Reliable Power Delivery and Analysis of Power-Supply Noise During Testing in Monolithic 3D ICs

Abhishek Koneru, Aida Todri-Sanial, Krishnendu Chakrabarty

Reliable Power Delivery and Analysis of Power-Supply Noise During Testing in Monolithic 3D ICs

Abhishek Koneru†, Aida Todri-Sanial‡, and Krishnendu Chakrabarty†

†Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
‡LIRMM-University of Montpellier II/CNRS, Montpellier, France

***Abstract*—Monolithic 3D (M3D) integration offers significant performance, power, and area benefits. However, the design of a reliable power-delivery network (PDN) is challenging for M3D ICs due to high power density and current demand per unit area. In addition, the higher susceptibility of interconnects to electromigration and stress migration increases the complexity of PDN design. We propose a framework to design a reliable PDN for M3D ICs using accurate electrical and reliability models. We leverage genetic programming to explore the design space to optimize the resources dedicated for power delivery in order to achieve reliable operation. We also analyze power-supply noise (PSN) during scan-based testing and compare it with that observed during functional operation. We quantify the impact of PSN during scan-based testing on yield loss. Our results show that the PDN design obtained using the proposed approach significantly increases the reliability of at least 40% of the wire segments in the PDN. In addition, the proposed PDN design reduces the worst-case power-supply droop by 52.5% compared to a baseline PDN. The yield loss due to power-supply droop for the proposed design is also significantly lower compared to a baseline PDN.**

I. Introduction

Monolithic 3D (M3D) integration is an emerging technology in which sequential integration of transistor tiers enables high-density vertical interconnects, known as monolithic inter-tier vias (MIVs). The size and pitch of an MIV are typically one to two orders of magnitude smaller than those of a through-silicon via (TSV) [1]. Therefore, M3D integration can result in reduced area and higher performance compared to 3D die stacking.

Despite the above benefits, challenges related to physical design, power delivery, reliability and test need to be addressed before M3D can be adopted by industry [2, 3]. Reliable power delivery in M3D ICs is a major concern due to high power density and current demand per unit area. In addition, the high susceptibility of interconnects in M3D ICs to electromigration (EM) and stress migration (SM) increase the complexity of power-delivery network (PDN) design.

Although several recent studies on power delivery for M3D ICs have been reported, the reliability of PDN in M3D has not been explored. In [4], system-level modeling, and time-domain and frequency-domain analysis of PDN for M3D ICs were carried out. In [5], the impact of PDN on full-chip wire length, routability, power, and thermal effects in M3D ICs was studied. In order to minimize voltage droop in tiers farther away from the I/O pins, a partitioning method that assigns power-hungry blocks to the tier closest to the I/O pins was proposed in [6]. However, there is no previous work that optimizes the PDN to address reliability challenges.

Power-supply noise (PSN) during testing is another major concern for M3D ICs since it can cause defect-free chips to fail on the tester, i.e., yield loss. PSN is typically higher in magnitude during testing when compared to functional operation due to high switching activity during scan shift or capture [7]. Several power-aware test solutions have been proposed to carry out testing without violating specified power budgets [7, 8]. However, there is no previous work that analyzes PSN during test in M3D ICs.

In this paper, we propose a framework to design a reliable PDN for M3D ICs. This framework relies on accurate electrical and reliability models to guide PDN optimization. We also analyze PSN during testing and propose a statistical model for quantifying the impact of PSN on yield loss. The main contributions of this paper are as follows:

1) We use genetic programming (GP) to explore the design space to optimize the PDN in order to achieve reliable operation. The objective is to minimize the voltage droop on power-supply rails while satisfying constraints on resilience to EM and SM.
2) We analyze PSN during functional operation and compare it with that observed during testing. We also show that the worst-case PSN in the proposed PDN is significantly lower compared to a baseline PDN design (in both functional and test modes).
3) We quantify the impact of PSN on yield loss and show that the proposed PDN design reduces yield loss.

In order to evaluate the proposed framework, we carry out the complete M3D design flow and PDN optimization for a processor benchmarks, namely Leon3. We also place and route a baseline design with uniformly spaced minimum-width rails in each metal layer. Our results show that the PDN design obtained using the proposed approach significantly increases the reliability of at least 40% of wire segments in the PDN. We also carried out dynamic simulation to obtain the worst-case power-supply droop when a functional workload is executed on Leon3 for the proposed PDN and the baseline. We observed that the worst-case voltage droop is 52.5% lower for the proposed PDN compared to the baseline PDN.

The rest of the paper is organized as follows. Section II provides an overview of M3D technology and related prior work. Section III presents the electrical and reliability models for the PDN. The proposed PDN optimization flow in Section IV. We report simulation results in Section V. Finally, Section VI concludes the paper.

II. Background

Fig. 1 presents an overview of a typical PDN for M3D ICs. Significant resources from the uppermost metal layers in each tier are dedicated for power delivery depending on the number of supply voltages used in the design. In addition, the uppermost metal layer in the bottom tier is used to design landing pads for power vias from metal-1 (M1) in the top tier. Significant routing blockages can occur in the uppermost metal layer of the bottom tier if it is heavily used for power delivery. Because of these blockages, the signal MIV count and wire length get impacted, thereby adversely affecting performance.

Another challenge associated with power delivery is voltage droop in the PDN, which can lead to performance degradation

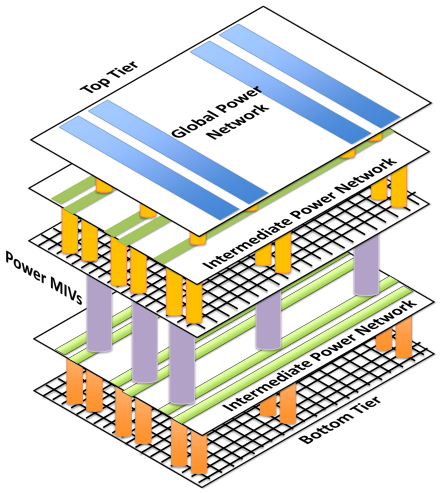

Fig. 1: Overview of a typical PDN for M3D ICs.

and failures during test and functional operation. There are two sources of voltage droop in the PDN: (i) IR drop caused by instantaneous current through the resistance of the PDN, and (ii) Ldi/dt drop caused by rapid changes in the current flowing through the PDN. In order to ensure reliable chip operation, the PDN should be designed such that the voltage droop due to IR drop or Ldi/dt drop does not exceed 10% of V_{DD} in the worst case [9].

EM and SM are critical challenges associated with the reliability of the PDN in M3D ICs. M3D ICs are more susceptible to EM compared to conventional ICs due to higher current density and thermo-mechanical stress [5]. Power-supply rails are more prone to EM compared to signal wires due to large unidirectional currents. In addition, the current flowing in the power-supply rails of non-bottom tiers is greater than that for the bottom tier since power is supplied from the top tiers, thereby making them more susceptible to EM. Therefore, the gradual transport of metal atoms in power-supply rails leads to the formation of voids or hill-locks, thereby increasing the resistance of these rails.

A. Power Delivery and Reliability Challenges

SM refers to the diffusion of vacancies due to thermo-mechanical stress. For an M3D IC, a high thermal budget step (around 500 ^{o}C) is used to activate dopants in the non-bottom layers [1]. Such high-temperature processing leaves the copper BEOL in the bottom layer with a large mechanical stress due to a mismatch in coefficient of thermal expansion of the materials involved. The stress can relax with time leading to the formation of voids and any residual stress can exacerbate EM. Since significant resources from the uppermost metal layers in the bottom layer are dedicated for power delivery, SM can impact the reliability of the PDN.

B. Related Prior Work on M3D Power Delivery

An M3D IC can be partitioned at the core-level, block-level, gate-level, and transistor-level. Significant research efforts are being directed towards gate-level design partitioning since transistor-level design partitioning requires extensive redesign of standard cells [6, 10] and block-level design partitioning does not fully exploit the benefits of M3D. In [11], a complete RTL-to-GDSII flow for gate-level M3D was proposed and the reduction in power compared to 2D designs was shown. However, the power delivery challenges were not considered in that work. The impact of tungsten BEOL in the bottom tier was evaluated with and without PDN in [10]. However, the primary focus of that work was to develop a physical-design flow and evaluate the impact of tungsten BEOL power and performance.

A system-level PDN model for analyzing IR drop and Ldi/dt drop in M3D ICs was presented in [4]. In addition, frequency-domain and time-domain analysis was carried out for a full-chip die model. However, this work only describes some guidelines for PDN design and does not address reliable power delivery. A methodology to optimize signal, power, and thermal interconnects in TSV-based 3D IC based on the design of experiments was proposed in [12]. However, this work considers only inter-die interconnects during optimization, and critical challenges associated with the reliability of PDN such as EM and SM were not considered. For M3D ICs, in addition to optimization of inter-tier interconnects, tier-specific optimization needs to carried out to address EM and SM challenges.

III. PDN Model

PDN design depends on the placement and routing of standard cells since power MIVs cannot be placed over active areas. This can lead to irregular power MIV placement, thereby causing significant IR drop in the bottom tier and increasing the susceptibility of non-bottom tiers to EM. In addition, thermo-mechanical stress in the uppermost metal layer during fabrication can lead to voids or exacerbate EM, thereby impacting PDN reliability. In order to evaluate these issues, we need to model the PDN and its reliability.

A. Electrical Model

The electrical model of the die-level PDN consists of the parasitics of the metal wires that constitute the PDN in the top and bottom tiers, and the parasitics of power MIVs. The die-level model can be extended with models for C4 bumps, package, and PCB to obtain a system-level model for PDN. The C4 bumps and power lines in the package and PCB are modeled as a series connection of resistors and inductors. From [13] and [4], we obtain the values of these resistors and inductors: (i) R_{C4} = 1 mΩ and L_{C4} = 10 pH; (ii) R_{PKG} = 10 mΩ and L_{PKG} = 100 pH; (iii) R_{PCB} = 5 mΩ and L_{PCB} = 1 μH.

B. EM Model

Failures due to EM and SM occur due to the formation of voids in a voltage-rail segment. These voids arise over time; therefore, EM signoff has to be carried out as a part of PDN design in order to ensure reliable operation over the lifetime. We obtain the worst-case current density for each voltage-rail segment using simulations and estimate the mean time to failure (MTTF) based on the current-density values. To facilitate our discussion of the EM model, we use the following notations: (i) T is the chip operational temperature; (ii) σ_T is the thermal stress in a metal line confined in inter-metal dielectric when it is cooled from zero-stress temperature T_{ZS} to T; (iii) L is the length of the metal line; (iv) D_a is the effective metal atomic diffusivity; (v) eZ is the effective charge of migrating metal atoms; (vi) B is the effective bulk elasticity modulus; (vii) Ω is the metal atomic lattice volume; (viii) ρ is the metal resistivity; (ix) j is the current density; (x) k_B is the Boltzmann constant.

The time for void nucleation due to EM in a metal line can be expressed as a function of the stress σ_T induced in that line [14]. The void nucleation time is defined as an instant

in time when the stress at the cathode end of the metal line reaches critical stress (σ_{crit}) and expressed as:

$$t_{nuc} = \frac{L^2 k_B T}{2 D_a B \Omega} \ln \left\{ \frac{\frac{eZ\rho jL}{2\Omega}}{\sigma_T + \frac{eZ\rho jL}{2\Omega} - \sigma_{crit}} \right\} \quad (1)$$

where $D_a = D_0 exp(-(E_D - \Omega^* \sigma_{crit})/k_B T)$. Here, E_D is the activation energy for vacancy diffusion and D_0 is the exponential prefactor. Note that this equation is employed only if the Blech limit [15] is not satisfied, i.e., $(j \times L) \geq \frac{\Omega \delta \sigma^i}{eZ\rho}$.

For each voltage-rail segment in the bottom tier, we obtain the worst-case current density using Monte-Carlo simulation and estimate the MTTF using (1). We use (1) since metal lines develop stress due to high-temperature processing of the top layer. Table I presents the values of the parameters used in (1). We estimate the MTTF of the PDN in the bottom tier as the MTTF of the weakest link in that tier.

For the top tier, we estimate the MTTF using Black's equation (shown below) since (1) is applicable only when an additional stress is induced. Therefore, we express the MTTF for a metal line in the top tier as: $\text{MTTF}_{\text{use}} = \text{MTTF}_{\text{stress}}(\frac{j_{stress}}{j_{use}})^{-n} exp\{E_a(T_{stress} - T_{use})/k_B T_{use} T_{stress}\})$, where $\text{MTTF}_{\text{stress}}$ is the MTTF in the accelerated-stressed condition. Here, T_{stress} is 600 K, j_{stress} is 3 MA/cm^2, and E_a is 0.86 eV [14]. The current density in a voltage-rail segment is obtained using simulation and the MTTF of the top tier is estimated as the MTTF of the weakest link in that tier.

TABLE I: Parameters used in the EM model.

Parameter	Value
E_D	0.65 eV
σ_{crit}	47.5 MPa
σ_T	46 MPa
Z	10
B	7.6×10^9 Pa
D_0	6.7×10^{-12} m^2/s

IV. PDN Optimization

We employ GP to intelligently explore the design space to optimize the resources dedicated for power delivery in order to minimize the impact on signal routing and ensure reliable operation of PDN. GP-based design space exploration (DSE) has been shown to be effective for architectural-level design optimizations [16]. Given a sample of collected performance metrics, approximation functions to predict the global performance of all candidate designs are built. As design spaces grow, it is more efficient to employ predictive approximation functions rather than a standard search algorithm. We employ GP to create appropriate approximation functions, and fit them to the collected performance metrics. Predictive models are built in GP-based DSE using the simulation results of a small sample of designs picked in advance.

The first step in the proposed optimization algorithm is to identify key features ϕ to represent a PDN and set the design space D. For the PDN-optimization problem, design features such as MIV count, MIV distribution, and resources for power delivery in the each tier can be used. The appropriate ranges for each design feature can be defined to set D. A small subset d of the design space is chosen and the performance metrics θ are extracted for that subset of design points. For the PDN-optimization problem, performance metrics such as maximum IR drop, MTTF, and total wire length can be used.

Input: Features ϕ for PDN designs in D, design space D, subset of design space d, performance metrics θ for PDN designs in D, function to evaluate quality of approximation Q, fitness function ψ, GP parameters χ and μ, maximum iterations Max
Output: Best approximation function f_{opt} for the mapping between features ϕ and performance metrics θ
Initialization: initial approximation function F_0, training set $Z \leftarrow \emptyset$, iteration $k \leftarrow 0$, an intial set of N randomly-generated approximation function F_0
1: ***Generate training data:*** For each design PDN_i in d, add $(\phi(PDN_i), \theta(PDN_i))$ to Z
2: **for** each approximation function f_i in F_0 **do**
3: Evaluate $Q(f_i)$ and $\psi(f_i)$
4: **end for**
5: **while** $\max\{\psi(f_0), \psi(f_1), \ldots, \psi(f_N)\}$ does not converge, where $f_i \in F_k$ or $k < Max$ **do**
6: Select $(1-\chi)$ approximation functions from F_k and insert in F_{k+1}
7: Select χ approximation functions from F_k, pair them up, produce an offspring, and insert in F_{k+1}
8: Select μ approximation functions from F_{k+1} and randomly mutate them
9: **for** each approximation function f_i in F_{k+1} **do**
10: Evaluate $Q(f_i)$ and $\psi(f_i)$
11: **end for**
12: $k \leftarrow k+1$
13: **end while**
14: **return** $\max\{\psi(f_0), \psi(f_1), \ldots, \psi(f_N)\}$, where $f_i \in F_k$

Fig. 2: Algorithm for PDN optimization using GP-based DSE.

GP is then used to find the best approximation function for the mapping between features ϕ and performance metrics θ. The algorithm is structured such that accurate candidate approximation functions are more likely to survive and be recombined. The candidate approximation functions are polynomial equations represented as expression trees. An individual node in a tree represents a design feature. The accuracy of a candidate approximation function is decided based on the quality of approximation Q and fitness function ψ.

$$Q(f_i) = \frac{\sum_{n=1}^{|d|} |f_i(\phi(PDN_n) - \theta(PDN_n)|}{\sum_{n=1}^{|d|} |f_i(\phi(PDN_n)} \quad (2)$$

$$\psi(f_i) = Q(f_i) - c * p^2 \quad (3)$$

The quality $Q(f_i)$ of approximation function f_i is obtained by calculating the sum of absolute differences between the approximation function's predictions and the collected performance metrics for all sample design points. Equation (3) presents the fitness function, where c is a constant used to control the size penalty and p is the number of tuning parameters. The fitness function ψ is used to penalize lengthy expression trees and keep the equations more compact.

V. Simulation Results

A. Simulation Setup

Simulations were performed on the Leon3 processor benchmark to evaluate the proposed technique. This benchmark was implemented using the Arizona State Predictive PDK (ASAP) 7 nm technology library. We used Cadence RTL compiler to carry out synthesis and Cadence Innovus to carry out placement and routing. We used six metal layers of the metal stack from ASAP 7 nm library for routing the benchmark.

We implemented the PDN for Leon3 before placed and routed the standard cells using Cadence Innovus. We then carried out RC extraction for that benchmark. We use the extracted RC model to carry out PDN IR-drop and reliability analysis using Cadence Voltus. We used a Python script to extrapolate the current density values obtained from Voltus to MTTF. This Python script implements the model described in Section IV.B. We implemented the GP-based DSE algorithm in Python. The code was run on a 64-bit Linux Server with a quad-core Intel Xeon 2.53 GHz CPU and 12 GB memory.

B. Design Flow

First, the design is partitioned into two tiers using a modified version of Shrunk-2D [17]. We start with a placed and routed 2D design that satisfies timing constraints. From this design, only the placement information is retained. Next, we scale the placement coordinates of each standard cell in the design by $1/\sqrt{2}$ since the footprint of an M3D design is half the footprint of the corresponding 2D design. This results in the overlap of standard cells. The standard cells are then partitioned into two tiers to remove the overlap by defining regular partitioning grids, and performing an area-balanced global min-cut [17]. We scale the placement coordinates instead of standard cell dimensions, as proposed in [17], since this method does not require modification to standard cell and technology LEF, which can be time-consuming and error-prone.

In order to carry out PDN design and signal MIV insertion using existing tools, we duplicate the metal stack in the technology LEF [18]. The original metal layers now constitute the metal stack for the bottom tier, and the duplicated metal layers constitute the metal stack for the top tier. In addition, the standard cell LEF is also duplicated, and the pins in the duplicated LEF are modified to the top tier metal layers. Next, we build the PDN with the full metal stack before inserting signal MIVs. This avoids the placement of signal MIVs at the location of power MIVs.

We then determine the location of signal MIVs by routing the design with a duplicated metal stack, and obtaining the location of vias that connect the uppermost metal layer of bottom tier with lowermost metal layer of top tier. More details about signal MIV insertion can be found in [18]. Note that duplication of the metal stack will lead to cell overlap during placement, but there will be no overlap during routing. However, we can reuse the placement of standard cells obtained from the 2D baseline. Next, we carry out initial routing, timing characterization, clock tree optimization, and timing-driven routing for each tier using a netlist corresponding to that tier and signal MIV locations. We then obtain placed and routed designs for each tier. These designs are then merged to a final M3D design, and timing analysis is subsequently carried out.

C. PDN Optimization

We considered four features to represent the PDN design space: (i) power and ground MIV count; (ii) maximum power and ground MIV area per placement tile; (iii) percentage of routing resources in the top tier that is allocated for power delivery; (iv) percentage of routing resources in the bottom tier that is allocated for power delivery.

We next define the design space. We restrict the power and ground MIV count to be no more than 5% and no less than 1% of the total number of MIVs between two tiers. We restrict the maximum power and ground MIV area to be no more than 5% and no less than 1% of the area of a single placement tile. We also restrict the total routing resources allocated for power delivery in the top and bottom tier to be between 5% and 25% of the total routing resources. We have chosen these ranges based on the values reported for these parameters in prior work [4–6].

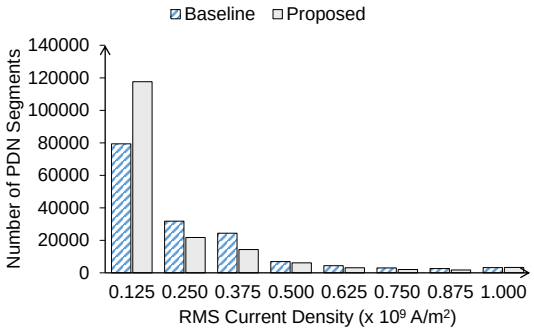

Fig. 3: Current-density histogram for baseline and proposed PDN designs for Leon3 benchmark.

We run simulations for 81 PDNs designs. These designs are obtained by setting the above four features to their min, max, and median values. For each design, we obtain the MTTF. Since existing EDA tools do not support 3D placement, we use the physical design flow described in Section III. The dimensions of the voltage-supply rails used in this paper are similar to the ones defined in the ASAP 7 nm technology library. Due to limitations associated with the license of Cadence Innovus available to us, the LEFs and QRC techfile are scaled by a factor of four.

Using the above features and MTTF values for 81 PDN designs as input, we run the genetic-programming algorithm to obtain the approximation function. For the Leon3 benchmark, we obtained the features of the optimal PDN design as: (i) MIV count: 2.4%, (ii) MIV density: 3.8%, (iii) total routing resources for power delivery in the top tier: 14.7%, and (iv) total routing resources for power delivery in the bottom tier: 12.3%.

D. PDN Reliability

We compare the reliability of the PDN design obtained using the proposed approach with a baseline PDN designed by uniformly spaced minimum-width rails in each metal layer. We then carry out signoff-quality RC extraction for the baseline and the proposed PDN designs. We use the extracted RC model to carry out PDN reliability analysis using Cadence Voltus. The thermal stress-aware electromigration model developed for M3D ICs is given as input to Voltus when we carry out reliability analysis.

Fig. 3 presents the current density histogram for the baseline and proposed PDN designs. We observe that the proposed PDN has a higher percentage of wire segments with current density less than 0.125×10^9 A/m^2 compared to the baseline PDN. Therefore, the PDN design obtained using the proposed approach significantly increases the time-to-failure of at least 40% of the wire segments in the PDN.

E. PSN Analysis

We also developed a framework to carry out dynamic simulation using Cadence Voltus. This framework relies on simulation waveform dumps of realistic workloads. In the proposed framework, we carry out a dynamic simulation of PDN

to obtain the voltage at every node in the design for a specified time window in the workload. Due to tool limitations, the length of the simulation window is limited to 1000 time steps. We executed three workloads from the MiBench benchmark suite, namely basicmath, qsort, and crc32 on Leon3. We obtained the signal activity from the execution of each workload in the form of a simulation dump (VCD file) generated by performing a post-synthesis simulation in ModelSim. We read the simulation waveform dump during dynamic simulation to obtain the switching activity for the gates in the design.

In order to analyze PSN during testing, we also carry out dynamic simulation of transition-delay patterns generated using the netlist obtained after the place-and-route step. We write out patterns in STIL format from test-pattern generation tool. We then convert the STIL patterns to Verilog testbench and carry out gate-level simulation using that testbench. We dump a VCD file when we carry out the gate-level simulation. This VCD file contains the switching activity for gates in the design while executing the pattern set.

Fig. 4 presents the supply voltage at the gate with maximum droop for three MiBench workloads. We observe that the worst-case power-supply drop is 59 mV, 46 mV, and 62 mV for qsort, basicmath, and crc32, respectively. Fig. 5 compares the supply voltage at the gate with maximum droop for the proposed PDN design and the baseline when basicmath is executed on Leon3. We observe that the worst-case voltage droop is 97 mV for the baseline PDN. Therefore, the proposed PDN design reduces the worst-case voltage droop by 52.5%.

Fig. 6 presents the supply voltage at the gate with maximum droop during scan shift and capture phases of test application. In this figure, the power budget during scan shift or capture specifies the number of flip-flops that can toggle at the same time. As expected, the worst-case power-supply droop reduces as we reduce the power budget for both scan shift and capture. We also observe that this is more significant during scan shift. This is expected since high switching activity during scan shift can cause a significant droop in supply voltage. On the other hand, we do not observe a significant change during capture; Leon3 is a single-clock design and the number of flip-flops toggling per pattern is much lower than the specified budget.

Fig. 7 compares the supply voltage at the gate with maximum droop for the proposed PDN design with the baseline when test patterns generated using 60% power budget are applied. The worst-case voltage droop is 73 mV and 59 mV for the baseline PDN during scan shift and capture, respectively. On the other hand, the worst-case voltage droop is only 55 mV and 41 mV for the proposed PDN design during scan shift and capture, respectively. Therefore, the proposed PDN design also reduces the worst-case voltage droop during testing.

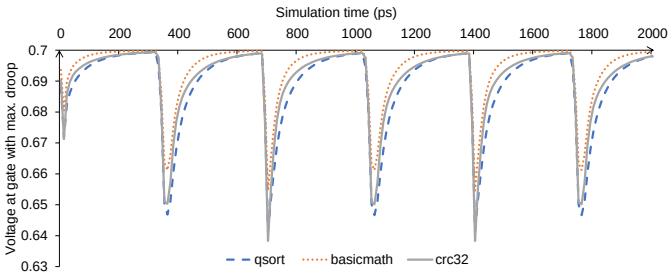

Fig. 4: Plots of voltage at gate with maximum droop obtained for different workloads.

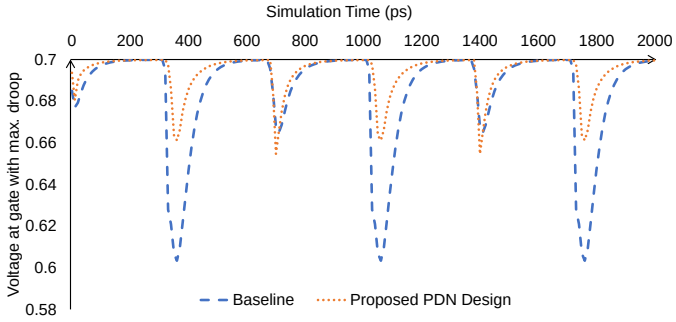

Fig. 5: Plots of voltage at gate with maximum droop obtained for the proposed PDN design and baseline when basicmath is executed.

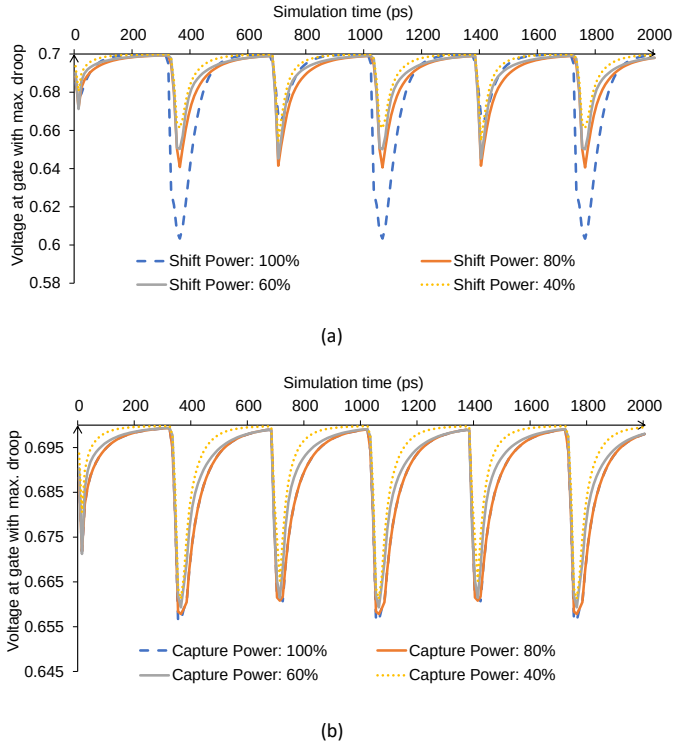

Fig. 6: Plots of voltage at gate with maximum droop obtained for different power budgets during: (a) scan shift; (b) capture.

F. *Analysis of Yield Loss*

We next quantify the impact of PSN on yield for our PDN design and the baseline. We use a statistical model to quantify the robustness of the design to PSN during scan-based testing. A design is robust to PSN during testing if all the scan flip-flops in the design are robust to PSN. An incorrect value captured at a scan flip-flop can cause the chip to fail at the tester.

The robustness of a flip-flop can be assumed to consist of two statistically independent events: (i) robustness to PSN during scan shift; (ii) robustness to PSN during capture. Therefore, the probability that an incorrect value is captured at a scan flip-flop FF_i, $P'(FF_i)$, can be defined as: $P'(FF_i) = 1 - P_S(FF_i) \times P_C(FF_i)$, where $P_S(FF_i)$ (or $P_C(FF_i)$) is the probability that a correct value is captured at the scan flip-flop FF_i during scan shift (or capture).

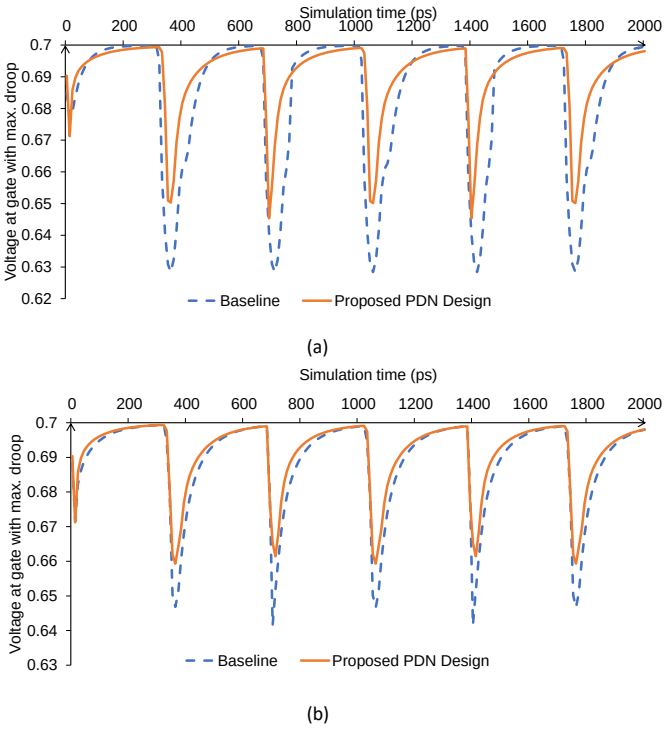

Fig. 7: Plots of voltage at gate with maximum droop obtained for the proposed PDN design and baseline for patterns generated using 60% power budget during: (a) scan shift; (b) capture.

To facilitate our discussion of the statistical model, we use the following terms: (i) shift path starts at the Q output of one flip-flop and ends at the SI input of the next flip-flop in the scan chain; (ii) capture/data path starts at the Q output of one flip-flop and ends at D input of the next flip-flop. Let us consider the design presented in Fig. 8(a). For FF_4 in this design, the path from Q of FF_3 to SI of FF_4 is the shift path for FF_4, and the paths from Q of FF_1 and FF_2 to D of FF_4 are capture paths for FF_4. There can be multiple capture paths for a flip-flop. However, there is only one shift path.

Under the assumption that all the gates in the shift path get the same supply voltage v, we can obtain the delay distribution of the shift path ending at each flip-flop in the design. Since delay of a path varies linearly with the supply voltage [19], the delay distribution for the shift/capture path ending at FF_i can be approximated as shown in Fig. 8(b). Let $D_S(v)$ denote the delay distribution of the shift path ending at FF_i. Let T_Smgn denote the timing margin on the shift path and V_{crit} denote the supply voltage at which the timing margin on

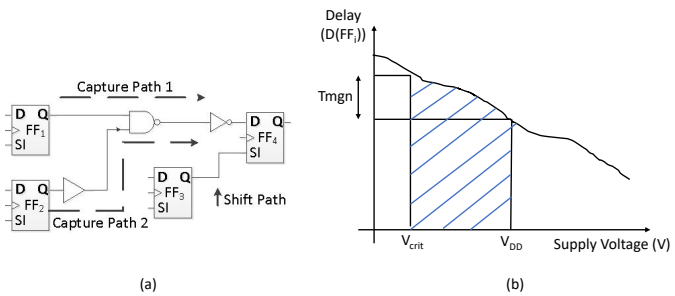

Fig. 8: Illustration of: (a) shift and capture path for FF_4; (b) delay distribution for shift or capture path.

that path becomes zero. The probability that a correct value can be captured at FF_i during scan shift can be expressed as: $P_S(FF_i) = \int_{V_{crit}}^{V_{Supply}} D_S(v)$. d(v).

Let $D_C(v)$ denote the delay distribution of the capture path with the smallest slack for FF_i, where v is supply voltage to all the gates in the fan-in cone of FF_i. Let T_Cmgn denote the timing margin on that path and V_{crit} denote the supply voltage at which the timing margin on that path becomes zero. The probability that a correct value is captured at FF_i can be expressed as: $P_C(FF_i) = \int_{V_{crit}}^{V_{Supply}} D_C(v)d(v)$.

From the above equations, we express the probability that an incorrect values can be captured at scan flip-flop FF_i as:

$$P'(FF_i) = 1 - (\int_{V_{crit}}^{V_{Supply}} D_S(v)d(v)) \times (\int_{V_{crit}}^{V_{Supply}} D_C(v)d(v)), \quad (4)$$

For a design with N scan flip-flops, we define the yield loss, YL, as: $YL = \sum_{i=1}^{N} P'(FF_i)$. This metric indicates the likelihood of a good chip failing on the tester. Smaller values of YL are clearly desirable.

In order to obtain YL for the proposed design and the baseline, we first obtain the timing of all the paths in the design using Synopsys PrimeTime. We also obtain the supply voltage at every flip-flop for the proposed design and baseline using Cadence Voltus and assume that all the gates in the shift and capture paths observe the same supply voltage. We then obtain the probability that an incorrect value is captured during shift and capture for all flip-flops using (**??**). We then obtain YL using (4) for the proposed design and the baseline as 8124.6 and 12100.2, respectively. We therefore conclude that the proposed PDN design reduces yield loss significantly.

VI. Conclusions

We have presented an approach to carry out design-space exploration for reliable power delivery in M3D ICs using GP. The proposed framework relies on accurate electrical and reliability models to guide PDN design and optimization. We have also analyzed the PSN during testing and compared it with that observed during functional operation. We have quantified the impact of PSN during testing on yield loss. Our results have shown that the proposed PDN design significantly increases the reliability of at least 40% of the wire segments in the PDN. In addition, our results have shown that the proposed PDN design significantly reduces the worst-case voltage droop and yield loss due to PSN compared to a baseline PDN.

References

[1] P. Batude et al. 3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS. *IEEE J. Emerging and Selected Topics in Circuits and Sys.*, 2(4):714–722, Dec 2012.

[2] K. Chang et al. Design Automation and Testing of Monolithic 3D ICs: Opportunities, Challenges, and Solutions: (Invited paper). In *ICCAD*, pages 805–810, Nov 2017.

[3] A. Koneru et al. Impact of Electrostatic Coupling and Wafer-Bonding Defects on Delay Testing of Monolithic 3D Integrated Circuits. *J. Emerg. Technol. Comput. Syst.*, 2017.

[4] K. Chang et al. Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs. In *ISLPED*, July 2017.

[5] S. K. Samal et al. Full Chip Impact Study of Power Delivery Network Designs in Gate-Level Monolithic 3-D ICs. *TCAD*, 36(6):992–1003, June 2017.

[6] S. Panth et al. Tier-partitioning for Power Delivery vs Cooling Tradeoff in 3D VLSI for Mobile Applications. In *DAC*, pages 92:1–92:6, 2015.

[7] S. Ravi. Power-Aware Test: Challenges and Solutions. In *ITC*, pages 1–10, Oct 2007.
[8] P. Girard et al. *Power-Aware Testing and Test Strategies for Low Power Devices*. Springer, 2009.
[9] N. James et al. Comparison of Split-Versus Connected-Core Supplies in the POWER6 Microprocessor. In *International Solid-State Circuits Conference*, Feb 2007.
[10] O. Billoint et al. A comprehensive study of Monolithic 3D cell on cell design using commercial 2D tool. In *DATE*, 2015.
[11] S. Panth et al. Design and CAD Methodologies for Low Power Gate-Level Monolithic 3D ICs. In *ISLPED*, pages 171–176, Aug 2014.
[12] Y. J. Lee and S. K. Lim. Co-Optimization and Analysis of Signal, Power, and Thermal Interconnects in 3-D ICs. *IEEE Trans. CAD*, 30(11):1635–1648, Nov 2011.
[13] S. Das et al. Modeling and characterization of the system-level Power Delivery Network for a dual-core ARM Cortex-A57 cluster in 28nm CMOS. In *ISLPED*, pages 146–151, July 2015.
[14] X. Huang et al. Physics-Based Electromigration Models and Full-Chip Assessment for Power Grid Networks. *TCAD*, 35(11):1848–1861, Nov 2016.
[15] I. A. Blech. Electromigration in thin aluminum films on titanium nitride. *Journal of Applied Physics*, 47(4):1203–1208, 1976.
[16] H. Cook and K. Skadron. Predictive design space exploration using genetically programmed response surfaces. In *DAC*, 2008.
[17] S. Panth et al. Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs. *TCAD*, 36(10):1716–1724, Oct 2017.
[18] S. Panth et al. Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs. *IEEE Trans. CAD*, 34(4):540–553, April 2015.
[19] A. Ramalingam et al. Robust analytical gate delay modeling for low voltage circuits. In *ASPDAC*, pages 6 pp.–, Jan 2006.