High-Performance Logic-on-Memory Monolithic 3D IC Designs for Arm Cortex-A Processors

Lingjun Zhu, Lennart Bamberg, Sai Surya Kiran Pentapati, Kyungwook Chang, Francky Catthoor, *Fellow, IEEE*, Dragomir Milojevic, Manu Komalan, Brian Cline, Saurabh Sinha, Xiaoqing Xu, Alberto Garcia-Ortiz, *Senior Member, IEEE*, and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*—Monolithic 3D IC (M3D) is a promising solution to improve the performance and energy-efficiency of modern processors. But, designers are faced with challenges in design tools and methodologies, especially for power and thermal verifications. We develop a new physical design flow that optimally places and routes cache modules in one tier and logic gates in the other. Our tool also builds high-quality clock and power delivery networks targeting logic-on-memory M3D designs. Lastly, we develop a sign-off analysis tool flow to evaluate power, performance, area (PPA), thermal, and voltage-drop quality for given M3D designs. Using our complete RTL-to-GDS tool flow, we design commercial quality 2D and M3D implementation of Arm Cortex-A7 and Cortex-A53 processors in a commercial 28nm technology. Experimental results show that our 3D processors offer 20% (A7) and 21% (A53) performance gain, compared with their 2D commercial counterparts. The voltage-drop degradation of our 3D Cortex-A7 and Cortex-A53 processors is less than 3% of the supply voltage, while temperature increase is 10.71°C and 13.04°C, respectively.**

***Index Terms*—Monolithic 3D, physical design, power delivery network, power integrity, thermal analysis.**

I. Introduction

As transistor scaling is approaching the physical limit, 3D integration becomes a promising solution to prolong the continuous improvement in performance and energy efficiency of integrated circuits (ICs), as predicted by Moore's law [1]. There are different types of 3D integration technology: Through-Silicon Via based 3D (TSV 3D) [2], Face-to-Face bonded 3D (F2F 3D) [3], and Monolithic 3D (M3D) ICs. Since the first two alternatives rely on mechanical stacking and bonding of pre-fabricated 2D dies, they only allow a rather limited inter-tier interconnect density. In contrast, monolithic 3D ICs are fabricated sequentially, which enables up to 100 million/mm^2 3D interconnections with the lowest parasitics, and therefore allows us to fully exploit the potential gain of 3D integration [4].

In M3D ICs, the top tier and bottom tier are fabricated sequentially [5]. Monolithic inter-tier vias (MIVs) are used for vertical interconnection in M3D ICs, which have a much smaller size (typically $< 100nm$) compared to the conventional TSVs. Therefore, M3D allows much higher 3D interconnection density, up to 100 million$/mm^2$ [4]. Since the MIV size is similar to the regular vias, each MIV introduces less than $1fF$ parasitic capacitance to the 3D ICs. With this high density and small parasitics, MIVs enables a large number of vertical interconnection in 3D ICs. However, placement and routing considering the 3D space are challenging tasks for very-large-scale integration (VLSI) system and no commercial EDA tools directly support this. On the other hand, the fabrication process of M3D ICs requires a low-temperature transistor manufacturing process [5] or a heat-tolerant interconnect technology [6], which has a significant impact on the performance and energy consumption of the resulting M3D ICs, and also needs to be considered in the physical implementation and analysis flow.

In recent years, multiple physical design methodologies have been proposed to implement gate-level monolithic 3D ICs. Most of these flows utilize commercial electronic design automation (EDA) tools to ensure the layout quality, and trick the tools into implement 3D ICs by shrinking cell size [7] or extending the 2D plane [8], [9]. Although these approaches have shown performance and power benefits in many circuits, they have a few drawbacks: (1) the design quality degrades for the 3D ICs with significant memory area occupation; (2) the number of MIVs highly depends on partitioning algorithms; (3) the performance degradation caused by M3D fabrication process is not considered by these flows; (4) sign-off verification approaches such as workload-based power and thermal analysis are not fully supported by these flows.

Logic-on-memory stacking is another promising 3D integration scheme. Unlike the traditional gate-level 3D ICs, it separates the standard cells and memory blocks into different tiers, removing the compatibility constraint between these two components, which helps to boost the performance and energy efficiency of the memory blocks. Furthermore, logic-on-memory stacking avoids long interconnections from/to the memory blocks, which provides the potential to improve the memory throughput and system performance drastically. In addition, it is shown in this work that logic-on-memory stacking mitigates the power-delivery and thermal issues in M3D ICs, and thus paves the road to commercialization of M3D integration. Some of these benefits have been explored by multiple studies [10]–[12]. However, there is no existing physical design flow to directly implement logic-on-memory

Lingjun Zhu, Sai Surya Kiran Pentapati, Kyungwook Chang, and Sung Kyu Lim are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. E-Mail: {lingjun, sai.pentapati, k.chang}@gatech.edu; limsk@ece.gatech.edu

Lennart Bamberg and Alberto Garcia-Ortiz are with the Institute of Electrodynamics and Microelectronics (ITEM) of the University of Bremen, Germany. E-Mail: {bamberg, agarcia}@item.uni-bremen.de

Francky Catthoor, Dragomir Milojevic, and Manu Komalan are with IMEC, Belgium. E-Mail: {Francky.Catthoor, Dragomir.Milojevic.ext, Manu.Perumkunnil}@imec.be

Brian Cline, Saurabh Sinha, and Xiaoqing Xu are with Arm Inc., Austin, TX. E-Mail: {Brian.Cline, Saurabh.Sinha, Xiaoqing.Xu}@arm.com

M3D ICs with optimized placement and routing, and fully explore its potential from power and thermal perspectives.

Power and thermal integrity have always been major challenges for 3D ICs. Studies show that 3D ICs suffer from larger static IR-drop [13] due to smaller footprint and higher temperature due to higher power density [14]. Therefore, it is important to consider the power delivery network (PDN) structure in M3D IC implementation and verify the power and thermal integrity for design sign-off.

In this paper, we present a physical design methodology for logic-on-memory M3D ICs, which includes optimized placement and routing (P&R) using commercial 2D EDA tools, a PDN design special for logic-on-memory integration, and power/thermal sign-off verification at the same time. We implement the high-performance Arm® Cortex®-A7 and Cortex®-A53 processors using TSMC 28nm technology as a case study. Workload-based power, voltage-drop, and thermal analysis are performed to verify the performance and reliability of the M3D ICs. The results show that the M3D ICs implemented with the proposed flow have 20% and 21% performance gain for the 2 benchmark processors respectively, while they still provide 2% energy saving, compared to the 2D counterparts. In addition, the voltage-drop of our M3D ICs is within 10% of the supply voltage (V_{dd}), and temperature increase is lower than 15°C.

The main contributions of this paper are: (1) an implementation flow for memory-stacked 3D ICs is presented; (2) in-depth analysis on the performance of commercial 32-bit and 64-bit processor is provided; (3) the potential benefits of memory stacking technology on power and thermal integrity are explored.

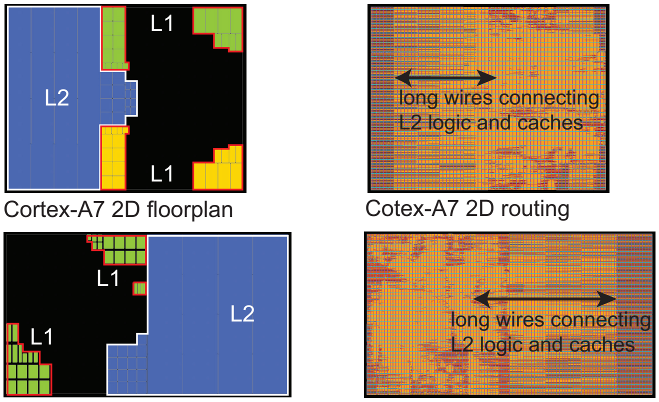
II. 2D Cortex-A Processor Designs

In this section, the motivation of this study is demonstrated with an analysis of the 2D Cortex-A processors. Memory plays a significant role in modern processor designs. With the emerging of machine learning, computer version, and other memory-intensive applications, the demands on memory capacity and bandwidth keep increasing. On the other hand, large memory blocks have proposed challenges for conventional 2D physical design flow: the large size of memory blocks makes floorplan very critical, while the interconnections between the logic and memory pins often become too long to optimize.

We implement the Cortex-A7 and Cortex-A53 processors with the conventional 2D flow and TSMC 28nm technology. We use TSMC 28nm technology (CLN28HPM) and the high-performance kit (HPK) to achieve high target frequency in these processors, which also allows us to extend the designs to 3D because we have access to the back-end-of-line (BEOL) technology files. The Cortex-A7 processor consists of 2 CPU cores, 32kB L1 instruction and data caches, and 512kB L2 cache. The Cortex-A53 processor consists of 1 CPU core, 32kB L1 instruction and data caches, and 1024kB L2 cache. The architectural configurations of the processors are summarized in Table I. We configure the processors this way to demonstrate the potential limits of 2D ICs in multi-core systems and designs with large memory blocks. The 2D ICs are implemented with 6 metal layers for signal routing and 2 extra metal layers (M7, M8) for the PDN.

TABLE I: Architectural configurations of our Arm Cortex-A processor benchmarks. NEON is an advanced Single Instruction Multiple Data (SIMD) architecture extension, and FPU is the floating point unit.

	Cortex-A7	Cortex-A53
architecture	Armv7-A	Armv8-A
instruction	32-bit	32/64-bit
# core	2	1
L1 cache (kB)	32	32
L2 cache (kB)	512	1024
L2 latency	2 cycles	1 cycle
NEON	present	present
FPU	present	present


L1
L2
L1
long wires connecting L2 logic and caches
L1
L2
L1
long wires connecting L2 logic and caches

Cortex-A7 2D floorplan

Cotex-A7 2D routing

Cortex-A53 2D floorplan

Cortex-A53 2D routing

Fig. 1: Cortex-A7 and Cortex-A53 2D layouts using TSMC 28nm.

Fig. 1 shows the 2D floorplan and layouts after P&R. The 2D floorplan is designed based on the reference flow provided by Arm. The memory blocks occupy more than 50% of the silicon area in 2D ICs, and thus their placement is critical for system performance. In this floorplan, the large memory blocks are clustered together to create a continuous empty area, which is beneficial for register-to-register paths, but results in longer register-to-memory paths. For example, the nets connecting the L2 cache blocks and the L2 controller logic are very long, and they heavily use the wires on the topmost routing layer (M6) as shown in the layouts. These global interconnections often generate considerable RC delay, limiting the performance of the whole system. With logic-on-memory stacking technology, these long 2D interconnections can be replaced by short 3D vertical interconnections, leading to a drastic improvement in timing.

Moreover, the gap between memory blocks cannot be too large, otherwise it will waste the silicon area and enlarge the critical path; it cannot be too small either to avoid routing congestion and signal integrity issues. Moreover, the clock tree becomes unbalanced with the existence of the memory blocks and the distance between the clock pins increases due to the large footprint. As a result, the quality of clock-tree-synthesis (CTS) tends to degrade, leading to large clock latency and skew.

In addition, moving large memory blocks to another tier also brings potential benefits in terms of power and thermal in-

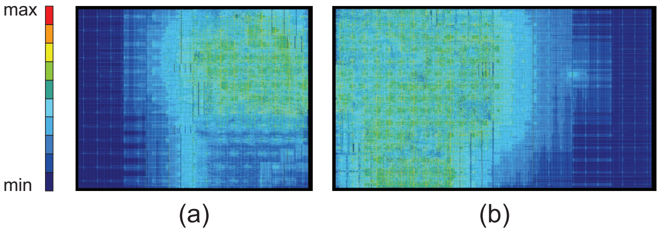


Fig. 2: The current map of wires and vias in (a) Cortex-A7 2D IC, and (b) Cortex-A53 2D IC.

TABLE II: Comparison between the Macro-3D flow [17] and our proposed flow.

	Macro-3D	proposed flow
target technology	face-to-face 3D	monolithic 3D
integration scheme	memory-on-logic	logic-on-memory
3D net #	less than 5,000	more than 300,000
PDN implementation	not supported	supported
power analysis	static	dynamic
votage-drop analysis	not supported	supported
thermal analysis	not supported	supported
benchmark design	open-source core	Arm processors

tegrity. Considering real workload in a processor, the switching activity of most memory blocks is often much lower than the logic blocks. Therefore, memory blocks have smaller current demand and power density. Fig. 2 shows the current map of the 2D ICs under workload-based analysis. We observe that the current demand is higher on the logic side but lower on the memory side. As a result, the power bumps on the memory side does not deliver current effectively. In logic-on-memory 3D ICs, we make use of the different switching activity and current demand of memory and logic components to overcome the drawbacks of high voltage-drop and temperature in other 3D ICs.

III. Limitations of Existing 3D P&R Flows

Shrunk-2D [7] and Compact-2D [8] are the state-of-the-art physical design flows for gate-level 3D ICs. They both place standard cells on a 2D plane, while Shrunk-2D shrinks the standard cell size by $\times\sqrt{2}$ but Compact-2D enlarges the 2D plane by $\times\sqrt{2}$. Although they have shown great performance and power benefits in a large set of benchmark designs, they still have some important drawbacks: (1) Complexity in handling memory blocks. Both Shrunk-2D and Compact-2D replace memory blocks with partial blockages to guide the placement of standard cells, but the placer does not always honor the blockage and thus the unevenly placed standard cells will overlap and cause timing degradation after tier partitioning; (2) Approximation in die-by-die routing. Both approaches route the top tier and bottom tier separately, so the timing constraints and load parasitics of the other die need to be estimated, which leads to accumulated errors, lower routing quality, and longer runtime.

Cascade-2D [9] does not have the limitations in macro handling, die-by-die routing. However, the number of MIVs is limited in this flow due to the use of anchor cells and dummy wires. In addition, it does not support PDN generation and analysis, and thus does not result in commercial-grade 3D layouts.

The authors of [15] proposed a floorplanner for hybrid monolithic 3D ICs, which shows significant area saving and power reduction in the OpenSPARC T2 processor. However, this approach cannot be directly used to implement logic-on-memory M3D ICs, because it handles gate-level monolithic logic and block-level monolithic memory on different tiers separately. Also, the floorplanner does not consider the potential power delivery challenges in the monolithic 3D ICs

The authors of [16] presented a transistor-level monolithic 3D design that shows great footprint and wirelength reduction. However, the transistor-level M3D requires significantly redesign of the standard cell library, and it also introduces thermal and power delivery challenges due to the tight coupling between the two tiers, which are have not been addressed in [16].

IV. Logic-on-Memory Cortex-A Processor Designs

A. *Overview of Our Logic-on-Memory Design Flow*

We develop the physical design methodologies for the logic-on-memory M3D ICs based on the Macro-3D flow proposed in [17]. Compared to the original Macro-3D flow, our methodology provides much higher 3D interconnection density, supports PDN implementation and power/thermal integrity analysis, and therefore results in more reliable 3D layouts for commercial-grade processors. These differences are summarized in Table II. We implement the logic-on-memory M3D ICs with modified technology files and commercial tools, and then verify the 3D designs from timing, power, and thermal perspectives.

Another methodology target at logic-on-memory M3D ICs is proposed in [18]. Compared to this method, our flow does not require a pre-partitioned netlist. Therefore, it allows the designers to freely change the floorplan and memory placement on the memory tier during the physical design stages, without rerunning the synthesis. We verify the proposed flow with the commercial Arm processor cores, and compare the design metrics with commercial-level 2D implementations. In addition, our flow incorporates power delivery network in the design and enables workload-based power delivery and thermal simulation in M3D ICs, which is not supported by [18] but critical to implement reliable monolithic 3D ICs.

In this paper, we consider 2-tier logic-on-memory M3D ICs with 6 routing metal layers on each tier: one tier only consists of memory blocks (namely the memory tier); the other tier can consist of both standard cells and memory blocks (namely the logic tier). In M3D ICs, MIVs need to go through the silicon substrate of one tier. If we insert MIVs in the memory tier, it is hard to place MIVs freely because memory blocks are large in size and MIVs are not allowed to overlap with them, which also makes it difficult to access the memory pins. Therefore, we add MIVs only in the logic tier.

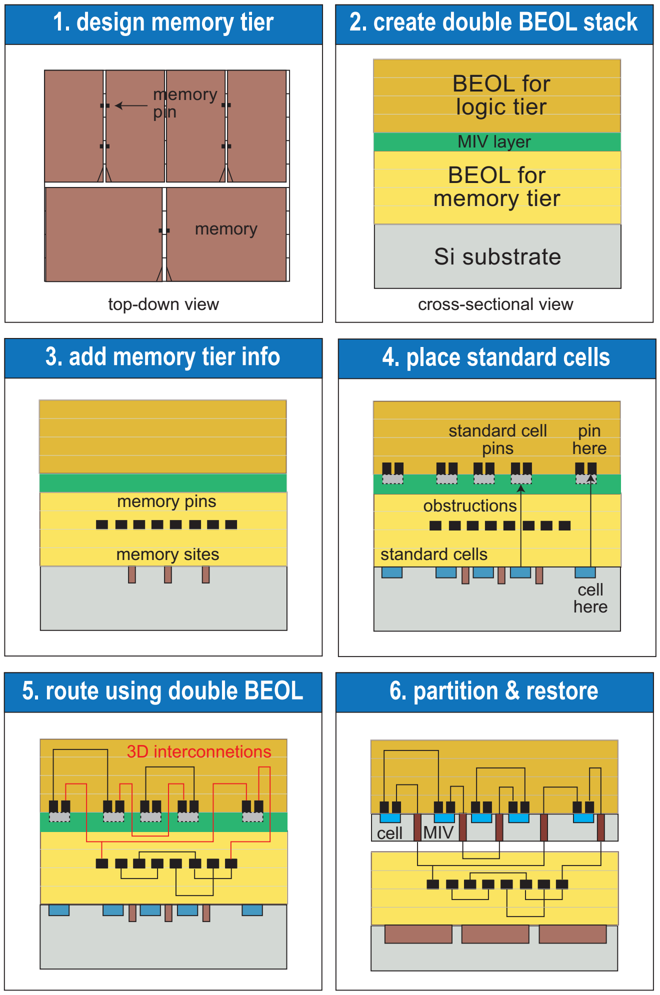


Fig. 3: Our P&R flow for logic-on-memory monolithic 3D IC layouts. Step 1 is done manually, step 2, 3, and 6 using our custom scripts, and step 4 and 5 using commercial tools for 2D ICs.

B. 3D Technology File Creation

The fundamental idea of this P&R flow is to shrink the memory blocks on the bottom tier, place them on the same 2D plane with the standard cells, make sure the cell and memory pins at the exact locations, and then use the commercial 2D tools to complete P&R The overview of the 3D P&R flow is shown in Fig. 3. The design exchange file (DEF) and library exchange file (LEF) of conventional 2D flows cannot represent the logic-on-memory 3D ICs directly. Therefore, we create technology files with double back-end-of-line (BEOL) stack to describe the structure of M3D ICs, as shown in Fig. 3. That is, we duplicate the 6 BEOL layer (M1-M6) information in the 2D LEF, remove the original top metal layer (M7-M8), generate the double BEOL stack with 12 layers (M1-M12), and create the MIV layer and its via cells.

In the modified LEF, the M1-M6 layers belong to the memory tier, while the M7-M12 layers belong to the logic tier, and the MIV layer is in between of M6 and M7. The size of MIVs is set to $70nm \times 70nm$, while the height is set to $140nm$. The RC parasitic technology file is also characterized for the double BEOL stack, based on the dimensions and electrical properties of each layer. This technology file enables the 2D RC extraction engine to extract the RC parasitics in M3D ICs, including the RC of the MIVs. In the final M3D ICs, the average capacitance of a single MIV is 0.13 fF, while the resistance is 9.06 Ω, similar to the values reported in previous research [19]. With this setup, a router for 2D ICs is able to complete routing in a 3D space and optimize the locations of MIVs as if they are regular vias.

After that, we modify LEFs for the memory blocks to create the memory sites, which have the minimum width and height allowed by the technology node. For each memory block, we define one filler cell with the size of the memory site but have the same pin and blockage locations as the original memory block. The size of the filler cell is so small that it can be placed on the same 2D plane with little impact on the placement of standard cells. For all the standard cells, we move their pins to the higher metal layers (M7-M12) to reflect the exact positions after tier partitioning. Moreover, to prevent MIVs from penetrating standard cells, we create an extra obstruction for each standard cell on the MIV layer with the same size as the cell itself. The placer and router are able to honor these obstructions and avoid placing the standard cells, signal MIVs or P/G MIVs at the same (x, y) location.

C. 3D Tier Degradation and Characterizations

In M3D ICs, the two tiers are fabricated sequentially, while the conventional high-temperature annealing procedure can potentially cause damage to the transistors and copper interconnects on the other tier, and lead to performance degradation. Two methods have been proposed to mitigate this issue [20]: 1) utilize a low-temperature process on the top tier; 2) employee heat-tolerate material, such as tungsten, for the interconnects on the bottom tier. In our logic-on-memory 3D ICs, there are fewer signal interconnects on the bottom memory tier, while the standard-cell transistors on the top logic tier have much greater impact on the performance of the M3D IC. Therefore, we adopt the second method to mitigate the issue and use tungsten for interconnects on the bottom tier from M1 to M6 (the material from M7 to M12 is still copper).

We characterize the parasitics of this heterogeneous BEOL stack based on the interconnect degradation model proposed in [20]. Using the same dimensions as the original BEOL stack in the 28-nm technology, we calculate the size-dependent resistivity for tungsten and regenerate the parasitic technology file using Cadence$^{\circledR}$ Quantus^{TM} QRC. Fig. 4 shows the changes of resistivity in different layers of the BEOL stack. The resistivity of tungsten is around 2.5× of copper, which can introduce higher wire resistance and delay into the design. The tungsten resistivity and the high-temperature process on the top tier also have an impact on the memory latency and power. In other to reflect this impact, we modify the cell delay and internal power tables for the memory blocks on the bottom tier by applying a scaling factor.

D. 3D Physical Design

Once the necessary technology files are created, we design the floorplan for the logic tier and memory tier separately. After that, we replace the memory blocks on the memory

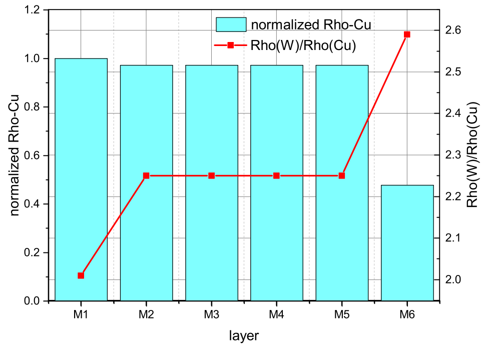


Fig. 4: The resistivity values of copper and tungsten for various metal layers in the BEOL stack.

tier with filler cells while making sure the memory pins remain at the same location, and then add the memory tier information to the 3D floorplan. With the 3D floorplan and modified technology files, we place and route the logic-on-memory M3D design directly with Cadence Innovus^{TM}, the 2D commercial engine. The standard cells and memory blocks are placed on the same 2D plane, but their pins are on different BEOL layers. This allows the placer to optimize the locations for standard cells considering the exact locations of the memory pins, and enable the router to establish the high-density 3D connections. Since the MIV size is much smaller than the standard cells, there is still enough space for the router to find the optimized locations for signal MIVs. Therefore, the generated layouts are effectively optimized with the commercial engine.

The clock distribution network (CDN) design in the M3D ICs is similar to the 2D ICs, because all the standard cells, including the clock gates, flip-flops, and buffers are placed on the logic tiers. The memory and the standard cells belong to the same clock domain, while only a small number of clock nets need to be connected to the clock pin of the SRAM blocks on the memory tier. We adopt the H-tree shape for the CDN, and utilize the Innovus clock tree synthesis (CTS) engine to implement the CDN in the M3D ICs. Thanks to the smaller footprint area, the shorter vertical interconnects, and the removal of memory obstructions, we can achieve a more balanced CDN in the M3D ICs with a shorter clock wirelength and lower averaged latency compared with the 2D CDN.

E. Logic-on-Memory Power Delivery Network Design

Before performing signal routing, we design the PDN for the logic-on-memory 3D ICs in the 3D space. Power bumps are placed on the topmost metal layer of the logic tier (M12). During the power plan stage, power and ground (P/G) rails are created from M2 to M6 and from M8 to M12 for logic tier and memory tier, respectively. Meanwhile, P/G via arrays are generated at the intersections of power rails, as shown in Fig. 5. Therefore, the locations of P/G MIVs are fixed and will not be affected by standard cell placement, which ensures the shortest power delivery path and overcomes the drawback of irregular P/G MIV placement in gate-level M3D ICs [13].

In addition, we design an unbalanced PDN for the logic tier and memory tier. As the logic tier has higher current demand and the current needs to go through the logic tier PDN to reach the memory tier, the PDN of the logic tier is more critical for the power integrity of the entire chip. Therefore, we reduce the power rail width and density on the memory tier and use wider and denser power rails on the logic tier. Table III shows the details about power bumps and PDN usage on each layer. Comparing with the 2D ICs, the 3D ICs have 50% fewer power bumps due to the smaller footprint area, which is a disadvantage for power delivery. Moreover, the 3D ICs do not have extra top metal layers for PDN, so the routing resources on the logic tier are also limited by PDN. However, with the unbalanced PDN structure, more routing resources of the memory tier are saved for signal routing. Therefore, we are able to ensure the power integrity in the M3D ICs without adding extra cost or causing routability problems.

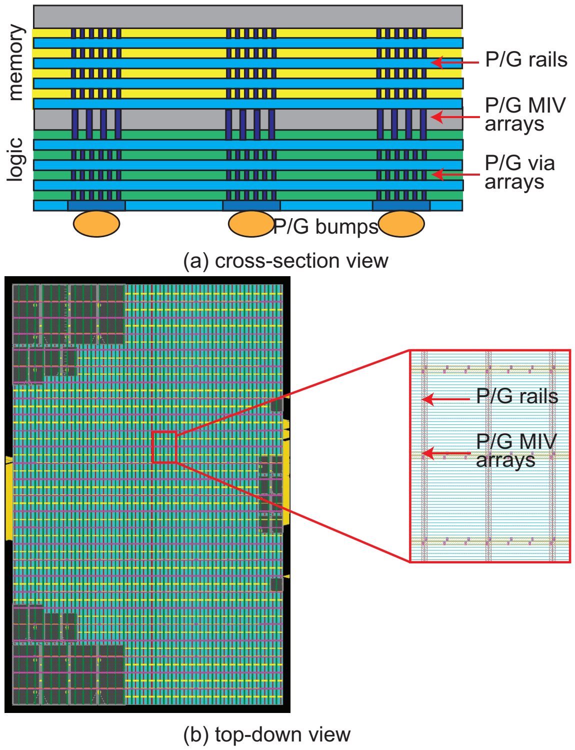
TABLE III: Power bump number and PDN routing track usage in 2D and 3D designs.

	Cortex-A7		Cortex-A53	
	2D	3D	2D	3D
footprint	1.00	0.50	1.00	0.50
bump #	218	104	385	190
layer-wise PDN usage				
M1	0.00%	0.00%	0.00%	0.00%
M2	21.23%	0.00%	21.45%	0.00%
M3	0.00%	5.80%	0.00%	15.13%
M4	0.00%	7.36%	0.00%	4.98%
M5	38.10%	15.64%	37.93%	15.75%
M6	39.17%	15.43%	39.43%	15.29%
M7	22.23%	0.00%	22.27%	0.00%
M8	15.79%	21.60%	14.91%	21.50%
M9	-	38.09%	-	37.61%
M10	-	23.73%	-	23.34%
M11	-	21.90%	-	22.05%
M12	-	16.09%	-	15.63%

F. Dynamic Power, voltage-drop, and Thermal Analysis

We perform workload-based dynamic analysis to verify the power, voltage-drop, and thermal integrity of the 3D ICs. Fig. 6 shows the workload-based analysis flow of the M3D ICs. The main difference of the proposed flow compared with other 3D flow is that it allows us to use commerical power and voltage-drop analysis tool on the M3D ICs directly, with the customized 3D technology files, which ensure the reliability of the sign-off analysis procedures. We use the Dhrystone benchmark program to evaluate the workload-based power consumption and voltage drop of the processors. For thermal analysis, we use the Max Power benchmark provided by Arm to verify the temperature of the M3D ICs at the high-power consumption scenario. With these benchmark programs, we perform simulation on the gate-level netlist and generate the Value-Change Dump (VCD) file, which records the voltage-change events for all the signals at the specific time window.

Using the VCD file, we perform time-based power analysis with Synopsys® PrimeTime® PX and vector-based rail analysis with ANSYS® RedHawk^{TM}. Since the power consumption and voltage drop do not depend on the geometries or vertical positions of the cells or memory blocks (the resistive effects of MIVs are already considered during RC extraction for the



(a) cross-section view

(b) top-down view

Fig. 5: Our power delivery network structure.

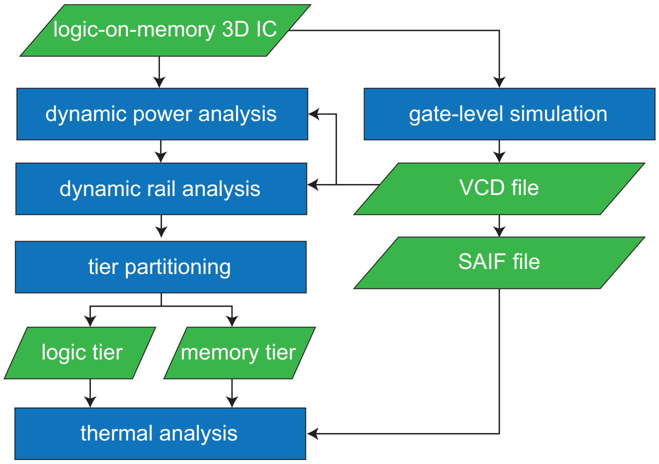


Fig. 6: The flowchart of workload-based power, voltage-drop and thermal analysis for the logic-on-memory 3D ICs.

M3D ICs), we perform the analysis on the 3D designs directly before tier partitioning and restoring.

Fig. 7 shows the package model we use for thermal analysis. Since MIVs can only be placed on the logic tier, the order of layers in the face-to-back (F2B) bonding structure is fixed as: memory tier front-end-of-line (FEOL), memory tier BEOL, logic tier FEOL, logic tier BEOL. In order to connect to the C4 bumps on the package, the chip needs to be flipped for bonding. As a result, the logic tier is placed on the bottom of the package, and the memory tier is on the top. This structure has a negative impact on temperature because the power-intensive logic tier is placed far from the heat sink, but it provides significant benefits on timing and power delivery in the logic tier thanks to the short path to the package.

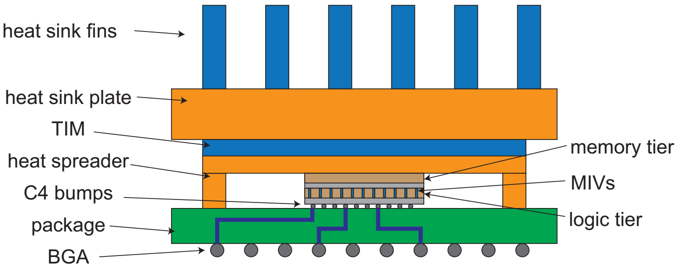


Fig. 7: Our package model with a air-cooling solution for thermal simulation of logic-on-memory M3D ICs.

For thermal analysis, we average the switching activity in the VCD file for each net and create the Switching Activity Interchange Format (SAIF) file to simulate a stable state. We use ANSYS® RedHawk-CTA™ tool to perform the workload-based thermal simulation. The tool requires us to generate the chip-thermal model (CTM) for each tier separately, because the geometries and vertical positions of cells have an impact on the heat generation and dissipation in the entire system. Therefore, we partition the M3D IC into two tiers, scale the memory blocks on the memory tier to the original size, annotate the switching activity of the nets in each tier with the block-level SAIF file, and then generate the CTM for each tier. We also design a dummy package substrate with power and signal routing, in order to simulate the thermal conductivity of the real package. The two CTMs are stacked onto the package model in a F2B bonding fashion. In addition, We assume an air-cooling solution with a heat spreader and fins for the thermal simulation.

During the thermal simulation, the RedHawk-CTA tool divides the design into a rectangular mesh on the XY plane and calculate the averaged power density and thermal conductivity within each tile. The highest mesh resolution is $5\mu m \times 5\mu m$. However, the MIV size is much smaller than the mesh resolution, so each individual MIV is not modeled directly in this flow, while it has an impact on the thermal conductivity of the tile. Therefore, the runtime and accuracy of the thermal simulation flow are affected by the mesh resolution. For Cortex-A7 and Cortex-A53, the resolution range from $5\mu m$ to $20\mu m$ can provide a good balance between accuracy and efficiency.

V. Experimental Results

A. *Timing and Power Analysis*

The Cortex-A7 and Cortex-A53 processors are implemented with the proposed flow using the same configurations and target frequencies for 2D ICs, as described in Section II. For Cortex-A7, we move 24 L2 cache blocks and 24 L1 cache blocks to the memory tier. For Cortex-A53, we move 16 L2 cache blocks to the memory tier. The memory block placement is arranged in a way that the 3D floorplan is the direct folding of the 2D floorplan along the horizontal direction. As a result, the M3D ICs have only half the footprint area of the 2D counterparts. Fig. 8 and Fig. 9 show the 3D floorplan and GDS layouts. The layout of the MIV layer on

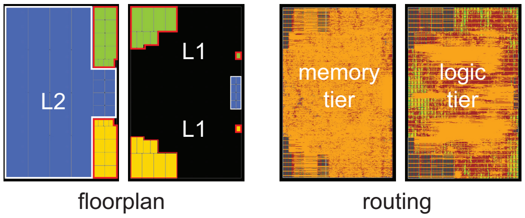


Fig. 8: Cortex-A7 logic-on-memory 3D layouts using TSMC 28nm.

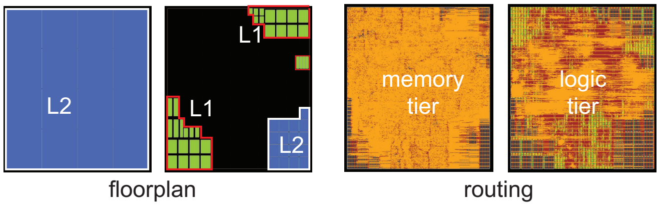


Fig. 9: Cortex-A53 logic-on-memory 3D layouts using TSMC 28nm.

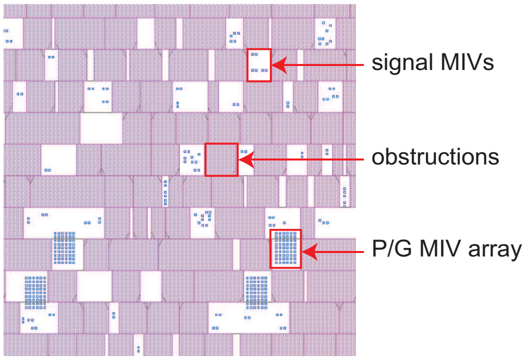
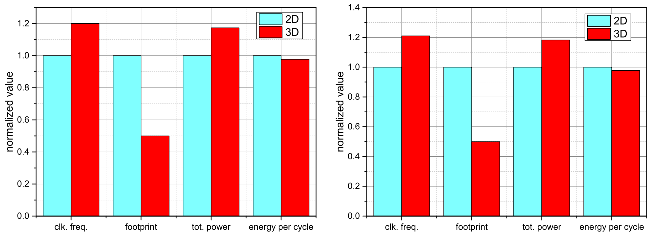


Fig. 10: Zoom-in layout of the MIV layer in our Cortex-A7 M3D.



(a) Cortex-A7 2D and M3D ICs (b) Cortex-A53 2D and M3D ICs

Fig. 11: PPA comparisons of the 2D and M3D ICs. All the M3D values are normalized with regard to the value of the 2D IC.

Fig. 10 shows that the P/G MIV arrays are placed regularly without overlapping with the standard cells, while the signal MIVs are inserted in the gaps of standard cells. There are 350k and 588k MIVs in the Cortex-A7 and Cortex-A53 M3D ICs, respectively, suggesting that this flow heavily utilizes the vertical interconnection capacity provided by M3D.

We perform static timing analysis with PrimeTime to evaluate the performance of the 2D and M3D ICs, and report the slacks at the typical corner (25°C). We sweep the target frequency to find the maximum frequencies at which all the timing paths meet the timing constraints. The timing constraints are also provided by Arm, including all the I/O delays and multi-cycle path specifications. And we assume the external clock latency is equivalent to the average of internal clock latency. Table IV and Fig. 11 shows the power, performance, area (PPA), voltage drop, and the thermal metrics of the 2D and M3D ICs at the maximum frequencies. According to the experimental results, the M3D ICs show 20% and 21% performance gain for Cortex-A7 and Cortex-A53, respectively. This timing improvement comes from 3D routing optimization, the smaller footprint, better clock tree, shorter register-to-memory and register-to-I/O paths in the logic-on-memory M3D ICs.

Table V shows the breakdown of critical path delays in the 2D and 3D designs. We also analyze the most critical register-to-register paths and show them in Fig. 12. In the M3D ICs, the critical path often consists of wires on both the logic tier and memory tier, even if the start and end points of the path are both on the logic tier. The reason is that the router tends to utilize the relatively empty routing space on the memory tier to mitigate the routing congestion on the logic tier. This inter-tier routing optimization leads to lower wire delay in M3D ICs. Compared with the conventional die-by-die routing method, the signal MIV locations are not predefined in this flow, but determined by the router and iteratively optimized during routing, which also explains the high MIV usage.

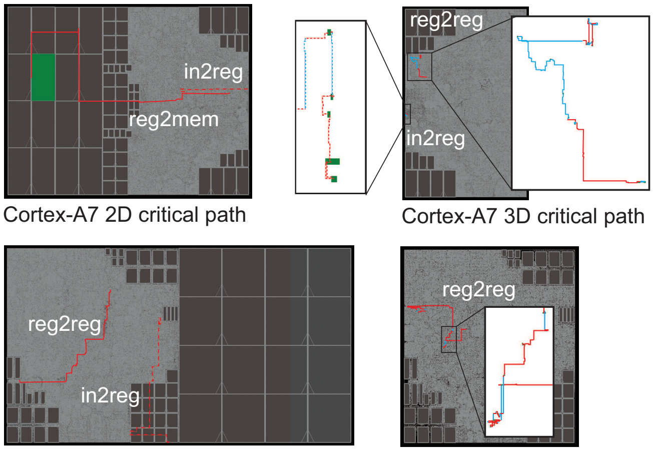
In addition, the smaller footprint of M3D ICs reduces the distance from the clock pin to registers. Also, with M3D integration, the long 2D nets connect the clock distribution network and the memory pins are replaced with short 3D nets, as shown in the Figure 13. Therefore, the maximum gate-to-gate wirelength in the CDN reduces by 13% and 50% in the Cortex-A7 and Cortex-A53 M3D ICs, compared with the 2D baselines, respectively. And the maximum depth of the CDNs is also reduced for both processors in 3D. This shows that the 3D CDN created by the proposed flow is more balanced and cost-efficient than the 2D ones. As a result, the 3D CDNs provide more than 35% lower averaged latency for the two processors (as shown in Table V), which significantly benefits the critical paths.

Besides, the long interconnections between registers and memory blocks are folded and shortened in the M3D ICs, as shown on Fig. 14. In the M3D ICs, the bottom-tier memory blocks have easy access to the standard cells on the top tier, while the top-tier memory blocks also have better connectivity due to reduced memory macros and obstructions surrounding them, and they have access to bottom die routing resources. As a result, the memory input and output latencies significantly reduce, compared to the 2D baselines, as shown in Table VI. Also, the decrease of memory net wirelength help to save the memory switching power. For example, the register-to-memory path in the Cortex-A7 2D IC is a timing bottleneck, but it becomes not critical in the M3D counterpart. Moreover, it is easier to access the top-level I/Os in the M3D ICs due to the absence of memory obstructions and smaller footprint, leading to shorter I/O-to-register paths. As a result, the worst-case input-to-register path delay reduces by 26% in Cortex-A7 by moving from 2D to 3D. Another example is in the Cortex-A53 2D IC, the most critical path is from an input port to a register; but in the 3D counterpart, the input-to-register path becomes less critical while the register-to-register

TABLE IV: Commercial 2D vs. our logic-on-memory 3D in terms of PPA, voltage-drop, and thermal integrity. We use TSMC 28nm technology. We normalize our data based on the 2D results for each benchmark, except that the worst slack is the percentage of the clock period, and voltage-drop is the percentage of the supply voltage of $0.9V$. Δ denotes the percentage difference between M3D and 2D. The green means M3D wins, and the red M3D loses.

Cortex-A7							
flow	2D	M3D	Δ	flow	2D	M3D	Δ
clk. freq.	1.00	1.20	20.07%	tot. power	1.00	1.17	17.39%
footprint	1.00	0.50	-50.00%	sw. power	0.28	0.34	20.12%
wirelength	1.00	1.00	-0.49%	int. power	0.55	0.66	21.30%
MIV count	0	349,978	-	leak. power	0.17	0.17	0.22%
density (%)	79.40	79.26	-0.18%	logic power	0.27	0.28	4.73%
worst slack (%)	0.00	0.11	-	seq. power	0.42	0.51	19.24%
total cap	1.00	1.00	0.01%	clk. power	0.21	0.27	27.07%
pin cap	0.43	0.42	-2.07%	macro power	0.10	0.12	23.45%
wire cap	0.57	0.58	1.55%	energy per cycle	1.00	0.98	-2.23%
volt. drop (%)	6.56	8.59	30.91%	temperature (°C)	59.28	69.99	18.07%
std. cell area	1.00	1.02	2.33%				

Cortex-A53							
flow	2D	M3D	Δ	flow	2D	M3D	Δ
clk. freq.	1.00	1.21	21.02%	tot. power	1.00	1.18	18.26%
footprint	1.00	0.50	-50.00%	sw. power	0.14	0.17	17.54%
wirelength	1.00	0.97	-3.43%	int. power	0.77	0.93	20.78%
MIV count	0	588,161	-	leak. power	0.09	0.09	-2.66%
density (%)	72.54	69.92	-3.61%	logic power	0.07	0.07	-3.10%
worst slack (%)	0.00	0.00	-	seq. power	0.30	0.36	20.28%
total cap	1.00	1.00	-1.49%	clk. power	0.17	0.20	18.13%
pin cap	0.43	0.42	-3.57%	macro power	0.46	0.55	20.31%
wire cap	0.57	0.58	-0.28%	energy per cycle	1.00	0.98	-2.28%
volt. drop (%)	7.29	7.71	5.83%	temperature (°C)	54.58	67.98	24.55%
std. cell area	1.00	1.02	1.71%				



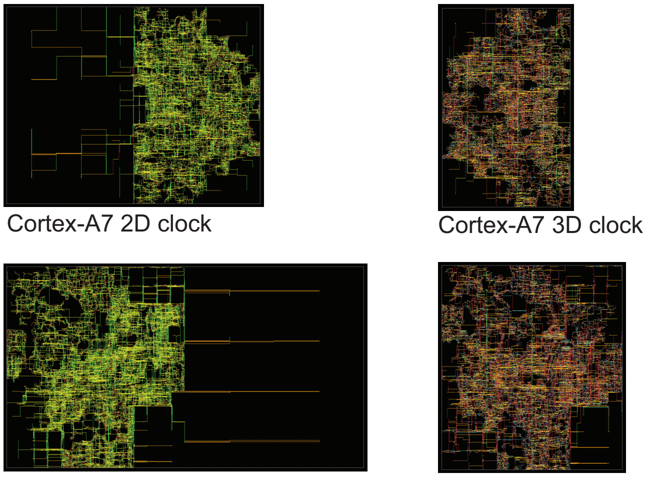
Cortex-A7 2D critical path

Cortex-A7 3D critical path

Cortex-A53 2D critical path

Cortex-A53 3D critical path

Fig. 12: Timing critical path comparisons. The green squares represent the memory block or the standard cells on the critical paths; the solid lines represent the wires on the register-to-register critical paths, while the dash lines represent the wires on input-to-register critical paths; the red lines represent the wires on M1-M6, while the blue lines represent the wires on M7-M12 in 3D ICs



Cortex-A7 2D clock

Cortex-A7 3D clock

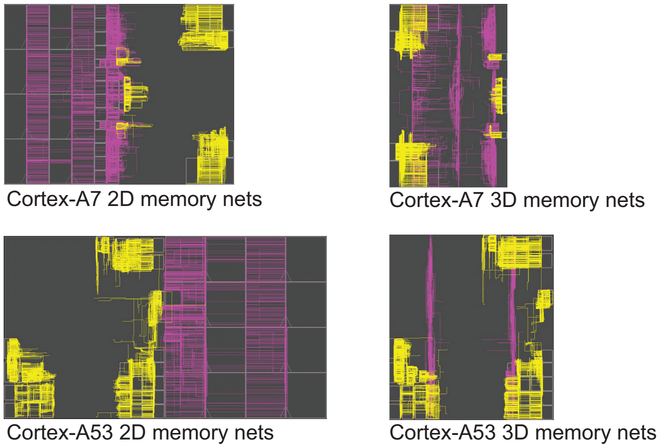
Cortex-A53 2D clock

Cortex-A53 3D clock

Fig. 13: Clock tree comparison. The 3D clock tree figures show clock nets on both the logic and memory tier projected on the 2D plane.

path determines the timing closure. In summary, the benefits of M3D ICs help remove the timing bottlenecks in 2D ICs and lead to the an over 20% performance boost for the two processors.

The right two columns on Table IV and the Fig. 15 show the results of workload-based power analysis with the Dhrystone benchmark. The power is also reported at the typical corner. The M3D ICs have 17.39% and 18.26% higher power consumption for Cortex-A7 and Cortex-A53, respectively. This is because they run at a higher frequency. Although the wirelength saving in logic-on-memory M3D ICs is not high, it also helps to reduce the RC parasitics and switching power. The timing benefits of M3D allow the M3D ICs to meet the timing constraints without too much buffering, which saves the logic power. The energy consumption of the 2D and 3D ICs is evaluated by energy per cycle, the product of the clock period and total power. Based on this metric, the 3D ICs have


Cortex-A7 2D memory nets

Cortex-A7 3D memory nets

Cortex-A53 2D memory nets

Cortex-A53 3D memory nets

Fig. 14: Memory net comparison. The magenta lines show the nets connected to the bottom-tier memory blocks; the yellow lines show the nets connected to the top-tier memory blocks. The figures of the 3D ICs show the memory nets of both tiers projected on the same 2D plane.

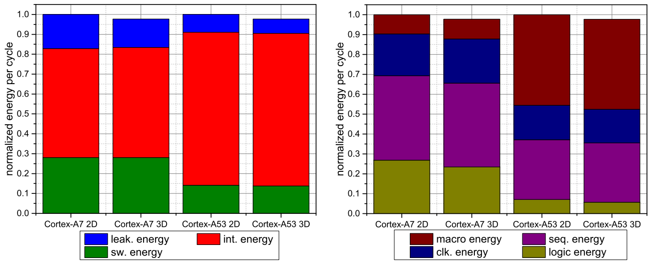
TABLE V: Clock tree and critical path in 2D and 3D designs. We normalize the latency and delay data. Our baseline is 2D clock period for each benchmark.

	Cortex-A7		Cortex-A53	
	2D	M3D	2D	M3D
clk. freq	1.00	1.20	1.00	1.21
Clock Tree				
max clock WL	1.00	0.87	1.00	0.50
max depth	32	27	33	26
avg latency	0.54	0.34	0.63	0.38
max skew	0.46	0.45	0.50	0.48

Critical Path				
type	in2reg	in2reg	in2reg	reg2reg
clock period	1.00	0.83	1.00	0.83
launch latency	0.54	0.34	0.63	0.52
capture latency	0.58	0.27	0.54	0.41
clock skew	-0.04	0.07	0.09	0.11
cell delay	0.43	0.22	0.22	0.54
wire delay	0.20	0.04	0.18	0.18
pin delay	0.40	0.50	0.50	0.00
path delay	1.03	0.76	0.91	0.72
setup time	0.01	0.01	0.01	0.00
slack	0.00	0.09	0.00	0.00

TABLE VI: Memory nets analysis in the 2D and 3D ICs. All the 3D power and latency numbers are normalized to the 2D baselines.

Cortex-A7			
flow	2D	3D	Δ
bottom-tier memory nets			
Switching Power	1.000	0.817	-18.3%
Max Input Latency	1.000	0.590	-41.0%
RMS Input Latency	1.000	0.800	-20.0%
Max Output Latency	1.000	0.561	-43.9%
RMS Output Latency	1.000	0.400	-60.0%
top-tier memory nets			
Switching Power	1.000	0.894	-10.6%
Max Input Latency	1.000	0.487	-51.3%
RMS Input Latency	1.000	0.806	-19.4%
Max Output Latency	1.000	0.475	-52.5%
RMS Output Latency	1.000	0.900	-10.0%
Cortex-A53			
flow	2D	3D	Δ
bottom-tier memory nets			
Switching Power	1.000	0.814	-22.9%
Max Input Latency	1.000	0.789	-21.1%
RMS Input Latency	1.000	0.759	-24.1%
Max Output Latency	1.000	0.372	-62.8%
RMS Output Latency	1.000	0.609	-39.1%
top-tier memory nets			
Switching Power	1.000	0.871	-12.9%
Max Input Latency	1.000	0.858	-14.2%
RMS Input Latency	1.000	0.758	-24.2%
Max Output Latency	1.000	1.000	0.0%
RMS Output Latency	1.000	0.765	-23.5%



(a) Energy breakdown based on power dissipation types.

(b) Energy breakdown based on components.

Fig. 15: Breakdown of energy consumption in the 2D and M3D ICs. All the M3D values are normalized with regard to the value of the corresponding 2D IC.

2.23% and 2.28% energy saving compared with 2D ICs for Cortex-A7 and Cortex-A53, respectively.

B. M3D Performance Degradation Analysis

However, the performance uplifts provided by the M3D ICs may degrade if we consider the process variation introduced by the M3D fabrication process. We can mitigate this issue by using tungsten interconnects on the bottom tier, but it still introduces additional wire delay and memory latency. To quantify this impact, we rerun parasitic extraction with the characterized technology file for the heterogeneous BEOL stack and perform timing analysis. Fig. 16 shows the critical path in Cortex-A7 and Cortex-A53 M3D ICs with the tungsten interconnects. The results show that the performance uplift in Cortex-A7 reduces from 20% to 13%, while the gain in Cortex-A53 decreases from 21% to 20%, respectively. Cortex-A7 suffers from larger performance degradation because the M3D ICs utilize more space on the memory tier for routing, which leads to increasing wire resistance. On the other hand, for Cortex-A53, the impact of tungsten resistivity is smaller, since the critical path is mainly on the logic tier and does not involve the memory blocks.

C. Voltage-Drop Analysis

The Workload-based voltage-drop analysis is performed to verify the power integrity of the 2D and M3D ICs. Fig. 17 shows the resulting voltage-drop map. According to our experimental results, the worst-instance voltage-drops in both the 2D and M3D ICs are within 10% of the supply voltage ($V_{dd} = 0.9V$), while the voltage-drop degradation of M3D ICs is just 2.03% of V_{dd} and 0.42% of V_{dd} compared to 2D ICs

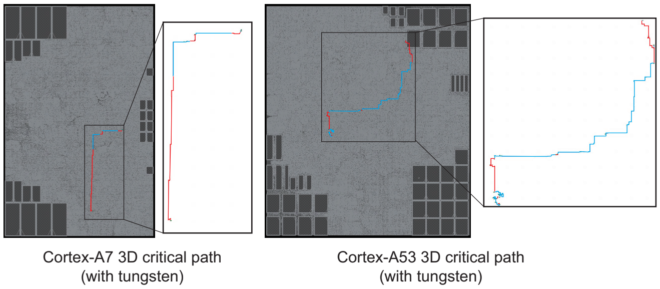


Fig. 16: Timing critical in the M3D ICs with tungsten interconnects on the memory tier; the red lines represent the wires on M1-M6, while the blue lines represent the wires on M7-M12 in 3D ICs

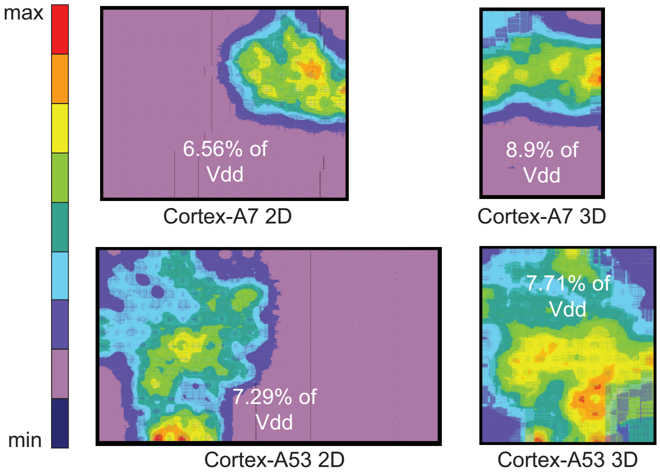
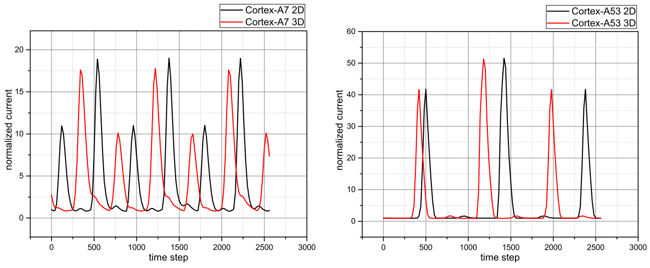


Fig. 17: Voltage-drop analysis results.

for Cortex-A7 and Cortex-A53. That is because the proposed logic-on-memory M3D flow overcomes multiple drawbacks in previous PDN designs for M3D ICs: (1) shorter resistive path from power sources to sinks: in our M3D ICs, the logic tier is close to the power source, while memory blocks on the bottom tier have power pins on higher metal layers, and therefore the worst-case power path length in 3D is similar to that in 2D; (2) no irregular power MIV placement: in the proposed flow, power MIV locations are predetermined at the power plan stage and are not affected by the standard cell placement. Thus our M3D ICs do not have extra voltage drop caused by power routing detour [13]; (3) optimized PDN dimensions considering switching activity: the memory blocks on the bottom tier have much lower switching activity and draw less current, and we use wider and denser power rails on logic tier to compromise this unbalance, so that the additional voltage drop in 3D is minimized; (4) reduction of current: with the optimized standard cell placement, the M3D ICs consume less current with the same workload and higher frequency. For example, the peak current in the Cortex-A7 M3D IC is 6.64% lower than the 2D IC, as shown in Fig. 18. As a result, the logic-on-memory stacked M3D ICs implemented with the proposed flow meet the power integrity requirements.

D. Thermal Analysis

We also perform thermal analysis to verify the reliability of the 3D ICs under the real workload. Fig. 19 shows the temperature map of the device layer of each design, with a mesh resolution equal to $5\mu m$. The results show that the maximum temperature of the M3D ICs are just 10.71°C and 13.40°C higher than the 2D counterparts for Cortex-A7 and Cortex-A53, respectively. On the one hand, M3D ICs run at higher frequencies and have higher power consumption, resulting in a higher heat generation rate. In addition, the memory tier is placed on the top of the logic tier after chip flipping, which impedes the heat dissipation from the logic tier to the environment. On the other hand, we observe that in both the 2D and M3D ICs, the logic blocks consume more power and have higher temperatures, while the memory blocks, especially the L2 caches, are relatively cold. For example, in the Cortex-A7 M3D temperature map, the logic tier has multiple hot spots in the two CPU core regions, while the memory tier only has a few small regions with slightly higher temperatures, which belong to the L1 caches. With this unbalanced power and heat distribution, logic-on-memory M3D only slightly increases the power density after stacking and it is less likely to create extra hot spots. Therefore, logic-on-memory M3D mitigates the thermal integrity issues in 3D ICs.



(a) Current waveform of the Cortex-A7 2D and M3D ICs.

(b) Current waveform of the Cortex-A53 2D and M3D ICs.

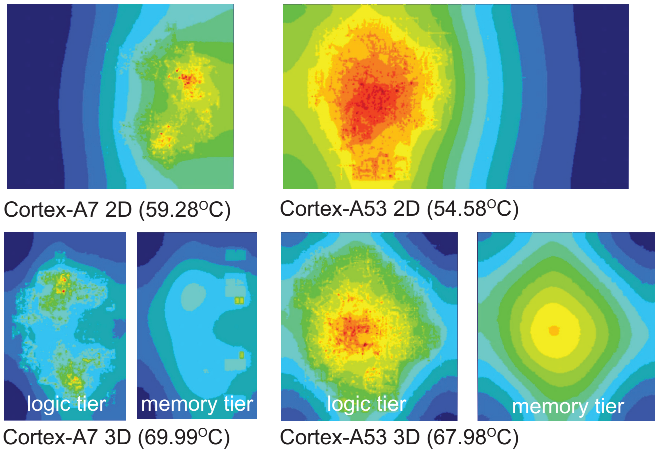
Fig. 18: Current waveform comparison. The current values are normalized to the 2D baselines.

TABLE VII: Runtime and accuracy of the thermal analysis flow in Cortex-A7 with various mesh resolution. The relative error refers to the error of the maximum junction temperature compared with the value obtained with a $5\mu m$ mesh.

mesh resolution (μm)	runtime (s)	relative error (%)
5	204	0.00%
10	68	1.64%
20	59	1.64%
40	56	2.40%
80	55	2.79%

Using the thermal analysis results with a $5\mu m$ mesh as the baseline, we analyze the trend of runtime and accuracy in Cortex-A7 with various mesh resolutions, as shown in Table VII. With a lower resolution, the runtime of the flow is improved, but the relative error of the maximum junction temperature also increases. Therefore, it is important to determine an appropriate mesh resolution for thermal simulation in large-scale M3D ICs to balance the runtime and accuracy.

We consider two what-if scenarios to investigate the thermal properties of the M3D ICs. First, we swap the position of the logic tier and memory tier and assume the performance and power consumption of the design do not change, in order to



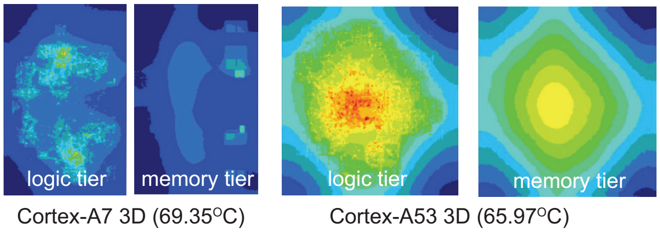
Cortex-A7 2D (59.28°C)

Cortex-A53 2D (54.58°C)

Cortex-A7 3D (69.99°C)

Cortex-A53 3D (67.98°C)

Fig. 19: Thermal analysis results. In M3D ICs, the logic tier is hotter.



Cortex-A7 3D (69.35°C)

Cortex-A53 3D (65.97°C)

Fig. 20: Thermal analysis results in M3D ICs after swapping the position of the logic tier and the memory tier.

analyze the impact of die stacking order on the temperature. Fig. 20 shows the temperature maps in the M3D ICs after swapping the tiers. The maximum junction temperature decreases by 0.6°C in Cortex-A7, and by 2.0°C in Cortex-A53, respectively. The differences are not significant, because the total power density of the M3D IC remains the same, and the C4 bumps and the metal traces in the package also forms an effective heat dissipation path from the logic tier to the PCB in the original 3D stack. Therefore, the order of die stacking has only a small impact on temperature under the specific workload and air-cooling solution. However, the results can be different if more advanced cooling techniques such as water cooling are applied, because the heat dissipation through the heat sink will be more effective.

Second, we analyze the impact of MIV density on the temperature in the M3D ICs. We find that although the number of MIVs is large ($> 300k$), the area occupied by MIVs is relatively small (¡0.2% in both Cortex-A7 and Cortex-A53) compared with the entire floorplan in the M3D ICs, due to the small size of individual MIVs. As a result, the impact of MIVs on the vertical thermal conductivity is nearly negligible. Assuming the power consumption does not change, we perform another experiment on the M3D ICs by removing all the MIVs and rerunning the thermal simulation. The results show that the temperature changes are within 0.1° in the M3D ICs after the removal. Therefore, the impact of MIV density on temperature is not obvious in these M3D ICs.

VI. Conclusions

A physical design flow for logic-on-memory M3D ICs is proposed in this paper. Using the double BEOL stack and MIV obstructions, it allows the commercial 2D tools to optimize the P&R of the M3D ICs, leading to 20% and 21% performance improvement in the Arm Cortex-A7 and Cortex-A53 processors. Workload-based power analysis shows that the logic-on-memory M3D designs have around 2% energy saving. They also mitigate the power and thermal integrity issues in 3D ICs, and have less than 3% of V_{dd} voltage-drop overhead and lower than 15°C temperature increase, compared to the 2D counterparts.

References

[1] I. L. Markov, "Limits on fundamental limits to computation," Nature, vol. 512, no. 7513, pp. 147–154, 2014.

[2] J. H. Lau, "Evolution, Challenge, and Outlook of TSV, 3D IC Integration and 3D Silicon Integration," in 2011 international symposium on advanced packaging materials (APM). IEEE, 2011, pp. 462–488.

[3] Z. Li, Y. Li, and J. Xie, "Design and Package Technology Development of Face-to-Face Die Stacking as a Low Cost Alternative for 3D IC Integration," in 2014 IEEE 64th Electronic Components and Technology Conference (ECTC). IEEE, 2014, pp. 338–341.

[4] M. Vinet, P. Batude, C. Fenouillet-Beranger et al., "Monolithic 3D Integration: A Powerful Alternative to Classical 2D Scaling," in 2014 SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S). IEEE, 2014, pp. 1–3.

[5] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, J.-M. Hartmann, L. Sanchez, L. Baud, V. Carron et al., "Advances in 3D CMOS Sequential Integration," in 2009 IEEE International Electron Devices Meeting (IEDM). IEEE, 2009, pp. 1–4.

[6] P. Batude et al., "3-d sequential integration: A key enabling technology for heterogeneous co-integration of new function with cmos," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 2, no. 4, pp. 714–722, 2012.

[7] S. Panth, K. Samadi, Y. Du et al., "Shrunk-2-D: A Physical Design Methodology to Build Commercial-quality Monolithic 3-D ICs," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 10, pp. 1716–1724, 2017.

[8] 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 Proceedings of the 2018 International Symposium on Physical Design. ACM, 2018, pp. 90–97.

[9] K. Chang, S. Sinha, B. Cline et al., "Cascade2D: A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools," in Proceedings of the 35th International Conference on Computer-Aided Design. ACM, 2016, p. 130.

[10] D. H. Kim, K. Athikulwongse, M. B. Healy et al., "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," IEEE Transactions on Computers, vol. 64, no. 1, pp. 112–125, 2013.

[11] M. M. Shulaker, T. F. Wu et al., "Monolithic 3D Integration of Logic and Memory: Carbon Nanotube FETs, Resistive RAM, and Silicon FETs," in IEEE International Electron Devices Meeting, 2014.

[12] S. S. K. Pentapati, L. Zhu, L. Bamberg et al., "A Logic-on-Memory Processor-System Design with Monolithic 3D Technology," IEEE Micro, 2019.

[13] K. Chang, S. Das, S. Sinha et al., "Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs," in 2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED). IEEE, 2017, pp. 1–6.

[14] S. K. Samal, S. Panth, K. Samadi et al., "Fast and Accurate Thermal Modeling and Optimization for Monolithic 3D ICs," in 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC). IEEE, 2014, pp. 1–6.

[15] A. Guler and N. K. Jha, "Hybrid monolithic 3-d ic floorplanner," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, no. 10, pp. 1868–1880, 2018.

[16] J. Shi et al., "On the design of ultra-high density 14nm finfet based transistor-level monolithic 3d ics," in 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). IEEE, 2016, pp. 449–454.

[17] L. Bamberg et al., "Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs," in 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2020.

[18] S. Thuries et al., "M3d-adtco: monolithic 3d architecture, design and technology co-optimization for high energy efficient 3d ic," in 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2020, pp. 1740–1745.

[19] S. K. Samal, D. Nayak, M. Ichihashi et al., "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET Technology," in 2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S). IEEE, 2016, pp. 1–2.

[20] S. Panth et al., "Tier degradation of monolithic 3-d ics: A power performance study at different technology nodes," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 8, pp. 1265–1273, 2017.

[21] H. Wei et al., "Cooling three-dimensional integrated circuits using power delivery networks," in 2012 International Electron Devices Meeting. IEEE, 2012, pp. 14–2.

VII. Novelty Statement

This paper is not an extension of any published papers, though there is an overlap with previous work. In our research, we have developed a complete EDA tool flow for logic-on-memory monolithic 3D ICs from physical implementation to sign-off verifications. The idea to project the memory pins to the logic tier and complete 3D routing with 2D tools is similar to the 2020 DATE paper, "Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs". However, in our flow, we consider the face-to-back bonding structure of M3D ICs, address the MIV placement problems with cell obstructions, achieve much higher 3D connection density, and enable voltage-drop and thermal analysis. The detailed comparisons between the proposed flow and the previous work are also described in Section III and IV. Experimental results show that the M3D processors implemented with our flow provide more than 20% performance uplift compared with their 2D counterparts, while the voltage-drop is within 10% of the supply voltage and the temperature increase is lower than 15°C.

Except for the overlap described above, the contents of this manuscript are entirely new and original.

Appendix A
Response to Reviewers

At first, we would like to thank the editors and reviewers for their service. We carefully addressed the reviewers' comments in the revised manuscript as attached. We also highlighted the changes in the manuscript with the blue, red, and green colors regarding Reviewer 1, 2, and 3's comments, respectively. Our response and changes are summarized as bellow.

A. Reviewer 1

Comment 1: The process variation as a result of M3D fabrication process, results in severe energy and latency penalties. This is not discussed in the paper and how the current flow incorporates it. The result of the process variant M3D has to be compared against the 2D counterpart to give a realistic comparison in the performance.

Response: Thank you very much for the suggestion. We now consider the performance degradation caused by M3D fabrication process in our implementation and analysis flow and mitigate the impacts with tungsten-based heterogeneous 3D interconnects. As described in **Section IV-C**, two methods have been proposed to mitigate this issue [20]: 1) utilize a low-temperature process on the top tier; 2) employee heat-tolerate material, such as tungsten, for the interconnects on the bottom tier. In our logic-on-memory 3D ICs, there are fewer signal interconnects on the bottom memory tier, while the standard-cell transistors on the top logic tier have a much greater impact on the performance of the M3D IC. Therefore, we adopt the second method to mitigate the issue. Tungsten has a much higher melting point ($> 3500K$) and allows high-temperature annealing for the transistors on the logic die. With this heterogeneous 3D technology, the standard cells on the logic die have no performance degradation, while the interconnects on the memory die introduces higher wire delay and the SRAMs suffers from larger latency due to the higher resistivity of tungsten. Our experimental results in **Section V-B** show that the performance gain in M3D ICs reduces from 20% to 14% in Cortex-A7 and from 21% to 20% in Cortex-A53.

Comment 2: Further, the technology aspect of the transistor is not discussed clearly. I would like to see the transistor and the interconnect characteristics and the potential transistor performance degradation between tiers, that needs to be modeled into the flow.

Response: Thank you very much for the comment. We describe the procedure to characterize the interconnects and memory devices in **Section IV-C**. First, we characterize the parasitics of this heterogeneous BEOL stack based on the interconnect degradation model proposed in [20]. Using the same dimensions as the original BEOL stack in the 28-nm technology, we calculate the size-dependent resistivity for tungsten and regenerate the parasitic technology file using Cadence® Quantus™ QRC. Fig. 4 shows the changes of resistivity in different layers of the BEOL stack. The resistivity of tungsten is around 2.5× of copper, which can introduce higher wire resistance and delay into the design. The tungsten resistivity and the high-temperature process on the top tier also have an impact on memory latency and power. In other to reflect this impact, we modify the cell delay and internal power tables for the memory blocks on the bottom tier by applying a scaling factor.

Comment 3: One of my main concern is that the content of the paper mostly discusses the issues in the current methodologies with comparison and the entire crux of the paper is limited to Sections IV and V only. The content needs to be expanded comprehensively to be accepted as a Regular paper.

Response: Thank you very much for the comment. We have included the in-depth analysis of performance degradation caused by M3D fabrication, a detailed description of the 3D thermal model, and the guidelines for thermal and voltage drop analysis in the **Section III, IV and V**, as suggested by the comments.

Comment 4: Also, please provide graphical/analytical plots of power, performance, and latency analysis from all the experiments conducted. It is very hard to read the data across from all the tables.

Response: Thank you very much for the comment. We have added Figure 11 to provide a graphical representation for the PPA comparisons between the 2D and M3D ICs; we use Figure 15 to visually demonstrate the energy consumption breakdown in the 2D and M3D ICs. Thank you again for the valuable suggestions.

B. Reviewer 2

Comment 1: Overall, although the experimentation results are impressive, from the academic perspective, the work lacks of novelty on the basic idea. It is more like an engineering work to adapt the 2D commercial vendor tools to 3D scenario with some tricks. One suggestion is that authors could think about what the basic or fundamental change to conventional physical design flow brought by M3D technology. By the way, if the work can use the open-source EDA tools, it will be more easy for other researchers to follow and reproduce your experimentation.

Response: Thank you very much for the comments. In this paper, we focus on the performance improvement for logic-on-memory monolithic 3D ICs with the commercial CPU cores as benchmark designs. Therefore, we trick the 2D commercial tool into implementing M3D ICs in order to achieve industry-level layout quality. In addition, we also try to incorporate the fundamental changes introduced by the M3D technology into our implementation and analysis flow, including (1) the ultra-high 3D interconnect density; (2) the tight coupling between multiple tiers from the timing, clocking, power delivery, and thermal perspectives; (3) the potential performance degradation caused by the M3D fabrication process. We find that the unique structure of logic-on-memory stacking provides a shortcut to address these changes with 2D commercial tools and the proposed memory pin projection and BEOL modeling approaches. We think the academic novelties of this work is also demonstrated by the in-depth analysis of the performance, power delivery, and thermal benefits provided by the logic-on-memory M3D ICs.

On the other hand, we agree that using open-source EDA tools will help other researchers follow this work. And for future studies, an open-source true-3D placer will be necessary to allow standard cells to be placed on both tiers and explore the benefits of logic-on-logic M3D stacking, while it is beyond the scope of this transaction paper. Thank you again for these valuable suggestions.

Comment 2: The design in this work assumed the logic die was located in the bottom tier which is not reasonable in my opinion. As the heat sink is usually mounted on the top of the chip, it will be more reasonable to locate the logic die on the top tier which is near to the heat sink to remove heat more effectively. Are there any special reasons that the authors made this special structure?

Response: Thank you very much for the comment. In the M3D ICs, MIVs need to go through the silicon substrate of one tier. If we insert MIVs in the memory tier, it is hard to place MIVs freely because memory blocks are large in size and MIVs are not allowed to overlap with them. Therefore, we add MIVs only in the logic tier. Since MIVs can only be placed on the logic tier, the order of layers in the F2B bonding structure is fixed as: memory tier front-end-of-line (FEOL), memory tier BEOL, logic tier FEOL, logic tier BEOL. In order to connect to the C4 bumps on the package, the chip needs to be flipped for bonding. Thus, the logic tier is placed on the bottom in the package, and the memory tier is on the top. This structure has a negative impact on temperature because the power-intensive logic tier is placed far from the heat sink, but it provides significant benefits on timing and power delivery in the logic tier thanks to the short path to the package.

As described in the **Section V-D**, we did an what-if study to analyze the impact of the die stacking order. We swap the position of the logic tier and memory tier and assume the performance and power consumption of the design do not change. Fig. 20 shows the temperature maps in the M3D ICs after swapping the tiers. The maximum junction temperature decreases by 0.6°C in Cortex-A7, and by 2.0°C in Cortex-A53, respectively. The differences are not significant, because the total power density of the M3D IC remains the same, and the C4 bumps and the metal traces in the package also forms an effective heat dissipation path from the logic tier to the PCB in the original 3D stack. Therefore, the order of die stacking has only a small impact on temperature under the specific workload and air-cooling solution. However, the results can be different if more advanced cooling techniques such as water cooling are applied, because the heat dissipation through the heat sink will be more effective.

Comment 3: The clock network and power grid design are important for M3D IC designs. But the proposed physical design flow lacks of novelty on these two issues. Do logic and memory portion share the same clock domain? Which topology of clock network did you use? Does the power grid design have any impact on the thermal dissipation?

Response: Thank you very much for the comment. We incorporate the clock network and power grid design as part of our implementation and analysis flow, and improve the quality of the clock network and power grid based on the structure of the logic-on-memory M3D ICs. As discussed in the **Section IV-D**, the clock distribution network (CDN) design in the M3D ICs is similar to the 2D ICs, because all the standard cells, including the clock gates, flip-flops, and buffers are placed on the logic tiers. The memory and the standard cells belong to the same clock domain, so only a small number of clock nets need to be connected to the clock pin of the SRAM blocks on the memory tier. We adopt the H-tree shape for the CDN, and utilize the Innovus clock tree synthesis (CTS) engine to implement the CDN. Thanks to the smaller footprint area, the shorter vertical interconnects, and the removal of memory obstructions, we can achieve a more balanced CDN in the M3D ICs with a shorter clock wirelength and lower averaged latency compared with the 2D CDN. Our experiment results in **Section V-A** show that the 3D CDN provides 37% and 40% lower average latency in the Cortex-A7 and Cortex-A53 design, compared with the 2D ones, respectively.

As described in **Section IV-E**, we design an unbalanced power grid for the logic tier and memory tier to compensate for the different current consumption on the two tier. This also has an impact on the heat dissipation and the maximum temperature in the M3D ICs. Since the power rails on the logic tier are denser and wider, it improves the thermal conductivity of the BEOL layer and helps the heat dissipation from the power-intensive logic tier to the environment, similar to the method proposed in [21]. On the other hand, the power grid dimensions are limited by the routing congestion and voltage drop constraints, and thus cannot be solely optimized for cooling. Building additional grids for cooling and co-optimizing the single routing, power delivery, and cooling grids in M3D ICs can be an interesting topic for future studies, while it is beyond the scope of this transaction paper.

Comment 4: Additionally, note that you used tens of thousands of MIVs in the chip, does such high MIV density have any impact on the temperature dissipation? From your experimental result, it seems that MIVs have little help for thermal removal. Is it correct?

Response: Thank you very much for the comment. Yes, we find that the impact of MIVs on thermal removal is negligible in the M3D ICs. As discussed in **Section V-D**, though the number of MIVs is large ($> 300k$), the area occupied by MIVs is relatively small (¡0.2% in both Cortex-A7 and Cortex-A53) compared with the entire floorplan in the M3D ICs, due to the small size of individual MIVs. As a result, the impact of MIVs on the vertical thermal conductivity is nearly negligible. Assuming the power consumption does not change, we perform another experiment on the M3D ICs by removing all the MIVs and rerunning the thermal simulation. The results show that the temperature changes are within $0.1°$ in the M3D ICs after the removal. Therefore, the impact of MIV density on temperature is not obvious in these M3D ICs.

Comment 5: The experimentation results only compared 3D and 2D baseline. Note that there are several literatures talking about the M3D physical design flow like [Guler TVLSI18, Shi ISVLSI16, etc.]. How about your results when compared with those methods?

Response: Thank you very much for the comment. We have added the comparisons and discussions regarding the literatures in **Section III**. Thank you again for the valuable suggestions.

C. Reviewer 3

Comment 1: Regarding power & thermal analysis. In section IV.E, there is no real innovation besides the fact of using existing tools (APACHE RedHawk), and integrate these analyses in the overall design flow. Any limitation here? Any difficulty in terms of flow/modeling/etc? Lastly, an interesting discussion is provided on section V.B on power & thermal analysis at the end.

Response: Thank you very much for the comment. One challenge of the flow is that the tool requires us to generate the chip-thermal model (CTM) for each tier separately, because the geometries and vertical positions of cells have an impact on the heat generation and dissipation in the entire system. Therefore, we have to partition the M3D IC into two tiers, scale the memory blocks on the memory tier to the original size, annotate the switching activity of the nets in each tier with the block-level SAIF file, and then generate the CTM for each tier.

A limitation of the thermal simulation flow is that there is a trade-off between accuracy and efficiency. During the thermal simulation, the RedHawk-CTA tool divides the design into a rectangular mesh on the XY plane and calculate the averaged power density and thermal conductivity within each tile. The highest mesh resolution is $5\mu m \times 5\mu m$. However, the MIV size is much smaller than the mesh resolution, so each individual MIV is not modeled directly in this flow, while it has an impact on the thermal conductivity of the tile. Therefore, the runtime and accuracy of the thermal simulation flow is affected by the mesh resolution. For Cortex-A7 and Cortex-A53, the resolution range from $5\mu m$ to $20\mu m$ can provide a good balance between accuracy and efficiency. The trend of runtime and accuracy of the flow with various mesh resolutions in Cortex-A7 M3D ICs is presented in Table VII.

Comment 2: In table IV, a full summary of all results is given. Please also include the overall 2D & 3D area numbers

Response: Thank you very much for the comment. We have included the overall 2D and 3D area numbers in Table IV.

Comment 3: regarding state of art analysis, a detailed comparison is done with author previous work (CASCADE 2D, SHRUNK 2D, their DATE'20 paper, etc), but there is missing recent and similar work on M3D automatic Place & Route design flow, also targetting Memory-on-Logic partitionning. Please see :

S.Thuries et al., "M3D-ADTCO: Monolithic 3D Architecture, Design and Technology Co-Optimization for High Energy Efficient 3D IC", DATE'2020

A very similar approach was proposed, including PDKIT creation, library preparation (lef formats, etc), floorplanning, 2 layer routing in a single session, obstruction handling for proper MIV creation/obstruction, power grid creation, etc. One main difference is that for the authors' proposal, the netlist need to be partionned in two netlists, for 2 steps floorplanning, while it was not necessary for mentionned above paper. Please compare in details your work with this recent paper.

Response: Thank you very much for the comment. We have added a detailed comparison between our work and the mentioned paper in Section IV.

Comment 4: In many place of the paper, it is mentionned, "for the first time, we present a physical ...", which is thus not true, seen this previous paper. Please fix the wording accordingly.

Response: Thank you very much for the comment. We have modified the corresponding statement in **Section I**.

Comment 5: Minor remarks :

- in the abstract, a gain of 20.07% and 21.02% are mentionned. Please round the result at single digit, 20% and 21%. System architecture cannot be that much accurate !

- for the overall design flow, as a last phase, the GDS is split into 2 separated files. Why ? in M3D a single circuit is fabricated, with a full set of masks including both layers, one would not expect to split the GDS into 2 set of files.

Response: Thank you very much for the comment. For the performance uplift values, we have modified the double-digit numbers in the abstract, **Section I, V-A, and VI**.

The M3D IC splitting procedure is mainly used to generate separated chip-thermal models for thermal simulation, as described in the response to the first comment. We agree that it is not necessary to split the M3D GDS files after P&R. We have removed the statement in **Section IV-D**. Thank you again for all the valuable and detailed suggestions.