3D IC Tier Partitioning of Memory Macros: PPA vs. Thermal Tradeoffs

Lingjun Zhu
Georgia Institute of Technology
United States
lingjun@gatech.edu

Nesara Eranna Bethur
Georgia Institute of Technology
United States
nbethur3@gatech.edu

Yi-Chen Lu
Georgia Institute of Technology
United States
yclu@gatech.edu

Youngsang Cho
Samsung Electronics, Republic of Korea
ys0228.cho@samsung.com

Yunhyeok Im
Samsung Electronics, Republic of Korea
imim@samsung.com

Sung Kyu Lim
Georgia Institute of Technology
United States
limsk@ece.gatech.edu

ABSTRACT

Micro-bump and hybrid bonding technologies have enabled 3D ICs and provided remarkable performance gain, but the memory macro partitioning problem also becomes more complicated due to the limited 3D connection density. In this paper, we evaluate and quantify the impacts of various macro partitioning on the performance and temperature in commercial-grade 3D ICs. In addition, we propose a set of partitioning guidelines and a quick constraint-graph-based approach to create floorplans for logic-on-memory 3D ICs. Experimental results show that the optimized macro partitioning can help improve the performance of logic-on-memory 3D ICs by up to 15%, at the cost of 8°C temperature increase. Assuming air cooling, our simulation shows the 3D ICs are thermally sustainable with 97°C maximum temperature.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**; **Partitioning and floorplanning**.

ACM Reference Format:
Lingjun Zhu, Nesara Eranna Bethur, Yi-Chen Lu, Youngsang Cho, Yunhyeok Im, and Sung Kyu Lim. 2022. 3D IC Tier Partitioning of Memory Macros: PPA vs. Thermal Tradeoffs. In *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED '22), August 1–3, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3531437.3539724

1 INTRODUCTION

Micro-bump technology provides a low-cost solution for the 3D integration of electronic systems and thus paves the way for 3D IC's applications in commercial products. This technology uses micron-level solder balls consisting of intermetallic compounds to establish the 3D connection, and the bonding pitch can be lower than 10μm [2]. Through-silicon vias (TSVs) are also required to connect signals and deliver power from the package to the 3D stack. On the other hand, hybrid bonding is another 3D technology that provides even higher bonding density (with lower than 1μm bump pitch). With these high-density 3D interconnects, hybrid-bonding 3D enables various types of 3D stacking.

3D ICs continue to face challenges in physical designs and thermal management. Due to the additional z-dimension and 3D connection density limit, the partitioning, floorplanning, placement, and routing (P&R) problems become more complicated with the 3D technology. In addition, 3D ICs tend to suffer from high temperature and small thermal headroom because of the high power density and low heat dissipation rate caused by die stacking. Existing physical design flows, such as Pin-3D [4] or Macro-3D [7], address the P&R problems by modifying technology files and tricking the commercial 2D P&R engine into implementing 3D ICs, but they did not provide solutions for macro partitioning or placement in 3D space.

Memory macro partitioning is an important step in the physical design of 2D and 3D ICs. As the functionality of today's digital systems keeps improving, more and more physical IPs, including memory blocks, are integrated into the systems, and they are regarded as macros during the physical design stages. These macros can occupy significant design area and consume considerable power, so their partitioning and placement are crucial for the power, performance, and area (PPA) of the entire system.

Tier partitioning and floorplanning of memory macros become more complicated in 3D because of the constraints of 3D footprint area and connection density, in addition to the impacts on temperature. Previous studies such as [6] have proposed automatic approaches to create a thermal-aware 3D floorplan, but the impact of macro-placement on PPA is not considered, and thermal impacts are evaluated with architectural-level thermal simulator rather than sign-off layouts.

In this paper, we study PPA vs. thermal tradeoffs targeting 3D ICs built with micro-bump and hybrid-bonding technologies. Especially, we focus on the impact and proper ways to tier partition memory macros to strike a balance between PPA and thermal. We use one open-source RISC-V-based CPU and a commercial 64-bit CPU as benchmarks. The main contributions of this work are to quantify the impacts of memory macro tier partitioning on the performance and temperature of micro-bump and hybrid bonding 3D ICs with real-world benchmarks and layout-level simulation and to propose guidelines for logic-on-memory tier partitioning.

ISLPED '22, August 1–3, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9354-6/22/08…$15.00
https://doi.org/10.1145/3531437.3539724

Table 1: Design and technology setup used in this paper.

Benchmark design				
Technology	TSMC 28nm			
Benchmark	Processor A (RISC-V)		Processor B (commercial CPU)	
# of cores/tiles	25		4	
L2 cache size (MB)	256kB (per tile)		2MB (shared)	
# of memory macros	28 (per tile)		185 (in total)	
3D technology				
Bonding orientation	face-to-face bonding			
Bonding technology	micro-bump		hybrid-bonding	
Bonding pitch (µm)	25	10	5	1
Max signal bump # for A	613	3828	15313	382813
Max signal bump # for B	3453	21582	86328	2158200

2 DESIGN AND ANALYSIS METHODOLOGIES

2.1 Design Setup

In this section, we set up commercial-grade 3D ICs with state-of-the-art physical design flows for our design comparisons. Two processor benchmark designs are used for the experiments: Processor A is the RISC-V-based processor OpenPiton [1]; Processor B is the 64-bit Arm® Cortex®-A53 quad-core CPU. The benchmark designs are implemented with TSMC 28nm technology. The OpenPiton processor is first implemented at the single-core level, then tiled up to create a 25-tile system for thermal analysis. Table 1 shows the detailed settings used for the benchmark designs and technology.

We consider 4 different 3D bump pitch settings in this paper: the 25µm and 10µm pitches represent the micro-bump and hybrid bonding technology that is currently available for mass production, while the 5µm and 1µm pitches project the future opportunities for micro-bump and hybrid-bonding 3D, respectively. With a smaller bump pitch, the maximum bump number increases given a fixed footprint area, and thus it enables more 3D connections in the design. We assume 50% of the bumps are used for power delivery, and the remaining 50% are used for signal communication.

2.2 Physical Design and Thermal Analysis

We consider the logic-on-memory (LoM) partitioning for the physical implementation, in which standard cells are placed on the top die only, while memory blocks can be placed on both die. The state-of-the-art Macro-3D [7] flow and Pin-3D [4] flow are adopted for the 3D IC physical implementation. These flows use modified technology files and double metal stack based on the TSMC 28nm PDK and employ the commercial P&R tool, Cadence® Innovus™, to optimize the 3D ICs, and thus result in commercial-grade 3D layouts. After the physical design, the 3D layout is split into top and bottom die layouts, with corresponding constraints, parasitics, and cell libraries.

For thermal analysis, we use ANSYS® RedHawk™ to generate the Chip-Thermal Model (CTM) for each die. The CTM consists of a tile-based power density map and metal density map (tile size is 10µm×10µm), and thus it can accurately represent the on-chip heat generation and dissipation in the 3D stack. We then use ANSYS Chip-Package System (CPS) to stack the CTMs, create a 3D thermal model for simulation, and generate temperature maps for the 3D ICs. Using this flow, we are able to implement 3D ICs with various memory macro partitioning and obtain PPA and temperature reports with layout-level granularity.

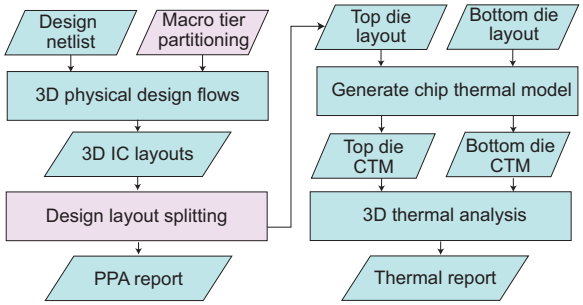


Figure 1: Overview of our 3D IC physical design flow.

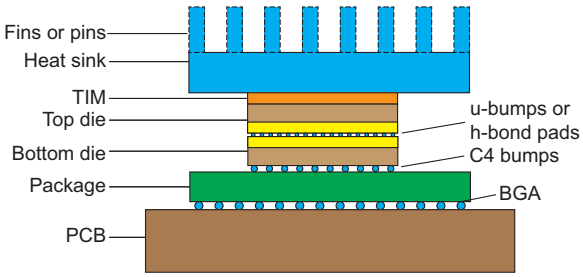


Figure 2: Our comprehensive 3D IC model for thermal analysis that includes chip, package, and board elements.

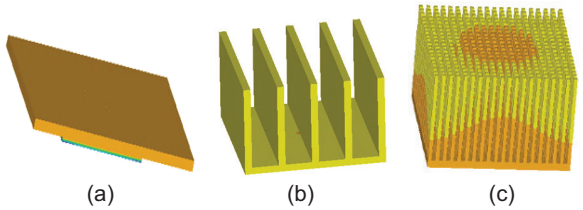


Figure 3: Heat sink technologies used in this work. (a) metal plate only, (b) 5 fins, (c) 400 pins.

2.3 Thermal Model and Cooling

In order to reflect the thermal impacts of memory macros in a realistic scenario, we create a complete 3D thermal model in CPS, as shown in Figure 2. This thermal model includes the top and bottom dies of a 3D IC, the micro bumps or hybrid-bonding pads between the two dies, the TSVs and C4 bumps that connect the bottom die to the package, and the dummy package, BGA, and PCB board model. All the bumps (including C4 bumps and micro bumps) are modeled as individual objects during the thermal simulation. As a result, their impacts on thermal conductivity are included in

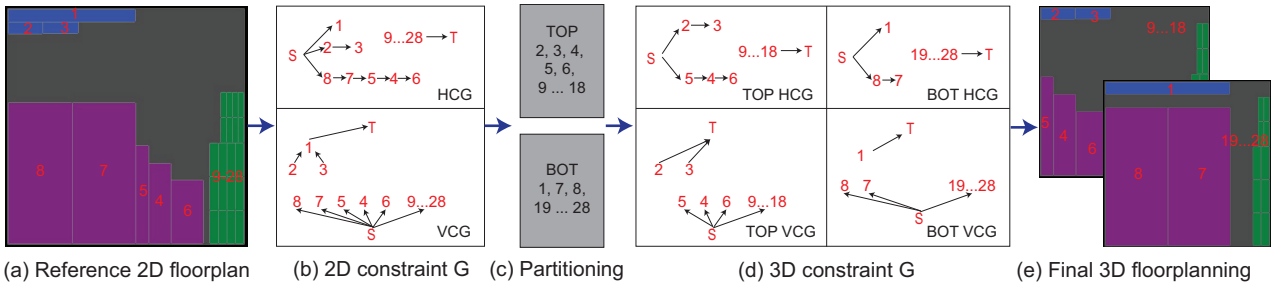


Figure 4: Example of memory partitioning and floorplan generation process.

the analysis results. Moreover, this model allows us to evaluate the effects of air cooling techniques in Section 4.2.

Heat sink design is important, especially for 3D ICs, because it helps improve heat dissipation towards the up direction and reduces the overall chip temperature. We consider three different heat sink designs in this study. Figure 3 (a) demonstrates a heat sink plate without pin or fin structures. Since this has the lowest surface area and sub-optimally uses the airflow, the chip temperature reduction is minimal, but it also has a small volume and can fit into mobile products. Figure 3 (b) has a heat sink plate with attached fins. The fins are long metal plates that increase the surface area and utilize the airflow better in one direction. A more sophisticated heat sink design is shown in Figure 3 (c), which involves a heat sink plate with pin structures. These pins are long metal rods placed in a grid array fashion. In addition to the benefits of fin structure, this pin array utilizes the airflow in both directions and further increases the surface area. We examine the impacts of different cooling techniques and heat sink settings on the 3D IC temperature via thermal simulations.

3 MEMORY PARTITIONING METHODOLOGY

We adopt a method to quickly generate logic-on-memory partitioning and placement for various 3D designs. This method is based on constraint graphs, similar to the sequence-pair approach for 2D floorplan [3]. Figure 4 shows an example of memory macro partitioning and floorplan generation in an OpenPiton tile. In this approach, we consider only the memory macro partitioning and placement, but not other IPs or functional blocks. The 2D floorplan is created according to the design manual of OpenPiton and Cortex-A53. In the reference 2D floorplan, we notice that the memory macros are clustered together at the edge or corner of the floorplan. The reason is that the designer expects to create a large and continuous space for standard cell placement, and minimize the memory-to-memory distance for closed-coupled memory macros as well, which helps improve the P&R quality of the final design.

Based on this observation, we can extract the horizontal constraint graph (HCG) and vertical constraint graph (VCG) for the memory macros from the reference 2D floorplan, as shown in Figure 4 (b). The weight of the edge is equal to the width of originating block in the HCG, while it is equal to the height in the VCG. Unlike the conventional constraint graphs for 2D floorplanning, these HCG and VCG contain only memory macros. And they are not necessarily connected graphs since some memory macros are clustered at one corner of the floorplan (e.g., Macro 1, 2, 3) and have no relative position constraint with other macros. Therefore, the resulting constraint graphs may consist of multiple disjointed trees or graphs.

Using these constraint graphs, we can partition the memory macros and construct a 3D floorplan that simultaneously satisfies the physical, geometrical, and logical constraints. First, we manually select a set of memory macros to be partitioned into the bottom die (called S_0), and all the remaining memory macros (called S_1) will be placed on the logic die. Then, we create the constraint graphs for each die based on the original 2D constraint graphs. For the top die, all the nodes in the bottom die (S_0) are removed from HCGs and VCGs, and we add an edge between each of its parents and its children.

Similarly, the HCG and VCG for the bottom die are created by removing nodes in S_1. Then, the 3D memory macro floorplan can be constructed based on the constraint graphs for each die. That is, the x and y coordinate of each macro can be determined by calculating the longest distance from the source (representing the left or the bottom edge) to the macro or from the macro to the terminal (representing the right edge or upper edge). In the generated 3D memory macro floorplan, the tightly-coupled memory macros are still clustered together, and the result is ready to be input into our 3D P&R flow. This approach facilitates our design comparison with various memory macro partitioning.

4 DESIGN AND ANALYSIS RESULTS

4.1 Partitioning Impact on PPA

We create a set of memory macro partitions and results using the constraint-graph-based method for processors A and B, respectively. Figure 5 shows the overview of the various memory macro partitioning and floorplan for each design. Processor A, for example, has 5 different 3D memory macro partitioning and floorplan (FP1-5). We calculate the memory occupancy rate based on the ratio of the total memory macro area on the memory die to the footprint area. From floorplan 1 (FP1) to floorplan 5 (FP5), the memory die occupancy rate increases from 62.4% to 84%. That is, we partition more memory blocks, including the L1 and L2 caches, into the memory die. As a result, the empty area on the logic die for standard cell placement becomes larger, but it also requires more 3D nets to connect the memory pins on the memory die. We perform P&R

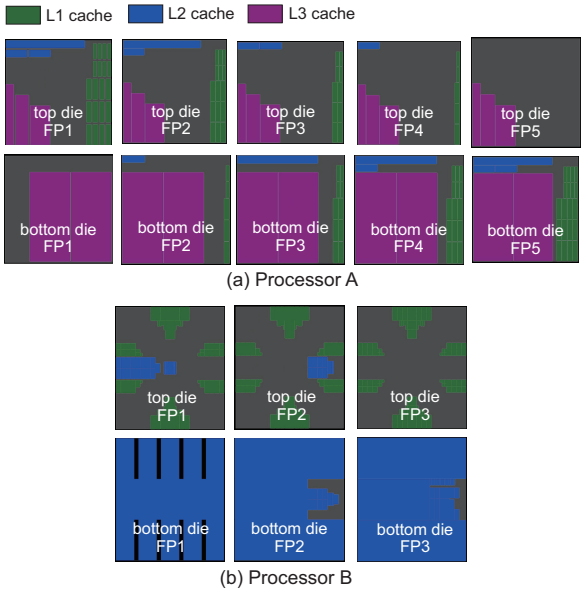


Figure 5: Memory macro partitioning and floorplan solutions.

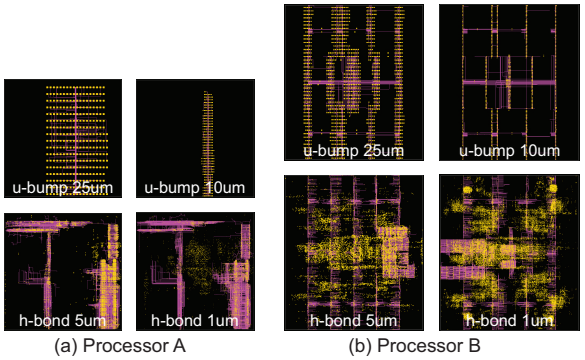


Figure 6: Layouts of the bumps and pads in the 3D ICs.

implementation based on these various partitioning and floorplans with the 4 different 3D technology settings.

Table 2 summarizes the trend of the number of 3D memory nets and the maximum frequency of the design with increasing memory occupancy. The F_{max} metric represents the maximum frequency that each design achieves. Due to the limited number of 3D connections, some floorplans cannot be used when the bump pitch is large. For example, the floorplan 5 (FP5) for processor A cannot be implemented with the micro-bump 3D technology when the bump pitch is 25µm because it requires 6264 3D nets to connect to the memory pins on the memory die, while there are only 613 signal bumps available. As a result, the 3D technology with smaller bump pitches provides a large solution space with more flexibility in partitioning and floorplanning for 3D ICs.

The results show that the 3D ICs with a smaller bump pitch tend to have better performance. When the bump pitch is fixed, the memory macro partitioning and memory die occupancy rate have significant impacts on the maximum achievable frequency. As shown in Table 2, the maximum frequency of the 10-µm and 25-µm micro-bump 3D ICs is achieved with floorplan 1 for both processors A and B. Floorplan 1 (FP1) has the lowest memory die occupancy. However, the 5-µm and 1-µm hybrid-bonding 3D ICs achieve their F_{max} when the memory die occupancy is the highest, which means most of the memory macros are partitioned into the memory die. These differences are mainly caused by the various 3D bump pitches, 3D wirelength overhead, and routing optimization.

Figure 6 shows the layouts of the signal bumps and 3D nets in the 3D ICs with various bump pitches. The yellow dots represent the micro-bumps or hybrid-bonding pads, and the magenta lines denote the 3D nets which connect the memory pins. Table 3 provides detailed wirelength and clock comparisons for the max-performance designs with different 3D technology settings. As demonstrated in the figure, when the bump pitch is large in the 3D ICs, the wirelength overhead of the 3D nets becomes non-negligible. For example, in the 25µm-pitch micro-bump 3D design, the micro bumps are spread over a large area, and the 3D nets need to traverse a long distance to access the micro bumps. As a result, the total memory wire length increases, and 3D routing quality degrades, which then leads to lower performance.

This also explains why increasing the memory die occupancy does not help improve the performance in the micro-bump 3D design. The reason is that the negative impacts of additional micro-bump and 3D wirelength overhead compensate for the benefits of a larger empty area on the logic die. On the contrary, the hybrid bonding 3D ICs have much smaller bump pitches, and that allows the P&R flow to find the optimized location for the hybrid bonding pads, which, in turn, helps reduce the wirelength overhead of additional 3D nets. The results show that with 1-µm hybrid-bonding 3D, the optimized memory macro partitioning can help improve the performance by 12% and 15% for processors A and B, respectively.

The timing critical paths of the max-performance 3D ICs are shown in Figure 7. As shown in the figure, the critical paths in the micro-bump 3D design with larger bump pitches predominately consist of long register-to-I/O global interconnects. Instead, for the hybrid-bonding 3D ICs with smaller pitches, the critical paths consist more of short register-to-register or register-to-memory paths, and local interconnect, such as the critical paths of hybrid 1-µm Processor A and B design. That is because the high-density hybrid-bonding 3D allows more memory macros to be partitioned into the memory die, which, in turn, reduces the placement and routing obstruction on the logic die and benefits the long global interconnect optimization.

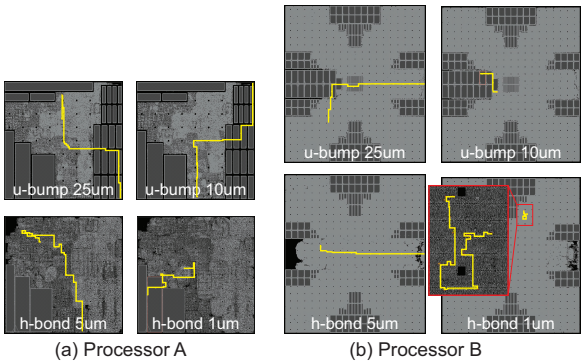
In addition, with a large number of 3D nets available, the local register-to-register and register-to-memory interconnect also benefit from the shared routing resources on the memory die. As a result, the overall routing quality is improved with the smaller bump pitches. Table 4 provides a detailed breakdown of the timing critical paths in Processor A. Compared to the micro-bump 3D ICs, the hybrid-bonding 3D ICs have similar total wirelength on the critical path, but the maximum wirelength between each cell along the critical path is smaller. This demonstrates that with the smaller bump pitches and more memory macros on the top die, the logic-on-memory 3D ICs can achieve better signal buffering

Table 2: Impact of memory partitioning on performance. The processor B frequency numbers are normalized based on the frequency of a 2D baseline design. "u-bump" is short for micro-bump 3D, and "h-bond" stands for hybrid-bonding 3D. The green color highlights the maximum frequency achieved with each 3D technology setting. The percentage shows the change of F_{max} compared to the baseline partitioning (FP1).

Partition	3D net #	Occupancy (%)		Fmax (MHz)			
		memory die	logic die	u-bump 25um	u-bump 10um	h-bond 5um	h-bond 1um
Processor A							
FP1	489	62.4%	38.9%	682.8	703.3	759.3	755.0
FP2	2130	67.1%	34.1%	x	577.6 (-18%)	759.1 (+ 0%)	801.2 (+ 6%)
FP3	3571	74.6%	26.6%	x	593.0 (-16%)	777.9 (+ 2%)	782.8 (+ 4%)
FP4	5030	79.4%	21.9%	x	x	791.1 (+ 4%)	750.0 (- 1%)
FP5	6264	84.1%	17.1%	x	x	801.9 (+ 6%)	844.3 (+12%)
Processor B (normalized to 2D IC results)							
FP1	1817	88.5%	26.2%	1.02	1.19	1.35	1.51
FP2	3120	92.3%	22.4%	x	0.72 (-40%)	1.29 (- 5%)	1.50 (- 0%)
FP3	4221	95.7%	19.0%	x	0.72 (-39%)	1.41 (+ 4%)	1.74 (+15%)

Table 3: Detailed wirelength (WL) and clock comparisons of the max-performance 3D ICs. The red color highlights the WL and timing bottlenecks, while the green color highlights the improvements (same as Table 4).

3D technology	u-bump 3D		h-bond 3D	
bump pitch (μm)	25	10	5	1
Processor A				
F_{max} (MHz)	682.8	703.3	801.9	844.3
# of signal bumps	313	313	4319	11413
3D net WL (m)	0.80	0.74	0.56	0.51
Max. Clock WL (μm)	778.3	778.3	272.2	270.1
Max. clock skew (ns)	0.71	0.77	0.21	0.22
Processor B (normalized to 2D IC results)				
F_{max} (MHz)	1.02	1.19	1.41	1.74
# of signal bumps	1188	1188	4319	65200
3D net WL (m)	1.51	1.43	1.35	1.33
Max. clock WL (μm)	324.3	309.4	309.7	335.2
Max. clock skew (ns)	0.38	0.43	0.33	0.39



(a) Processor A (b) Processor B

Figure 7: Critical paths of the max-performance 3D ICs.

on the critical paths, and thus it helps optimize timing in these max-performance designs.

Table 4: Critical path timing breakdown for Processor A.

Processor A				
3D technology	micro-bump		hybrid-bonding	
bump pitch (um)	25	10	5	1
type	reg2out	reg2out	reg2out	reg2mem
total WL (um)	1564.70	1401.56	1517.05	1673.92
max WL (um)	379.99	323.58	196.26	279.61
launch latency (ns)	0.38	0.54	0.33	0.50
capture latency (ns)	0.00	0.00	0.00	0.18
clock skew (ns)	0.38	0.54	0.33	0.32
path delay (ns)	0.46	0.38	0.46	0.40
cell delay (ns)	0.34	0.27	0.24	0.33
wire delay (ns)	0.12	0.11	0.22	0.07
pin delay (ns)	0.63	0.50	0.45	0.00
setup time (ns)	0.00	0.00	0.00	0.46

4.2 Partitioning Impact on Thermal

We first evaluate the air cooling technique and heat sink designs on our 3D ICs. We take the 3D ICs with the max-performance as a benchmark and perform thermal simulations with various air speed and heat sink structures. The CTMs of the 3D ICs are generated at the maximum frequency, assuming the defaulted switching activity is equal to 0.05 during the steady-state thermal simulation, which can reflect an average workload of the benchmark CPU designs. And the ambient temperature is set to 25°C.

Table 5 summarizes the simulation results. For the 25-tile OpenPiton design, forced air convection with 4-m/s air speed is necessary to cool down the system. However, the air cooling solution is not very effective with the plate-only and fin-based heat sink, and the maximum temperature (T_{max}) is far beyond the thermal threshold. On the other hand, with a heat sink and 400 pins, T_{max} of the 25-tile system is significantly reduced to 84.7°C, and the junction-to-ambient thermal resistance (R_{ja}) is 6.8°C/W. This result aligns with the R_{ja} value of forced air convection reported in [5].

For the quad-core Processor B, forced air cooling and a pin-based heat sink also effectively cool the system. But to simulate a mobile

Table 5: Impact of cooling techniques and heat sink design.

Air speed	Heat sink structure	T_{max} (°C)	R_{ja} (°C/W)
Processor A (25-tile)			
4 (m/s)	Plate only	199.9	21.0
4 (m/s)	Plate and 5 fins	194.6	20.3
4 (m/s)	Plate and 400 pins	84.7	6.8
Processor B (quad-core)			
0 (m/s)	Plate only	96.9	28.0
0 (m/s)	Plate and 5 fins	90.1	25.4
4 (m/s)	Plate and 400 pins	43.1	7.0

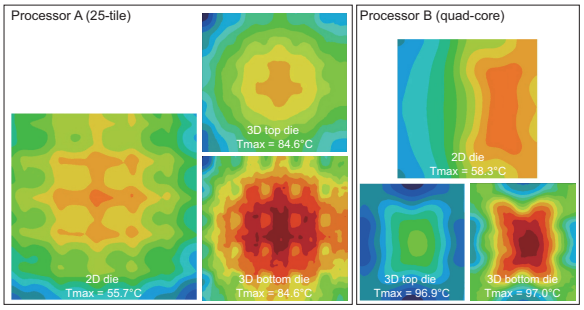


Figure 8: Thermal maps of the max-performance 2D ICs and hybrid-bonding 3D ICs with 1μm bump pitch (not to scale).

setting, we assume natural convection (air speed = 0 m/s) and heat sink plate only for Processor B in the following experiments.

Figure 8 shows the temperature map in the max-performance design (hybrid bonding with 1-μm pitch) for processors A and B, using forced air cooling and natural convection, respectively. According to these maps, the thermal hotspots are generated in between the top die and bottom die, near the 3D bonding interface. This is because the heat generated by both dies tends to be trapped in the face-to-face bonded BEOL layers. Therefore, the 3D bonding interface can have a significant impact on vertical heat dissipation. Despite the top die having a higher power generation rate, the temperature difference between the two dies is small (<1.5°C). The reason is that the top die is attached to the heat sink, and the air-cooling solution removes heat from its upper surface effectively.

The trend of maximum temperature with various 3D bump pitches and macro partitioning is shown in Figure 9. Compared with the 2D baseline, the 3D ICs have at least 14°C higher temperature and 18% higher thermal resistance (R_{ja}) because of their higher power density and lower heat dissipation rate. In addition, we observe that the maximum temperature in the 3D ICs increases with the decreasing bump pitch and increasing memory die occupancy rate, mainly caused by the increase of maximum frequency. R_{ja} remains similar with various memory macro partitioning since it is affected more by the cooling techniques and heat sink designs. Therefore, the trade-off between performance and temperature is important to implement 3D ICs for practical applications: by using a smaller 3D bump pitch and increasing the memory die occupancy, we can achieve up to 15% F_{max} improvements, but it is at the cost of 8°C higher temperature and smaller thermal headroom.

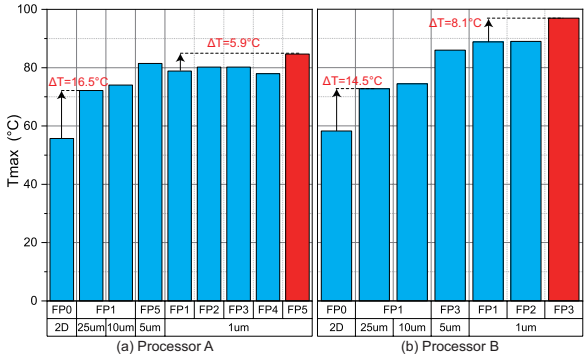


Figure 9: Trend of junction temperature with various 3D bump pitches and partitioning (simulated at F_{max}).

4.3 Logic-on-Memory Partitioning Guidelines

Based on our experimental results, we propose the following guidelines for memory macro partitioning in logic-on-memory 3D ICs: 1) Partition large memory macros with fewer pins to the memory die first. This helps remove the large memory obstructions on the logic die and reduce the wirelength overhead introduced by the bump pitch. 2) Partition smaller memory macros with shared nets into the memory die. By doing this, the number of 3D bumps required to connect these memory macros can be minimized. Our results show that up to 15% performance improvement can be achieved with hybrid-bonding 3D using this method.

5 CONCLUSIONS

In this paper, we perform a comprehensive design comparison on memory macro tier partitioning for 3D ICs. We use a constraint-graph-based method to generate various memory placement and analyze the PPA and thermal performance of the resulting design at the layout level. Our results show that better memory macro partitioning can help improve the performance of logic-on-memory 3D ICs by 15%, but the solution space is constrained by the 3D bump pitch. Also, the performance improvement can introduce 8° temperature increase, and thus dedicated heat sink design needs to be adopted to maintain the thermal sustainability of the system.

REFERENCES

[1] Jonathan Balkind et al. 2016. OpenPiton: An open source manycore research framework. *ACM SIGPLAN Notices* 51, 4 (2016), 217–232.
[2] B. Majeed et al. 2014. Microbumping technology for hybrid IR detectors, 10um pitch and beyond. In *Electronics Packaging Technology Conf.* 453–457. https://doi.org/10.1109/EPTC.2014.7028336
[3] Hiroshi Murata et al. 1996. VLSI module placement based on rectangle-packing by the sequence-pair. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems* 15, 12 (1996), 1518–1524.
[4] Sai Surya Kiran Pentapati et al. 2020. Pin-3D: a physical synthesis and post-layout optimization flow for heterogeneous monolithic 3D ICs. In *Proc. IEEE Int. Conf. on Computer-Aided Design*. 1–9.
[5] Delong Qiu et al. 2017. Experimental and numerical study of 3D stacked dies under forced air cooling and water immersion cooling. *Microelectronics Reliability* 74 (2017), 34–43.
[6] Zongqing Ren et al. 2020. Thermal TSV optimization and hierarchical floorplanning for 3-D integrated circuits. *IEEE Trans. on Components, Packaging, and Manufacturing Tech.* 10, 4 (2020), 599–610.
[7] Lingjun Zhu et al. 2021. High-Performance Logic-on-Memory Monolithic 3-D IC Designs for Arm Cortex-A Processors. *IEEE Trans. on VLSI Systems* 29, 6 (2021), 1152–1163.