDN. Ideally, if the current loads from all the voltage domains are perfectly balanced, then the input voltage is evenly divided with no supply voltage noise fluctuation. Voltage stacking’s theoretical peak power delivery efficiency under balanced power activity is close to 100%, making it an attractive solution. However, in real applications, voltage stacking is seriously limited by its exacerbated supply voltage noise caused by the current imbalance between the serially-stacked voltage domains [16]. This limitation prevents wide adoption in practical systems that require consistent and reliable operation. In this paper, we systematically investigate the feasibility and potential benefits of applying voltage stacking to a graphic processing unit (GPU) processor to improve its power delivery efficiency.

II. BACKGROUND AND RELATED WORK

A. Power Delivery System

The power delivery system (PDS) in modern processors consists of a step-down voltage regulation module (VRM) on the motherboard; sockets, off-chip decoupling capacitors and electrical connections at the board, package, and chip levels in the form of PCB traces; and socket bumps and C4 bumps, where undesirable parasitic resistance and inductance reside. The decoupling capacitors (C) and the parasitic resistance (R) and inductance (L) along the connection path form the electrical model of the PDN in a computing system with a conventional PDS. To study the power delivery efficiency and system reliability, it is usually sufficient to assume the output of board level VRM is an ideal voltage source, and we adopt this convention in this work.

In a conventional setting, voltage conversion using a step-down VRM is necessary because the voltage level at the board is higher than the digital supply of a processor. Yet due to inherent inefficiency of step-down VRMs, energy is lost during the voltage conversion. Resistive parasitics along the PDN path also contribute to energy loss and incur voltage drop across the resistance, which is known as IR-drop. These two major efficiency losses can approach 20% or more in advanced technology nodes and under peak power operations.

Moreover, because of the non-ideal effect of the parasitic RLC network, electrons cannot be delivered instantaneously from the VRM output to immediately satisfy the fast changing current loads of various on-chip components. This lag results in on-chip voltage fluctuations and causes supply voltage noise reliability issues during operation.

B. Voltage Stacking

In voltage stacking (VS), the step-down VRM can be eliminated by serially stacking the voltage domains. It can be intuitively understood as allowing electron charges to recycle through the stacking layers in series. In addition to eliminating step-down conversion loss, voltage stacking lowers the PDN loss due to resistive parasitics, because in a N-layer voltage-stacked system, the PDN path current is reduced by \( N \times \), which corresponds to \( N^2 \times \) reduction in power loss. These efficiency improvements have been demonstrated in prior work [17], [18].

A theoretical peak PDE close to 100% can be achieved using voltage stacking [18] when all the stacking layers have balanced activities, and hence the same transient current demands. In practice, though, applying voltage stacking in real computing systems, where activity mismatches abound both spatially and temporally, proves to be challenging. As has been shown in previous studies, such activity mismatches can cause severe voltage fluctuations in a voltage-stacked system [8], [19]–[23]. The aggravated noise problem remains one of the most obstinate obstacles preventing voltage stacking adoption in the mainstream.

C. Supply Voltage Noise

Due to its impact on system reliability, supply voltage noise has been diligently studied and characterized for conventional single-layer PDS in single-core [24], [25], multi-core [26], [27], and manycore GPU processors [21], [28]–[31]. While circuit techniques such as load line compensation are effective at taming IR-drop induced noise [32], dynamic \( \frac{Ldi}{dt} \) noise, and resonance noise in particular, are more dominant and harder to tackle [33], [34], and often demand a cross-layer solution. However, a voltage-stacked manycore processor experiences more serious and complex supply voltage noise behavior due to the interactions between the cores that can lead to constructive or destructive noise composition. This aggravated supply voltage noise prevents the wide adoption of voltage stacking in mainstream computing systems, despite its higher power delivery efficiency. Up until now, it has been only intuitively understood that the supply voltage noise in voltage stacking is from an imbalanced workload, and a systematic supply voltage noise study of manycore processors with multilayer voltage-stacked PDS is still lacking.

D. Power Delivery Efficiency

The underlying physical mechanism 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 can be broken down into three parts:

First, energy is lost in voltage conversion to step down the supply voltage [9]. We define the conversion efficiency of a voltage regulator \( \eta_{VR} \) as the ratio between the power it delivers at the voltage regulator output over the power it consumes at the input. \( \eta_{VR} \) is usually a function of the step-down conversion ratio. 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 [35].

The second part of the energy loss occurs in the power delivery network mostly because of heat dissipation when current runs through the parasitic resistance that exists along the path of the PDN, which is related to the IR-drop component of supply voltage noise [10], [11]:

\[
\eta_{PDN} = \frac{R_{core}(V_{core})}{(R_{PDN} + R_{core}(V_{core}))}
\]
where $R_{PDN}$ represents the total parasitic resistance contributed by the PDN, and $R_{core}$ represents the equivalent resistive impedance of the computational load as a function of $V_{core}$. The definition of $R_{core}$ suggests that $R_{core} = V_{core}/I_{core}$. For fixed $V_{core}$ value, $R_{core}$ is a measure of the power rating.

The final 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 and sustain fault-free operation [36], [37]. We can express this component as $\eta_{\Delta V}$:

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

$P_{core}$ 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 above analysis, the full power delivery system efficiency can be expressed as

$$\eta_{PDS} = \frac{P_{src}(V_{src})}{P_{src}} = \eta_{VR} \cdot \eta_{PDN} \cdot \eta_{\Delta V} \quad (3)$$

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

Voltage stacking can reduce the power loss in step-down voltage regulator by eliminating supply voltage conversion and power loss in power delivery network parasitic resistance by reducing path current. Besides, previous works [21], [38] further prove that voltage stacking can also diminish the voltage margin to further improve power delivery efficiency. In this paper, for a fair comparison, we assume voltage stacking has a same voltage margin with conventional single-layer power delivery system and mainly focus on its improvements in conversion efficiency ($\eta_{VR}$) and PDN efficiency ($\eta_{PDN}$).

### E. Related Work

Proof-of-concept circuits [15], [39] and silicon prototypes [14], [17], [18], [40], [41] have been presented previously to explore voltage stacking using low-power microcontrollers, along with design methodology for floorplanning and placement [42], [43]. These pioneering works demonstrate the feasibility of voltage stacking, but they are often limited to simple assembly of uncorrelated cores with low power density. Interlayer current imbalance has been discussed qualitatively [16] as a contributor to the supply voltage noise in voltage-stacked systems, but without rigorous quantitative derivation of worst-case conditions. To overcome supply voltage noise, most voltage stacking prototypes [14], [17], [18], [40], [41] resort to employing charge-recycling integrated voltage regulators (CR-IVR) to actively balance the current mismatches. However, the overhead and trade-offs from CR-IVR require further discussion and should be reduced.

Built upon these early prototypes, a number of novel approaches have been proposed to take advantage of voltage stacking under different scenarios, such as 3D-IC with varying TSV, on-chip decoupling capacitance, and package parameters [44], [45]; optimal system partitioning to unfold CPU cores [20], [38]; and GPU systems with supercapacitors [46] operating under near-threshold voltages [21], [38]. CoreUnfolding [20] is a novel method that voltage stacking can be used within each core. However, it is highly invasive as it requires separating function units inside the core to balance the groups of units. Voltage-Stacked GPUs [47], [48] models the power grid of the voltage-stacked GPU as a linear dynamic system to derive the power control strategy for supply noise guarantee, but the architecture-level power control scheme sacrifices the system performance for reliability.

### III. System Configuration

In this paper, we use the GPU system with a NVIDIA Fermi architecture as a representative manycore processor. Table I lists the configuration details of Fermi architecture. The Fermi architecture GPU has 16 streaming multiprocessor (SM) cores [49]. We use a $4 \times 4$ voltage stacking structure: 16 SM cores are stacked in 4 layers and each layer has 4 SM cores, as 4V is generally available on the board and SM cores require 1V. Our analysis and solutions are not limited to this $4 \times 4$ voltage stacking configuration and can be applied to other manycore processors with arbitrary voltage stacking configurations.

### A. Power Grid Routing and PDN Modeling of VS

Voltage stacking can be implemented in both 2D and 3D-IC chips, but for a fair comparison with conventional power delivery methods, we focus on voltage stacking implementation in a 2D planar technology. To properly isolate the transistors in each voltage layer from the global substrate, voltage stacking often relies on advanced process technology such as triple wells or Silicon on Insulator (SOI) to establish local body
biasing voltages [16], [22], [23], [50], [51]. A hierarchical structure is used in 2D power routing, as shown in Fig. 2(a). The top metal layers are for global power grid which connects cores or modules. The next layers are local power grids connecting the function blocks such as ALU and Reg. Finally local power grids in the bottom metal layers connect to the logic gates. As illustrated by the power/ground routing scheme in Fig. 2(b) and 2(c) topologically stacking the voltage domains on a 2D chip can be achieved with minimal modifications by re-routing the top metal layers from parallel connections to series connections, leaving the local power/ground grids in the lower metals and the physical floorplans of the underlying blocks largely intact. Assuming this minimally-invasive routing method, we derive the voltage stacking PDN model shown in Fig. 3 based on the typical RLC circuits and parameters introduced previously to study GPU manycore processors [28], [51]. Note that there is parasitic resistance ($R_{ij}$) between the vertically-connected cores (modeled by current sources), as depicted in Fig. 3. Our study focuses on the SM core power grid to clearly demonstrate the benefit of voltage stacking and evaluate the proposed hybrid regulation methodology, since its peak and average powers account for 80% and 93% of the total GPU chip power consumption [52]. Similar scheme can also be adapted to other on-chip components like SRAM in a voltage-stacked configuration [39].

B. Communication Across Layers

A voltage-stacked system suffers inherently complex communications across different voltage layers: instructions and data are communicated among memory, cache, and core registers, which are in different voltage layers. In the GPU system, SM cores do not directly communicate with each other, and the cross-layer communication mainly happens between SM cores and Memory Partition Units through an interconnection network. There are two interconnection networks with butterfly topology and 22 nodes: one for traffic from SM cores to Memory Partitions, and one for traffic from Memory Partitions back to SM cores. Cross layer communication requires extra level shifters added to the interconnection network. Several level shifter designs are suitable for a stacked architecture, such as capacitive-coupling-based (conventional) [53], two-stage cross-coupled (TSCC) [54]–[58], Wilson current mirror (WCM) [55], [57], [59], stacked Wilson current mirror (Stacked) [60], switched-capacitor (Tong) [22] and modified switched-cap (Mod-Tong). Tested by Ebrahimi [61] with input signal at 1GHz, Tong has the best energy-delay trade-off.

Fig. 3: Power delivery network (PDN) of a 2x4 voltage-stacked manycore processor

IV. SUPPLY VOLTAGE NOISE ANALYSIS

A. Supply Voltage Noise Characterization

Unlike previous empirical approaches [28], [31], we develop an analytical modeling framework to study and characterize supply voltage noise responses in voltage stacking PDN, especially in the presence of both correlated and uncorrelated core activities. The cornerstone of our analytical approach lies upon the decomposition and superposition principles in the fundamental circuit theory.

1) Noise Decomposition & Superposition: Since the basic electrical model of voltage stacking PDN consists of only linear components, including the RLC and ideal voltage and current sources, the superposition principle in linear systems generally holds, allowing us to decompose the core current to different components to reveal their distinctive characteristics. Without loss of generality, let us assume a voltage-stacked system that consists of $N_L$ vertically-stacked layers with $N_V$ cores on each layer. For example, Fig. 3 shows a $N_L = 2$ and $N_V = 4$ voltage-stacked system. The cores that align vertically are defined as a voltage stack. To facilitate later analysis, we adopt the s-domain expressions for current sources and give the following definitions:

$$I_{c,ij}^{core}(s) = I_G(s) + I_{ST}^{core}(s) + I_{g,ij}^{s}(s)$$

$$I_G(s) = \sum_{i=1}^{N_L} \sum_{j=1}^{N_V} I_{c,ij}^{core}(s)$$

$$I_{ST}^{core}(s) = \sum_{i,j} I_{c,ij}^{core}(s) - I_G(s)$$

$$I_{g,ij}^{s}(s) = \frac{(N_L - 1) I_{c,ij}^{core}(s) - \sum_{k \neq j} I_{c,ik}^{core}(s)}{N_L}$$

where $I_{c,ij}^{core}(s)$ is the current contributed by the core in the $i$th stack and the $j$th layer. It is decomposed into three components: $I_G(s)$, $I_{ST}^{core}(s)$, and $I_{g,ij}^{s}(s)$, in Eq. (4) - (7). $I_G(s)$ represents the global current component shared by all the cores, $I_{ST}^{core}(s)$ represents the common current components shared by the cores in the $i$th stack, and $I_{g,ij}^{s}(s)$ is the residual current components after removing the global and per-stack common terms. Now, the supply voltage noise at the core (in the $i$th stack and the $j$th layer) can be expressed by the current components working on their respective effective impedances, $Z_{eff,i}$, $Z_{eff,i,j}$, and $Z_{eff,i,i,j}$, and $\Delta V_{core,ij}$, $\Delta V_{ST,core,ij}$, and $\Delta V_{n,core,ij}$, as described in Eq. (8) and Fig. 4.

$$\Delta V_{core,ij} = \Delta V_{core,ij}^{G} + \Delta V_{core,ij}^{ST} + \Delta V_{core,ij}^{R} = I_{G}^{Z_{eff,i}} + I_{ST}^{Z_{eff,i,j}} + \sum_{k=1}^{N_L} \sum_{j \neq k} I_{g,ik}^{s} Z_{eff,i,j}$$

To illustrate how the decomposition and superposition in Eq. (4) - Eq. (8) help us analyze and characterize supply voltage noise effects in voltage stacking, we use a simplified RLC network of a $2 \times 2$ voltage stacking PDN, as shown in Fig. 4.
2) Global Uniform Current: Since \( I^G(s) \) is a uniform component across all the cores, the effective network can then be transformed by removing the path between equal-potential nodes and merging the parallel components as in Fig. 5(b) according to our \( 2 \times 2 \) example. We can derive the supply voltage noise caused by \( I^G \) with an analytical expression for a general \( N_s \times N_v \) network:

\[
\Delta V_{i,j}^G = I^G_{i,j} Z_{eff,i}^G = I^G_{i,j} \frac{Z_{CA} Z_{S}}{N_v} \frac{N_L}{N_L} Z_{off}/|Z_C| \quad (9)
\]

Due to the uniform nature of the global current, all cores share the same common mode, \( Z_{eff,i}^G \), and thus the same \( \Delta V_{core,i,j} \). Eq. 9 also applies to the case when \( N_L = 1 \), which is a conventional single-layer PDN. From Eq. 9 and the typical impedance profile of \( Z_{eff}^G \) shown in Fig. 5(c), we can see that in a \( N_L \times N_v \) voltage stacking PDN, \( \Delta V_{core,i,j}^G \) peaks at the dominant resonant frequency of \( Z_{off} \), similar to the conventional single-layer, but its magnitude is reduced by \( N_L \times \) when stacked.

3) Local Uniform Through-stack Current: Following our definition of \( I^{ST}(s) \), we can see that since \( \sum_{n=1}^{N_v} I^{ST}_{i,n}(s) = 0 \), there is no current going through \( Z_{off} \) according to Kirchhoff’s Current Law (KCL) and the entire branch can be eliminated. The linear circuit network is again transformed to a simpler form as in Fig. 3(d). For example, in our \( 2 \times 2 \) example, we can derive \( \Delta V_{core,i,j}^{ST} \), for \( i = 1, 2 \) and \( j = 1, 2 \) respectively, as a function of the unit current stimulus \( I^{ST}_{i,1} \) and complex impedances in the form of \( Z_{L} \) and \( Z_{C} \):

\[
\Delta V_{core,i,j}^{ST} = I^{ST}_{i,1} Z_{eff,i}^{ST} = I^{ST}_{i,1} \frac{1}{N_L} |Z_C|/|Z_L| \quad (10)
\]

where \( \Delta V_{core,i,j}^{ST} \) represents the supply voltage noise induced by \( I^{ST}_{i,1} \), the common components shared by all the cores in the \( i^{th} \) stack. All cores in the \( i^{th} \) stack share the same common-mode \( \Delta V_{core,i,j}^{ST} \) disturbance. The resulting expression suggests that on the first order, the combined effect of all the \( I^{ST} \) exerts differential voltage fluctuations between the vertical stacks, and it is further voltage divided across the cores in the same stack, as illustrated in Fig. 5(d). The dividing ratio depends on the ratio of \( Z_L/Z_C \), and in its high-frequency limit asymptotically approaches \( Z_L/N_L \). The analytical results of the local uniform through-stack current again suggest that by moving from single-layer to multi-layer, the supply voltage noise experienced at each core level and contributed by this current component is reduced by \( N_L \times \) on average.

4) Residual Per-Core Differential Current: On closer inspection of Eq. 7, \( I_{i,j}^R \) can be rearranged as the summation of differential currents in the form of \( I_{core,i,j}^{R} - I_{core,j}^{R} \), where \( k \neq j \). The summation suggests that the remaining voltage noise effect, unaccounted for by the global and the local terms, \( \Delta V^G \) and \( \Delta V^{ST} \), are induced by the aggregated differential currents. This differential current represents the mismatched part of current between cores which will not only cause voltage noise at itself but also cause noise at other cores. For example, at \( core(i,j) \), the noise from residual current is from its own residual current and other cores’ residual current:

\[
\Delta V_{i,j}^{R} = I_{i,j}^{R} Z_{eff,i}^{R} + \sum_{m \neq j} \sum_{n=1}^{N_v} I_{n,m}^{R} Z_{eff,m}^{R} \quad (11)
\]

where \( I_{i,j}^{R} Z_{eff,i}^{R} \) is the supply voltage noise caused by its own residual current, and \( \sum_{m \neq i} \sum_{n=1}^{N_v} I_{n,m}^{R} Z_{eff,m}^{R} \) is the supply voltage noise caused by residual current from other cores. Most importantly, this type of residual per-core differential current is unique to voltage stacking, since these terms simply vanish when \( N_L = 1 \).

B. Dominating Supply Voltage Noise

Based on the above system configuration, we characterize the effective impedances, \( Z_{eff,i}^{R} \), \( Z_{eff,i}^{ST} \), and \( Z_{eff,i}^{G} \) of each current component defined in Eq. 8. The effective impedance for \( core(1,1) \) is shown in Fig. 6. Due to location symmetry, the effective impedances of other cores are similar to \( core(1,1) \). We divide the frequency range into low frequency (< 10 MHz), medium frequency (10 MHz – 50 MHz), and high frequency (> 50 MHz). From the effective impedance curve, we can see that both \( Z_{eff,i}^{R} \) at low frequency and \( Z_{eff,i}^{G} \) at high frequency (especially at resonance), have relatively large magnitudes. The corresponding low frequency residual current components and high frequency (resonance) global current components that excite these effective impedances can thus cause large supply voltage noise, and we identify them as the dominant causes of voltage noise in voltage stacking.

C. Worst-Case Supply Voltage Noise

Identifying the root cause of noise is not sufficient for rigorous reliability analysis. We must also consider what core activity conditions can result in the worst-case supply voltage noise. Understanding the condition and the magnitude of worst-case would help us determine the necessary and sufficient noise mitigation strategy to guarantee reliable operation in real-world voltage-stacked systems.
Fig. 6: Effective impedance of current components

After characterizing $Z_{Z_{eff}}$, $Z_{ST_{eff}}$, and $Z_{R_{eff}}$, and establishing the relationship between $\Delta V_{core}$, as a function of these impedances, searching for the load current conditions that would result in worst-case supply voltage noise can now be performed in the frequency domain. We formulate it as an optimization problem of finding the optimal frequency distribution of each core current $I_{i,j}^{core}$ to maximize their combined effects $\Delta V_{i,j}^{m,n}$ on $core(m,n)$. This optimization can be solved as a linear programming problem, and the process is described in Algorithm 1. The optimization variables are each core current distribution at different frequency range $I_{i,j}^{core}(s)$. The optimization objective function is the supply voltage noise $\Delta V_{i,j}^{m,n}$ at $core(m,n)$ and the constraints are from voltage noise decomposition Eq. (4) - (7) and peak GPU SM core power, as shown in Table I. This linear optimization formulation with a general constraint of max power/current allows us to search the vast space of arbitrary synthetic core current stimuli from all possible activity combinations, including the effects cause by clock gating and power gating, and therefore can quantitatively represent the worst-case supply voltage noise for rigorous reliability analysis.

Algorithm 1 Maximize supply voltage noise

Objective Function:

$\Delta V_{core,ij} = \Delta V_{core,ij}^G + \Delta V_{core,ij}^{ST} + \Delta V_{core,ij}^{R}$ in Eq. (8)

Subject to:

1. $\forall i,j; \ 0 \leq I_{i,j}^{core}(s)$
2. $\forall i,j; \ I_{i,j}^{core}(t) = F^{-1}(I_{i,j}^{core}(s)) \leq$ peak current (14A)
3. $\forall i,j; \ I_{i,j}^{core}(s), 0 \leq s \leq$ clock frequency (700MHz)
4. Eq. (4) - (7): current decomposition rules

The numerical solution of the linear programming problem based on the GPU configurations in Table I gives us a glimpse of the core current distribution and combination that act together and cause the largest supply voltage fluctuation at $core(m,n)$, as shown in Table II. The currents, $I_{i,j}^{core}(s)$, are distributed at low frequency with major components of residual currents, while the currents, $I_{i,j}^{core}(s)$, are distributed at the resonant frequency of $Z_{R_{eff}}$ with major components the global currents. This worst-case scenario is plausible in real GPU applications shown in Fig. 1 when the $I_{i,j}^{core}(s)$ are alternating between idle and Sine/Cosine special function instructions (SF Inst) at the resonant frequency, while $I_{i,j}^{core}(s)$ are at peak power executing Sine/Cosine special function instructions (SF Inst). We compare the worst-case noise derived by our optimization algorithm with three other scenarios based only on heuristics: (1) all cores have low frequency residual currents, (2) all cores have high frequency global currents, and (3) all cores have randomly distributed currents. From the supply voltage noise histograms in Fig. 8 we can see that the worst-case rigorously derived by our method is more severe than the heuristic ones, and therefore is more representative as a stressmark for supply voltage noise reliability analysis.

V. NOISE MITIGATION BY HYBRID REGULATION

A. Hybrid Regulation Framework

To combat elevated and hard-to-predict supply voltage noise and guarantee reliable operation in spite of worst-case conditions in voltage-stacked manycore processors, we explore a hybrid voltage regulation mechanism using both on-chip charge-recycling integrated voltage regulators (CR-IVRs) and an off-chip charge-recycling voltage regulator module (CR-VRM). Fig. 9 shows the framework of the proposed hybrid regulated voltage stacking using either switched-capacitor or low dropout voltage regulators. This hybrid approach takes advantage of the unique merits of on-chip and off-chip voltage regulators and simultaneously avoids their individual defects.

Unlike step-down voltage regulators converting supply voltage, charge-recycling voltage regulators move extra charge between different layers to balance current and maintain a stable voltage of each layer. Because the direction and amplitude of extra charge keeps changing with core workload conditions, charge-recycling voltage regulators should support bidirectional fast switching current. Voltage regulators, such as low drop-out voltage regulators and switched capacitor voltage regulators, can be used as charge-recycling voltage regulators, while inductor based voltage regulators, such as buck converters, do not support bidirectional fast switching of current movement and incur extra Ldi/dt noise. Multi output switched capacitor (SC) voltage regulators are the most widely used charge-recycling voltage regulators, because they have higher power efficiency, but they require each layer to have the same voltage. Previous work has demonstrated a multi-output switched-capacitor integrated voltage regulator [40] that balances the layer currents in voltage-stacked systems.

TABLE II: Freq. Distribution of decomposed core current

<table>
<thead>
<tr>
<th>Core Current</th>
<th>Frequency</th>
<th>Major Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{i,j}^{core}(s)$</td>
<td>low frequency ($&lt; 10 MHz$)</td>
<td>residual current</td>
</tr>
<tr>
<td>$I_{i,j}^{core}(s)$</td>
<td>high frequency ($&gt; 50 MHz$)</td>
<td>global current</td>
</tr>
</tbody>
</table>

Fig. 7: An example instruction trace contributing to worst-case supply noise
Although low drop-out voltage regulators have lower power efficiency, they do not force each layer to have exactly the same voltage, and hence are more suitable to support dynamic voltage and frequency scaling in voltage stacking.

B. Centralized and Distributed Integrated Voltage Regulator

Located closer to the point-of-load, on-chip integrated voltage regulators enjoy fast regulation response, but have limited on-die area and capacity, making them suitable for reducing high-frequency noise of smaller magnitude. According to the analysis in Section IV, one of the dominant causes of worst-case supply voltage noise is high frequency global currents. This noise can be mitigated by on-chip CR-IVRs.

By moving charges across the stacking layers, the CR-IVR effectively behaves as an additional parallel impedance connected to the original effective impedance $Z_{\text{eff}}^G$. It thus reduces the supply voltage noise caused by global current: $\Delta V_{\text{cor},i,j}^G = jf^{G}[Z_{\text{eff}}^G/(Z_{CR-IVR}^G + Z_{CR-IVR-path}^G)]$  \((12)\)

Here, $Z_{CR-IVR}^G$ is the impedance of the on-chip charge-recycling voltage regulator, and $Z_{CR-IVR-path}^G$ is the parasitic impedance of the on-chip power grid between the core and voltage regulator. By deploying CR-IVR with the desired impedance, $\Delta V_{\text{cor},i,j}^G$ from global current $f^{G}$ can be effectively mitigated. The effective impedance of a multi-output switched-

---

**Fig. 8:** Histogram comparison between analytically derived worst case and other heuristic core activation patterns

**Fig. 9:** Hybrid voltage regulation based on distributed on-chip CR-IVRs and off-chip CR-VRM

**Fig. 10:** Voltage distribution among the 16 SMs during execution of the backp benchmark

**Fig. 11:** Voltage distribution among the 16 SMs during execution of the blackscholes benchmark
The implementation overhead of the distributed design is mainly consisting of IVR directly to the SMs in each layer, with each sub-IVR divided into four equal sub-IVRs and connects each sub-IVR to the original central-IVR. The distributed CR-IVR divides the original central-IVR and without CR-IVR, reveals that the regulation effect is significantly mitigated.

The location dependence is now completely removed and the same regulation effect is achieved across the board. A optimal design parameters [63] of distributed CR-IVR are shown in Table IV.

C. Off-Chip Charge-Recycling VR

Compared with CR-IVR, off-chip CR-VRMs have slower response times, but they offer better efficiency [64], [65] and do not consume expensive die area. It is important to note that although on-chip CR-IVRs can be designed to provide similar regulating capacity as an off-chip counterpart, they incur large area overhead, sometimes exceeding the total area of the logic cores, making them impractical in real systems. Therefore, off-chip CR-VRM is a better and more economical choice for regulating supply voltage noise at low frequency. Similarly, the addition of the CR-VRM results in an effective parallel impedance connected with the original $Z_{eff(i,j)}$ through the C4 pad, package, and PCB. In this case, the supply voltage noise caused by residual current becomes

$$\Delta V_{core(i,j)} = \sum_{i,j} \sum_{i,j} R_{eff(i,j)} (Z_{VRM} + Z_{PCB})$$

where $Z_{CR-VRM}$ is the impedance of the off-chip charge-recycling voltage regulator module; $Z_{CR-VRM-path}$ includes the parasitic impedances of not only the on-chip power grid but also the C4 pads, package, and PCB board between the CR-VRM and the cores. A design optimization similar to that for CR-IVR is applied to arrive at an optimal set of design parameters, as summarized in Table IV. The new effective impedance of the residual current after employing off-chip CR-IVR and off-chip CR-VRM is shown in Fig. 12 With reduced effective impedance, the supply voltage noise, $\Delta V_{core(i,j)}$ and $\Delta V_{core(i,j)}$, are significantly mitigated.

D. Charge-Recycling VR Power Loss

In voltage stacking, most of the current goes through the stacked layers, the occasional residual current is absorbed by decoupling capacitors, and only the accumulated residual current components goes through the CR-IVR or CR-VRM.
For the accumulated residual current, we call it imbalanced current. When imbalanced current is recycled by charge-recycling VR, power losses in these VRs are unavoidable.

1) Switched-capacitor charge-recycling VR: A switched capacitor voltage regulator delivers current to a synchronous digital system, whose frequency is determined by the clock frequency, set by the minimum voltage over a clock period. Power loss in the voltage ripple over the minimum voltage is the intrinsic switched-capacitor loss [66], which is

\[ P_{\text{intrinsic}} = I_{\text{imbalance}} \frac{\Delta V}{2} = \frac{I_{\text{imbalance}}^2}{M_{\text{cap}} C_{\text{fly}} f_{\text{sw}}} \]  

(16)

Switching conductance Loss: Also, the finite conductance of the transistor switch has a series power loss:

\[ P_{\text{R}_{\text{sw}}} = N I_{\text{imbalance}}^2 R_{\text{sw}} D \]  

(17)

Plate Parasitic Capacitance Loss: In steady-state operation, both the top and the bottom plates experience approximately equal voltage swings, and parasitic capacitance causes extra power loss:

\[ P_{\text{plate-}\text{cap}} = M_{\text{bott}} V^2 C_{\text{plate}} f_{\text{sw}} \]  

(18)

Switching Parasitic Capacitance Loss: The loss in voltage swings at the switch transistor parasitic capacitance, which can be expressed as

\[ P_{\text{sw-cap}} = N C_{\text{sw}} V^2 f_{\text{sw}} \]  

(19)

Among these losses, intrinsic switched-capacitor loss and switch conduct loss are the main components [63].

2) Low Drop-out Charge-Recycling VR: Low drop-out voltage regulators suffer from three power losses [67]:

Switch Conduct Loss: The main cause of loss in low drop-out voltage regulators is the power dissipated as heat on the transistor switch resistance. It is highly dependent on the difference between the input and output voltage:

\[ P_{\text{R}_{\text{sw}}} = I_{\text{imbalance}}^2 R_{\text{sw}} \]  

(20)

Switching Parasitic Capacitance Loss: The gate parasitic capacitance switching loss is similar to the loss happens in switched capacitor charge recycle voltage regulator.

Control Logic Loss: In LDO, a feedback control logic circuitry is used to control the voltage at the reference value. The loss is due to the current flowing through the operational amplifier, the resistive voltage divider, and the voltage reference generator in the control logic circuitry; the sum of these currents is called quiescent current. The quiescent current can be reduced by optimizing the components, making it negligible when compared to the load current consumption.

E. Hybrid Regulated VS Power Delivery Efficiency

The power delivery efficiency of hybrid regulated manycore voltage stacking can be described as

\[ \eta_{\text{PDS}} = \frac{P_{\text{core}}}{P_{\text{core}} + P_{\text{PDN}} + P_{\text{CR-IVR}} + P_{\text{CR-VRM}}} \]  

(21)

\[ \frac{I_{\text{core}} V_{\text{core}} + (I_{\text{core}}/N)^2 R_{\text{PDN}} + P_{\text{CR-IVR}} + P_{\text{CR-VRM}}}{I_{\text{core}} V_{\text{core}}} \]

where, \( P_{\text{core}} \) is the power consumed by cores, and \( P_{\text{PDN}} \) is the power loss in the parasitic resistance along the power delivery network. \( P_{\text{CR-IVR}} \) and \( P_{\text{CR-VRM}} \) are the power loss in CR-IVR and CR-VRM, derived in Eq. (16) – (20).

In the ideal case, if there is no residual current, the power losses in CR-IVR and CR-VRM are negligible and the power delivery efficiency can approach nearly 100%. In the normal case, most of the current goes through the stacked layers, and only the accumulated residual current components goes through the CR-IVR or CR-VRM where it introduces a small amount of power loss. In the worst-case, when cores in one layer are powered off and cores in the other layers are powered on, all the current consumed by cores is imbalanced current and goes through the CR-IVR and CR-VRM, causing significant power loss. However, this worst case seldom happens, and the system on average achieves high power delivery efficiency. Consequently, the hybrid regulation scheme not only mitigates the supply voltage noise but also maintains the high power delivery efficiency.

VI. ADVANCED POWER MANAGEMENT

We consider a well-designed power delivery system should be compatible with advanced high-level power management techniques. Among them, the most common techniques are dynamic voltage and frequency scaling (DVFS) and power gating. In this section, we will discuss applying DVFS and power gating together with the proposed hybrid regulation in voltage-stacked GPU systems.

A. Dynamic Voltage and Frequency Scaling

Dynamic voltage and frequency scaling (DVFS) adjusts the supply voltage and frequency of a voltage domain to boost performance or save power. In voltage stacking each layer can be divided into different voltage domains and the division should be consistent across different layers to maintain the stacked power delivery. In this paper, we assume the SMs in each layer share one voltage domain in the proposed GPU manycore voltage-stacked system. When DVFS is applied in voltage stacking, each layer (voltage domain) may have a different voltage. The low drop-out voltage regulator will be used to recycle the imbalance current because LDO can support that each layer has its own voltage. We will use LDO hybrid regulations in following voltage stacking DVFS analysis and evaluations. When one voltage domain needs to change to a different supply voltage, the low drop-out voltage regulator can change the reference voltage and conversion ratio to adjust the voltage of each layer [17], [19], [68].

Compared to the original voltage stacking, DVFS maintains the stacked power delivery but may brings more frequent current imbalance. As the current imbalance introduced by DVFS does not go beyond the worst cases studied in Section IV where one layer is totally powered off, the proposed hybrid regulation can effectively guarantee the system’s stability under DVFS operation. Although the amplitude of the imbalanced current does not exceed the worst case, the extra current imbalance introduced by DVFS will cause more power loss in CR-IVR and CR-VRM. Compared with the original voltage stacking, part of efficiency benefit will be sacrificed under DVFS.
### Algorithm 2 Power Saving from DVFS and Power Gating

**Input Variables:**
- DVFS / power gating command: $f_{i,j}^{\text{core}} / P_{i,j}^{\text{core-gate}}$

**Output Variables:**
- Power estimation of each Core: $P_{i,j}^{\text{core}}$

**Steps:**
1. Replace $I_{i,j}^{\text{core}}$ with $f_{i,j}^{\text{core}} / P_{i,j}^{\text{core-gate}}$ in Eq. (4) - (7).
2. Calculate residual frequency / gated power: $f_{i,j}^{\text{residual}} / P_{i,j}^{\text{resid}}$.
3. Residual current can be known as:
   
   $I_{i,j}^{\text{residual}} = \frac{\alpha C V f_{i,j}^{\text{residual}}}{f_{i,j}^{\text{gate}}}$

4. Calculate VR loss $P_{i,j}^{\text{VR}} = P_{i,j}^{\text{CR-IVR}} + P_{i,j}^{\text{CR-VRM}}$ in Eq. (16) - (20).
5. Return power estimation:
   
   $P_{i,j}^{\text{core}} = \frac{\alpha C V^2 f_{i,j}^{\text{residual}}}{f_{i,j}^{\text{gate}}} - P_{i,j}^{\text{CR-IVR}} - P_{i,j}^{\text{CR-VRM}}$

**B. Power Gating**

Power gating turns off the circuitry inside a core or the core itself for a while when not in use. Power gating introduces current imbalance and also causes supply voltage noise. The most severe imbalance happens when one layer is totally powered off while other layers are working. This scenario is already captured by supply voltage noise worst case analysis in Section IV and supply voltage can be also guaranteed by the proposed hybrid regulation as described in Section V. Similar to DVFS, the extra imbalanced current introduced by power gating will cause more power loss from CR-IVR and CR-VRM thus degrading efficiency gains.

**C. Power Management Hypervisor in Voltage Stacking**

The DVFS, power gating and other power management techniques optimize the power and performance tradeoffs based on the commands from software operating system. At the software commands level, power management techniques should taken voltage stacking into consideration and many techniques such as fast thread migration [69] can balance the workload before current imbalance happens. To make the correct decision at the software level, the power management techniques first need to know the potential power benefit and performance loss and then find the proper tradeoff point. To estimate the potential power benefit that each core can earn, we introduce a power management technique estimator for the software level power management as described in Algorithm 2. The estimator can evaluate the potential net power consumption of each core considering the extra power loss in power delivery system.

At the hardware power delivery system, we provide a power delivery efficiency guaranteed power management hypervisor of DVFS or power gating instructions for voltage stacking. According to Section IV and V the power loss in voltage stacking comes from the accumulated residual current component going through charge-recycling voltage regulators. The hypervisor guarantees the power delivery efficiency by limiting the maximum allowed residual current, described in Algorithm 3. In the hypervisor, the residual current of each core under DVFS and power gating is calculated with Eq. (4) - (7). The residual current threshold $\Delta I_{i,j}$ is given to limit the residual current $I_{i,j}^{\text{residual}}$ and guarantee power loss in power delivery system. Then each core whose residual current exceeds the threshold $\Delta I_{i,j}$ and power loss in power delivery system are limited within desired range.

**Algorithm 3 Power Management Hypervisor in VS**

**Input Variables:**
- Commands in conventional system: $f_{i,j}^{\text{core}} / P_{i,j}^{\text{core}}$

**Output Variables:**
- Commands for Voltage Stacking: $f_{i,j}^{\text{core}} / P_{i,j}^{\text{core}}$

**Steps:**
1. Replace $I_{i,j}^{\text{core}}$ with $\alpha C V f_{i,j}^{\text{core}}$ in Eq. (4) - (7).
2. Calculate residual current: $I_{i,j}^{\text{residual}} - \Delta I_{i,j}$
3. Dynamic Voltage Frequency Scaling:
   
   for $i, j = 1, 2, 3, 4$
do 
   if $|I_{i,j}^{\text{residual}}| > |\Delta I_{i,j}|$
   then
   $I_{i,j}^{\text{core}} = I_{i,j}^{\text{core}} - (P_{i,j}^{\text{residual}} - \Delta I_{i,j})$
   else
   $I_{i,j}^{\text{core}} = \alpha C V$
   
   Return DVFS commands: $f_{i,j}^{\text{core}}$

4. Power Gating:
   
   for $i, j = 1, 2, 3, 4$
do 
   if $|I_{i,j}^{\text{residual}}| > |\Delta I_{i,j}|$
   then
   $I_{i,j}^{\text{core}} = I_{i,j}^{\text{core}} - (P_{i,j}^{\text{residual}} - \Delta I_{i,j})$
   else
   $P_{i,j}^{\text{core}} = P_{i,j}^{\text{residual}}$
   
   Return power gating commands: $P_{i,j}^{\text{core}}$

### Table IV: LDO Regulator Parameters

<table>
<thead>
<tr>
<th>Design Parameters</th>
<th>CR-IVR</th>
<th>CR-VRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of VR</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Switch frequency</td>
<td>50 MHz</td>
<td>2 MHz</td>
</tr>
<tr>
<td>Total capacitor per VR</td>
<td>1.1uF</td>
<td>600uF</td>
</tr>
<tr>
<td>Capacitor density</td>
<td>50nF/mm²</td>
<td>0.2uF/mm²</td>
</tr>
<tr>
<td>Switch on resistance</td>
<td>130Ω·um</td>
<td>37600Ω·um</td>
</tr>
<tr>
<td>Area per VR</td>
<td>22.0mm²(Die)</td>
<td>3.1cm²(Board)</td>
</tr>
</tbody>
</table>

VII. System Evaluation

In this section, we evaluate the hybrid regulated GPU manycore voltage-stacked system in terms of supply voltage noise, power delivery efficiency, advanced power management compatibility, and finally compare it with other power delivery systems. We develop an hybrid simulation infrastructure that combines SPICE3 [70] and GPGPU-Sim 3.1.1 (with GPUWatch) [71], [72]. SPICE3 simulates the circuit transient response of the full voltage-stacked power delivery system and the charge-recycling voltage regulators as illustrated in Fig. 9 and GPGPU-Sim 3.1.1 simulates the GPU architecture level system specified in Table I. We use ten representative benchmarks that cover a wide range of scientific and computational domains from two benchmark suites, five from Rodinia 2.0 [73] and five from NVIDIA CUDA SDK [74].

A. Supply Voltage Noise Evaluation

We first evaluate the supply voltage noise across real GPU benchmarks and the worst case derived by Algorithm 1. As
Fig. 13: Evaluation of the supply voltage noise in hybrid regulated voltage stacking system

(a) Supply voltage noise comparison between SC/LDO hybrid regulated and default voltage stacking

(b) Worst supply noise distribution

Table V: SM Core DVFS Frequency and Voltage Pairs

<table>
<thead>
<tr>
<th>Core freq. (MHz)</th>
<th>700</th>
<th>650</th>
<th>600</th>
<th>550</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core voltage (V)</td>
<td>1</td>
<td>0.95</td>
<td>0.91</td>
<td>0.87</td>
<td>0.46</td>
</tr>
</tbody>
</table>

regulator in voltage stacking outperforms the step-down voltage regulator in the single-layer PDS because the former only needs to shuffle the accumulated imbalanced part, usually within 20% of the layer power, whereas the latter delivers the total power. For example, in benchmark Transpose, only 11.8% and 2.9% of current are imbalanced current that goes through CR-IVR and CR-VRM respectively, and causes 3.7% and 1.1% of power loss in switched capacitor CR-IVR and CR-VRM respectively.

C. Compatibility with Advanced Power Management

First we leverage the common and classic DVFS algorithm proposed in \cite{77} to explore per-core DVFS on a voltage-stacked GPU system, which monitors and predicts the application status (compute bound/memory bound) to adjust each core and memory frequency. The SM core frequency and voltage pairs are shown in Table V. In conventional single-layer power delivery system, each core has its own frequency and voltage. In voltage stacking, the cores in each layer share a voltage domain and the highest voltage and frequency from the cores in one layer is used as the voltage and frequency for this layer. We evaluate DVFS on the voltage stacking and compare with DVFS on conventional single-layer power delivery system in Fig. 15. Although DVFS on voltage stacking causes more power loss than normal execution on voltage stacking, but it still has a higher power delivery efficiency than on conventional power delivery system at most benchmarks except Transpose. This is because GPU benefits the single instruction multiple thread (SIMT) architecture causing a synchronized activity and synchronized DVFS commands for the cores during most of time. Besides, shown in the right bars in Fig. 15, the power delivery efficiency guided hypervisor in Algorithm 3 can further prevent the aggravated power loss in CR-IVR and CR-VRM by limiting the occasional current imbalance from DVFS. Power delivery efficiency guided hypervisor can help voltage stacking achieve a near 90% power delivery efficiency under DVFS operations. The normalized energy consumption across benchmarks of conventional single-layer system, DVFS on conventional single-layer system and DVFS on power delivery efficiency guided voltage stacking is shown in Fig. 16. On conventional single-layer system, DVFS can reduce the energy consumption of cores across most benchmarks. On
TABLE VI: Power delivery system comparison

<table>
<thead>
<tr>
<th>Power Delivery System</th>
<th>Efficiency</th>
<th>Die Area</th>
<th>Reliable</th>
<th>Compatibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-layer VRM [46]</td>
<td>79.9%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-layer IVR [76]</td>
<td>85.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS IVR [18]</td>
<td>92%</td>
<td>88.3 mm²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS IVR (worst) [18]</td>
<td>92%</td>
<td>912 mm²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS Hybrid (this paper)</td>
<td>93.5%</td>
<td>99.2 mm²</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...and further analysis...
worst supply voltage is derived as an optimization problem. Based on the characteristics of supply voltage noise, a hybrid regulation, with distributed on-chip and a off-chip charge recycle voltage regulators, is proposed to effectively mitigate supply voltage noise. The supply voltage noise is guaranteed within a safe range even under the worst case. Also, the proposed hybrid regulated voltage-stacked system can not only be compatible with other power management techniques like DVFS and power gating but also maintains a high power delivery efficiency. Compared with conventional power delivery system, the proposed hybrid regulated voltage-stacked system achieves a 13.6% improvement of power delivery efficiency.

ACKNOWLEDGMENT

The research described in this paper was partly supported by NSF CPS grant CNS-1739643, NSF award CCF-1528045, Semiconductor Research Corporation (SRC) task 2810.003 through the University of Texas at Dallas Texas Analog Center of Excellence (TxA CE), and the National Natural Science Foundation of China (NSFC) 61702328. We are also grateful to the reviewers for their constructive feedback.

REFERENCES


