Power Delivery Solutions and PPA Impacts in Micro-Bump and Hybrid-Bonding 3D ICs

Lingjun Zhu, *Graduate Student Member, IEEE*, Chanmin Jo, and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*— Face-to-face bonded 3D integration has been shown to provide remarkable performance and power benefits for next-generation computing systems, but power delivery remains a challenge for its applications. In this article, we design industrial-level 3D processors with robust power delivery networks based on two types of 3D integrated circuit (3D IC) fabrication processes, micro-bumping 3D and hybrid-bonding 3D. Considering various process nodes and 3D bonding pitches, we develop a hierarchical physical design flow to build 3D ICs with various power delivery configurations and quantify the impacts of 3D bonding technology on the performance and power integrity. Our experimental results show that fine-pitch hybrid-bonding 3D ICs achieve up to 76% performance improvement or 17-mV IR drop reduction compared with the micro-bump 3D counterparts. Our in-depth analyses on critical paths show that intertier metal sharing is crucial for signal interconnects and power delivery in 3D ICs. In addition, we propose a decoupling capacitance sharing approach based on metal–insulator–metal (MIM) capacitors, which effectively reduces the dynamic voltage drop in micro-bump 3D ICs by up to 77 mV.**

***Index Terms*— 3D integrated circuit (3D IC), heterogeneous integration, physical design, power delivery network (PDN), power integrity.**

I. Introduction

THE emerging 3D integrated circuit (3D IC) technology has enabled system and IP integration beyond the capacity of traditional 2D ICs. It provides a cost-efficient solution to integrate more on-chip components and improve the performance of electronic systems after the slowdown of transistor scaling predicted by Moore's law. Commercial products, such as Intel's Lakefield and AMD's EPYC, have shown the performance and power benefits of 3D integration through mass production. In addition, recent developments in micro-bump and hybrid-bonding technology are proposing new possibilities and challenges for 3D IC performance improvement and power integrity.

In micro-bump 3D technology, two silicon dies are stacked together in a face-to-face (F2F) bonded fashion. The micro bumps consist of copper pillars and intermetallic compound (IMC) solder balls. The pitches of the micro-bump 3D ICs vary from 50 to 10 μm, as reported by recent studies [1]. The advantages of micro-bump 3D include its relatively low manufacturing cost and alignment requirement. However, the large size, pitch, and parasitics of micro bumps can limit the performance boost and power delivery reliability in 3D ICs.

Hybrid bonding is another practical approach to establishing F2F bonding in 3D ICs. It utilizes Cu–Cu direct bonding to create high-density 3D intertier connections. The bonding pitches of the hybrid-bonding pads can be lower than 1 μm [2]. These bonding pads introduce small parasitics into the system and provide more flexibility for signal routing and power delivery but also propose challenges for routing optimization and alignment. However, the impacts of the various 3D technology and bonding pitches on system performance and power integrity have not been studied thoroughly.

Power delivery network (PDN) design has been a challenge for 3D IC applications. Due to the stacking structure, 3D ICs have high power density and different on-chip power delivery paths compared with 2D ICs. Various studies [3], [4] have been done to analyze and compare the power delivery characteristics in face-to-back bonded 3D ICs using through-silicon via (TSV)-based 3D or monolithic 3D (M3D) technology. On the other hand, these studies have not quantified or addressed the challenges in F2F bonded 3D PDNs.

For 3D ICs, it is critical to perform technology, design, physical IP, and package co-optimization in order to achieve a balance between performance, power, area (PPA), and power integrity. Zhu et al. [5] proposed a method to quickly verify the 3D PDN design at the early design stage, but they have not considered the impacts of various 3D technology or evaluated the effects on the PPA of final layouts.

In this article, we use a commercial-level CPU design as the benchmark and build 3D ICs with various technology assumptions. For the large-scale designs and advanced process nodes, we propose a hierarchical flow to enable the physical design and PDN optimization. Based on these designs, we evaluate and compare the impacts of various PDN designs and 3D technology on the design PPA and power integrity. The main contribution of this work is to identify and quantify the power delivery challenges in micro-bump and hybrid-bonding 3D ICs, as well as propose a set of design guidelines to implement

Manuscript received 7 August 2022; revised 7 October 2022; accepted 15 October 2022. Date of publication 9 November 2022; date of current version 18 January 2023. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) ERI 3DSOC Program under Award HR001118C0096, in part by Semiconductor Research Corporation under Grant GRC Task 2929, and in part by Samsung Electronics, Inc. Recommended for publication by Associate Editor Z. Zhou upon evaluation of reviewers' comments. *(Corresponding author: Lingjun Zhu.)*

Lingjun Zhu and Sung Kyu Lim are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: lingjun@gatech.edu; limsk@ece.gatech.edu).

Chanmin Jo is with Samsung Electronics, Suwon 16677, South Korea (e-mail: chanmin.jo@samsung.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCPMT.2022.3221025.

Digital Object Identifier 10.1109/TCPMT.2022.3221025

robust 3D PDNs with regard to technology nodes, metal stack, and design characteristics.

This article is organized as follows. Section II introduces the related works on 3D PDN design and verification. Section III describes the design setup and 2D physical design methods to build the benchmarks. Section IV proposes the methodology for 3D PDN modeling and implementation. Section V demonstrates the experimental results of the 3D ICs and also provides in-depth analyses regarding the trend of PPA and power integrity in the 3D ICs with various technology assumptions. Section VI summarizes our observation and design guidelines for the robust 3D PDN design. Section VII concludes this article.

II. Related Works

Several physical design flows have been proposed to build micro-bump 3D and hybrid-bonding 3D ICs. Ku et al. [6] proposed a method to implement and optimize F2F-bonded 3D ICs based on commercial tools and scaled interconnect RC parasitics, which can cause 3D timing degradation due to RC estimation errors. Xu et al. [7] developed an approach to represent the 3D ICs in a 2D design database by placing the two tiers side by side and using anchor cells to optimize the 3D placement and routing (P&R) iteratively. However, the anchor cells cannot be used for power delivery, and the side-by-side representation cannot reflect the topology of 3D PDNs. Bamberg et al. [8] and Agnesina et al. [9] presented the design flows for memory-on-logic and logic-on-logic 3D ICs, respectively, but PDNs have not been included in the the designs. Kim et al. [10] implemented micro-bump 3D ICs with customized micro-bump placement, they did not consider power delivery, and the customized approach cannot be used when the bump number is large.

A large number of works have been done on PDN design and optimization in 2D ICs, including heuristic approaches [11], mathematical optimization-based approaches [12], and machine learning-based approaches [13]. These methods cannot be applied to F2F bonded 3D ICs directly since they have not considered the impacts of different 3D bonding technologies and changes in the PDN structure.

Chang et al. [4] provided a comprehensive study on the PDN analysis and optimization for M3D ICs. They modeled system-level M3D PDNs with three benchmark circuits, evaluated both the static and dynamic characteristics of the PDNs using a practical package model, and proposed two M3D PDN optimization approaches based on cell repositioning and asymmetric PDNs. Their results show that M3D suffers from high static IR drop due to extra metal layers and irregular power monolithic intertier via (MIV) placement. However, these conclusions cannot be directly applied to micro-bump and hybrid-bonding 3D ICs. This is because, unlike the latter two, the M3D ICs are face-to-back bonded, where the power delivery is connected from the package to the top metal layer of the top die, and the top die transistors can block the power MIV placement since it needs to penetrate the silicon substrate. Therefore, the power delivery challenges need to be evaluated and addressed separately for the F2F bonded 3D ICs.

TABLE I

Resistance of Unit-Length Wires and Default Vias in the Commercial 28- and 16-nm Metal Stacks. The Values Are Normalized Based on the M8 Wire Resistance at 28 nm

Unit resistance				
Components	Unit-length wire		Default via	
Process	28nm	16nm	28nm	16nm
M1	139.5	452.6	153.8	219.4
M2	174.2	566.6	153.8	219.4
M3	174.3	405.5	153.8	219.4
M4	174.3	297.9	153.8	183.1
M5	174.3	246.4	153.8	183.1
M6	174.0	246.4	153.8	183.1
M7	1.0	249.3	190.6	5.1
M8	1.0	1.1	-	-

A power delivery architecture was proposed in [5] for F2F bonded 3D ICs. The authors modeled power delivery scenarios in a multicore 3D system based on current regions, considering the impacts of various TSV technology. Then, they evaluated the voltage drop in the system using early rail analysis (ERA). However, they did not analyze the impacts of such 3D PDN designs on the PPAC of the final design since they have not implemented the design after P&R. In addition, the ERA cannot provide the details of voltage drop paths and potential bottlenecks because the standard cells and other physical IPs have not been placed. Therefore, more works need to be done to quantify the PPAC and voltage drop impacts of 3D PDNs and to optimize the PDN designs.

III. 2D Baseline Designs

A. Technology Setup

In this section, we describe the design and technology setup for our benchmark designs. We select two commercial process nodes for our design implementation: 28 and 16 nm. These process nodes have been widely used in 2D IC manufacturing and are proven to provide a high yield. However, few studies have been done to understand the impacts of the process nodes on the PDN quality of 3D ICs. In fact, the 28- and 16-nm nodes have very different metal stack specifications and parasitics, which affects the impedance of 2D and 3D PDN significantly.

Table I shows the comparison of wire and via resistance between 28- and 16-nm metal stacks. In the 28-nm metal stack, the bottommost six layers (M1–M6) are Mx layers, and the topmost two layers are Mz layers. The wires on Mx layers are much thicker and wider than those on Mx layers, and thus, they can provide much smaller resistance, while in the 16-nm metal stack, Mx layers include seven layers (M1–M7) with a variety of thicknesses. More importantly, on the lower Mx layers (M1–M3), the 16-nm metal stack shows more than 200% higher wire resistance and 25% higher via resistance. This is because the metal interconnects scale down in the advanced node, and the lower metal layer structure is specially designed for the FinFET technology. However, this resistance increase can cause challenges for PDN design since wire and via resistance contributes to a significant portion of the on-chip PDN impedance. Therefore, both 2D and 3D PDNs need to be

TABLE II

Design Setup for Our Cortex-A53 Processor. NEON Is an Advanced Single Instruction Multiple Data (SIMD) Architecture Extension, and FPU Is the Floating-Point Unit

Benchmark		
Design	Cortex-A53	
# core	4	
Instruction	32/64-bit	
Architecture	Armv8-A	
L1 cache	32 kB (per_core)	
L2 cache	2 MB (shared)	
L2 latency	1 cycle	
NEON	present	
FPU	present	
2D design setup		
Process	28nm	16nm
Contact poly pitch (nm)	117	88
Metal layer #	6M	6M
Footprint area	1.00	0.50
P/G C4 bump #	900	460
IR drop (mV)	10.39	14.25
2D PDN configuration		
Layer	M2, M5-M6	M1, M5-M6
Pitch (CPP)	32	32
Pitch (um)	3.744	2.816
Critical dimension (um)	1.0	1.0
Max track usage (%)	54%	71%

designed with regard to specific technology nodes and metal stack settings.

B. Design Configuration and Implementation

We use the Arm Cortex-A53 CPU as the benchmark design for the study. The Cortex-A53 CPU is a 32-/64-bit processor based on the Armv8 architecture. We configure the system with four CPU cores and a 2-MB L2 cache in order to explore the power delivery challenges in a practical multicore computing system. The detailed design specifications are shown in Table II.

We first implement the 2D Cortex-A53 CPU designs in both 28- and 16-nm process nodes as a baseline. Six metal layers are used in both 28- and 16-nm 2D for signal routing. The footprint area of the 16-nm 2D design is exactly 50% of the 28-nm 2D with similar core utilization, which shows the trend of area scaling down in the advanced technology nodes.

We build the basic PDNs in these 2D designs as a baseline for our 3D IC power delivery study. Fig. 1(a) shows the structure of the 2D PDN. The power source is connected to the chip through C4 bumps. We assume that the C4 bump pitch is 100 μm, and the bumps are attached to the topmost metal layers of each design. The power and ground (P/G) rails are mainly inserted on the topmost two metal layers (M5 and M6) to distribute the power supply. The default P/G rail pitch is set to 32 times the contact poly pitch (CPP) to satisfy the power integrity constraint based on the results reported in [14]. For 28-nm standard cells, the P/G pins are located on M2, while for 16-nm cells, the P/G pins are on M1 instead. The P/G pins of the SRAM blocks are on M4 for both process nodes. We create P/G rails and via pillars to connect the power grid to these pins.

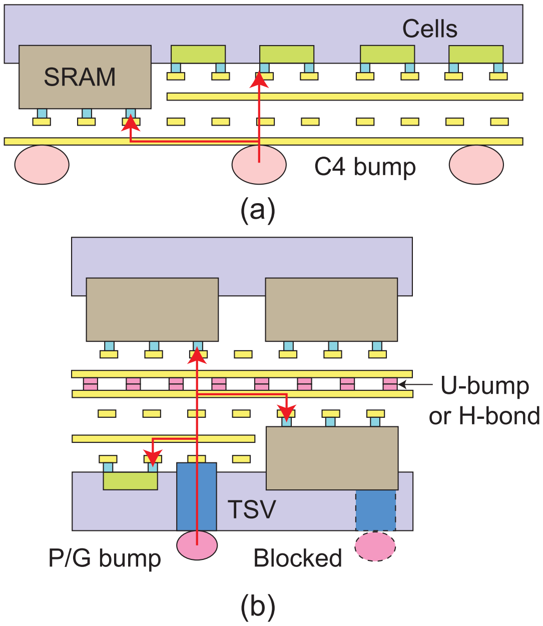


Fig. 1. Structure of the (a) 2D and (b) F2F bonded 3D PDNs. The red arrows show the potential power delivery paths.

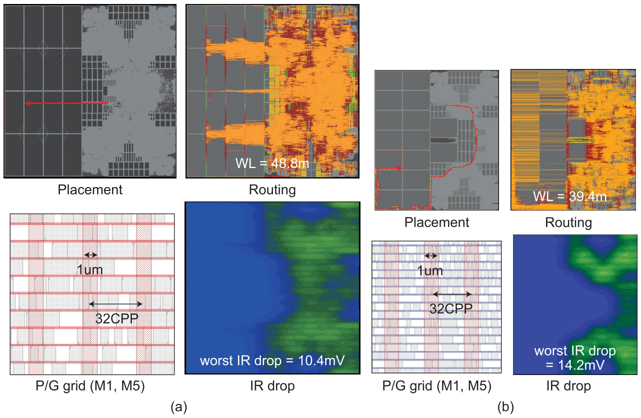


Fig. 2. Baseline 2D designs with (a) 28- and (b) 16-nm process. PDN is not shown in routing layouts for clarity. The IR drop map is generated by static IR analysis with 0.05 switching activity.

Fig. 2 shows the layouts and IR drop maps of the final 2D designs. The critical timing paths are highlighted with red arrows on the placement layouts. The critical paths are long register-to-memory paths in the 28- and 16-nm 2D designs. This demonstrates the potential to improve the performance of the system with 3D integration, which can help optimize these long global interconnect.

On the other hand, the 2D designs show marginal IR drop (<15 mV) and thus can be a baseline for 3D PDN design and evaluation. The least-resistive power delivery path to the IR drop hotspot is directly from the nearest C4 bump, going through the upper metal layers, via pillars, and down to the P/G pins on lower metal layers, as shown by the red arrows in Fig. 1(a). The results also show that the 16-nm design suffers from higher IR drop due to the more resistive metal layers. In addition, the 3D power delivery path can be very different from 2D [as shown in Fig. 1(b)] because of the C4 bump location, additional TSVs, metal stack structure, and so on.

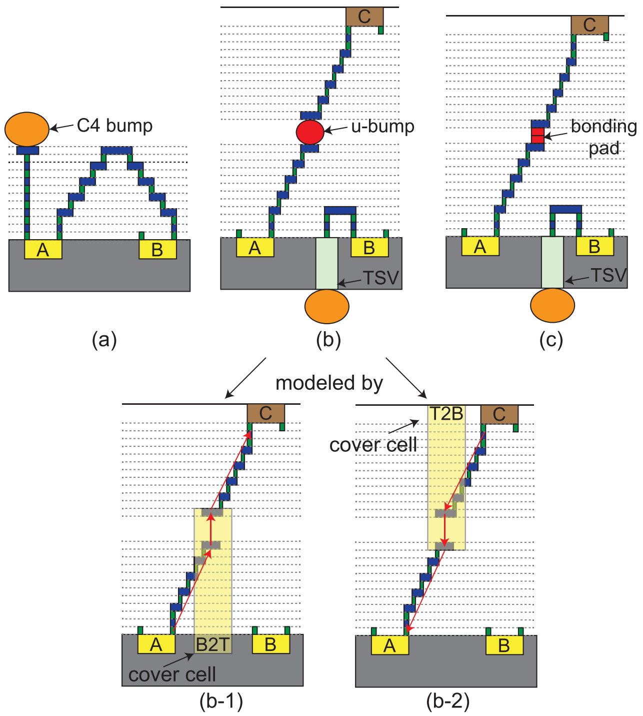


Fig. 3. 2D and 3D metal stack and signal interconnect structure (not to scale) of 28-nm process. The silicon substrate of the top tier is not shown for simplicity. (a) 2D. (b) Micro bump. (b-1) A drives C. (b-2) C drives A. (c) Hybrid bonding.

These differences propose challenges for 3D PDN designs in the advanced process nodes and motivate this study on power delivery bottlenecks and solutions.

IV. 3D Modeling and Implementation

In this section, we discuss our approaches to setting up 3D technology, modeling the 3D ICs with various bonding pitches, and designing a robust 3D PDN with regard to the process, metal stack, and design specifications. We consider memory-on-logic 3D integration in this study, in which large memory blocks are partitioned to the top tier, while standard cells and small memory blocks are partitioned to the bottom tier.

A. 3D Technology Modeling

The main idea of the 3D implementation flow is to build and optimize 3D ICs with commercial electronic design design automation (EDA) tools and generate high-quality design layouts. First, to represent the various 3D we create mock-up technology files for the EDA tools based on foundry data and simulation results. We generate the Library the Library Exchange Format (LEF) files for the 3D metal stack in both 28- and 16-nm process nodes. For process, we use six metal layers for the top tier and six metal layers for the bottom tier. For the 16-nm process, we assign six metal layers to the bottom tier and seven metal layers to the top tier in order to compensate for the *RC* the *RC* increase on Mx layers and increase the routing resources on the top tier. The 2D and 3D metal stack structure of 28-nm process is shown in Fig. 3(a)–(c). For parasitic extraction, we create the 3D process description files [Cadence Interconnect Technology (ICT)] to represent the 3D metal stack based files provided by the foundry. The thickness of the metal and dielectric layers in 16 nm is scaled from the 28-nm data 28-nm data since those values are confidential. After that, we extract the parasitics of the entire 3D metal stack Cadence Quantus QRC, which also includes the intertier parasitics between the top metal layers of the two tiers.

TABLE III

Dimensions and Parasitics of the Bumps/Pads in the 2D/3D ICs. The Small Inductance Values Are Omitted for the Bumps/Pads

Components	C4	TSV	u-bump		h-bond	
Dimensions						
Pitch (um)	100	10	25	10	5	1
Diameter (um)	80	7	12.5	5	2.5	0.5
Height (um)	80	70	12.5	5	0.17	0.17
Parasitics						
R (mΩ)	13	47	40	99	17	97
C (fF)	38	86	34	3	0.1	0.07
L (pH)	-	48	-	-	-	-

As shown in the cross-sectional view in Fig. 1, the C4 bumps can be directly attached to the top metal layer of the 2D IC. However, in the F2F 3D ICs, TSVs are needed to connect the C4 bumps to the bottom tier. We consider the via-middle process for TSVs in this study, so the TSVs are attached to the bottom metal layer (M1) of the bottom tier. The diameter and height of the TSVs are set to 7 and 70 μm, respectively. Also, we create a 10 × 10 μm keep-out zone (KoZ) for each TSV. The *RLC* parasitics of the TSV are estimated to be 47 mΩ, 48 pH, and 86 fF based on the empirical formula proposed in [15].

In addition, the intertier connections in the 3D ICs are established using micro bumps and hybrid-bonding pads. The dimensions and pitches of these bumps and pads can vary based on the 3D stacking technology, while they do not scale scale down according to the process node. In this article, we consider four different 3D connection settings: 1) bump with 25-μm pitch; 2) micro bump with 10-μm pitch; 3) hybrid bonding with 5-μm pitch; and 4) hybrid bonding bonding with 1-μm pitch. Settings 1) and 3) are based on the currently available 3D fabrication process for production, while 2) and 4) are the projection of future processes with smaller bump/pad sizes and pitches. Using ANSYS[1] HFSS, we simulate the electrical properties of these bumps/pads and extract their *RLC* parasitics for the 3D model. Table III summarizes the bump/pad dimensions and parasitics. The simulation results show that the smaller bumps/pads tend to provide smaller capacitance but also introduce larger resistance into the design, which can affect the intertier signal integrity and power delivery in 3D ICs.

The size and pitches of the hybrid-bonding pads are close to the regular vias in the process nodes, so we add them into the 3D metal stack as a special via between the topmost metal layers of the two tiers. The bonding pitch is also added to the technology LEF file as a design rule for the special via layer.

[1] Registered trademark.

By doing that, the pads can be placed automatically by the routing engine.

On the other hand, in micro-bump 3D, the bumps are much larger, and they need to be placed on a regular grid with a specific pitch constraint. Otherwise, the resulting design layouts cannot be manufactured or stacked properly. This regular placement cannot be handled by the traditional 2D routing engine automatically. Therefore, to model the signal intertier connections in micro-bump 3D, we create a special type of cells called "cover cells." Similar to the anchor cells presented in [7], the cover cells have two pins to represent the 3D connections, but they work for for the 3D metal stacks. Fig. 3(b-1) and (b-2) shows two cover cells: the bottom-to-top (B2T) is placed on the bottom tier and it represents the connection when the bottom-tier cell (cell A) drives the top-tier cell (cell C); the top-to-bottom (T2B) cell is placed on the top tier and it represents the connection when the top-tier cell (cell C) drives the bottom-tier cell (cell A). These cover cells do not have placement blockage or routing obstructions, so they have no direct impacts on P&R. In addition, we add the cover cells into the standard cell library (Synopsys Liberty format) and create a timing table to reflect the RC delay introduced by the micro bumps.

B. Hierarchical 3D Physical Design Flow

We enable the placement, routing, and PDN design for large-scale multicore systems with a bottom-up hierarchical implementation flow. The motivation is that the existing flattened design flows, such as Macro-3D [8] and Pin-3D [16], are not able to handle large-scale designs in advanced technology nodes. Especially with the routing congestion introduced by PDNs, the routing stage in these flattened design flows can last excessively long. A previous study on the hierarchical design flow, Hier-3D [9], however, has addressed the runtime problem but has not considered PDN implementation and the impacts of various 3D technologies. Therefore, we develop this physical design flow to enable PDN insertion in a hierarchical 3D IC design.

The overview of this design flow is shown in Fig. 4 and as described in the following.

1) Partition the top-level design into multiple blocks.
2) Build block-level designs with 2D or 3D technology settings.
3) Extract geometric and timing information to create block-level abstraction.
4) Use block abstraction for top-level 3D physical design.
5) Assemble the 3D IC by instantiating the block-level designs in the top level.
6) Perform timing and IR drop analysis on the assembled top-level design.

Note that the blocks can be 2D or 3D blocks. The 2D blocks contain instances only on one tier, and their routing space does not include the 3D interface. The 3D blocks may contain instances from both tiers and include the 3D interface (micro bumps or hybrid-bonding pads) in their routing space. The 2D and 3D blocks need to implemented with different technology files, but they can all be instantiated and assembled in the top-level design with the complete 3D technology setup and back-end-of-line (BEOL) stacks. Using this flow, we can implement a large-scale 3D system with a shorter runtime. The block-level designs can also be instantiated and reused multiple times.

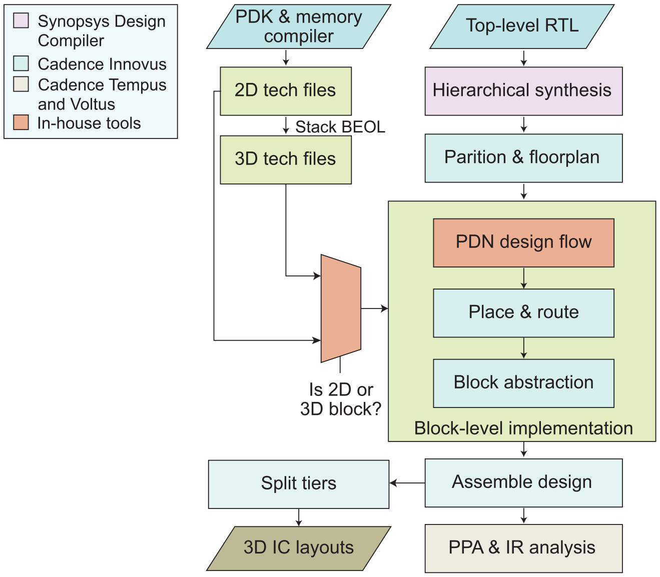


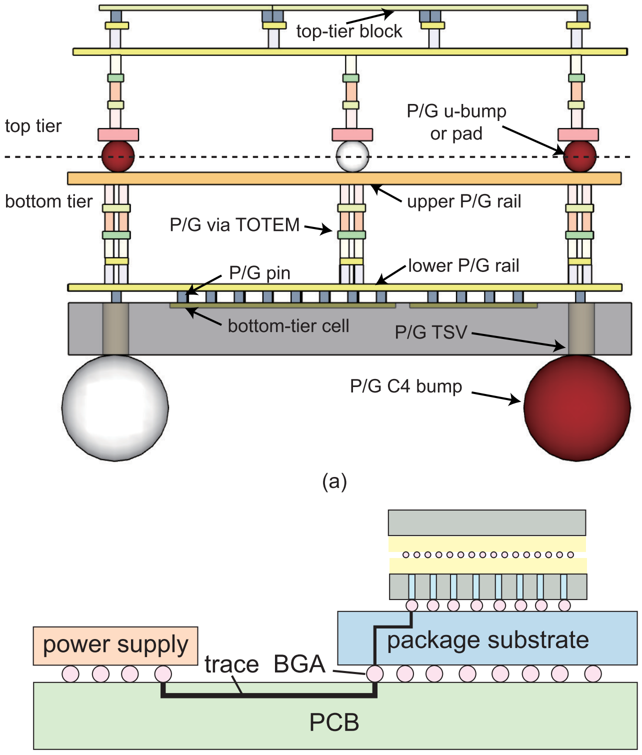
Fig. 4. Hierarchical physical design flow for our 3D ICs. The details of the "PDN design flow" are discussed in Section IV-C.

The block-level design implementation procedures (highlighted by the light green color) are based on Cadence Innovus and the Macro-3D flow proposed in [8], but significant changes have been made to enable micro-bump 3D and PDN, using our 3D technology modeling and PDN generation flow.

First, we partition the large memory blocks into the top tier of the 3D ICs and project the memory pins to the corresponding layer on the 3D metal stack. For the Cortex-A53 benchmark, the 16 L2 data cache blocks are selected for partitioning since their large size contributes to the huge footprint and long interconnects in 2D. We use the same partitioning and floorplan for micro-bump and hybrid-bonding 3D ICs as a fair comparison.

For micro-bump 3D, the signal bumps are inserted after the partitioning and floorplan. We insert the B2T or T2B cover cells on each intertier net based on the driving directions of the net. After that, the cover cells are placed on the bump manufacturing grid point closest to the driver instance in order to minimize the 3D interconnect overhead. The cover cells are fixed during standard cell placement to ensure that they are always on the grid. Then, the cover cell pins can be can be routed to the target instance pins during the signal routing stage. For hybrid-bonding 3D, the signal bonding pads automatically inserted as regular vias during the routing stage, and the pitch constraint is honored as part of the design rules.

The clock distribution network (CDN) synthesis is similar to 2D since all the clock gates and buffers are placed on the bottom tier. However, the bump pitches and RC parasitics have impacts on the quality and delays of 3D CDNs. After CDN synthesis and routing, we can perform timing, power, and rail analysis on the 3D design database before splitting the tiers



(a)

(b)

Fig. 5. Cross-sectional view of the 3D PDN structure. P/G is short for "power or ground." (a) On-chip PDN. (b) Off-chip PDN.

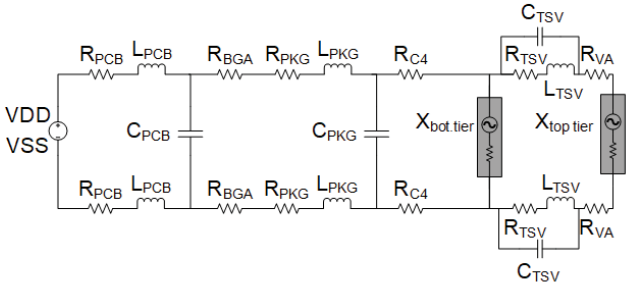


Fig. 6. Parasitic *RLC* model for our power integrity and voltage drop analysis.

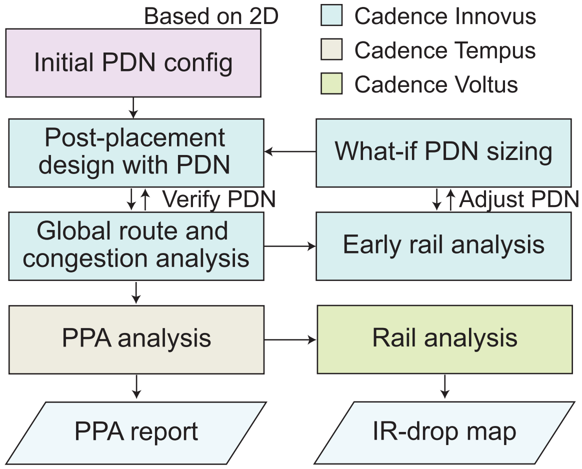


Fig. 7. PDN design flow for our block-level designs. This flow is used as a step in our hierarchical physical design flow in Fig. 4.

to generate the PPA report and IR drop map. This is because the 3D technology model provides a direct representation of the timing and power delivery characteristics of the 3D ICs.

Finally, the 3D design database is divided at the 3D interface into two IC layouts for the top tier and bottom tier. For micro-bump 3D, we need to replace the cover cells with real micro bumps at the exact location. After that, the 3D IC layouts are ready for manufacturing and stacking.

C. 3D PDN and Rail Analysis

Fig. 5 shows the basic structure of the 3D PDNs in our F2F-bonded 3D ICs. We assume that the power supply is connected to the 3D ICs through the printed circuit board (PCB) traces, ball grid array (BGA), and the routing wires in the package substrate. Those components belong to the off-chip PDN, and they are not not modeled directly in our 3D design. Instead, the parasitics of these components are estimated based on the size and empirical values from [17].

On the other hand, the on-chip PDN consists of the power or ground (P/G) C4 bumps, TSVs, rails, totems (via pillars), and micro bumps or hybrid-bonding pads. The P/G TSVs deliver power to the bottommost metal layer of the bottom tier. The P/G rails on the lower metal layers connect those TSVs and route to all the instance P/G pins on the bottom tier. On top of that, we stack multicut vias and short wire segments for multiple layers to create via totems. Those totems can efficiently deliver power to the upper layers. On the 3D interface, micro bumps or hybrid-bonding pads are placed at the intersection of the top- and bottom-tier P/G rails. The P/G rail pitch (VDD-VSS pitch) is set to 2× of the bump/pad size to avoid short circuits between bumps and rails. On the top tier, a similar via totem and rail structure is used to deliver power to the top-tier blocks. Fig. 6 shows the equivalent *RLC* circuit, including both the off-chip and on-chip PDNs.

Although the 3D PDN structure is straightforward, it is challenging to determine the optimal PDN dimensions, including the P/G rail width and pitch. This is because the runtime of the entire P&R process is too long. We develop a method to design 3D PDN and optimize it iteratively, as shown in Fig. 7. We first start with the PDN dimension similar to the baseline 2D IC and implement the design until the placement stage. After that, we perform global routing on the placement layout and analyze the routing congestion. If congestion is too high, we need to reduce the power rail density and width on the specific layer and return to the PDN design stage. Otherwise, we can proceed to the ERA stage.

The ERA mode is provided by Cadence Voltus to evaluate the power integrity at the early design stage. We first generate a 3D power map based on the placement layout. With the power map and the current 3D PDN design as inputs, the Voltus ERA engine can perform rail analysis and generate IR drop maps. It also enables what-if analysis, which allows us to scale the resistance of the power grid on a specific layer without rebuilding the entire PDN. By performing what-if PDN sizing and ERA repeatedly, we can determine the PDN dimensions to satisfy the IR drop constraint. Then, we can

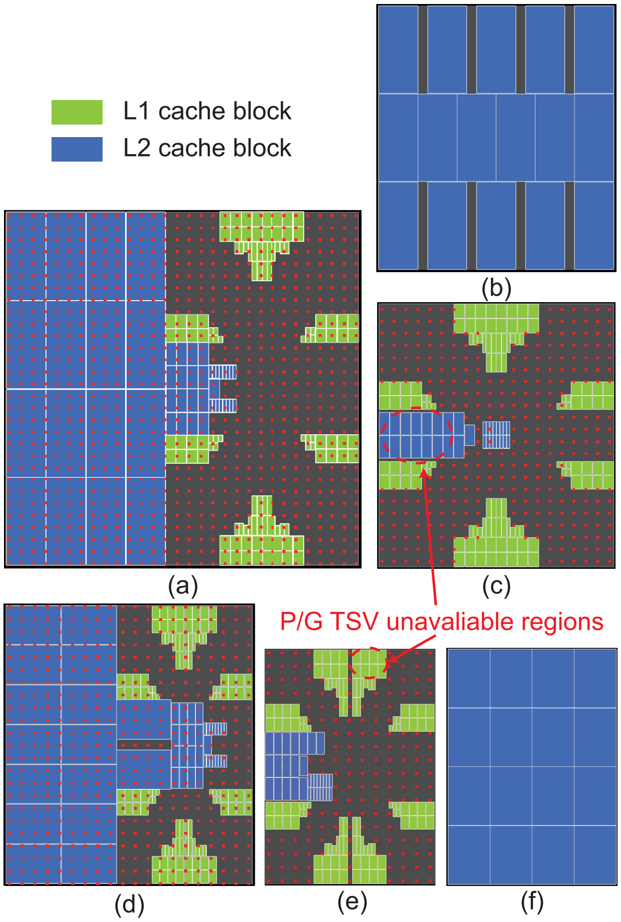


Fig. 8. Partitioning, floorplan, and C4 bump assignment of (a) 28-nm 2D, (b) 28-nm 3D top tier, (c) 28-nm 3D bottom tier, (d) 16-nm 2D, (e) 16-nm 3D bottom tier, and (f) 16-nm 3D top tier. The red dots represent the locations of P/G C4 bumps and TSVs.

regenerate 3D PDN using those dimensions and continue the design implementation.

After routing the 3D IC, the sign-off-level rail analysis is performed to evaluate the power integrity using the parasitic RLC model shown in Fig. 6. Unlike ERA, the impacts of signal routing and switching power are included in this analysis. In this study, we use 0.05 as the default switching activity for primary inputs and registers in the benchmark design. This switching activity setting can reflect the workload of the Cortex-A53 CPU in realistic applications [14]. Based on this assumption, we perform a static and dynamic rail analysis to generate the final IR drop map.

V. Experimental Results

A. Placement and Routing Results

In this section, we demonstrate the P&R results of 2D, micro-bump 3D, and hybrid-bonding 3D ICs, and then evaluate their design quality in detail.

First, we show the 3D partitioning and floorplan results of our 28- and 16-nm ICs in Fig. 8. The same 3D floorplan is used for both micro-bump and hybrid-bonding 3D ICs for fair comparisons. In the 3D ICs, the 16 L2 data cache blocks (SRAMs) are partitioned into the top tier, and the resulting 3D footprint area is exactly 50% of the 2D footprint area in both process nodes. We place the large SRAMs on the top tier so that the current-hungry logic gates are placed on the bottom tier, which is closer to the current source.

TABLE IV
C4 Bump, Micro Bump, and Hybrid-Bonding Pad Usage in Our 2D and 3D ICs

	2D	u-bump 3D		h-bond 3D	
Bump pitch (μm)	-	25	10	5	1
28nm					
Total C4 bump #	900	440			
Blocked C4 bump #	0	112			
P/G u-bump/pad #	-	1716	10628	86240	86240
Signal u-bump/pad #	-	1188	1188	1770	63260
16nm					
Total C4 bump #	460	234			
Blocked C4 bump #	0	58			
P/G u-bump/pad #	-	438	2772	10905	139889
Signal u-bump/pad #	-	1188	1188	37943	251932

However, the SRAM blocks on the bottom tier can block power delivery from P/G C4 bumps and TSVs because the conventional SRAM block does not allow P/G TSVs to go through the block. It is hard to avoid this issue by breaking large SRAMs into small blocks, manipulating the SRAM placement, or placing P/G TSVs in the gaps. The reason is that small SRAM instances introduce nonnegligible timing overhead, and the SRAM obstructions can block the power rails and create long resistive paths. As shown in Table IV, 112 and 58 C4 bumps are blocked in the 28- and 16-nm 3D ICs, respectively. The number of effective C4 bumps is 63% lower than the 2D baseline. It is possible to adjust the SRAM peripheral circuit to enable feed-through TSVs, as proposed in [5], but this requires a significant redesign of SRAM IPs. Therefore, we need to consider the impacts of these P/G TSV unavailable regions in 3D PDN design and quantify their impacts on power integrity.

Fig. 9 shows the final graphic design system (GDS) layouts of our 3D ICs with various bump/pad pitches. The majority of the signal routing is located on the bottom tier because the number of interconnects among standard cells is much larger than the number of logic-to-memory interconnects. For micro-bump 3D, the signal micro bumps are only used to connect the memory pins, so the bump usage is low, and the routing on the top tier is minimized. With decreasing bump pitches (e.g., from 25 to 10 μm), the 3D wirelength overhead is reduced since the bumps and signal pins get closer [as shown in Fig. 9(b) and (f)]. The reduced wirelength can lead to smaller intertier wire delays and better timing. For hybrid-bonding 3D, however, some signal bonding pads are used to establish 3D intertier interconnects for instances on both tiers. This intertier metal sharing explains why the routing wirelength increases in hybrid-bonding 3D with smaller pad pitches [as shown in Fig. 9(d) and (h)], as more signal pins are used in 3D intertier connections to mitigate the routing congestion on the logic tier.

The final micro bump and hybrid-bonding pad usage in our 3D ICs is shown in Fig. 10 and Table IV. We do not

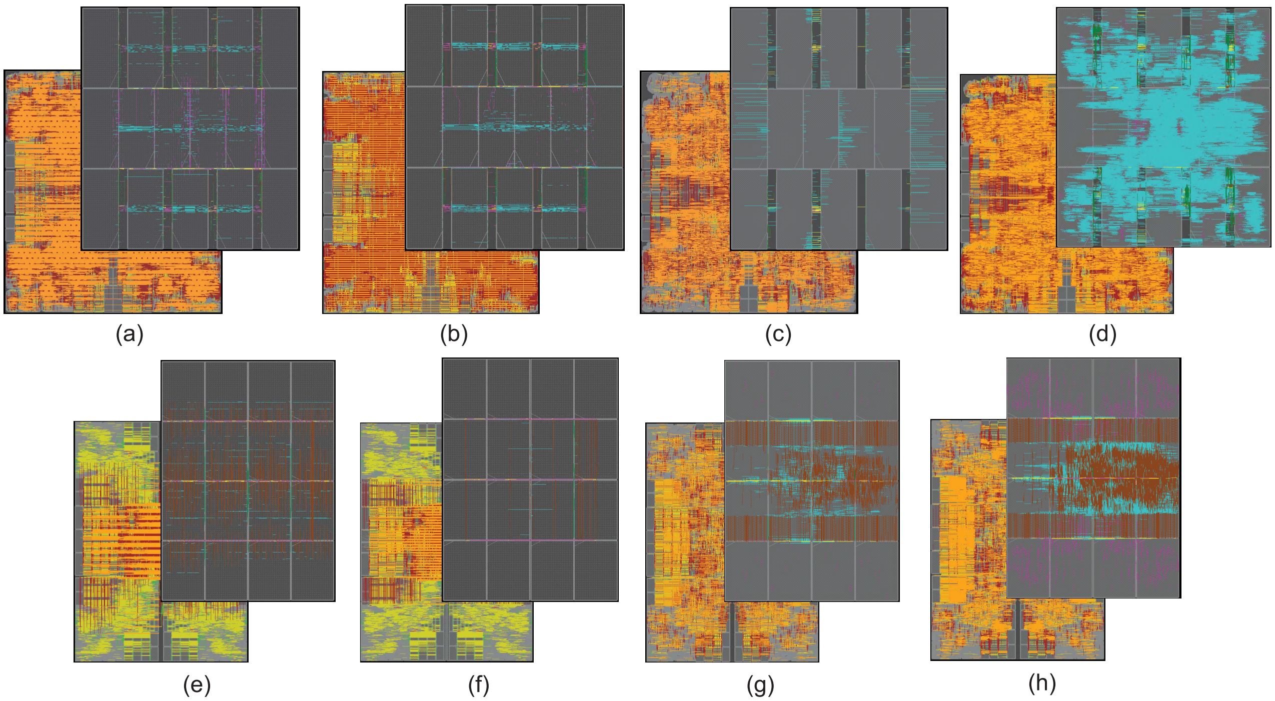


Fig. 9. GDS layouts of the max-performance 28-nm designs: (a) u-bump 25 μm, (b) u-bump 10 μm, (c) h-bond 5 μm, and (d) h-bond 1 μm. GDS layouts of the max-performance 16-nm designs: (e) u-bump 25 μm, (f) u-bump 10 μm, (g) h-bond 5 μm, and (h) h-bond 1 μm. The layouts are not to scale, and the P/G rails are not shown for clarity.

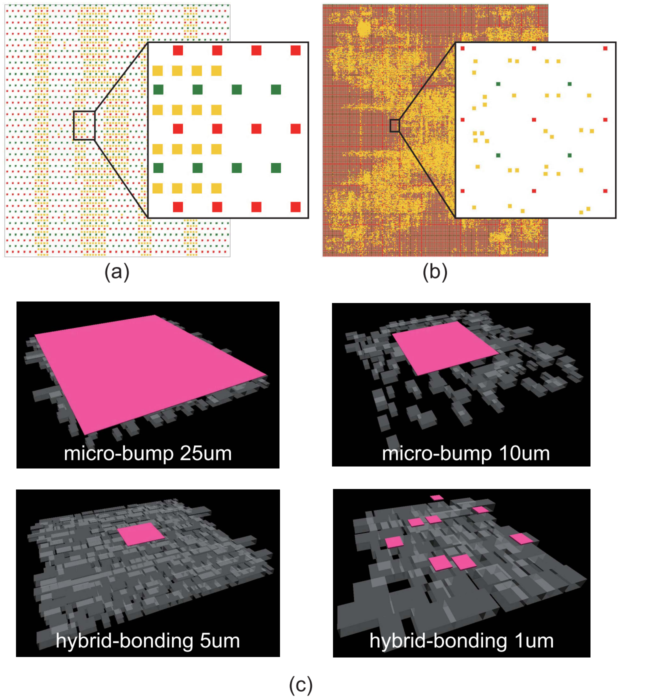


Fig. 10. Top-down view of (a) micro bumps and (b) hybrid-bonding pads in our 3D ICs. The red, green, and gold dots represent the power, ground, and signal bumps/pads, respectively. (c) 3D views of the bumps/pads compared with the standard cells.

consider the signal C4 bump usage because we assume that this CPU design is part of a large system-on-chip (SoC), and the I/O ports will be connected to other IP blocks at the edge. According to the table, the small bump/pad pitches allow more micro bumps or bonding pads to be used for signal routing and power delivery, and their impacts on timing and power delivery will be evaluated in the following.

B. Max-Performance Comparisons

We perform the max-performance analysis on the various 2D and 3D ICs at 28- and 16-nm process nodes, respectively, that is, we sweep the target frequency during P&R to search for the maximum achievable frequency ($F_{\max}$) for each technology configuration. The objective is to understand the impacts of 3D technology on the PPA metrics of these designs.

We compare the PPA metrics of the 2D and 3D ICs and summarize the comparisons in Table V. The results show that the maximum frequency of the 3D ICs tends to increase with the decreasing bump pitches. The hybrid-bonding hybrid-bonding 3D IC with 1-μm bump pitch shows the best performance, with 25% and 76% higher frequency than the 25-μm pitch micro-bump 3D ICs in 28- and 16-nm process nodes, respectively. This trend is caused by lower intertier routing overhead, more metal resource sharing, and a more balanced clock tree in the 3D ICs with smaller pitches, which we will be analyzed in detail.

We highlight the interconnects to SRAM blocks (memory nets) of 28-nm 3D ICs in Fig. 11. The magenta lines represent the memory nets connected to the top-tier SRAM. In micro-bump 3D ICs [Fig. 11(a) and (b)], the magenta nets need to be routed to the nearest micro bump first and then connected to the SRAM pins on the top tier, which introduces intertier routing overhead. As a result, the maximum memory access latency in the micro-bump 3D ICs is larger than or close to the 2D baseline. The hybrid-bonding 3D ICs, on the other

TABLE V

MAX-PERFORMANCE COMPARISONS OF PPA METRICS IN OUR 2D AND 3D ICs. THE AREA, WIRELENGTH, FREQUENCY, AND POWER VALUES ARE NORMALIZED BASED ON THE 2D BASELINE VALUES

28nm					
Configuration	2D	u-bump 3D		h-bond 3D	
Bump pitch (um)	-	25	10	5	1
Footprint	1.00	0.50	0.50	0.50	0.50
Metal stack	6M	M6BM6T	M6BM6T	M6BM6T	M6BM6T
Core util. (%)	83.0	78.9	77.8	82.2	81.7
Fmax	1.00	1.27	1.27	1.48	1.58
WL (um)	48.79	47.65	46.53	50.03	48.76
Total power	1.00	1.29	1.27	1.60	1.67
16nm					
Configuration	2D	u-bump 3D		h-bond 3D	
Bump pitch (um)	-	25	10	5	1
Footprint	1.00	0.50	0.50	0.50	0.50
Metal stack	6M	M7BM6T	M7BM6T	M7BM6T	M7BM6T
Core util. (%)	85.69%	85.73%	85.73%	88.56%	87.61%
Fmax	1.00	1.10	1.55	1.86	1.94
WL (um)	39.39	36.53	36.32	38.58	37.70
Total power	1.00	1.09	1.48	2.12	2.03

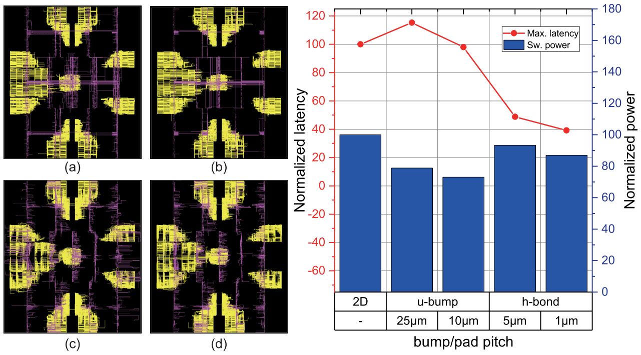


Fig. 11. Memory nets in the 28-nm 3D ICs (projected to the bottom tier): (a) u-bump 25 μm, (b) u-bump 10 μm, (c) h-bond 5 μm, and (d) h-bond 1 μm. The yellow and magenta lines represent the memory nets on the bottom tier and top tier, respectively. The latency and power values are normalized to the 2D baseline.

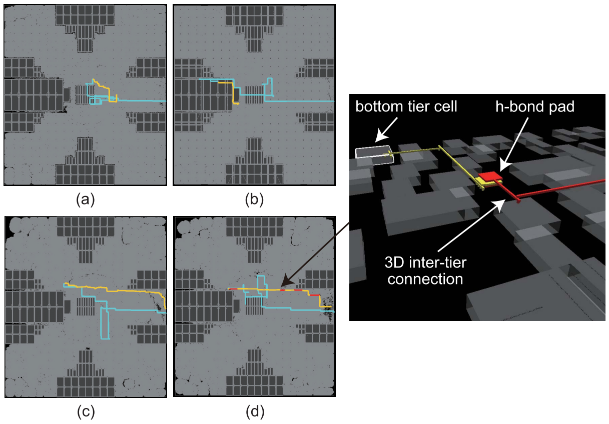


Fig. 12. Critical timing paths of the 28-nm 3D ICs (projected to the bottom tier): (a) u-bump 25 μm, (b) u-bump 10 μm, (c) h-bond 5 μm, and (d) h-bond 1 μm. The yellow and red lines show the critical path interconnects on the bottom tier and top tier, respectively. The cyan lines represent the clock launch path.

hand, provide up to 61% reduction in the maximum memory access latency since the routing overhead is minimized with the small bump pitches. Also, the memory switching power is reduced by at least 6% in all 3D ICs despite the increasing frequency because the average wirelength is reduced with the 3D interconnects compared with the 2D baseline.

Fig. 12 shows the critical timing paths in our 28-nm 3D ICs. Compared with the 2D baseline, the 3D ICs have a smaller footprint area, and 3D intertier connections help reduce the long logic-to-memory interconnects. However, among the 3D ICs, we find that the critical path delay decreases with the increasing signal bump/pad density, especially for the hybrid-bonding 3D ICs. The reason is that the high bump/pad density allows the router to use intertier 3D wires to optimize signal routing with a small wirelength overhead, that is, when routing congestion occurs on the bottom tier, the less-congested top tier can be used to establish the interconnect, as shown in the 3D view in Fig. 12. This metal layer sharing is only possible when the bump/pad pitch is small; otherwise, the overhead of cell-to-bump routing can compensate for the benefits.

Moreover, the small 3D bump/pad pitches benefit the clock tree. In the micro-bump 3D design, the clock routing wirelength to the SRAM clock pins on the top tier is long due to the overhead introduced by the bump pitches. As a result, the clock launch latency can become significantly large (even larger than 2D), which contributes to the negative slacks. With decreasing bump pitches, the clock launch latency can be remarkably reduced (as shown in Fig. 13). Together with the decreasing path delay, the small bump pitches help 3D ICs achieve better timing easily.

In addition, with PDN insertion, the bottom tier tends to become more congested, which adds to the importance of intertier metal sharing. However, the P/G rails on the interface layers can block the signal intertier connections. In other words, PDN insertion makes the 3D intertier routing problem more challenging.

C. Iso-Performance Comparisons

We then perform the iso-performance analysis to evaluate the power delivery quality of these 2D and 3D ICs. The 2D

TABLE VI

ISO-PERFORMANCE COMPARISONS OF PPA METRICS IN OUR 16-nm 2D AND 3D ICS. THE AREA, FREQUENCY, AND POWER VALUES ARE NORMALIZED BASED ON THE 2D BASELINE VALUES. STATIC IR DROP REPRESENTS THE WORST INSTANCE IR DROP (VDD-VSS) IN EACH DESIGN

Configuration	2D	u-bump 3D		h-bond 3D	
Bump pitch (um)	-	25	10	5	1
Footprint	1.00	0.50	0.50	0.50	0.50
Metal stack	6M	M7BM6T	M7BM6T	M7BM6T	M7BM6T
Core util. (%)	80.07%	67.80%	67.79%	59.07%	58.76%
Target freq.	1.00	1.00	1.00	1.00	1.00
WL (um)	37.69	35.97	35.75	34.10	32.30
Total power	1.00	0.90	0.90	0.78	0.75
PDN area (mm^2)	3.26	3.33	3.34	3.33	3.34
Static IR drop (mV)	14.25	41.48	42.91	27.54	24.43
Hotspot location	2D	Bottom	Top	Bottom	Top

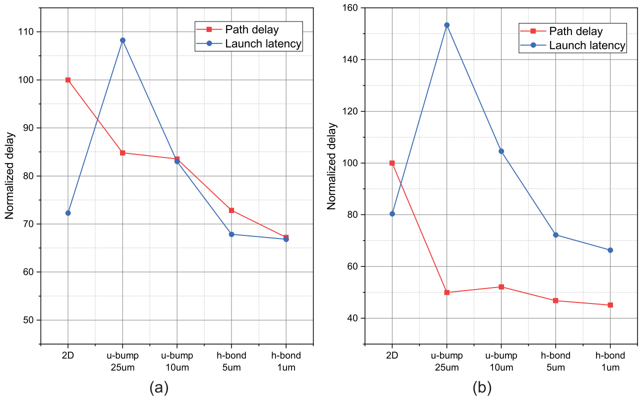


Fig. 13. Trend of delays in 2D and 3D ICs with reducing bump/pad pitches: (a) 28 and (b) 16 nm. The delay values are normalized based on the 2D critical path delay.

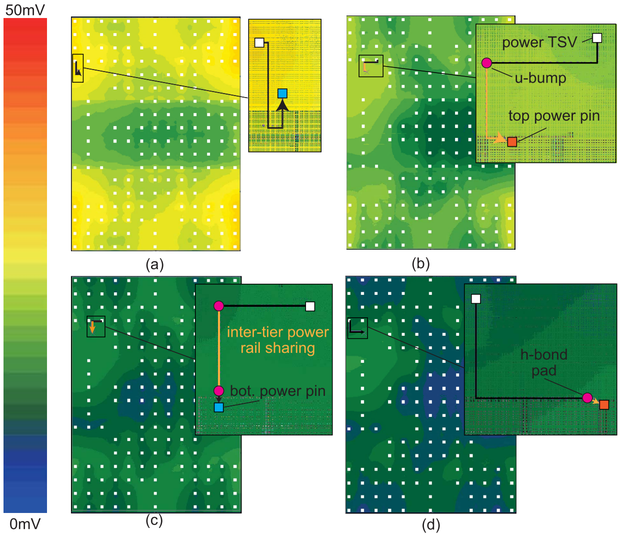


Fig. 14. IR drop maps of our 16-nm 3D ICs: (a) u-bump 25 μm, (b) u-bump 10 μm, (c) h-bond 5 μm, and (d) h-bond 1 μm. The white squares represent the P/G C4 bump and TSV locations. The black and orange lines show the LRP to the IR drop hotspots on the bottom tier and top tier, respectively.

and 3D ICs are redesigned at the same target frequency ($F_{\max}$ of 2D). We use Cadence Voltus to perform power simulation and IR drop analysis on these iso-performance designs to evaluate their power integrity.

The PPA and IR drop metrics of the iso-performance designs are summarized in Table VI. Overall, the results show that with similar PDN metal usage, the 3D ICs suffer from higher IR drop (up to 43 mV) compared to 2D, 2D, especially for micro-bump 3D ICs with large bump pitches. However, decreasing bump/pad pitches can help reduce the IR drop remarkably by up to 17 mV in the 16-nm process. To understand this trend, we use the 16-nm 3D ICs as examples, analyze the IR drop hotspots, break down the IR drop into layers, and identify the 3D power delivery bottlenecks.

Fig. 14 shows the IR drop maps of our 16-nm 3D ICs. The rail analysis is done before tier partitioning, so both the bottom- and top-tier PDNs are included in these maps. The results show that the IR drop hotspots are near the P/G TSV unavailable regions for the 3D ICs. As shown in the 3D view on Fig. 15, the power delivery path needs to take a detour through P/G TSVs outside the region to reach the target instance. This power delivery detouring increases the resistive path length and can cause a power delivery bottleneck in F2F-bonded 3D ICs.

We analyze the least-resistive paths (LRPs) to the IR drop hotspots in the 3D ICs. For the four 16-nm 3D ICs, the

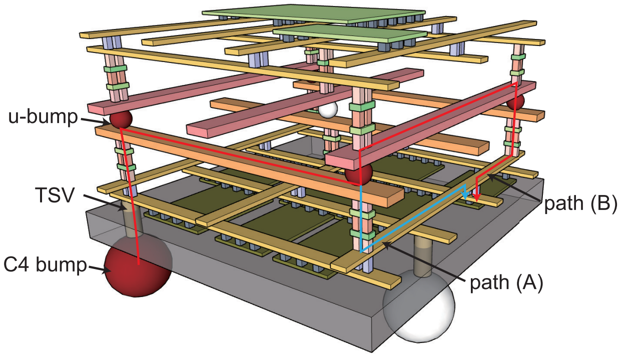


Fig. 15. 3D view of an F2F bonded 3D PDN structure.

maximum IR drop can occur on the top tier because the power delivery to the top tier needs to go through the entire entire 3D PDN stack and the P/G micro bumps or hybrid-bonding pads, as shown in Fig. 14(b) and (d). also occur on the bottom tier since the logic-intensive bottom tier has higher power density, as shown in Fig. 14(a) and (c). In Fig. 16, we show the layer-based voltage drop breakdown of these LRPs. The results

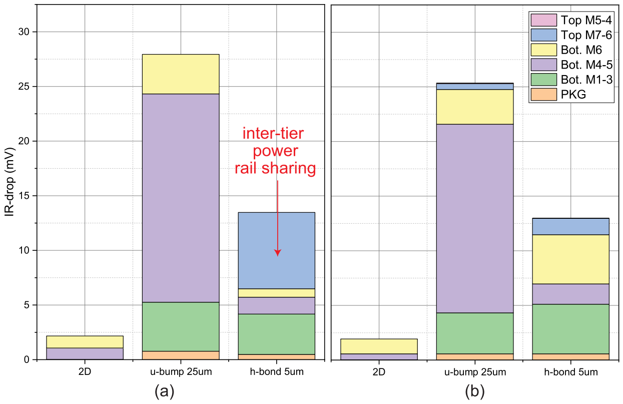


Fig. 16. Layer-based breakdown of the power delivery path to (a) L2 data RAM (on the top tier of 3D partitioning) and (b) L2 tag RAM (on the bottom tier of 3D partitioning). The IR drop on via layers is counted to the IR drop on its bottom metal layer.

show that the 3D interface layers (M7 of the bottom tier and M6 of the top tier) play an important role in the total IR drop. Because of the TSV unavailable regions, the current needs to travel a long distance on the *XY* plane from the C4 bump to reach the target block power pin. The lower metal layers (Mx) are too resistive for these long-distance power delivery, so the LRP goes through the via totem, the interface layer, and the intertier bumps/pads to reach the target pin. As a result, the intertier power delivery is significantly affected by the interface layer PDN density and bump pitches.

On the other hand, the interface layers also have an impact on the power delivery within the bottom tier. The power TSVs are connected to M1, the same layer on which the cell power pins are located. However, the LRP does not go directly on the Mx layers to reach the target cell pin because the rails are too resistive. Instead, it can go up and down. The red arrows in Fig. 15 show two potential power delivery paths from the power TSV to a cell power pin on the bottom tier: path (A) uses only the bottom tier power rails; path (B) uses the intertier bump/pad to reach the top tier and travel on the top-tier power rails and goes down through another bump/pad to reach the target pin. Path (B) takes a detour on the top tier and uses the top-tier PDN metal resources for bottom-tier power delivery, so we call it "intertier power rail sharing."

When the bump/pad pitch and parasitics are small and the power rails on the interface layer are dense, the overhead of this intertier detour can be minimized. As a result, path (B) can present a lower impedance compared with path (A). For example, in Fig. 14(d), the LRP to the bottom-tier hotspot consists of two h-bond pads and a segment of top-tier power rails. In other cases, the existence of intertier power delivery paths also helps distribute the current and reduce IR drop to the bottom-tier target. As a result, we observe that the IR drop to the bottom-tier hotspots reduces with the decreasing bump/pad pitch in Fig. 16, similar to the top-tier IR drop trend.

The trend of IR drop reduction with decreasing bump/pad pitches is also contributed by power saving in 3D ICs. In Table VII, we break down the total power consumption of each design based on power types and components that consume the power. The results show that the hybrid-bonding 3D ICs have much smaller total power compared with the micro-bump 3D counterparts, which provide up to 16.0% total power saving at the same working frequency. This power saving is mainly contributed by dynamic and combinational power reduction. The hybrid-bonding 3D ICs provide significant timing benefits, as discussed in Section V-B. As a result, they can meet the timing constraint easily at this lower target frequency, and thus fewer buffers or inverters are inserted to optimize the timing paths. This low total power consumption leads to a small average current in the hybrid-bonding 3D ICs, which, in turn, helps reduce the IR drop and improve their power integrity.

TABLE VII

Power Breakdown of the Iso-Performance Design by Types and Components. The Power Values Are Normalized Based on the 2D Baseline Total Power

Configuration	2D	u-bump 3D		h-bond 3D	
Bump pitch (um)	-	25	10	5	1
Total	100.0	89.6	89.7	77.9	75.3
Type					
Internal	55.4	49.6	49.8	44.8	43.2
Switching	36.8	32.8	32.8	30.0	29.1
Leakage	7.9	7.2	7.2	3.2	3.0
Components					
Sequential	24.5	22.2	22.4	22.1	22.4
Macro	19.3	17.1	17.1	17.2	17.2
Combinational	33.6	29.2	29.2	18.2	17.3
Clock	22.6	21.1	21.0	20.4	18.4

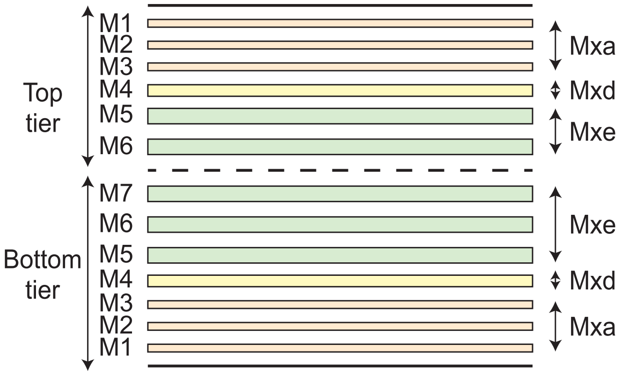


Fig. 17. Layer assignment in our 16-nm 3D BEOL stack.

In conclusion, we find that the TSV unavailable regions cause power delivery bottleneck in F2F-bonded 3D ICs, while small bump pitches, dense power rail on interface layers, and intertier metal sharing can help mitigate this issue for IR drop hotspots on both tiers.

D. Mitigate Static IR-Drop

Based on our observations, we improve the 3D PDNs using the iterative PDN reconfiguration flow to reduce the static IR drop. For the 16-nm 3D ICs, we first increase the power rail density on the interface layers (Mxe), then on the bottom layers (Mxa), and finally on the intermediate layers (Mxd). The Mx layer assignment is visualized in the cross-sectional view in Fig. 17. The metal usage of these five 3D PDN designs is shown in Table VIII. The M1–M4 layers of the top tier

TABLE VIII

DIMENSIONS AND PARASITICS OF THE BUMPS/PADS IN THE 2D/3D ICS. THE SMALL INDUCTANCE VALUES ARE OMITTED FOR THE BUMPS/PADS

PDN config.		PDN1	PDN2	PDN3	PDN4	PDN5
Bottom	M1	29%	30%	30%	28%	28%
	M2	0%	0%	0%	0%	0%
	M3	45%	49%	66%	49%	66%
	M4	30%	50%	66%	50%	66%
	M5	0%	0%	0%	50%	67%
	M6	0%	0%	0%	50%	67%
	M7	40%	80%	80%	80%	80%
Top	M6	40%	80%	80%	80%	80%

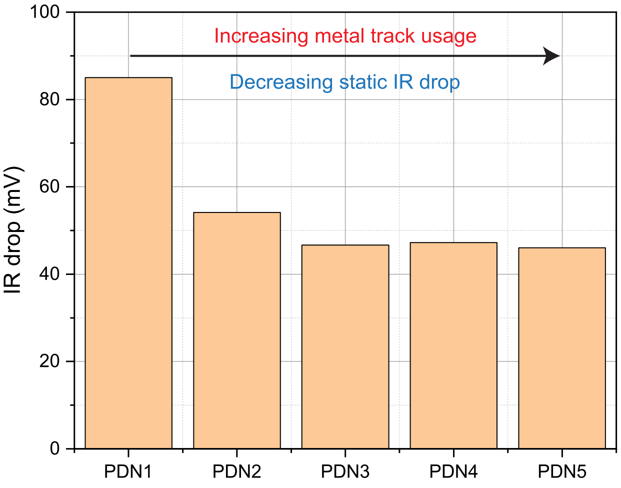


Fig. 18. Trend of IR drop in 16-nm h-bond 3D ICs with various PDN designs.

are occupied by the SRAM blocks, so no power rails can be inserted there.

We rebuild the h-bond 3D IC with a 10-μm bump pitch to evaluate the impacts on PPA and IR drops. Fig. 18 shows the trend of IR drop with various PDN configurations. According to the results, IR drop reduces significantly with the up-sized PDNs, especially from PDN1 to PDN2 (IR drop reduction is 36%). The reason is that the denser power rails on the interface layers create more intra-tier and intertier power delivery paths with low impedance and help power delivery to both top and bottom tiers. Among PDN2, 3, and 4, the IR drop differences are small because the impacts of lower layers and intermediate layers have saturated. However, this PDN reconfiguration cannot be directly applied to the micro-bump 3D ICs because the large size of the micro bumps (5 or 12.5 μm) can block such a dense PDN on the interface layers. Therefore, the micro-bump 3D ICs are still faced with challenges of intertier power delivery.

E. Mitigate Dynamic Voltage Drop

Furthermore, we explore the power integrity challenges in 3D ICs through dynamic rail analysis. In this analysis, the dynamic vector-less power simulation is performed to generate the power and current waveforms for each design based on the same switching activity assumption as the static analysis. Then, the waveforms are used in time-based dynamic rail analysis to capture the worst case voltage drop in a fixed 10-ns time window. As a result, both the steady-state voltage drop (IR) and the transient voltage drop ($L \cdot dI/dt$) are included in the analysis. We evaluate the reliability of the various 3D PDNs using this dynamic analysis and propose solutions to mitigate the voltage drop problem in this section.

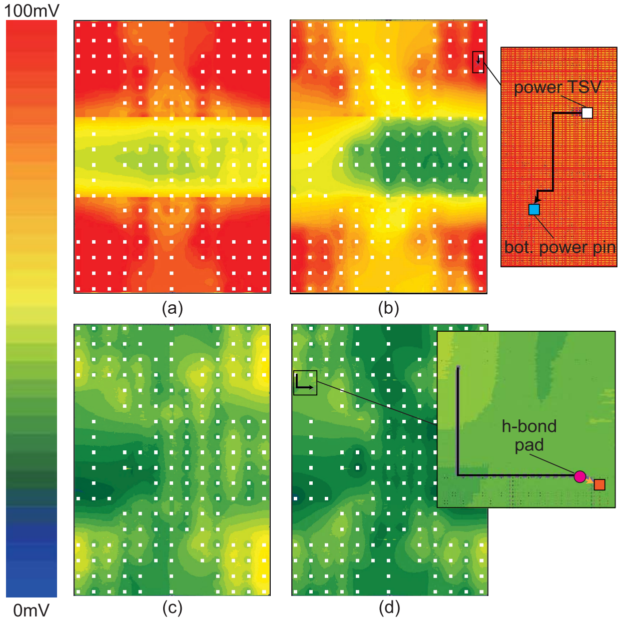


Fig. 19. Dynamic voltage drop maps of the 16-nm 3D ICs: (a) u-bump 25 μm, (b) u-bump 10 μm, (c) h-bond 5 μm, and (d) h-bond 1 μm. The white squares represent the P/G C4 bump and TSV locations.

The dynamic voltage drop maps of our 16-nm 3D ICs are shown in Fig. 19, and the design metrics related to the dynamic voltage drop are summarized in Table IX. According to these results, micro-bump 3D ICs suffer from excessive dynamic voltage drop (117 mV higher than the 2D baseline), and the hotspots are often located in the CPU core regions on the bottom tier. The reason is that these regions have high averaged switching activity, which leads to rapid current fluctuations and a large $L \cdot dI/dt$ drop. The cell and memory capacitance, together with the grid capacitance, can act as decoupling capacitance (decap) to mitigate the dynamic voltage drop. drop. However, in micro-bump 3D ICs, the top-tier power grid is connected to the bottom tier through a small number micro bumps (limited by the bump pitches). As a result, the decoupling effects of capacitance on the top tier are less effective.

To address these dynamic power integrity challenges, we propose a decoupling-capacitance-sharing approach based on the metal–insulator–metal (MIM) capacitors. The MIM capacitors are a class of capacitors that consist of insulators between metals, and they can be embedded into the BEOL layers. They can provide a high capacitance density (up to 40fF/μm^2 [18]) and have been proven to be effective in mitigating voltage drop issues in 2D ICs [19].

For our 3D ICs, we notice that the utilization of the BEOL layers on the top tier is low, especially in micro-bump 3D [as shown in Fig. 9(e) and (f)] since these layers are only used for signal routing and power delivery to the SRAM pins. Therefore, we can use these empty spaces on the top tier to

TABLE IX

DYNAMIC ANALYSIS RESULTS OF OUR ISO-PERFORMANCE 2D AND 3D ICS. THE FREQUENCY, CURRENT, AND POWER VALUES ARE NORMALIZED BASED ON THE 2D BASELINE VALUES. DYNAMIC IR DROP REPRESENTS THE WORST CASE INSTANCE IR DROP (VDD-VSS) IN EACH DESIGN

Configuration	2D	u-bump 3D		h-bond 3D	
Bump pitch (um)	-	25	10	5	1
Target freq.	1.00	1.00	1.00	1.00	1.00
Dynamic power	1.00	0.96	0.96	0.87	0.84
Peak current	1.00	0.87	0.87	0.62	0.60
Dynamic IR drop (mV)	69.62	187.14	184.59	90.50	78.65
Max. RLRP (Ohm)	112.09	298.94	238.52	157.10	241.71
Hotspot location	2D	Bottom	Bottom	Bottom	Top

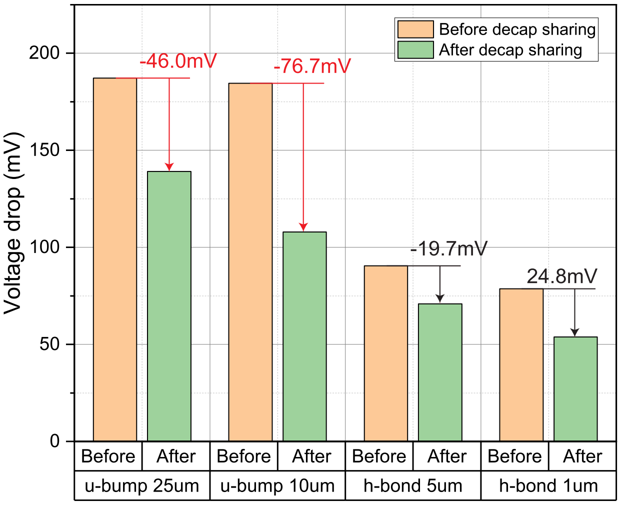


Fig. 20. Worst case dynamic voltage drop before and after enabling decap sharing in the 16-nm 3D ICs.

insert MIM capacitors and mitigate the dynamic voltage drop on the bottom tier. We call this approach "decap sharing" since it allows the top tier to share the decoupling capacitance resources with the more congested bottom tier.

We assume that the MIM capacitance density is 40 fF/μm^2 and then insert as many MIMs as possible on M5–M7 layers of the top tier in our 3D ICs based on a greedy approach. After that, we rerun the dynamic rail analysis to analyze the impacts on the worst case voltage drop. Fig. 20 shows the voltage drop differences before and after applying this approach. The results show that decap sharing can remarkably reduce the dynamic voltage drop in micro-bump 3D ICs by up to 77 mV, but it is more effective with smaller bump pitches such as 10 μm. The reason is that the high-density 3D connections can help the transient current go through the 3D interface and reach the decoupling capacitance on the top tier, which reduces the $L \cdot dI/dt$ drop generated by this current. However, we think that decoupling capacitance sharing is still an effective method to manage the dynamic voltage drop in 3D ICs. Together with the power rail sharing for static IR drop, it enables better utilization of metal resources in 3D ICs to combat the challenges in power delivery.

VI. OBSERVATIONS AND GUIDELINES

We observe that despite the great frequency improvement, F2F bonded 3D ICs have higher static IR drop and dynamic voltage drop when the PDN metal usage is similar to the 2D baseline. The causes include: 1) 3D ICs have a smaller footprint and SRAM blocks (or other physical IPs) can block TSVs at the bottom tier, which reduces the effective C4 bump number by more than 50%; 2) the TSV unavailable regions cause power delivery path to detour; 3) power sources are connected to the high-resistance Mx layers first; 4) power delivery path to the top tier is long and needs to go through the entire metal stack; and 5) the decoupling capacitance on the top tier cannot effectively mitigate the dynamic voltage drop on the bottom tier due to the few 3D connections.

Among the 3D ICs, we find that the maximum frequency improves with decreasing bump/pad pitches, especially in hybrid-bonding 3D ICs, due to low 3D routing overhead and more metal sharing for signal interconnects. In addition, the fine-pitch 3D technology also helps reduce IR drop. It provides more P/G bumps/pads and denser 3D interconnects, which, in turn, facilitates intertier metal sharing for power delivery and decoupling capacitance.

To design a robust 3D PDN, we propose the following guidelines.

1) Avoid placing large physical IP blocks on the bottom tier by prioritizing them for the top tier during partitioning in order to overcome the power delivery bottleneck caused by TSV unavailable regions.
2) Reduce the 3D bump/pad pitch, and insert as many P/G bumps as possible because the fine-pitch 3D connections have been shown to improve both signal routing and power integrity.
3) Increase the power rail density on the interface layers since it helps power delivery to both tiers through metal resource sharing.
4) Enable decoupling capacitance sharing with MIM capacitors to mitigate the dynamic voltage drop.

VII. CONCLUSION

Using Cortex-A53 as a benchmark and TSMC 28 and 16 nm as process nodes, we model 2D, micro-bump, and hybrid-bonding 3D ICs with PDNs in this study and evaluate their PPA and power delivery characteristics in this study. Our experimental results show that compared with 2D baselines, the 3D ICs suffer from increased static IR drop (up to 43 mV) in advanced technology nodes, caused by the fewer C4 bumps, long resistive path, and TSV unavailable regions. Fine-pitch 3D technology is beneficial for both signal routing and power delivery. It can help improve the maximum frequency by up to 76% and reduce IR drop by up to 17 mV if we decrease the bump pitch from 25 μm (micro-bump 3D) to 1 μm (hybrid-bonding 3D). The MIM-based decoupling capacitance method can effectively reduce the dynamic voltage drop in micro-bump 3D by up to 77 mV. This study demonstrates the potential and importance of 3D technology scaling with finer pitch and paves the way for future 3D IC power delivery studies.

Acknowledgment

The authors would like to thank Dr. Sai Surya Kiran Pentapati and Dr. Jinwoo Kim for their great help with this article.

References

[1] J. Derakhshandeh et al., "Die to wafer 3D stacking for below 10μm pitch microbumps," in *Proc. IEEE Int. 3D Syst. Integr. Conf.*, Nov. 2016, pp. 1–4.

[2] S.-W. Kim et al., "Ultra-fine pitch 3D integration using face-to-face hybrid wafer bonding combined with a via-middle through-silicon-via process," in *Proc. IEEE 66th Electron. Compon. Technol. Conf. (ECTC)*, May 2016, pp. 1179–1185.

[3] J. Shi, M. Li, and C. A. Moritz, "Power-delivery network in 3D ICs: Monolithic 3D vs. skybridge 3D CMOS," in *Proc. IEEE/ACM Int. Symp. Nanosc. Architectures (NANOARCH)*, Jul. 2017, pp. 73–78.

[4] K. Chang, S. Das, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "System-level power delivery network analysis and optimization for monolithic 3-D ICs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 4, pp. 888–898, Apr. 2019.

[5] L. Zhu et al., "Power delivery and thermal-aware arm-based multi-tier 3D architecture," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2021, pp. 1–6.

[6] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A physical design methodology to build commercial-quality face-to-face-bonded 3D ICs," in *Proc. Int. Symp. Phys. Design*, Mar. 2018, pp. 90–97.

[7] X. Xu, M. Bhargava, S. Moore, S. Sinha, and B. Cline, "Enhanced 3D implementation of an Arm® Cortex®-A microprocessor," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2019, pp. 1–6.

[8] L. Bamberg, A. Garcia-Ortiz, L. Zhu, S. Pentapati, D. E. Shim, and S. K. Lim, "Macro-3D: A physical design methodology for face-to-face-stacked heterogeneous 3D ICs," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2020.

[9] A. Agnesina et al., "Hier-3D: A hierarchical physical design methodology for face-to-face-bonded 3D ICs," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design*, Aug. 2022, pp. 1–6.

[10] J. Kim, L. Zhu, H. M. Torun, M. Swaminathan, and S. K. Lim, "Micro-bumping, hybrid bonding, or monolithic? A PPA study for heterogeneous 3D IC options," in *Proc. 58th ACM/IEEE Design Autom. Conf. (DAC)*, Dec. 2021, pp. 1189–1194.

[11] S. Pan and B. Achkir, "Optimization of power delivery network design for multiple supply voltages," in *Proc. IEEE Int. Symp. Electromagn. Compat.*, Aug. 2013, pp. 333–337.

[12] A. E. Engin, "Efficient sensitivity calculations for optimization of power delivery network impedance," *IEEE Trans. Electromagn. Compat.*, vol. 52, no. 2, pp. 332–339, May 2010.

[13] S. Sourav, A. Roy, Y. Cao, and S. Pandey, "Machine learning framework for power delivery network modelling," in *Proc. IEEE Int. Symp. Electromagn. Compat. Signal/Power Integrity (EMCSI)*, Jul. 2020, pp. 10–15.

[14] D. Prasad et al., "Buried power rails and back-side power grids: Arm® CPU power delivery network design beyond 5 nm," in *IEDM Tech. Dig.*, Dec. 2019, pp. 1–19.

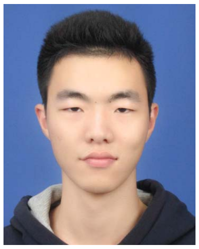
[15] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ICs," *IEEE Trans. Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan. 2010.

[16] S. S. K. Pentapati et al., "A logic-on-memory processor-system design with monolithic 3D technology," *IEEE Micro*, vol. 39, no. 6, pp. 38–45, Nov. 2019.

[17] I. P. Vaisband et al., *On-Chip Power Delivery and Management*. Cham, Switzerland: Springer, 2016.

[18] K. Fischer et al., "Low-K interconnect stack with multi-layer air gap and tri-metal-insulator-metal capacitors for 14 nm high volume manufacturing," in *Proc. IEEE Int. Interconnect Technol. Conf. IEEE Mater. Adv. Metallization Conf. (IITC/MAM)*, May 2015, pp. 5–8.

[19] S. Moon, S. Nam, M. Kang, and Y. Lee, "Design and analysis of MIM capacitor on power integrity effects for HPC and network applications," in *Proc. IEEE Int. Symp. Electromagn. Compat., Signal Power Integrity (EMC SIPI)*, Jul. 2019, pp. 199–204.



Lingjun Zhu (Graduate Student Member, IEEE) received the B.S. degree in microelectronic science and engineering from Fudan University, Shanghai, China, in 2018, and the M.S. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2020, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

He was a Research Intern with Arm, Inc., Austin, TX, USA, in 2020. He was a Physical Design CAD Intern with Apple Inc., Santa Clara, CA, USA, in 2021. He was a Design Automation Engineer Intern with Intel Corporation, Hillsboro, OR, USA, in 2022. His current research interests include the physical design and electronic design automation (EDA) solutions for high-performance 3D integrated circuits (ICs) and advanced CMOS technology.

Chanmin Jo received the B.S. and M.S. degrees in electrical engineering from Hanyang University, Seoul, South Korea, in 2004 and 2006, respectively, with a focus on modeling and characterization for signal integrity of high-speed integrated circuits, integrated circuit (IC) interconnect, and IC packaging.

In 2006, he joined Samsung Electronics, Suwon, South Korea, where he was responsible for the development of modeling and analysis methodology for high-speed memory interfaces.

Sung Kyu Lim (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Computer Science, University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

In 2001, he joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, where he is currently a Motorola Solutions Foundation Professor. His research is featured as Research Highlight in the Communication of the ACM in January 2014. He is the author of the books *Practical Problems in VLSI Physical Design Automation* (Springer, 2008) and *Design for High Performance, Low Power, and Reliable 3D Integrated Circuits* (Springer, 2013). He has published more than 400 articles on 2.5-D and 3D integrated circuits (ICs). His research focus is on the architecture, design, test, and electronic design automation (EDA) solutions for 2.5-D and 3D ICs.

Dr. Lim received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006, the ACM SIGDA Distinguished Service Award in 2008, and the Best Paper Award from the IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY and the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS in 2022. He received the Best Paper Award from several conferences in EDA, including ACM/IEEE International Symposium on Low Power Electronics and Design in 2022. He joined the Defense Advanced Research Projects Agency (DARPA) in 2022 as a Program Manager at the Microsystems Technology Office (MTO).