Efficient and Reliable Power Delivery in Voltage-Stacked Manycore System with Hybrid Charge-Recycling Regulators

An Zou[1], Jingwen Leng[2], Xin He[1], Yazhou Zu[3], Vijay Janapa Reddi[3], Xuan Zhang[1]
[1]Washington University in St. Louis, [2]Shanghai Jiao Tong University
[3]The University of Texas at Austin

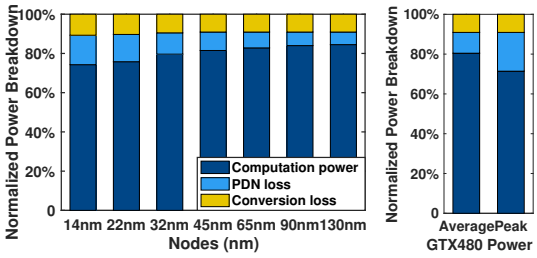
ABSTRACT

Voltage stacking (VS) fundamentally improves power delivery efficiency (PDE) by series-stacking multiple voltage domains to eliminate explicit step-down voltage conversion and reduce energy loss along the power delivery path. However, it suffers from aggravated supply noise, preventing its adoption in mainstream computing systems. In this paper, we investigate a practical approach to enabling efficient and reliable power delivery in voltage-stacked manycore systems that can ensure worst-case supply noise reliability without excessive costly over-design. We start by developing an analytical model to capture the essential noise behaviors in VS. It allows us to identify dominant noise contributor and derive the worst-case conditions. With this in-depth understanding, we propose a hybrid voltage regulation solution to effectively mitigate noise with worst-case guarantees. When evaluated with real-world benchmarks, our solution can achieve 93.8% power delivery efficiency, an improvement of 13.9% over the conventional baseline.

1 INTRODUCTION

Computers consume a non-trivial amount of the total electricity energy both globally and in the U.S [1]. A closer examination of the complete power delivery path reveals a provocative finding: transmitting and distributing electricity across tens or hundreds of miles in the grid to reach your power plug incurs only a 6% power loss [2], whereas *"the last centimeter"* from the PCB board to the microprocessor can waste more than 20% of the power [3, 4]. Two main power delivery losses contribute to the inefficiency: step-down voltage conversion loss when converting the higher voltage on the board to a lower supply voltage required by the microprocessor, and power delivery network (PDN) loss in transferring and distributing power from the off-chip source to various on-chip computing units [5, 6]. Generally speaking, both inefficiencies worsen with lower supply voltages, higher power densities, and higher power ratings, giving rise to the general trends depicted in Fig. 1.

Voltage stacking (VS), also known as charge recycling [7] or multi-story power delivery [8], is a novel technique to deliver power efficiently through a single high voltage source to multiple series-stacked voltage domains. The inherent voltage division among the voltage domains in series obviates the need for step-down voltage conversion and reduces the currents flowing through the PDN, resulting in improved power delivery efficiency (PDE). Ideally, if currents from all voltage domains are perfectly balanced, the input voltage will be evenly divided with no supply noise, and close to 100% PDE can be achieved [4, 6]. Unfortunately, under realistic loads, VS faces serious limitations due to exacerbated supply noise caused by current imbalance [4]. Although integrated switched-capacitor voltage regulators (SC-IVR) [9] can reduce supply noise in voltage-stacked systems, the large silicon area of the SC-IVR makes it impractical for high power-density applications.



(a) across technology nodes (b) GTX480 GPU

Figure 1: Power delivery efficiency breakdown and trends

Real-world power delivery systems require high efficiency, reliable operation against worst-case supply noise, and low implementation cost. A practical voltage stacking solution must satisfy all these requirements simultaneously. Towards this end, we propose a hybrid approach that combines the complementary effects of an off-chip charge-recycling voltage regulator module (CR-VRM) for slow and persistent current imbalances of large magnitudes and an on-chip distributed charge-recycling integrated voltage regulator (CR-IVR) to deal with fast transient current imbalances of smaller magnitudes. The proposed hybrid solution not only results in a 93.8% PDE, but also offers reliable operation with worst-case guarantee, and modest implementation costs geared towards practical and realistic system deployment.

Several key innovations and contributions are made in this paper:

- We formalize an analytical method to decompose noise-inducing current components based on superposition principle in linear circuits. The method fully captures correlated supply voltage noise from intra- and inter-layer core activities, employing a revised PDN circuit model that faithfully reflects the detailed power routing in VS systems.
- Based on the current decomposition method, we identify the principal causes of supply noise in voltage stacking as resonant global currents and low-frequency residual currents. We formulate a linear optimization algorithm to identify conditions that lead to worst-case supply noise in voltage-stacked manycore systems.

DAC '18, June 24–29, 2018, San Francisco, CA, USA

ACM ISBN 978-1-4503-5700-5/18/06…$15.00
https://doi.org/10.1145/3195970.3196037

- We propose a novel VS configuration with hybrid charge-recycling regulation using both on-chip distributed CR-IVR and off-chip CR-VRM to effectively mitigate supply noise. Supported by full-system simulation results, our solution achieves 93.8% PDE, outperforming conventional single layer system by 13.9%, also delivers reliable operation under worst-case conditions without negative performance impact.

2 BACKGROUND AND MOTIVATIONS

2.1 Power Delivery Subsystem

The conventional power delivery subsystem (PDS) in modern processors consists of a step-down voltage regulation module (VRM) on the motherboard, sockets, packages, off-chip decoupling capacitors and electrical connections at the board, package, and chip levels in the form of PCB traces, 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 the conventional PDS. Usually, voltage conversion using step-down VRM is necessary because the voltage level at the board is higher than the digital supply of a processor. Yet the inherent inefficiency of step-down VRMs means energy is lost during the voltage conversion. Resistive parasitics along the PDN path also contribute to energy loss, and incur a voltage drop across resistance, which is known as IR-drop. These two major efficiency loss are illustrated in Fig. 1 as percentages of power breakdown, suggesting that the their combined loss can approach more than 20% in advanced technology nodes and under peak power operations.

2.2 Voltage Stacking

In VS, the step-down VRM can be eliminated altogether by series-stacking voltage domains. In addition to eliminating step-down conversion loss, VS lowers the PDN resistive parasitics loss, because in a N-layer VS system, the PDN path current is reduced by $N\times$, which corresponds to a $N^2\times$ reduction in power loss. These efficiency improvements have been demonstrated in prior work [4, 6].

Although peak PDE close to 100% can be achieved using VS [4] under ideal conditions when all the stacking layers have balanced activities, and hence the same transient current demands, applying VS in real computing systems, where activity mismatches abound both spatially and temporally, proves to be challenging. Previous studies show that such activity mismatches can cause severe voltage fluctuations in a voltage-stacked system [4, 6]. The aggravated supply noise problem remains one of the most obstinate obstacles preventing VS adoption in the mainstream. Due to its impact on system reliability, supply noise[10] has been studied in the past for conventional single layer PDS in single-core, multi-core [11, 12], and manycore GPU systems [5, 13]. While circuit techniques[14] such as load line regulation are effective at taming IR-drop induced noise, dynamic Ldi/dt noise[15, 16] proves more dominant and harder to tackle. The interactions between the cores in manycore architectures lead to even more complex dynamic supply noise behaviors, which have only recently been characterized in GPUs [5, 17]. Yet, VS adds another vertical dimension to the complexity. To the best of our knowledge, apart from a few intuitive qualitative discussions [18, 19], there is no systematic quantitative noise characterization for multi-layer VS systems.

2.3 Related Work

Proof-of-concept circuits [7, 9] and silicon prototypes [4, 6, 8] have been presented to explore voltage stacking using low-power microcontrollers, along with design methodology for floorplanning and placement [20]. These pioneering works demonstrate the feasibility of voltage stacking, but are limited to simple assembly of uncorrelated cores with low power density. Inter-layer current imbalance has been discussed qualitatively as contributors to supply noise in VS systems, but without rigorous quantitative derivation of worst-case conditions. To overcome supply noise, most VS prototypes [4, 6] resort to employing integrated voltage regulators (IVR) to actively balance the current mismatches. Building on these early prototypes, a number of novel approaches have been proposed to take advantage of VS under different scenarios, such as 3D-IC with varying TSV, on-chip decoupling capacitance, and package parameters [21, 22]; optimal system partitioning to unfold CPU cores [18, 23]; and GPU systems with either supercapacitors [19] or operated under near-threshold voltages [23].

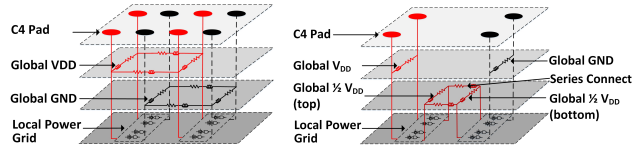
2.4 Motivations

Although existing research has explored many exciting opportunities and potential benefits of voltage stacking, none offers a practical path towards real-world manycore implementation. In addition to high efficiency, real systems emphasize reliability and implementation costs, but these two metrics have not received rigorous treatments in the previous work. For example, it is important to note that supply noise simulations across a subset of benchmarks do not provide sufficient and definitive evidence that the system would operate reliably under extreme worst-case conditions, yet many prior work relies on such optimistic assumption. Moreover, their proposed implementations can be unrealistic. When scaled to high-power density systems, the IVR area required to effectively balance the worst-case mismatched currents between the VS layers may exceed the core area, and technology remedies such as deep trench capacitors and on-chip supercapacitors [19] are far from being mature. Therefore, the most compelling yet unaddressed research task is to bridge the gap between ideal assumptions and practical implementations in order to firmly establish the credibility of VS in real manycore applications. In this paper, we tackle this crucial issue by proposing a hybrid approach to implementing charge-recycling regulators based on rigorous worst-case reliability analysis and without resorting to exotic process or technology.

3 MODELING METHODOLOGY

3.1 Power Routing-Aware PDN Model

The voltage stacking can be implemented in both 2D and 3D-IC chips, for a fair comparison with conventional power delivery methods, we focus on VS implementation in a 2D planar technology. As illustrated by the power/ground routing scheme in Fig. 2, 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



(a) conventional power grid routing (b) voltage stacking power grid routing

Figure 2: Power grid routing

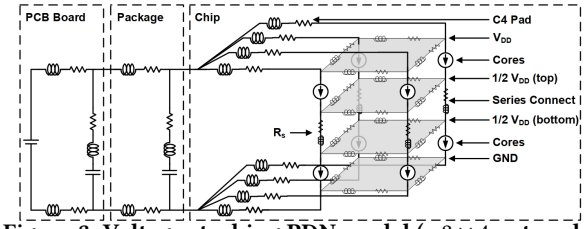


Figure 3: Voltage stacking PDN model (a 2×4 network).

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 can derive the corresponding PDN model for VS based on the typical RLC circuits and parameters introduced previously to study manycore system [5, 17]. Note that there are parasitic resistance (R_S) between the vertically-connected cores (modeled by current sources), as depicted in Fig. 3 in an example of a 2×4 VS PDN.

3.2 Analytical Supply Noise Model

Unlike previous empirical approaches [5, 17], we develop an analytical modeling framework to study and characterize voltage noise responses in VS PDN, especially in the presence of correlated/uncorrelated core activities.

3.2.1 Noise Decomposition & Superposition. Since the basic electrical model of VS PDN consists only of linear components, including RLC and ideal voltage and current sources, 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 stacking system that consists of N_L vertically-stacked layers with N_V cores on each layer. For example, Fig. 3 is a $N_L = 2$ and $N_V = 4$ VS 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:

$$I_{i,j}^{core}(s) = I^G(s) + I_i^{ST}(s) + I_{i,j}^R(s) \tag{1}$$

$$I^G(s) = \frac{\sum_{i=1}^{N_V}\sum_{j=1}^{N_L} I_{i,j}^{core}(s)}{N_V N_L} \tag{2}$$

$$I_i^{ST}(s) = \frac{\sum_{j=1}^{N_L} I_{i,j}^{core}(s)}{N_L} - I^G(s) \tag{3}$$

$$I_{i,j}^R(s) = \frac{(N_L-1)I_{i,j}^{core}(s) - \sum_{k=1,k\neq j}^{N_L} I_{i,k}^{core}(s)}{N_L} \tag{4}$$

where $I_{i,j}^{core}(s)$ is the current contributed by the core in the i^{th} stack and the j^{th} layer. It is decomposed into three components: $I^G(s)$, $I_i^{ST}(s)$ and $I_{i,j}^R(s)$, in Eq. (1) - (4). $I^G(s)$ represents the global current component shared by all the cores; $I_i^{ST}(s)$ represents the common current components shared by the cores in the i^{th} stack; and $I_{i,j}^R(s)$ is the residual current components after removing the global and per-stack common terms. Now, the voltage noise at the core (in the i^{th} stack and the j^{th} layer) can be expressed by superimposing the current components according to their respective effective impedances, Z_{eff}^G, $Z_{eff,i}^{ST}$, and $Z_{eff,i,j}^R$:

$$\Delta V_{core\,i,j} = \Delta V_{core\,i,j}^G + \Delta V_{core\,i,j}^{ST} + \Delta V_{core\,i,j}^R = I^G Z_{eff}^G + I_i^{ST} Z_{eff\,i}^{ST} + \sum_{i=1}^{N_V}\sum_{j=i}^{N_L} I_{i,j}^R Z_{eff_{i,j}}^R \tag{5}$$

To illustrate how the decomposition in Eq.(1) helps us analyze and characterize voltage noise effects in VS, we use a simplified RLC network of a 2×2 VS PDN, as shown in Fig. 4(a).

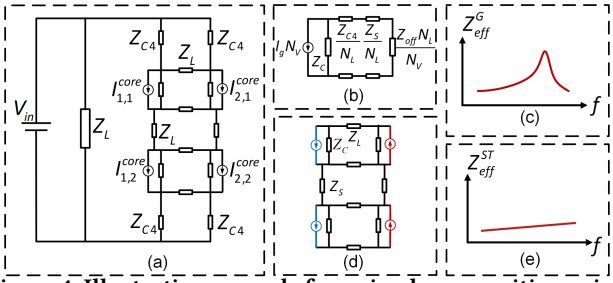


Figure 4: Illustrative example for noise decomposition using 2×2 voltage stacking network

3.2.2 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. 4(b) according to our 2×2 example. We can derive the voltage noise caused by I^G with an analytical expression for a general $N_L \times N_V$ network:[1].

$$\Delta V_{i,j}^G = I_{i,j}^G Z_{eff}^G = I_{i,j}^G \left(\frac{Z_{C4}}{N_L} + \frac{Z_S}{N_L} + \frac{N_V}{N_L} Z_{off}\right) // Z_C \tag{6}$$

Due to the uniform nature of the global current, all cores share the same common-mode, $\Delta V_{core\,i,j}^G$, and thus the same Z_{eff}^G. Eq.(6) also applies to the case when $N_L = 1$, which is a conventional single-layer PDN. From Eq. 6 and the typical impedance profile of Z_{eff}^G shown in Fig. 4(c), we can see that a $N_L \times N_V$ VS 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.

3.2.3 Local Uniform Through-stack Current. Following our definition of $I_i^{ST}(s)$, we can see that since $\sum_{i=1}^{N_V} I_i^{ST}(s) = 0$, 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. 4(d). For example, in our 2×2 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_i^{ST} and complex impedances in the form of Z_L and Z_C:

$$\Delta V_{core\,i,j}^{ST} = I_i^{ST} Z_{eff(i)}^{ST} = I_i^{ST} \frac{1}{N_L} [Z_C // Z_l] \tag{7}$$

where $\Delta V_{core\,i,j}^{ST}$ represents the voltage noise induced by I_i^{ST}, the common current 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_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. 4(d). The dividing ratio depends on the ratio of Z_L/Z_C, and in its high-frequency limit asymptotically approaches $1/N_L$. The analytical results of the local uniform through-stack current again suggest that by moving from single-layer to multi-layer, the voltage noise experienced at each core level and contributed by this current component is reduced by N_L times on average.

3.2.4 Residual Per-core Differential Current. On a closer inspection of Eq.(4), $I_{i,j}^R$ can be rearranged as the summation of differential currents in the form of $I_{i,j}^{core} - I_{i,k}^{core}$, $k \neq j$. The summation suggests that the remaining voltage noise effect, unaccounted for by the global and the local terms, ΔV^G and ΔV_i^{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

[1] symbol // is the circuit symbol for parallel connection

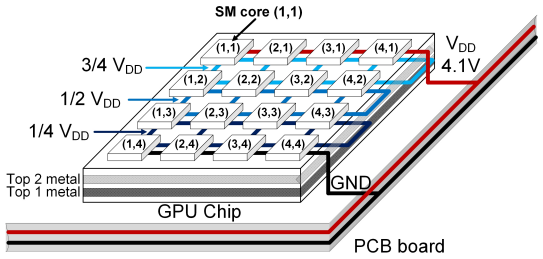


Figure 5: Layout and global power routing in VS GPU.

Table 1: VS GPU System Configuration

Configuration	Value	Configuration	Value
PCB supply voltage	4.1V	SM core supply voltage	1V
No. of SM cores	16	Clock frequency	700MHz
Voltage stacked layers	4	No. of SM cores per layer	4
Ave power per SM core	5W	Max power per SM core	14W
Threads per SM core	1536	Threads per warp	32
Registers per SM core	128KB	Shared memory	48KB

example, at $core(i, j)$, the noise from residual current is from its own residual current and other cores' residual current:

$$\Delta V^R_{core\,i,j} = I^R_{i,j} Z^R_{eff\,i,j} + \sum_{n\neq i}^{N_V} \sum_{m\neq j}^{N_L} I^R_{n,m} Z^R_{eff\,n,m} \quad (8)$$

where $I^R_{i,j} Z^R_{eff\,i,j}$ is the voltage noise caused by its own residual current, and $\sum_{n\neq i}^{N_V} \sum_{m\neq j}^{N_L} I^R_{n,m} Z^R_{eff\,n,m}$ is the voltage noise caused by residual current from other cores. Most importantly, the residual per-core differential current type is unique to voltage stacking, since these terms simply vanish when $N_L = 1$.

4 SUPPLY NOISE RELIABILITY

With the proposed analytical noise model, we are able to quantitatively study the root cause of supply noise in VS and identify the worst case conditions using a real manycore system.

4.1 System Configuration

We use a graphic processing unit (GPU) system of NVIDIA Fermi architecture as an example of a manycore system. Table I lists the system configurations. The Fermi architecture GPU has 16 streaming multiprocessor(SM) cores. We modify only the global power routing of the SM cores in the top metals from conventional single layer to 4×4 voltage stacking, without changing its physical layout. We number the 16 SM cores as (i, j): i is the stack number of the core, and j the layer number, as shown in Fig. Fig. 5. This voltage-stacked GPU system adopts the same off-chip and on-chip PDN equivalent circuit parameters as in GPUVolt[5, 24]. SPICE3 is used as the circuit simulator for calculating transient voltage noise.

4.2 Supply Noise Root Cause

Based on the above system configuration, we characterize its effective impedances, Z^G_{eff}, $Z^{ST}_{eff\,i}$, $Z^R_{eff\,i,j}$, of each current component defined in Eq. (5). 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 those of $core(1, 1)$. We divide the frequency range into low frequency (<10MHz), medium frequency (10MHz-50MHz), and high frequency (>50MHz). From these effective impedance curve, we can see that $Z^R_{eff\,i,j}$ at low frequency, and Z^G_{eff} 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 impedance can thus cause large supply noise, and we identify them as the dominant causes of supply noise in VS systems.

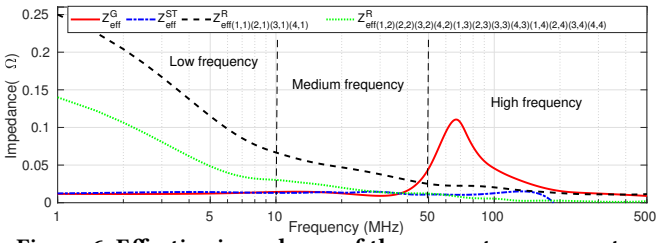


Figure 6: Effective impedance of the current components

4.3 Worst-case Noise Conditions

Identifying the noise root cause is not sufficient for rigorous reliability analysis. Furthermore, we consider what core activity conditions can result in the worst-case supply noise.

After characterizing Z^G_{eff}, $Z^{ST}_{eff\,i}$, $Z^R_{eff\,i,j}$ and establishing the relationship between ΔV_{core} as a function of these impedances, searching for the load current conditions that would result in worst-case supply noise can now be performed on the frequency domain. We formulate it as an optimization problem of finding the optimal frequency distribution of each core current $I^{core}\,i, j$ to maximize their combined effects on $\Delta V^{core}_{m,n}$. 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 $I^{core}_{i,j}(s)$. The optimization objective function is the voltage noise at $core(m, n)$ $\Delta V_{corem,n}$ and the constraints are from voltage noise decomposition Eq. (1) - (4) and peak GPU SM core power and as shown in Table 1.

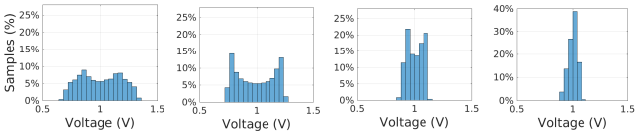
Algorithm 1 Maximize supply noise

Optimization Variables:
1: Each core current frequency distribution $I^{core}_{i,j}(s)$
Objective Function:
2: $\Delta V_{corem,n}$ in Eq. (5)
Subject to:
3: $\forall i, j; \quad 0 \leq I^{core}_{i,j}(s)$
4: $\forall i, j; \quad I^{core}_{i,j}(t) = \mathcal{F}^{-1}(I^{core}_{i,j}(s)) \leq$ peak power/current (14W/A)
4: $\forall i, j; \quad I^{core}_{i,j}(s), 0 \leq s \leq$ clock frequency (700MHz)
6: Eq. (1) - (4): current decomposition rules
End

The numerical solution of the linear programming problem based on the GPU configurations in Table 1 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 2. The currents, $I^{core}_{i,j=n}(s)$, are distributed at low frequency with major components of residual currents, while the currents, $I^{core}_{i,j\neq n}(s)$, are distributed at the resonant frequency of Z^G_{eff} with major components the global currents. This worst-case scenario is plausible in real GPU applications when the corresponding SM cores are alternating between NOP and Sine/Cosine special function instructions (SF Inst). We compare the worst-case noise derived by our optimization algorithm with three other current scenarios based on only heuristic understanding of the supply noise simulated in our VS GPU system: (1) all cores have low frequency residual currents, (2) all cores have high frequency global currents, and

Table 2: Core Current Frequency Distribution

Core Current	Frequency	Major Component
$I^{core}_{i,j=n}(s)$	low frequency	residual current
$I^{core}_{i,j\neq n}(s)$	high frequency	global current



(a) worst case (b) residual (c) global (d) random

Figure 7: Histograms of worst case and heuristic scenarios

(3) all cores have randomly distributed currents. From the supply noise histograms in Fig. 7, we can see that the worst case rigorously derived by our method is more severe than the heuristic ones, and more representative as stressmarks for supply reliability analysis.

5 SUPPLY NOISE MITIGATION

To combat the elevated and hard-to-predict supply noise and guarantee reliability in spite of worst-case conditions in voltage-stacked manycore systems, we explore a hybrid voltage regulation scheme using both on-chip charge-recycling integrated voltage regulators(CR-IVRs) and off-chip charge-recycling voltage regulator module(CR-VRM), shown in Fig. 8. This hybrid approach takes advantage of the unique merits of on-chip and off-chip voltage regulators and simultaneously avoids their individual shortcomings.

5.1 Distributed On-Chip Charge Recycling IVR

Located closer to the point-of-load, on-chip IVRs enjoy fast regulation response, but have limited capacity, making them suitable for reducing high-frequency noise of smaller magnitude. According to the analysis in Section 4, one of the dominant causes of worst-case supply noise due to high frequency global currents can be mitigated by on-chip CR-IVRs. Previous work has demonstrated multi-output switched-capacitor IVR [9] that balances the layer currents in VS systems, and we employ similar topology to implement our CR-IVRs. We disperse four distributed CR-IVR instead of one centralized one because such configuration has been proven to achieve better regulating effects [25, 26]. By moving charges across the stacking layers, the CR-IVR effectively behaves as an additional parallel impedance connected with the previous effective impedance Z^G_{eff}. It reduces the supply noise caused by global current as follows:

$$\Delta V^G_{corei,j} = I^G[Z^G_{eff}//Z^{CR-IVR}] \quad (9)$$

Here, Z^{CR-IVR} is the impedance of the distributed on-chip charge recycling voltage regulator. By deploying CR-IVR with desired impedance, $\Delta V^G_{corei,j}$ from global current ,I^G, can be effectively mitigated. We use a 4:1 multi-output switched-capacitor charge recycling voltage regulator, and its effective impedance can be expressed as

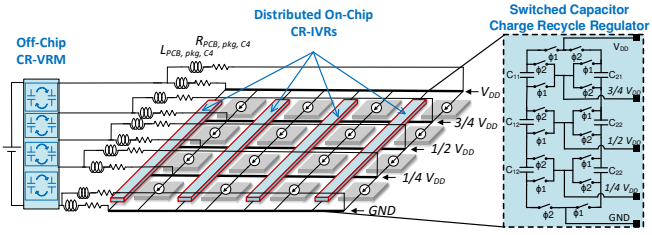


Figure 8: Hybrid voltage regulation based on distributed on-chip CR-IVRs and off-chip CR-VRM

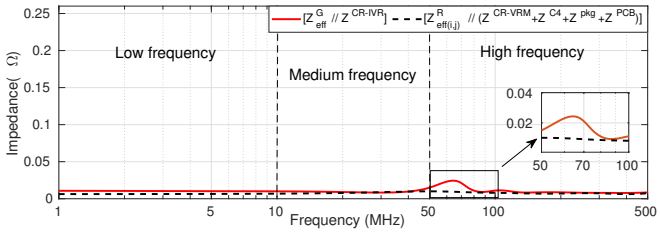


Figure 9: Effective impedance after deploying distributed on-chip CR-IVRs and off-chip CR-VRM

Table 3: Charge recycling voltage regulator design paras.

Design Parameters	On-Chip CR-IVR	Off-Chip CR-VRM
Number of VR	4	1
Switch frequency	$250MHz$	$500KHz$
Total capacitor per VR	$1.03uF$	$624uF$
Capacitor density	$50nF/mm^2$	$0.2uF/mm^2$
Switch on resistance	$130\Omega \cdot um$	$37600\Omega \cdot um$
Area per VR	$20.6mm^2$(Die)	$3.12cm^2$(Board)
Efficiency	77.7%	88.1%

$$Z^{CR-IVR} = \sqrt{Z^2_{SSL} + Z^2_{FSL}} \quad (10)$$

$$Z_{SSL} = \frac{1}{C_{total} f_{SW}} \left(\sum_1^n | a_{c,i} | \right)^2 \quad Z_{FSL} = \frac{G_{total}}{D_{cycle}} \left(\sum_1^n | a_{r,i} | \right)^2 \quad (11)$$

Here, C_{total} is the fly capacitance, G_{total} is the total switch conductance, f_{SW} is the switching frequency, and D_{cycle} is the duty cycle. Further, $a_{c,i}$ and $a_{r,i}$ are charge multiplier vectors [21, 27].

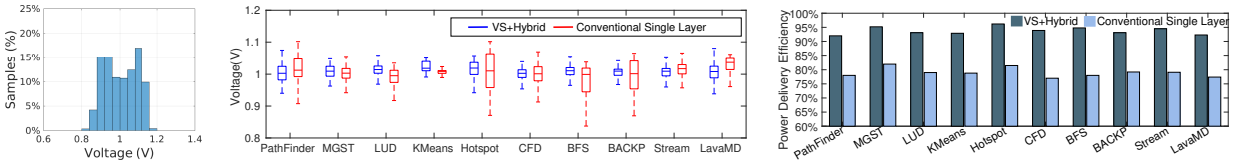
Many design parameter configurations can satisfy the desired CR-IVR impedance Z^{CR-IVR}. To obtain the optimal IVR efficiency, we follow the design methodology proposed by previous SC-IVR designs [9, 27], and arrive at an optimized CR-IVR. The design parameters are summarized in Table 3. After introducing the CR-IVR, the new effective impedance $[Z^G_{eff}//Z^{CR-IVR}]$ of global current I^G, is significantly reduced to mitigate noise caused by resonant global currents, as shown in Fig. 9.

5.2 Off-Chip Charge Recycling VRM

Compared with IVR, off-chip VRMs have slower response time, but they offer better efficiency [3] and do not consume expensive die area. It is important to note that although on-chip IVRs can be designed to provide similar regulating capacity as its 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 the supply noise at low frequency. Similarly, the addition of the CR-VRM results in an effective parallel impedance connected with the original $Z^R_{eff(i,j)}$ through the C4 pad, package, PCB. In this case, the supply noise caused by residual current becomes

$$\Delta V^R_{corei,j} = \sum_i^{N_V} \sum_j^{N_L} I^R_{i,j}[Z^R_{eff(i,j)}//(Z^{CR-VRM}+Z^{C4}+Z^{pkg}+Z^{PCB})] \quad (12)$$

Z^{CR-VRM} is the impedance of the off-chip charge recycling voltage regulator module; Z^{C4}, Z^{pkg}, and Z^{PCB} are the parasitic impedances of C4 pad, package and PCB board between CR-VRM and cores. Similar design optimization of CR-IVR is applied to arrive at an optimal set of design parameters, as summarized in Table 3. The new effective impedance of residual current after employing off-chip CR-VRM is shown in Fig. 9. With this reduced effective impedance, the supply noise, $\Delta V^R_{corei,j}$, is also significantly mitigated.



(a) worst noise distribution (b) supply noise comparison between VS+Hybrid and conventional single layer system across benchmarks (c) power delivery efficiency comparison between VS+Hybrid and conventional single layer system across benchmarks

Figure 10: Supply noise and power delivery efficiency evaluations of voltage stacking system with hybrid regulation

Table 4: Power delivery system comparison

Power Delivery Sys.	Eff.	Die Area	Reliable
Single layer+off-Chip VRM[19]	79.9%	N/A	√
Single layer+IVR[26]	85.8%	172.3mm^2	√
VS+CR-IVR (Worst)[4, 9]	92%	88.3 (912) mm^2	×(√)
VS+Hybrid	93.8%	82.4mm^2	√

6 FULL SYSTEM EVALUATION

In this section, we evaluate the performance of our proposed hybrid voltage regulation scheme in a full system setting. We first consider the worst-case reliability by using *Algorithm*1 to identify the worst supply noise magnitude and conditions when the previously-studied VS GPU system is equipped with the hybrid regulation solution. As demonstrated by the noise histogram in Fig. 10(a), after deploying hybrid regulation in the VS GPU system, the worst-case supply noise is limited within a range of $\pm 0.2V$, comparable with conventional single-layer PDS[2].

Next, we consider the noise behaviors of this VS GPU system under real world benchmarks and compare it with the conventional single layer GPU system, as shown in Fig. 10(b). Our results suggest that the hybrid voltage regulation scheme can also regulate voltage fluctuations in a voltage-stacked manycore system down to the same level achieved by the conventional single layer power delivery.

We also quantitatively evaluate the power delivery efficiency (PDE) across benchmarks in Fig. 10(c) and compare it with other existing and emerging PDS configurations in Table 4. Although charge-recycling voltage regulators with efficiencies of 77.7% and 88.1% are employed, the voltage-stacked system does not suffer large efficiency penalty, because most currents go through the vertically-stacked grid, without incurring energy loss at the regulators. Validated by benchmarks, the proposed VS system with hybrid regulation can achieve 93.8% power delivery efficiency on average. Compared with the conventional single layer PDS using off-chip VRM or integrated voltage regulators (IVRs), our solution improve full system efficiency by 13.9% and 8% respectively. Although previous VS+CR-IVR solution could provide similar efficiency but it is only tested with selected benchmarks and cannot guarantee the worst-case reliability. If the same design is scaled to handle the worst case, it would require a total of 912 mm^2 die area for the CR-IVR, which is 11× the area of our hybrid approach.

7 CONCLUSION

In this paper, we investigate a practical approach to enabling efficient and reliable power delivery in voltage-stacked manycore systems. Our hybrid regulation scheme leverages the complementary effects of fast on-chip CR-IVR and slow off-chip CR-VRM to mitigate the aggravated supply noise in VS and ensure reliability under even the worst-case noise conditions. Our solution not only achieves 93.8% power delivery efficiency, but also offers a low-cost realistic path of implementation with reliability guarantee, clearing the major hurdle for VS adoption in the mainstream.

[2]0.2 V is the voltage margin used for supply noise in NVIDA Fermi GPU[5, 28]

8 ACKNOWLEDGEMENT

This work was supported in part by NSF 1739643, NSF CCF-1528045, National Basic Research 973 Program of China 2015CB352403 and NSFC 61702328. We are also grateful to the reviewers for their constructive feedbacks.

REFERENCES

[1] EIA. Annual Energy Outlook 2016 with Projections to 2040.
[2] EIA. How much electricity is lost in transmission and distribution in the US?
[3] Xuan Wang et al. An analytical study of power delivery systems for many-core processors using on-chip and off-chip voltage regulators. *TCAD 2015.*
[4] Sae Kyu Lee et. al. A 16-core voltage-stacked system with adaptive clocking and an integrated switched-capacitor dc–dc converter. *TVLSI 2017.*
[5] Jingwen Leng et al. Gpuvolt: Modeling and characterizing voltage noise in gpu architectures. In *ISLPED 2014.*
[6] Kristof Blutman et al. A low-power microcontroller in a 40-nm cmos using charge recycling. *JSSC 2017.*
[7] Saravanan Rajapandian et al. Implicit dc-dc downconversion through charge-recycling. *IEEE JSSC 2015.*
[8] Pulkit Jain et al. A multi-story power delivery technique for 3d integrated circuits. In *ISLPED 2008.*
[9] Tao et al. A fully integrated reconfigurable switched-capacitor dc-dc converter with four stacked output channels for voltage stacking applications. *JSSC 2016.*
[10] Xuan Zhang et al. Supply-noise resilient adaptive clocking for battery-powered aerial microrobotic system-on-chip in 40nm cmos. In *CICC, 2013 IEEE.*
[11] Timothy N Miller et al. Vrsync: Characterizing and eliminating synchronization-induced voltage emergencies in many-core processors. computer arch. news.
[12] Xuan Zhang et al. Characterizing and evaluating voltage noise in multi-core near-threshold processors. In *ISLPED*, 2013.
[13] Jingwen Leng et al. Gpu voltage noise: Characterization and hierarchical smoothing of spatial and temporal voltage noise interference in gpu architectures. In *HPCA*. IEEE, 2015.
[14] Xuan Zhang et al. Evaluating adaptive clocking for supply-noise resilience in battery-powered aerial microrobotic system-on-chip. *IEEE TCAS I*, 2014.
[15] Vijay Janapa Reddi et al. Voltage smoothing: Characterizing and mitigating voltage noise in production processors via software-guided thread scheduling. In *MICRO*. IEEE, 2010.
[16] Vijay Janapa Reddi et al. Eliminating voltage emergencies via software-guided code transformations. *TACO 2010.*
[17] Renji Thomas et al. Emergpu: Understanding and mitigating resonance-induced voltage noise in gpu architectures. In *ISPASS 2016.*
[18] Ehsan K Ardestani et al. Managing mismatches in voltage stacking with coreun-folding. *TACO 2016.*
[19] Qixiang Zhang et al. Multi-story power distribution networks for gpus. date 16.
[20] Kristof Blutman et al. Floorplan and placement methodology for improved energy reduction in stacked power-domain design. In *ASP-DAC 2017.*
[21] Runjie Zhang et al. A cross-layer design exploration of charge-recycled power-delivery in many-layer 3d-ic. In *DAC 2015.*
[22] Kaushik Mazumdar et al. Breaking the power delivery wall using voltage stacking. In *GLSVLSI 2012.*
[23] Rafael T. Possignolo. Gpu ntc process variation compensation with voltage stacking. In *PACT 2015.*
[24] Jingwen Leng et al. Gpuwattch: enabling energy optimizations in gpgpus. In *ACM SIGARCH Computer Architecture News*, 2013.
[25] Pingqiang Zhou et al. Exploration of on-chip switched-capacitor dc-dc converter for multicore processors using a distributed power delivery network. cicc 2011.
[26] An Zou et al. Ivory: Early-stage design space exploration tool for integrated voltage regulators. In *DAC 2017.* ACM.
[27] Michael Douglas Seeman. *A design methodology for switched-capacitor DC-DC converters.* University of California, Berkeley, 2009.
[28] Jingwen Leng et al. Safe limits on voltage reduction efficiency in gpus: a direct measurement approach. In *MICRO*, 2015.