Transient Voltage Noise in Charge-Recycled Power Delivery Networks for Many-Layer 3D-IC

Runjie Zhang[†], Kaushik Mazumdar[‡], Brett H. Meyer[*], Ke Wang[†], Kevin Skadron[†], Mircea R. Stan[‡]
[†]Dept. of Computer Science, University of Virginia, Charlottesville, VA, USA
[*]Dept. of Elec. & Comp. Eng., McGill University, Montréal, QC, Canada
[‡]Dept. of Elec. & Comp. Eng., University of Virginia, Charlottesville, VA, USA
{runjie, km3sj, kewang, skadron, mircea}@virginia.edu, brett.meyer@mcgill.ca

***Abstract*—Aside from the benefits it brings, 3D-IC technology inevitably exacerbates the difficulty of power delivery with volumetrically increasing power consumption. Recent work managed to "recycle" current within the 3D stack by linking the different layers' supply/ground nets into a series connection. This charge-recycled (also known as voltage-stacked, or V-S) scheme provides a scalable solution for 3D-IC's power delivery because it supports an arbitrary number of layers with a constant off-chip current demand. Although prior work has studied the circuit implementation of a V-S power delivery network (PDN) and its current-reduction benefits, a whole-system evaluation of V-S PDNs' transient voltage noise and a noise comparison between the V-S PDN and the traditional PDN are missing. In this paper, we build a system-level model to examine voltage-stacked 3D-ICs' transient noise and explore the impact of different PDN design parameters and workload behaviors. Our results show that compared with the traditional PDN scheme, V-S provides stronger isolation for cross-layer noise interference, which in turn grants higher performance benefits for run-time noise mitigation techniques, such as dynamic margin adaptation. We observe that, compared with traditional PDNs, V-S PDNs provide up to 60% lower transient noise in the worst-case scenario. Furthermore, we show that V-S PDNs significantly reduce the packaging cost, because their noise is almost insensitive to the package impedance (e.g., a 300% impedance increase only raises worst-case noise by less than 0.3% Vdd).**

I. Introduction

Three-dimensional integrated circuits (3D-IC) make it possible to continue the historical trend of increasing device integration while maintaining high bandwidth, low latency and small form factor. Since the number of device layers in a 3D-IC stack is expected to grow, power density will inevitably increase. Unfortunately, the severity of the two major power-delivery-related reliability issues—supply voltage noise and electromigration-, or EM-induced power grid wearout—are directly related to the on-chip power density. Consequently, power delivery quality will become a limiting factor in the road towards many-layer 3D-ICs.

In response to the power delivery challenge caused by excessive current consumption, various research proposals [1]–[4] explored the idea of using a charge-recycled power delivery structure to support 3D-IC. Charge-recycling, or voltage-stacking (V-S), refers to power delivery that arranges multiple circuit blocks electrically in series. By connecting one block's ground net directly to the next one's power supply net, V-S power delivery network (PDN) "recycles" current between blocks. Blocks utilizing V-S PDN will share the same current, while their Vdd values are added. V-S provides a scalable solution for 3D-ICs' power delivery because by recycling current between layers, adding more layers to a 3D stack only requires higher off-chip supply voltage, while the current density within the PDN remains constant. This breaks the fundamental mismatch between 3D-IC's volumetric power dissipation and surface-limited (i.e., Controlled Collapse Chip Connection, or C4 array-based) power delivery.

With reduced current density in C4 bumps, through-silicon-vias (TSV), and on-chip wires, V-S significantly improves 3D-IC's robustness against EM-induced PDN wearout [5]. However, V-S PDNs are not guaranteed to have lower supply voltage noise compared with the traditional power delivery scheme, where all layers' power-supply and ground nets are connected with TSVs respectively: when the power consumption in the various layers are not perfectly matched, the voltages at the intermediate nodes in the V-S stack deviate from the nominal value. Therefore, explicit voltage regulation is required in V-S PDNs to compensate for the current-consumption mismatch between layers, and regulate voltages at the internal nodes. Based on circuit-level implementations and tests, prior research proposals have demonstrated the feasibility of using these explicit regulators in V-S PDNs [2,4]. However, the trade-off in voltage noise between V-S PDNs and traditional PDNs is not clear. To understand voltage noise in V-S PDNs under different workload conditions, to explore the impact of various PDN design parameters, and ultimately, to prove whether or when V-S PDNs have better noise quality than traditional PDNs, system-level modeling and analysis are required.

In this paper, we first design and validate a compact RC model for the voltage regulators in V-S PDNs. We then extend an open-source, system-level PDN model, VoltSpot version 1.0 [6], and integrate it with our regulator model, producing the first platform to enable whole-system, transient simulation for many-layer 3D-ICs' V-S PDN. This new version of VoltSpot has been released as version 2.0. Using an example low-power, ARM-based manycore 3D processor, we then compare the supply noise between voltage-stacked and traditional PDNs, and explore the impact of (a) cross-layer noise, (b) on-chip decoupling capacitance, and (c) package impedance. We observe that: 1. V-S provides stronger cross-layer noise isolation, increasing the effectiveness of run-time noise mitigation, and therefore system efficiency; 2. Under an area constraint for integrated capacitors, V-S provides up 60% lower worst-case noise amplitude; 3. V-S PDNs are less sensitive to package impedance. Consequently, we conclude that V-S achieves lower noise and lower cost compared with traditional 3D PDNs.

II. Background and Related Work

A. *Voltage Noise and Timing Margin in 3D-IC*

Supply voltage noise, which includes IR drop, LdI/dt, and LC resonance, refers to voltage fluctuation in the power-delivery network. Since transistor delay is directly proportional to source-to-drain potential differences [7], it is a common

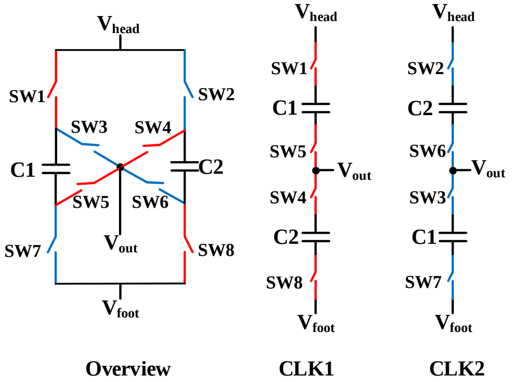


Fig. 1: A single cell of our 2:1 push-pull SC converter and its equivalent circuits in the two different clock phases.

design practice to assign a timing margin to critical paths to avoid noise-induced timing errors. Besides a design-time allocation that guards against the worst-case scenario, the timing margin can also be dynamically adjusted to improve system efficiency. For example, Lefurgy et al. [8] proposed to detect available timing margin at run-time with critical path monitors. Using digital phase-lock-loops, their scheme can rapidly change clock frequency to save energy during average-case execution (i.e., reduce margin) while guaranteeing functionality in the worst case (i.e., increase margin).

As more layers of active device layers are stacked together, the aggregate current demand increases, and the amplitude of the voltage noise grows proportionally with the layer count if the PDN impedance is kept constant [9]. To maintain a traditional PDN's robustness against voltage noise, 3D-IC designers will have to keep increasing timing margin (which degrades system performance with lower clock frequency), and/or reducing PDN impedance (which increases PDN cost with extra area overhead for on-chip decoupling capacitance or higher packaging complexity). Unfortunately, neither of these two approaches are scalable to many-layer 3D-ICs.

B. *Voltage Regulation in V-S PDN*

Although V-S significantly reduces 3D-IC's off-chip current demand [5], it introduces extra voltage noise caused by the workload imbalance between device layers. This is because, when layers are connected in series, the ratio of their effective resistances (which are inversely proportional to their power consumptions) directly affects the voltage levels of the intermediate nodes. Consequently, layers with higher power will experience greater voltage drops. To regulate this noise, prior work proposed using explicit regulators with V-S PDN [4,10]. Considering the rapid improvement of capacitive technology, we focus on switching-capacitor (SC) converters in this paper, due to their regulation efficiency [11].

Fig. 1 shows the detailed circuit structure of the V-S SC converter we adopt from the literature [4]. Each converter cell consists of two fly-capacitors (C1 and C2) and eight switches (SW1-8). By periodically interchanging the positions of the fly-caps (i.e., phase CLK1 and CLK2 in Fig. 1), the SC converter can either "source" or "sink" the charge difference between the stacked loads to regulate the voltage at its output. For a 2-layer system, this fixed 2:1 push-pull converter acts merely as a charge equalizer to assist the natural 2:1 voltage down-conversion of the stacked loads. For many-layer systems, we arrange the SC converters into a multi-output ladder structure to generate higher voltages. Similar to [5], we assume a fixed switching frequency for all converters to reduce design complexity.

C. *System-level Supply Voltage Noise Modeling*

In the past, researchers constructed system-level models to examine the supply voltage noise in both 2D ([6,12]) and 3D ([5,9]) chips. While prior work has demonstrated that stacking more layers of active silicon using the traditional PDN structure will monotonically increase on-chip noise [9], it is still not clear whether, or in which scenarios the V-S scheme provides better power delivery quality (in terms of transient noise) for 3D-ICs. To answer this question, we build a whole-system evaluation platform for V-S PDNs by designing a compact RC model for SC converters and integrating it with a full-chip power grid model.

The topic of SC converter modeling has been discussed in the past. However, prior work either focused on the traditional 2D-IC case without voltage stacking ([13]), or only studied the static noise (i.e., IR drop) of SC converters ([5]). To the best of our knowledge, ours is the first work to model transient voltage noise in SC-converter-supported V-S PDNs and compare V-S PDNs with traditional PDNs.

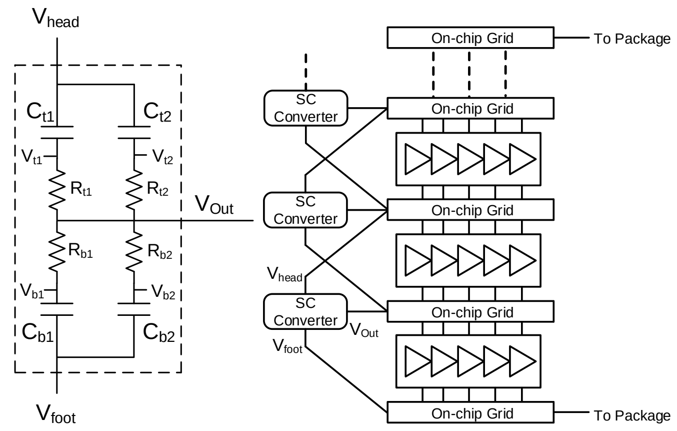
III. V-S PDN Modeling Methodologies

The power delivery networks of contemporary processors are usually large systems that contain up to several billion nodes, even in the context of 2D-IC. 3D integration and voltage stacking further increase the PDN's complexity with more device layers and new components such as TSVs and voltage regulators. For this reason, circuit-level simulations will be extremely computational-intensive and incapable of supporting whole-system design-space exploration studies. To enable a system-level study of V-S PDN's voltage noise, we design and validate a compact RC model for the SC converters and integrate it with a pre-RTL PDN model. This section discusses our modeling methodology and the validation results.

A. *A Transient Model for SC Converters*

Fig. 2a shows the compact RC circuit we use to model the interleaved SC converters. Each pair of top and bottom RC branches represent a cell of the converter that is controlled by a separate clock signal. At each clock edge, we exchange the position of the top and bottom fly-caps to model the switching activities of the converter cell. That is, we calculate $V_{head} - V_{t1}{}' = V_{b1} - V_{foot}$, where $V_{t1}{}'$ is the voltage value after the clock edge, while V_{b1} is the value before. Note that although we exchange the positions (i.e., electric charge) of the fly-caps at each clock edge, the resistance of each top and bottom branch is kept unmodified. This is because each time we "flip" the position of the fly-caps, we also change the set of switches to conduct the current (Fig. 1). Fortunately, the switches are designed in a symmetric way such that both the top and bottom RC branch in the two different clock phases have the same equivalent resistance [4]. Therefore, we can collapse the eight switches into two resistors (R_t represents SW1&5 and SW2&6, R_b represent SW4&8 and SW3&7) and reduce the model's complexity. From circuit simulations, we extract that $R_t = 4.208\Omega$, and $R_b = 4.68\Omega$ ($R_t \neq R_b$ because NMOS and PMOS have different channel resistances).

A common design technique to smooth the output ripple is to divide the single-cell converters into multiple sub-cells and interleave their switching clocks [4]. To model this structure, we simply instantiate a pair of top/bottom RC branches for each sub-cell, scale the capacitance values according to the number of total sub-cells, and shift the phase of each sub-cell's control clock. Fig. 2a illustrates an example model for a two-way interleaved SC converter. Similar to [4], we assume that all the sub-cells have identical structure, and therefore, the same RC values.

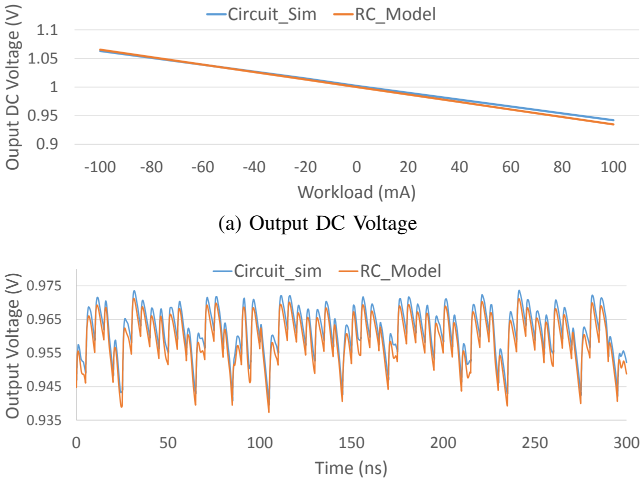


(a) Interleaved SC Converter (b) V-S PDN for Many-layer 3D-IC

Fig. 2: An RC model for the interleaved SC converters and a whole-system view of a V-S PDN in a many-layer 3D-IC.

B. Validation

We implement a 4-way interleaved, 2:1 push-pull SC converter in a commercial 28nm CMOS technology to validate our modeling methodology. It has an optimum switching frequency of 50MHz and a total capacitance of 8nF. Each SC converter can source/sink up to 100mA current to/from the load at a nominal voltage of 1V. Using the Cadence ADE environment and the Spectre simulator, we simulate this converter in a two-layer, voltage-stacked system (i.e., $V_{head} = 2V, V_{foot} = 0V$) and compare results against the output of our RC model.



(a) Output DC Voltage

(b) Output Transient Voltage Trace

Fig. 3: Validation results.

Fig. 3a shows the DC results comparison under constant workload conditions. Since the SC converter's output voltage is directly related to its output current, we attach an ideal current source directly to the V_{out} port and sweep the test cases from maximum sourcing (positive 100mA) to maximum sinking (negative 100mA). Under a constant workload, the output voltage shows a periodic rippling behavior caused by the converters' switching activities. Validation results show that with 1V Vdd, our model's maximum DC error 0.75%.

We also use a time-varying load current to validate our model. Fig. 3b shows the output voltage trace over 300 ns. The load current is sampled from Parsec 2.0 benchmark raytrace [14]; it induces an average current of 66.3mA in an ARM Cortex A9 core. Over the entire simulated time window, the output voltage trace of our model matches well with circuit simulation in term of DC component, AC amplitude, and slew rate. Overall, our model can capture the SC converter's transient output voltage with less than 72mV error at all times.

C. Whole-system Model

To study the interaction between the SC converters and the on-chip PDN grid, and to evaluate V-S PDNs' overall noise quality, we combine our SC converter model with an existing PDN model, VoltSpot [6]. VoltSpot uses a distributed RLC network to model the entire on-chip PDN metal stack, and a lumped RLC loop to model the chip package. Section IV-B will discuss the parameters we use and the modifications we made to VoltSpot in detail. Fig. 2b shows the structure of the whole-system model we build for many-layer V-S PDNs. For each SC converter, we connect its three ports (i.e., V_{head}, V_{out}, and V_{foot}) to three consecutive layers in the voltage-stacked power grids. We note that ideally, $V_{out} = (V_{head} + V_{foot})/2$, which indicates that any change in either V_{head} or V_{foot} will also affect the regulator's output voltage. Our model directly captures this inter-layer voltage dependency.

IV. Simulation Setup

A. Many-core 3D Processor Modeling

To study supply voltage noise in realistic 3D-IC design scenarios, we model an example many-core, many-layer 3D-IC based on a 40nm ARM Cortex A9 IP [15]. Using the architecture-level power and area model McPAT [16], we observe that when running at 1GHz with 1V supply voltage, each core has a peak power density of $172mW/mm^2$ (475 mW over 2.76 mm^2). Due to the power-efficient nature of these ARM processors, we can build our example many-layer 3D-IC without relying on aggressive, volumetric cooling solutions. With the help of pre-RTL flooplan tool ArchFP [17] and thermal model HotSpot [18], we evaluate the 3D stacks' maximum temperature and find that with a conventional air-cooling solution, we can stack up to eight layers of 16-core processors without violating the typical upper limit of 100 °C.

Although many-layer, especially many-logic-layer 3D-ICs, pose various fabrication challenges [3], the possibility of manufacturing 3D stacks economically has been exemplified by existing commercial products (e.g., the Micron hybrid memory cube with 4-8 layers [19]). To study the voltage noise in both short-term and long-term future 3D-ICs, and to evaluate how 3D scaling affects PDN design tradeoff, we build a series of example 3D systems with 2 to 8 layers. With 16 ARM cores per layer, the peak power consumption of these 3D processors ranges from 30.4W to 60.8W.

B. PDN Modeling

Besides integrating our SC converter model with VoltSpot, we also modify this 2D PDN model to support transient simulations for 3D-IC. Our major extension is an explicit resistor-inductor model for the TSVs. We adopt TSV parameters from prior work [20]. Similar to prior work [21], we ignore TSV capacitance in this paper, because it is usually orders of magnitude smaller than the on-chip and package decoupling

TABLE I: Primary PDN modeling parameters

Minimum C4 Pad Pitch (μm)	150
Single Pad Resistance/Inductance ($m\Omega/pH$)	10 / 7.2
Minimum TSV Pitch (μm)	10
Single TSV Resistance/Inductance ($m\Omega/pH$)	44.5 / 36.3
On-chip PDN's Pitch,Width,Thickness (μm)	810,400,720
Package Capacitance (μF)	7.3
Package Resistance/Inductance ($m\Omega/pH$)	0.054 / 10.8

capacitance. Other modeling parameters (Table I) are adopted from prior work [6].

By default, the VoltSpot version 1.0 utilizes ideal current sources to model the load (i.e., switching transistors). In order to model the voltage-stacked PDN organization, we replace the current sources with time-varying resistors. This is a necessary modification, because V-S PDN connects multiple layers of load in series, and using a resistive load model eliminates potential current source cutsets (if it exists, the solution is not unique) in the modeling circuit. The load resistance is calculated as $R = Vdd^2/Power$. This modification increases the model's computational complexity with more frequent LU-decomposition operations. This is because, unlike the original VoltSpot where the modeling circuit is time-invariant (only the current excitation changes), our model changes the load resistors over time to match the power consumption. To explore a broader design space within an affordable simulation time (e.g., 1 hour to simulate 1k cycles), we adopt the methodology from Huang et al. [21] and only simulate a "slice" of the entire 3D stack. Since each layer of our example 3D processor is a homogeneous 16-core ARM chip, we utilize the symmetry and simulate a reduced system of 2 cores per layer.

C. Workload Modeling

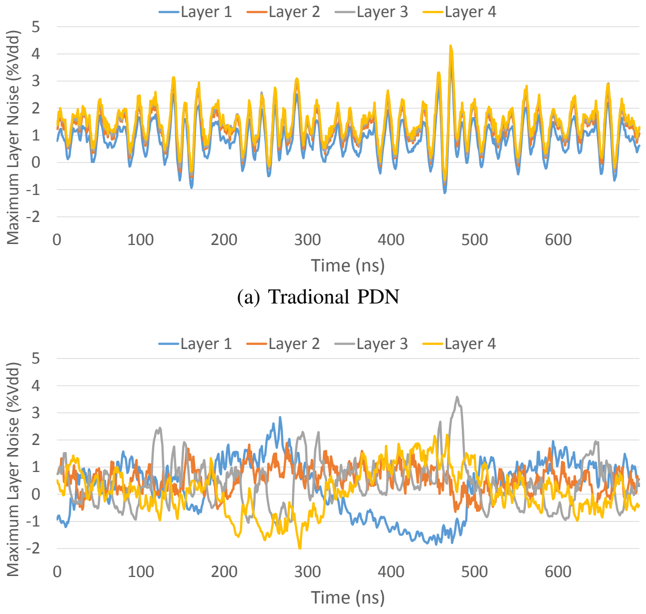
Using an integrated tool flow that combines McPAT with performance simulator Gem5 [22], we simulate the Parsec 2.0 benchmark suite [14] and extract dynamic power consumption traces to build realistic test cases for our noise study. Due to the limitation of PDN simulation's speed, we simulate 2k-cycle-long samples of power traces instead of whole-applications. To construct representative multi-layer workload behaviors, we first randomly collect a large number (i.e., 1000) of power samples from each benchmark, then profile each sample's average power consumption and maximum noise amplitude when running alone (on a 2D-IC). Section V gives more details about the workload we use in our study.

V. Results

A. Cross-layer Noise Interference

To study whether or how different layers' voltage variations affect each other in traditional and V-S PDNs, we pick one noisy workload and three less noisy ones from our sample pool and assign them to our 4-layer 3D processor. The first row in Table II shows each workload sample's maximum noise amplitude when running alone on a single-layer chip. Fig. 4 shows each layer's maximum voltage drop (%Vdd) over time.

In the traditional PDN (Fig. 4a), voltage noise in all layers is clearly highly correlated, a consequence of the layers' high-density, parallel interconnection. Supply voltage fluctuations in one layer affect the entire 3D stack through the vertical connections (i.e., TSVs). Conversely, the V-S PDN connects layers in series and regulates voltage levels with SC converters. Consequently, it breaks the inter-layer noise correlation (Fig. 4b). Table II shows each layer's maximum noise amplitude over the entire simulated time window. Compared with a 2D PDN, the traditional 3D PDN significantly reduces Task3's noise, because the decoupling capacitors (decap) on adjacent layers help to stabilize local voltage variation. However, other layers' voltage noise is also affected by Task3. In contrast, the V-S PDN isolates Task3's noise so that other layers have lower noise.



(a) Tradional PDN

(b) V-S PDN

Fig. 4: A plot of per-layer maximum noise amplitude over time. Only layer 3 has a noisy workload.

	Task1	Task2	Task3	Task4	Cross-layer
	Layer1	Layer2	Layer3	Layer4	Mean
Single-layer	4.0	3.0	10.9	2.8	N.A.
Traditional	3.7	4.2	4.2	4.3	4.1
V-S	2.8	1.9	3.6	2.3	2.7

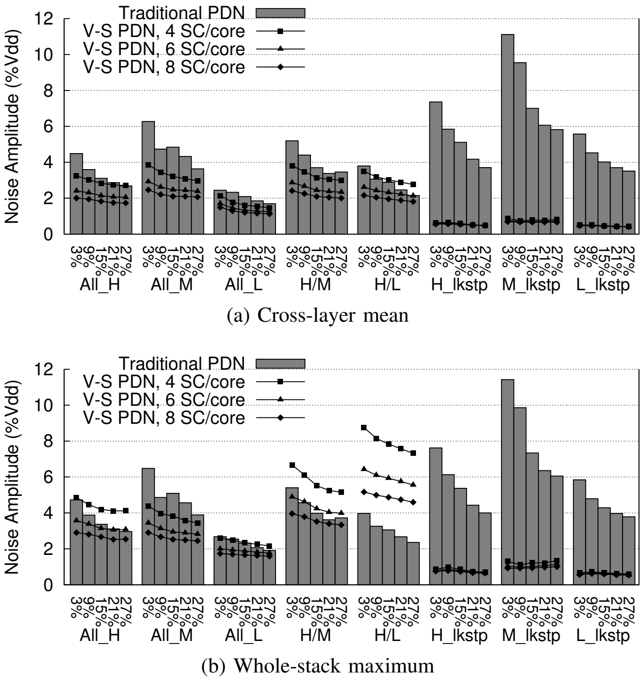
TABLE II: Maximum per-layer voltage noise (%Vdd) with different PDN schemes. The "cross-layer mean" value averages all layers' maximum noise amplitude.

With dynamic margin adaptation (see Sec. II-A, also reference [8]), each layer can adjust its timing margin according to its own maximum noise amplitude. Consequently, less noisy layers can run faster. Given the approximately linear relationship between noise amplitude and transistor delay, we assume that $x\%$ Vdd noise also requires an $x\%$ decrease in clock frequency. The last column in Table II shows the arithmetic mean of all four layers' maximum noise amplitude. This cross-layer mean metric shows the whole-stack's average slowdown when we use per-layer margin adaptation. By isolating the cross-layer noise interaction, V-S PDN can improve system performance with less slowdown. Since margin adaptation only slightly changes clock frequency (e.g., a few percent), we ignore its impact on processors' power consumption in this study.

B. Allocating On-Chip Capacitance: A Tradeoff Study

The on-chip integrated capacitors can serve as either explicit decap for both traditional and V-S PDNs, or as fly-caps for V-S PDNs' SC converters. Because of their high area overhead, the total amount of on-chip capacitance is usually limited. It is therefore important to understand the tradeoff between the allocation of explicit decap and SC converters in the V-S PDN before we compare the overall area overhead and voltage noise quality between the two schemes.

1) Workload selection: In order to understand 3D-ICs' voltage noise level under a wide range of workload conditions, we construct different scenarios to stress both traditional and V-S PDNs. Starting from our sample pool, we first sort all workloads by average power consumption and then select the top, medium, and bottom one-percentile samples as candidate-groups, categorized as high (H), medium (M), and low (L). Using these candidate groups, we build the following three classes of multi-layer workloads. The first class (All_H, All_M, and All_L) assigns different samples from the same group to different layers in the 3D-IC. The second class (H/M and H/L) selects samples from any two candidate groups and assigns them to the 3D stack in an interleaved fashion. This pattern is particularly stressful for V-S PDNs, because it forces all layers' SC converters to provide the same large amount of current, and the SC converters' output voltage drop is directly proportional to the load. In fact, the interleaved high-low (H/L) combination is the worst-case scenario for V-S PDNs. The last group (H_lkstp, M_lkstp, and L_lkstp) constructs a "lock-step" execution pattern by replicating the same workload to the entire stack. With all layers' power consumption changing simultaneously, this group will excite the largest LdI/dt and LC resonance voltage noise in the PDN. As an estimation for the worst-case scenario, we select the workloads with the highest single-layer noise within each H, M, and L candidate group.



(a) Cross-layer mean

(b) Whole-stack maximum

Fig. 5: A 4-layer 3D stack's voltage noise amplitude under different PDN configurations and workload conditions. The x-axis numbers within each data cluster represent the percentage of die area allocated for explicit decap. The size of each SC converter equals 3% of an ARM core.

2) Tradeoff study: Using our example 4-layer 3D processor, we simulated both V-S PDNs and traditional PDNs with different on-chip capacitance allocations. Fig. 5a shows the cross-layer-mean noise amplitude For both PDN schemes, we sweep the percentage of die area allocated for explicit decap (x-axis within each data group). For V-S PDNs, we assign different number of SC converters to each core (lines with different markers). We note that all SC converters have the same amount of capacitance and switching frequency. Using an advanced, high-density technology (e.g., trench capacitors [23]), each SC converter occupies $0.082mm^2$, which is 3% of an ARM core. Therefore, the V-S PDN's on-chip capacitance area equals $decap\_area + number\_SC\_percore * 3\%$.

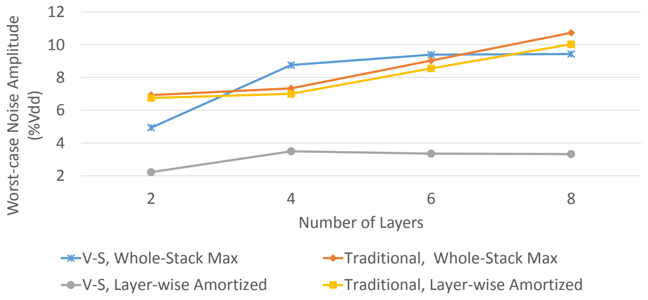


Fig. 6: 3D scaling's impact on worst-case voltage noise.

According to Fig. 5a, the V-S PDNs' overall noise is not as sensitive to the amount of explicit decap as traditional PDNs', especially in the lock-step scenarios, where the traditional PDN suffers from LC resonance. This is because the SC converters not only help to smooth local LdI/dt noise with the built-in fly-capacitors, they also isolate the on-chip PDN from the package RLC loop, so that the package LC resonance is greatly suppressed. Consequently, designers can significantly reduce the amount of explicit decap in V-S PDNs. If we compare two PDN designs with the same amount of on-chip area allocated for overall capacitance (i.e., a V-S PDN with 4 per-core converters and 3% decap allocation, and a traditional PDN with 15% decap allocation), we observe that under their respective cross-layer means, the V-S PDN's noise is significantly lower than the traditional PDN's. This means that if per-layer runtime margin adaptation is used, the performance loss will be significantly lower for V-S.

Fig. 5b shows the maximum noise amplitude observed in any layer for all test cases. The observation that the V-S PDN's cross-layer mean noise (Fig. 5a) is significantly lower than its global maximum noise (Fig. 5b) further proves the superior cross-layer noise isolation of V-S. This suggests that if a static worst-case noise margin is used, the V-S PDN will be worse. V-S PDN performance is only better when we utilize the per-layer dynamic margin adaptation.

C. Impact of 3D Scaling

To explore the effect of 3D scaling (i.e., stacking more layers) on both the V-S and traditional PDN's noise, we simulate our example 3D processors with two to eight layers, using the eight workload combinations. To make fair comparisons, we pick the design points described in Sec.V-B that allocate a 15% on-chip area for capacitors in both PDN schemes.

Fig. 6 plots all test cases' maximum noise amplitude (both whole-stack max and cross-layer mean) across all workload conditions. In general, stacking more layers together increases voltage noise in both types of PDNs. If a constant noise margin is applied to all layers at design time, this margin has to accommodate the worst-case whole-stack maximum noise. Consequently, the V-S structure requires smaller margin in 3D-ICs with 2 layers or more than 6 layers. With a per-layer dynamic margin adaptation technique enabled, the whole-stack's average margin will be no larger than the worst-case cross-layer mean value. As a result, V-S PDNs always require smaller timing margin, regardless of layer count. In the 8-layer 3D-IC, V-S PDN's noise is 60% lower than traditional PDNs'.

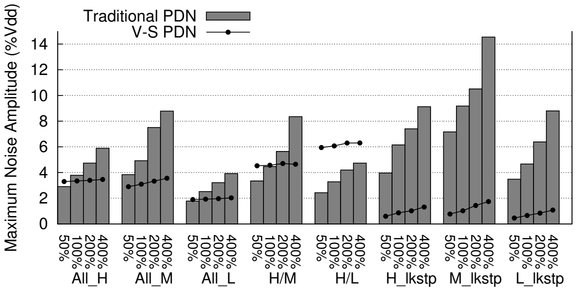


Fig. 7: Package impedance's impact on 3D-IC's whole-stack maximum noise. The x-axis of each data group shows the normalized package impedance.

One interesting observation is that a 2-layer V-S PDN's whole-stack maximum noise is significantly lower than the maximum noise of V-S PDNs with more layers. This is because in a 2-layer V-S PDN, the output voltage variations of the SC converters only affect one supply net (either foot-bounce or head-droop) of any layer while the other net is directly connected to the off-chip voltage source via C4 pads. As silicon layers are added, foot-bounce and head-droop can be added to the same layer, which significantly increases noise.

D. Impact of Package Impedance

Chip package impedance has a significant impact on the supply-voltage noise [24]. Although package designs with lower impedance can provide more current with lower noise, they usually have higher cost due to their increased complexity (e.g., more layers of power planes to reduce the package resistance and inductance, more package decap, etc.). To explore the impact of chip package quality, we simulate a series of package designs by applying different scaling factors to the resistance, inductance, and capacitance values of our PDN model's lumped package model. For example, to get a package with 200% impedance, we double the baseline package model's RL values (listed in Table I) and reduce the package capacitance by half. We note that this scaling factor does not change the package RLC loop's resonance frequency.

Fig. 7 illustrates how package impedance affects both V-S and traditional PDNs' noise in a 4-layer 3D processor. Compared with the traditional PDNs, the maximum noise in V-S PDNs is much less sensitive to the package quality. For example, a 300% impedance increase only raises the V-S PDN's worst-case noise by 0.23% Vdd. Since the V-S PDN reduces off-chip current significantly, package impedance contributes much less noise overall. By relaxing the constraint on package impedance, the V-S PDN is expected to reduce the cost of 3D-IC packaging.

VI. Conclusions

In this paper, we build a whole-system PDN model to: 1. Examine voltage-stacked 3D-ICs' transient noise under different workload conditions; 2. Compare voltage noise between V-S PDN and traditional PDN in the context of 3D scaling; 3. Explore the impact of various PDN design parameters. Our simulation results show that, compared with a traditional PDN, the V-S PDN provides stronger isolation for the cross-layer noise interference, but suffers higher noise in the particular case of highly imbalanced workloads. This is mitigated if dynamic, per-layer margin adaptation is used to respond to severe noise. If so, V-S PDN can better reduce timing margin and improve system performance. Without incurring extra on-chip area overhead for the integrated capacitors, the V-S PDN's cross-layer-mean noise amplitude under the worst-case scenario is up to 60% lower than the traditional PDN. Furthermore, we observe that the V-S PDN allows lower packaging cost for 3D-ICs. Overall, we demonstrate that the V-S PDN provides a low-noise, low-cost, and scalable solution to the challenges of 3D-ICs' power delivery.

Acknowledgments

This work is supported by NSF grant CNS-0916908, and DARPA MTO under contract no. HR0011-13-C-0022.

References

[1] J. Gu and C. H. Kim, "Multi-story power delivery for supply noise reduction and low voltage operation," in *ISLPED*, 2005.

[2] P. Jain, T.-H. Kim, J. Keane, and C. H. Kim, "A multi-story power delivery technique for 3d integrated circuits," in *ISLPED*, 2008.

[3] S. S. Sapatnekar, "Addressing thermal and power delivery bottlenecks in 3D circuits," in *ASP-DAC*, 2009.

[4] K. Mazumdar and M. Stan, "Breaking the power delivery wall using voltage stacking," in *GLSVLSI*, 2012.

[5] R. Zhang, K. Mazumdar, B. Meyer, K. Wang, K. Skadron, and M. Stan, "A cross-layer design exploration of charge-recycled power-delivery in many-layer 3D-IC," in *DAC*, 2015.

[6] R. Zhang, K. Wang, B. H. Meyer, M. R. Stan, and K. Skadron, "Architecture implications of pads as a scarce resource," in *ISCA*, 2014.

[7] M. Saint-Laurent and M. Swaminathan, "Impact of power-supply noise on timing in high-frequency microprocessors," *IEEE Transactions on Advanced Packaging*, vol. 27, no. 1, 2004.

[8] C. R. Lefurgy, A. J. Drake, M. S. Floyd, M. S. Allen-Ware, B. Brock, J. A. Tierno, and J. B. Carter, "Active management of timing guardband to save energy in POWER7," in *MICRO*, 2011.

[9] M. B. Healy and S. K. Lim, "Distributed TSV topology for 3-D power-supply networks," *IEEE Transactions on VLSI*, vol. 20, no. 11, 2012.

[10] S. Rajapandian, Z. Xu, and K. L. Shepard, "Implicit DC-DC downconversion through charge-recycling," *IEEE JSSC*, vol. 40, no. 4, 2005.

[11] M. Steyaert, T. Van Breussegem, H. Meyvaert, P. Callemeyn, and M. Wens, "DC-DC converters: From discrete towards fully integrated CMOS," in *ESSDERC*, 2011.

[12] M. S. Gupta, J. L. Oatley, R. Joseph, G. Wei, and D. M. Brooks, "Understanding voltage variations in chip multiprocessors using a distributed power-delivery network," in *DATE*, 2007.

[13] P. Zhou, D. Jiao, C. H. Kim, and S. S. Sapatnekar, "Exploration of on-chip switched-capacitor DC-DC converter for multicore processors using a distributed power delivery network," in *CICC*, 2011.

[14] C. Bienia, "Benchmarking modern multiprocessors," Ph.D. dissertation, Princeton University, Jan 2011.

[15] ARM, www.arm.com/products/processors/cortex-a/cortex-a9.php.

[16] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "The McPAT framework for multicore and manycore architectures: Simultaneously modeling power, area, and timing," *ACM TACO*, vol. 10, no. 1, 2013.

[17] G. Faust, R. Zhang, K. Skadron, M. Stan, and B. Meyer, "ArchFP: Rapid prototyping of pre-RTL floorplans," *VSLI-SoC*, 2012.

[18] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, D. Tarjan, and K. Sankaranarayanan, "Temperature-aware microarchitecture," in *ISCA*, 2003.

[19] Micron, www.micron.com/products/hybrid-memory-cube.

[20] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ICs," *IEEE Transactions on Electron Devices*, vol. 57, no. 1, 2010.

[21] G. Huang, M. Bakir, A. Naeemi, H. Chen, and J. D. Meindl, "Power delivery for 3D chip stacks: Physical modeling and design implication," in *EPEP*, 2007.

[22] N. Binkert *et al.*, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, Aug 2011.

[23] C. Pei *et al.*, "A novel, low-cost deep trench decoupling capacitor for high-performance, low-power bulk CMOS applications," in *ICSICT*, 2008.

[24] M. Popovich, A. V. Mezhiba, and E. G. Friedman, *Power Distribution Networks with On-Chip Decoupling Capacitors*. Springer, 2008.