A PPA Study for Heterogeneous 3-D IC Options: Monolithic, Hybrid Bonding, and Microbumping

Jinwoo Kim, Lingjun Zhu, *Member, IEEE*, Hakki Mert Torun, Madhavan Swaminathan, *Fellow, IEEE*, and Sung Kyu Lim, *Fellow, IEEE*

Abstract— In this article, we present three commercial-grade 3-D IC designs based on state-of-the-art design technologies, specifically microbumping (3-D die stacking), hybrid bonding (wafer-on-wafer bonding), and monolithic 3-D (M3D) ICs. To highlight tradeoffs present in these three designs, we perform analyses on power, performance, and area (PPA) and the clock tree. We also model the tier-to-tier interconnection in each 3-D IC methodology and analyze signal integrity (SI) to assess the reliability of each design. From our experiments using the OpenPiton benchmark, the hybrid bonding design shows the best timing improvement of 81.4% when compared to its 2-D counterpart, while microbumping shows the best reliability among 3-D IC designs. Moreover, we expand our study to the commercial processor architecture, which is Arm Cortex-A53, with the new set of 3-D integration options. In addition, we show the microbump assignment methodology to handle a large number of 3-D interconnections in the microbumping 3-D design. We also perform SI on the new set of 3-D intertier/interdie connections to discuss the reliability based on their physical dimensions. With a new benchmark design, the hybrid-bonding 3-D shows the best energy–delay-product (EDP) improvement, which is 25.8% compared to 2-D, and the largest eye-opening among 3-D integration options.

Index Terms— 3-D integrated chip (IC), electronic design automation (EDA), hybrid bonding, microbump, monolithic, power, performance, and area (PPA), signal integrity (SI).

I. Introduction

VARIOUS 3-D integration approaches have been proposed recently to cope with device scaling and heterogeneous integration challenges in modern electronics, including microbumping, hybrid bonding, and monolithic 3-D (M3D) ICs [1].

Most recently, Intel has introduced the Foveros technology that enables 3-D die stacking using microbump technology [2]. In microbumping 3-D ICs, two dies are stacked vertically with a dense array of microbumps in a face-to-face (F2F) fashion, which provides high yield and reliability. Moreover, microbonding 3-D ICs enable heterogeneous 3-D die stacking with a large flexibility in the technology selection and IP configurations.

Hybrid bonding technology enables a 3-D integration by using F2F bond pads to stack two predesigned 2-D wafers through the back-end-of-line (BEOL) layers [3]. As F2F bond pads are smaller than TSVs, hybrid bonding 3-D ICs also provide high-density vertical integration. Moreover, since already existing technologies are applied for hybrid bonding, 3-D integration exhibits a lower cost than M3D ICs.

M3D is an emerging technology that integrates device layers sequentially in the vertical direction [4]. Thanks to small monolithic intertier vias (MIVs), M3D offers the finest-grained integration. However, M3D suffers from low yield and high fabrication cost. Moreover, an unresolved challenge for M3D is the performance optimization of the top tier, which is processed at low temperatures to avoid the degradation of the bottom tier.

In this article, targeting commercial-grade 3-D IC designs, we conduct a comparative study of the state-of-the-art heterogeneous 3-D integration technologies aforementioned. Our contributions are as follows.

1) This work compares the three key heterogeneous 3-D integration approaches aforementioned for the first time. Our study is done using GDS layouts and sign-off quality simulations to convincingly quantify the power, performance, area (PPA) and signal integrity (SI) metrics.

2) We extend the state-of-the-art electronic design automation (EDA) tools built for 3-D ICs to obtain competitive designs. In addition, we developed a new flow to handle 3-D ICs that utilize the microbumping technology. We build our 2-D IC baseline designs with a leading commercial vendor tool to substantiate our 2-D versus 3-D IC comparisons.

3) We performed scalability analyses on a different tier/die partitioning and a different set of 3-D interconnections

in M3D and hybrid bonding 3-D designs to show the impact on PPA.

4) We expand our study to the commercial design benchmark to further perform the comparative analyses. We choose a new set of intertier/interdie connections according to the state-of-the-art mass productions.

5) Moreover, we propose the automatic microbump assignment methodology to handle a number of microbumps in the microbumping 3-D design. Through this proposed methodology, we assign the I/O pins to the corresponding microbumps to minimize the wirelength of the design.

6) Our study reveals useful PPA and SI tradeoffs among microbumping, hybrid bonding, and M3D integration technologies. We believe this study offers useful guidelines and pathfinding opportunities to system and circuit designers to make informed decisions to achieve the desired goals.

II. RELATED WORKS

Previous studies have proposed design flows and techniques to enable 3-D integration and improve the performance of 3-D ICs in the absence of commercial 3-D IC tools. Panth et al. [5] have proposed Shrunk-2-D (S2D) flow and Ku et al. [6] have shown Compact-2-D (C2D) flow, which enable M3D or F2F designs by shrinking physical dimensions and reducing the interconnect parasitics.

Macro-3-D has been proposed by Bamberg et al. [7] to overcome the drawbacks of S2D and C2D with memory-on-logic (MoL) stacking. In the Macro-3-D flow, they have shrunk the sizes of macroblocks to the minimum site size to address the overlap issue between the top and the bottom tiers. Lu et al. [8] have proposed a graph neural network framework for tier partitioning (TP-GNN) that is a GNN-based tier-partitioning framework to improve the performance of M3D designs. However, these studies are limited to improvements in 3-D design methodologies themselves.

Pentapati et al. [9] have presented a comparative study on 3-D IC designs. In their study, S2D, C2D, and Cascade-2-D are thoroughly compared in terms of the PPA benefit. Moreover, they have addressed various challenges including the fabrication process, the power delivery, and the thermal issues in 3-D IC designs. However, their study has covered only M3D ICs, not various 3-D integration options.

III. EXPERIMENTAL SETUP

We choose a commercial 28-nm technology node with high-k metal gates to perform the physical designs. Fig. 1 shows the vertical stack-ups of 3-D designs. As we adopt logic-on-memory partitioning in our 3-D design, the full metal stack is divided into logic and macrotiers/dies. In M3D IC and hybrid bonding 3-D IC designs, we duplicate the 2-D metal stack and form the doubled BEOL 3-D metal stacks to generate our two-tier designs as shown in Fig. 1(b) and (c). In the case of the microbumping 3-D IC design shown in Fig. 1(d), we use a 2-D metal stack for each die design and integrate those 2-D designs into a single 3-D design with a microbump model.

Fig. 2 represents the vertical view of intertier/interdie connections between logic and macrotiers/dies. In M3D, MIVs
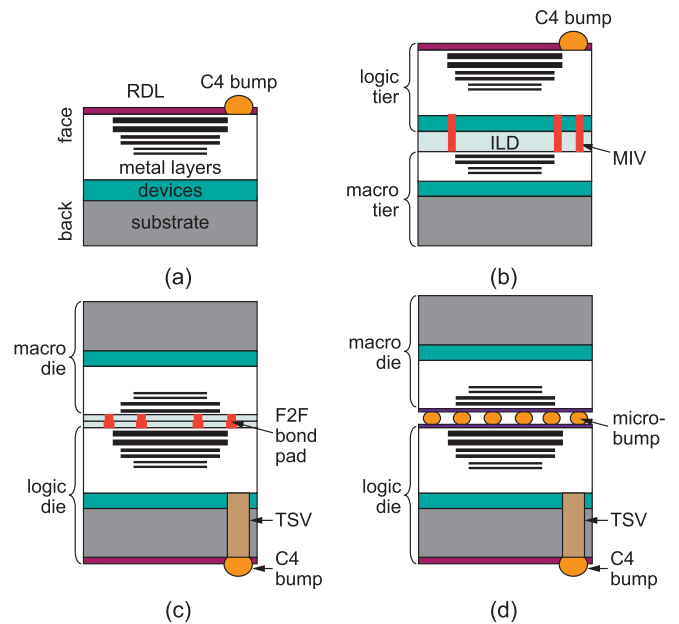

Fig. 1. Vertical stack-up of 2-D and 3-D integration options studied in this article. M3D is face-to-back bonding, while hybrid bonding and microbumping are F2F. (a) 2-D. (b) M3D. (c) Hybrid bonding 3-D. (d) Microbumping 3-D.
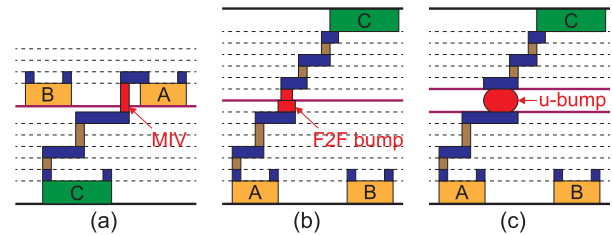

Fig. 2. Intertier/interdie interconnections of heterogeneous 3-D integration options. Logic gates in logic tier/die are marked as yellow, and macroblock as green. (a) Monolithic. (b) Hybrid bonding. (c) Microbumping.

provide connections from the pins of logic gates to the top metal layer of BEOL in the macrotier. In hybrid bonding 3-D, F2F bumps in logic and macrodies are bonded to provide the interdie interconnections. Microbumping 3-D uses microbumps for interdie connections. Target nets are connected to the corresponding bump pads in each die while pads are bumped through microbumps.

Table I shows the physical dimensions of vertical interconnections used in each 3-D IC technology. The MIVs used in M3D have the smallest size and pitch. The minimum pitch, size, and height of MIVs are chosen as 0.6 μm, 0.3 $\times$ 0.3 μm, and 0.1 μm, respectively. As the pitch of F2F bond pads in the hybrid bonding design are $<$1 μm [1], we include those as vias in the full metal stack with 1.0 μm of minimum pitch, 0.5 $\times$ 0.5 μm of size, and 0.17 μm of height based on 28-nm BEOL. In the microbumping design, we choose a microbump of 25-μm diameter and 50-μm pitch based on Intel's Foveros technology [2].

IV. HETEROGENEOUS 3-D IC DESIGN FLOWS

A. Partitioning of Memory Modules

We adopt logic-on-memory partitioning in our study [10]. In 3-D partitioning, there are two major 3-D partitioning

TABLE I
PHYSICAL DIMENSIONS OF INTERTIER/INTERDIE CONNECTIONS
ASSUMED IN THIS ARTICLE

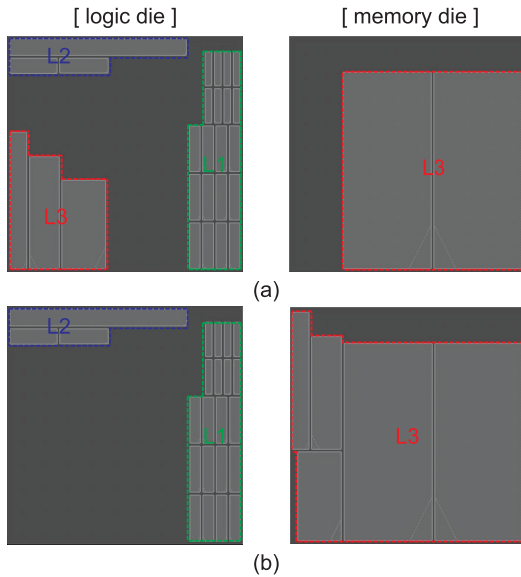|	monolithic	hybrid bonding	micro-bumping
Via/bump size	$0.3\mu m \times 0.3\mu m$	$0.5\mu m \times 0.5\mu m$	$25.0\mu m$
Via/bump pitch	$0.6\mu m$	$1.0\mu m$	$50.0\mu m$
Via/bump height	$0.1\mu m$	$0.17\mu m$	$25.0\mu m$

Fig. 3. Two partitioning and floorplanning options of the OpenPiton benchmark. We select Floorplan A in this article due to a limit on the maximum microbump count. (a) Floorplan A: 313 microbumps. (b) Floorplan B: 861 microbumps.

schemes, which are gate-level and logic-on-memory stacking. The gate-level partitioning has been widely used for 3-D IC benchmarks using Pseudo-3-D flows. However, the gate-level approach has shown performance degradation with benchmarks using a large memory. Since the logic-on-memory scheme avoids long connections between logic and memory blocks, the memory throughput and the system performance have improved significantly when compared to the gate-level 3-D design [10]. Therefore, we have chosen logic-on-memory stacking in this article.

We place the logic gates only in the logic tier, while the memory tier includes macroblocks only such as memory modules. Therefore, the tier partitioning of macroblocks is important because the partitioning result affects the number of vertical interconnections.

As the size of the microbump is larger than the MIV and F2F bond pads, the microbump counts in the microbonding design should be carefully considered to maintain a small form factor of the design. Fig. 3 shows the number of microbumps according to different floorplans of the OpenPiton architecture. Considering the footprint of the 3-D design as 0.88×0.88 μm, the maximum allowable number of microbumps is 400. Therefore, we decide to only assign L3 data cache in the memory die to minimize the bump counts as 313 as shown in Fig. 3(a). For fair comparisons, we use the same floorplan for all three designs.
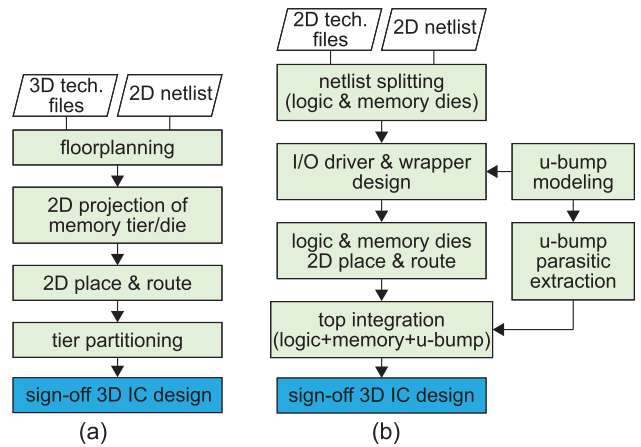

Fig. 4. Design flows used for monolithic [10], hybrid bonding [10], and microbonding 3-D IC designs. (a) M3D and hybrid bonding 3-D. (b) Microbumping 3-D.

B. Monolithic 3-D and Hybrid Bonding 3-D Design Flows

We design monolithic and hybrid bonding 3-D designs with the flow shown in Fig. 4(a) [10]. While both designs use the full 3-D metal stack, different 3-D technology files are used. In the M3D design, the 3-D technology file includes the MIV layer, while F2F bond pads are included as vias in the hybrid bonding 3-D design.

In the floorplanning stage, we generate 2-D floorplans for logic and memory tiers separately. As discussed in Section IV-A, we place L3 data cache blocks in the memory tier and other memory parts in the logic tier. We then project the floorplan of the memory tier to the logic tier and generate a single floorplan by using shrunk macroblocks. The conventional 2-D design tool accepts a single active layer per design. To avoid overlap issues between logic and memory tiers, we shrink memory blocks in the memory tier to the minimum size while maintaining routing blockages and pin locations. Therefore, logic gates can be freely placed on the logic tier with no placement blockage.

After merging two floorplans into one with the full 3-D metal stack, we perform 2-D place-and-route (P&R) using Cadence Innovus. As the tool considers the parasitics of the double-stacked BEOL and interlayer connections during the P&R stage, the final design is directly used to conduct various analyses by sign-off tools.

C. Microbumping 3-D Design Flow

Fig. 4(b) shows our design flow of microbumping 3-D ICs. Using ANSYS HFSS, we first perform the microbump modeling using physical dimensions presented in Table I. Then, we export the S-parameter of the microbump model and convert it to the equivalent circuit model. Finally, the equivalent model is used to generate the standard parasitic exchange format (SPEF) file for the sign-off PPA.

As logic and memory dies are designed separately, we generate netlists of logic and memory dies from the initial 2-D netlist considering the memory partitioning of Section IV-A. We then design the I/O driver for interdie connections, which contain microbumps in the path. Unlike the MIV or F2F bond
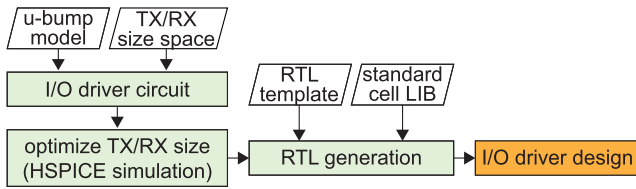
Fig. 5. I/O driver design and optimization flow used in our microbumping 3-D IC design of Fig. 4(c). We adopt Intel's AIB.

pad, the size of the microbump is significantly large, with a 25-μm diameter. Therefore, the I/O driver is necessary to transfer the signal properly through the microbump.

In the microbumping design, we adopt Intel's AIB and select the proper size of the transceiver with the microbump model shown in Fig. 5. For a wide range of TX/RX sizes and microbump models, we perform HSPICE simulations for the TX/RX pairs. Then, we calculate power–delay products (PDPs) and choose the pair with the minimum PDP. In these experiments, we choose TX and RX sizes as ×2 and ×1, respectively. The optimized I/O driver produces 23.1 μW of power and 20.2 ps of propagation delay, which are within the design limit.

With predesigned I/O drivers, we generate the I/O wrapper and finalize the netlist of each die. The netlists of logic and memory dies are fed to the 2-D P&R tool. In the P&R stage, we first place the microbump array and perform the I/O assignment. By setting proper output loads and input delays for I/O microbumps, we perform P&R to obtain the final design of each die. I/O drivers are treated as macroblocks and placed automatically by the tool. As we design logic and memory dies separately, we finally integrate those designs with microbumps and perform a sign-off analysis.

V. 3-D OPENPITON DESIGNS

In this experiment, we choose OpenPiton [11], a highly configurable open-source ISA as our benchmark architecture. A single OpenPiton chip integrates many tiles, where each tile consists of a 64-bit Ariane RISC-V core and three levels of caches. The L1 and L2 caches are private to each tile, while the L3 cache is coherently shared between tiles. A network-on-chip (NoC) in each tile arbitrates the communication between tiles. In our benchmark designs, we choose a single tile design with 8 kB of L1 instruction cache, 16 kB of L1 data cache, 16 kB of L2 cache, and 256 kB of L3 cache.

A. Power, Performance, and Area Comparison

Figs. 6 and 7 show GDS layouts of our 2-D and 3-D IC designs targeting 700-MHz operating frequency. Table II summarizes and compares the PPA results of different designs. In GDS layouts of the microbumping design, microbumps are marked in blue and I/O drivers in red. In 3-D IC designs, six metal layers are used in the logic tier/die, and four metal layers in the memory tier/die. As memory modules occupy only four metal layers, we can minimize the number of metal layers in the memory tier/die. However, the microbumping 3-D design uses one additional layer in the memory die due to microbump placement and routing.

TABLE II

PPA COMPARISONS BETWEEN 3-D IC DESIGNS. THE PERCENTAGE GAIN OVER THE 2-D DESIGN IS SHOWN IN (), WHERE NEGATIVE MEANS GAIN

	2D	monolithic	hybrid bonding	micro-bumping
Metal layer usage (logic+memory)	6	6+4	6+4	6+5
Area (mm^2)	1.51	0.77 (-49.4%)	0.77 (-49.4%)	0.77 (-49.4%)
Logic area (mm^2)	0.298	0.298	0.297	0.296
Total WL (mm)	7.701	6.253 (-18.8%)	6.317 (-18.0%)	7.169 (-6.9%)
MIV/bump count (#)	-	9,418	908	313
3D net count (#)	-	2,408	592	313
Target freq. (MHz)	700.0	700.0	700.0	700.0
WNS (ns)	-0.40	-0.10 (-74.7%)	-0.07 (-81.4%)	-0.41 (+3.8%)
Effective freq. (MHz)	547.1	653.3 (+19.4%)	664.8 (+21.5%)	542.5 (-0.8%)
Total power (mW)	205.33	196.62 (-4.3%)	199.80 (-2.7%)	212.58 (+3.5%)
Energy-delay-product ($nJ \cdot s$)	0.54	0.43 (-20.5%)	0.43 (-20.7%)	0.57 (+4.4%)

In all 3-D designs, the design areas have been reduced by −49.4% when compared to the 2-D counterpart while the areas of logic gates remain similar. The M3D design achieves an 18.8% total wirelength reduction while the hybrid bonding design exhibits an 18.0% reduction. As shown in Fig. 8, the microbumping design has longer wires than other 3-D options because the locations of microbumps are fixed, whereas MIVs and F2F bond pads are not. Therefore, the overall wirelength in microbumping 3-D is reduced by 6.9% compared to the 2-D design.

When comparing MIV/bump counts, the M3D design has around 10× more vertical connections than hybrid bonding. This is due to the *metal layer sharing*, which is the sharing of metal layers of the memory tier for logical connections of the logic tier to minimize the wirelength. As shown in Fig. 1, the logic gates are placed in the middle of double-stacked BEOL in M3D, while they are placed at the bottom in hybrid bonding. Therefore, the *metal layer sharing* is favored in the M3D design as the number of 3-D nets is 4.07× higher than hybrid bonding as shown in Fig. 9. The number of bumps in the microbumping design is fixed at 313 according to the memory partitioning.

In terms of timing closure, the hybrid bonding design shows 81.4% worst negative slack (WNS) improvement when compared to the 2-D design. As shown in Fig. 10 and Table III, the wirelength of the critical path in hybrid bonding is 40.7% shorter than 2-D, while monolithic shows 15.7% reduction. Moreover, hybrid bonding shows a 14.7% shorter clock launch delay than monolithic, leading to a 26.4% improvement in timing. In the case of microbumping, WNS has been increased by 3.8% when compared to the 2-D design even though the wirelength of the critical path is shorter than other integration options. Unlike other 3-D designs, the clock launch path in the microbumping design is formed across logic and memory dies with the microbump. Therefore, 1.27 ns of clock launch
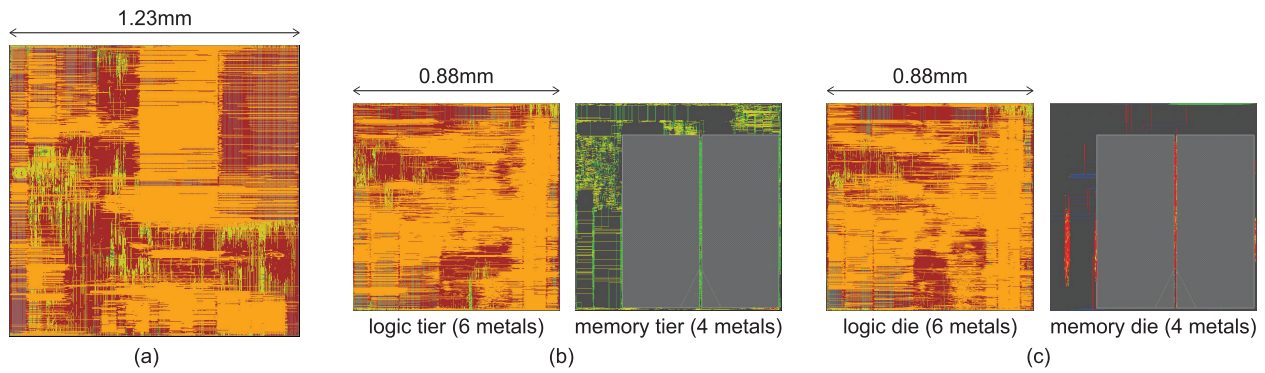
Fig. 6. GDS layouts of our 2-D, M3D, and hybrid bonding 3-D IC designs. (a) 2-D IC (six metals). (b) M3D IC. (c) Hybrid bonding 3-D IC.
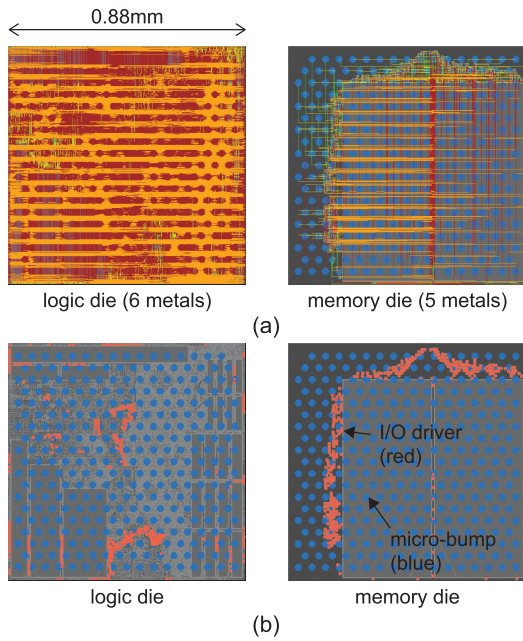

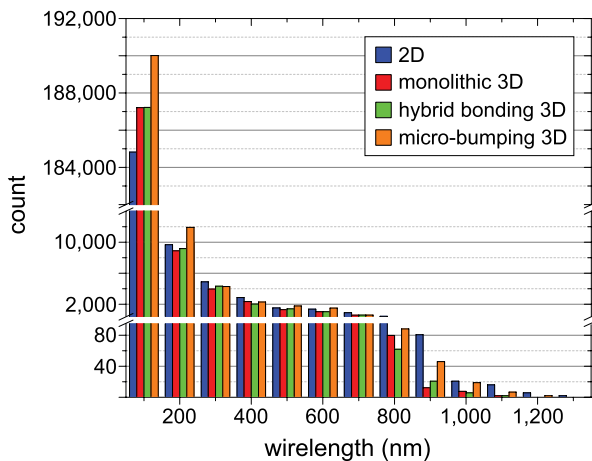Fig. 7. Microbumping 3-D design. (a) GDS layout. (b) I/O driver and microbump placement.

Fig. 8. Wirelength distribution in the 2-D and 3-D designs.

Fig. 9. Metal layer sharing in (a) M3D and (b) hybrid bonding 3-D designs. We highlight the logic nets in the memory tier/die.
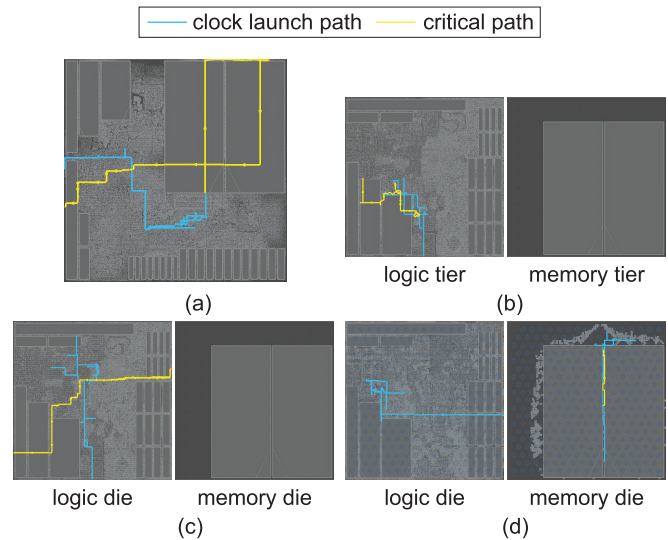
Fig. 10. Critical path and clock launch path in 2-D and 3-D designs. (a) 2-D. (b) M3D. (c) Hybrid bonding 3-D. (d) Microbumping 3-D.

delay with 3.71 mm of wirelength, which is 60.5% longer than the 2-D design, has led to the timing degradation in the microbumping 3-D design.

Fig. 11 shows the breakdown of the power consumption in 2-D and 3-D designs. Monolithic and hybrid bonding 3-D designs reduce power by 4.3% and 2.7%, while the microbumping design consumes 3.5% more power when compared to the 2-D design. The similar gate counts in the four designs lead to similar internal and leakage power consumptions considering the same temperature corner for each design. However, as the wirelength of monolithic and

TABLE III

CRITICAL PATH ANALYSIS OF 2-D AND 3-D DESIGNS. TARGET
CLOCK PERIOD IS 1.43 ns (= 700 MHz)

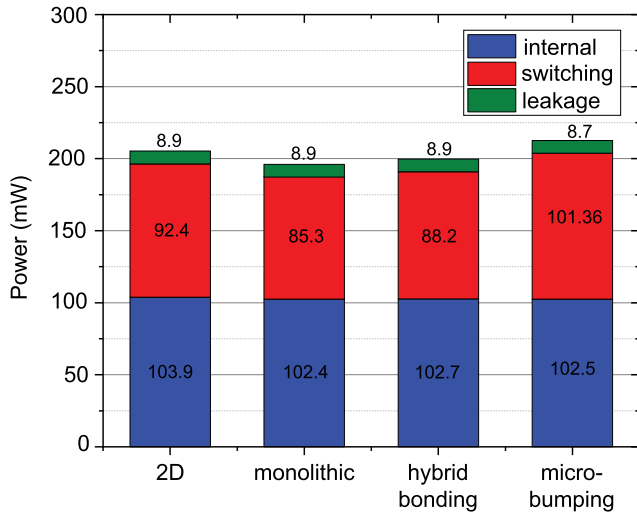|	2D	monolithic	hybrid bonding	micro-bumping
Worst negative slack (ns)	-0.40	-0.10 (-74.7%)	-0.07 (-81.4%)	-0.41 (+3.8%)
Critical path WL (mm)	3.12	2.63 (-15.9%)	1.85 (-40.9%)	1.25 (-59.9%)
Critical path delay (ns)	1.41	1.47 (+4.4%)	1.21 (-13.7%)	0.57 (-59.7%)
Clock launch path WL (mm)	2.18	1.64 (-24.7%)	1.78 (-18.4%)	3.71 (+70.1%)
Clock launch delay (ns)	0.42	0.34 (-20.1%)	0.29 (-31.2%)	1.27 (+200.9%)

Fig. 11. Breakdown of power consumption in all four designs.

Fig. 12. Clock tree layouts in our 2-D and heterogeneous 3-D designs. (a) 2-D. (b) M3D. (c) Hybrid bonding 3-D. (d) Microbumping 3-D.

TABLE IV

CLOCK TREE METRICS IN OUR 2-D AND HETEROGENEOUS 3-D DESIGNS

	2D	monolithic	hybrid bonding	micro-bumping
Buffer count (#)	3807	3489 (-8.35%)	3451 (-9.35%)	4245 (11.51%)
Clock WL (mm)	570.54	506.92 (-11.15%)	499.38 (-12.47%)	652.54 (14.37%)
Max. latency (ns)	0.68	0.42 (-39.04%)	0.44 (-35.38%)	0.85 (23.77%)
Max. skew (ns)	0.41	0.18 (-56.37%)	0.20 (-51.96%)	0.44 (7.75%)
Clock power (mW)	19.76	17.93 (-9.27%)	18.51 (-6.31%)	18.75 (-5.08%)

hybrid bonding designs have decreased, the switching powers have also reduced by 7.7% and 4.6%, respectively. In the microbumping design, the switching power has increased by 9.7% due to the microbump array between logic and memory dies. Even though the microbumping design has a 6.9% shorter wirelength, the parasitic of the microbump has mainly increased the switching power of the design.

B. Clock Tree Comparison

In this section, we compare the clock tree metrics in heterogeneous 3-D designs and propose guidelines for a robust clock tree design. Fig. 12 demonstrates the clock tree layouts in the various heterogeneous 3-D designs and Table IV shows the comparison of clock tree metrics. As shown in Fig. 12(a), the 2-D clock tree has long routing wires connecting the input clock port, clock gates, and clock pins of memory blocks, due to its large footprint size and obstructions of memory modules. However, some of these long 2-D clock nets are replaced with short 3-D vertical connections in the 3-D designs.

For the hybrid bonding 3-D design, the memory clock pins are all connected to F2F bumps directly and there is almost no clock net on the memory die. On the other hand, in the M3D design, the router utilizes the space available in the memory die to optimize the clock routing, which results

in a more balanced clock tree. As a result, monolithic and hybrid bonding 3-D designs provide a significant clock tree wirelength saving (>11%) compared to 2-D and require −8% fewer buffers to drive clock nets, which helps reduce the clock latency and power. M3D clock tree has the lowest skew, which enables high performance for heterogeneous 3-D systems.

The clock tree in microbumping 3-D shows inferior quality in clock wirelength and latency. One reason is that the large microbump pitch leads to longer routing wires between microbumps and clock pins, and microbumps themselves introduce nonnegligible *RC* delays. On the other hand, the clock tree of each die is implemented separately, which means that the tool cannot optimize the 3-D clock tree as a whole and the estimation of I/O delays introduce errors for clock tree balancing. These results suggest that the clock tree synthesis in microbumping 3-D designs needs to be done carefully with appropriate *RC* and I/O delay estimation, and iterative updates might be required to optimize the clock tree.

Clock trees also play an important role in full-chip power consumption due to the high switching activity of clock nets. Assuming no clock gating and a switching activity per cycle equal to 2 for all clock nets, we perform vector-less power analysis to evaluate the clock power in the heterogeneous 3-D design. The results show that the 3-D designs provide considerable clock power savings (up to 9.3%) compared with

0.88mm

[logic die] [memory die]

(a)

logic die (6 metals) memory die (4 metals)
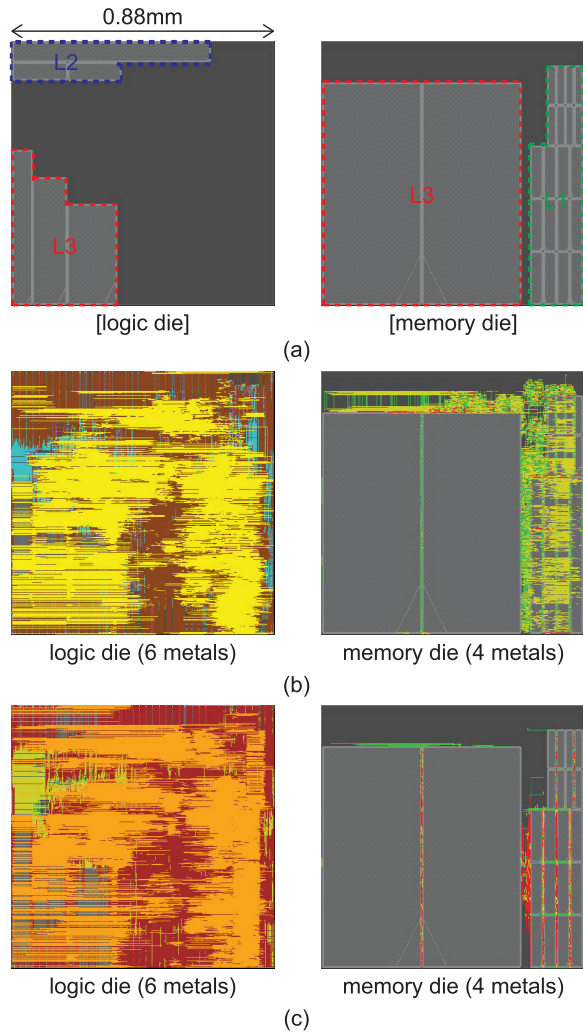
(b)

logic die (6 metals) memory die (4 metals)

(c)

Fig. 13. New partitioning result and GDS layouts of M3D and hybrid bonding 3-D. (a) New partitioning result. (b) GDS layouts of M3D. (c) GDS layouts of hybrid bonding 3-D.

2-D, and the best power reduction is provided by the M3D design because of the optimized 3-D clock tree.

VI. SCALABILITY OF 3-D IC DESIGNS

A. New Partitioning in Monolithic and Hybrid Bonding 3-D

In this section, we explore a new tier/die partitioning result in both M3D and hybrid bonding 3-D designs. As we have discussed in Section IV-A, the iso-partitioning analysis was done among all 3-D IC integration options since microbumping 3-D has a limited number of microbumps. However, as shown in Table I, MIV and F2F bond pads have submicrometer physical dimensions that provide denser 3-D interconnections. Therefore, more memory macros can be fit into the memory tier/die in M3D and hybrid bonding 3-D designs to take advantage of their fine-grained interconnections.

Fig. 13(a) shows our new tier/die partitioning result that is used in monolithic and hybrid bonding 3-D designs. In the new partitioning, we have kept the same footprint and moved L1 cache memories to the memory tier/die considering the architecture of OpenPiton [11]. Since L1 cache memories are included in the Ariane RISC-V core, the L1 cache has the

TABLE V
PPA RESULTS WITH OUR NEW TIER/DIE PARTITIONING. THE PERCENTAGE GAINS OVER BASELINE 3-D DESIGNS ARE SHOWN IN (), WHERE NEGATIVE MEANS GAIN

	monolithic	hybrid bonding
Logic area (mm^2)	0.292 (-2.0%)	0.292 (-1.6%)
Total WL (mm)	5.597 (-10.5%)	5.767 (-8.7%)
MIV/bump count (#)	19,680 (1.1×)	4,947 (4.5×)
Target freq. (MHz)	700.0	700.0
WNS (ns)	-0.06 (-40.4%)	-0.05 (-32.5%)
Effective freq. (MHz)	671.1 (-2.7%)	675.7 (-1.6%)
Total power (mW)	184.70 (-6.1%)	185.90 (-7.0%)
Energy-delay-product ($nJ \cdot s$)	0.410 (-4.6%)	0.407 (-5.3%)

majority of logic-to-memory connections. By relocating the L1 cache in the memory tier/die, we have shortened the wirelength of the existing partitioning as shown in Fig. 3(a). In new 3-D IC designs, we have chosen the same technology node, which includes six metal layers in the logic tier/die and four metal layers in the memory tier/die.

Table V summarizes the PPA comparison between two different partitioning results in M3D and hybrid bonding 3-D. Moreover, Fig. 13(b) and (c) shows their GDS layouts. The target frequency has remained the same at 700 MHz. As we have moved the L1 cache from the logic tier/die to the memory tier/die, the total wirelengths have been reduced by 10.5% and 8.7% in each monolithic and hybrid bonding 3-D designs. Moreover, MIV/F2F bump counts have been increased up to 4.5× to provide dense intertier/interdie connections when compared to the baseline designs.

The shorter wirelength has improved both timing and power consumption that leads to the improvement in terms of energy–delay product (EDP). Our new tier/die partitioning shows 40.4% and 32.5% of WNS reductions and 6.1% and 7.0% of total power savings in M3D and hybrid bonding 3-D, respectively. Up to 11.5% saving has been achieved in terms of the switching power the portion of which is largest in the power breakdown. Therefore, 4.6% and 5.3% of EDP improvements have been achieved in newly partitioned 3-D designs. Moreover, M3D with a new partitioning shows 4.7% EDP improvement when compared to the baseline hybrid bonding 3-D. This comparative analysis result indicates that monolithic and hybrid bonding 3-D designs show better performance in terms of PPA than microbumping 3-D with well-optimized 3-D design configurations such as tier/die partitioning.

B. 3-D Interconnect Scalability

In this section, we have chosen a new set of 3-D intertier/interdie connections as shown in Table VI to analyze the impact of 3-D interconnect dimensions on PPA. Considering the current mass production, we have chosen the 5.0-μm pitch of MIVs, 5.0-μm pitch of hybrid bonding pads, and 25.0-μm pitch of microbumps for hybrid bonding 3-D and microbumping 3-D designs, respectively.

Fig. 14 shows the GDS layouts of new 3-D OpenPiton designs, and the PPA analysis results are summarized in

TABLE VI

NEW SET OF INTERTIER/INTERDIE CONNECTIONS
USED IN CORTEX-A53 3-D DESIGNS

	monolithic	hybrid bonding	micro-bumping
Via/bump size	$0.25\mu m \times 0.25\mu m$	$2.5\mu m \times 2.5\mu m$	$12.5\mu m$
Via/bump pitch	$0.5\mu m$	$5.0\mu m$	$25.0\mu m$
Via/bump height	$0.1\mu m$	$0.17\mu m$	$12.5\mu m$

TABLE VII

PPA RESULTS OF 3-D OPENPITON DESIGNS WITH A NEW TIER/DIE
INTERCONNECTS. THE PERCENTAGE GAIN SHOWN IN () IS OVER THE
BASELINE DESIGN SHOWN IN TABLE II

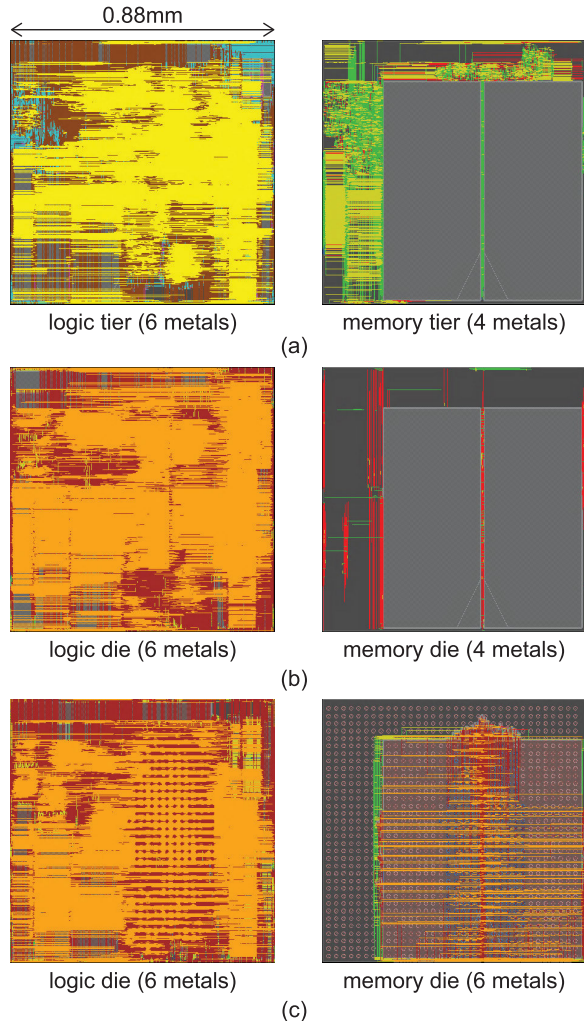|	monolithic	hybrid bonding	micro-bumping
Logic area (mm^2)	0.292	0.293	0.293
	(-2.0%)	(-1.3%)	(-0.9%)
Total WL (mm)	5.650	5.880	7.084
	(-9.6%)	(-6.9%)	(-1.2%)
MIV/bump count (#)	10,933	808	313
	$(1.2\times)$	$(0.9\times)$	$(1.0\times)$
Target freq. (MHz)	700.0	700.0	700.0
WNS (ns)	-0.09	-0.08	-0.38
Effective freq. (MHz)	657.9	662.3	552.5
	(+0.7%)	(-0.4%)	(+1.8%)
Total power (mW)	186.98	190.65	205.81
	(-4.9%)	(-4.6%)	(-3.2%)
Energy-delay-product	0.43	0.42	0.54
$(nJ \cdot s)$	(-1.5%)	(-4.1%)	(-5.0%)

Fig. 14.   GDS layouts of 3-D IC designs with a new set of 3-D interconnects in Table VI. (a) M3D. (b) Hybrid bonding 3-D. (c) Microbumping 3-D.

Table VII. Moreover, we have chosen the same floorplan shown in Fig. 3(a) for a fair comparison to the previous designs.

The overall PPA results are similar to the previous OpenPiton 3-D designs. In terms of the logic area, the maximum difference is 2.0% which is negligible since we have chosen the same partitioning in the previous designs. The wirelengths in all three 3-D designs have been reduced up to −9.6%. The finer pitch of MIV has enabled more *metal layer sharing* which leads to 1.2× of MIV count increase.

In terms of timing, the effective frequencies of new 3-D IC designs have remained similar by less than 2%. As shown in Fig. 10, the critical path in each 3-D design exists in the

same logic or memory tier/die. Even though the dimensions of 3-D interconnects have changed, the critical paths remain in the same tier/die. Therefore, the impacts on the timing due to changes in 3-D interconnects are minimal.

The total powers in M3D, hybrid bonding 3-D, and microbumping 3-D have reduced by 4.9%, 4.6%, and 3.2%, respectively. As the logic areas decreased, both internal power and leakage power were reduced in new 3-D OpenPiton designs. Moreover, the switching power has reduced due to the shorter wirelength, leading to the total power decrease. Finally, the EDPs have improved by 5% in the microbumping 3-D design due to the power improvement.

VII. 3-D CORTEX-A53 DESIGNS

In this section, we expand our comparative analyses to the commercial architecture. We chose the Arm Cortex-A53 processor, which is a high-efficiency processor that implements the Armv8-A architecture as our benchmark design. Our design benchmark consists of a single CPU core, 32-kB L1 instruction and data caches, and 1024-kB L2 cache memory. Arm Cortex-A53 also includes the NEON advanced single-instruction multiple-data (SIMD) engine and the floating-point unit (FPU).

A. Floorplanning and Microbump Assignment in Cortex-A53 Designs

For 3-D Cortex-A53 designs, we have also chosen a new set of 3-D intertier/interdie connections as shown in Table VI. For the technology node, we have chosen the same commercial 28-nm process design kit (PDK) as the previous experiments. The same design flows are chosen as shown in Fig. 4. During the floorplanning stage, we place 16 L2 cache modules in the memory die and the rest in the logic die as the MoL integration is chosen. Considering the architecture of Cortex-A53, we decide to move L2 cache modules to the memory die rather than L1 cache modules as shown in Fig. 15.

In Section IV-A, there were 313 microbumps in the final OpenPiton microbumping 3-D design. Therefore,
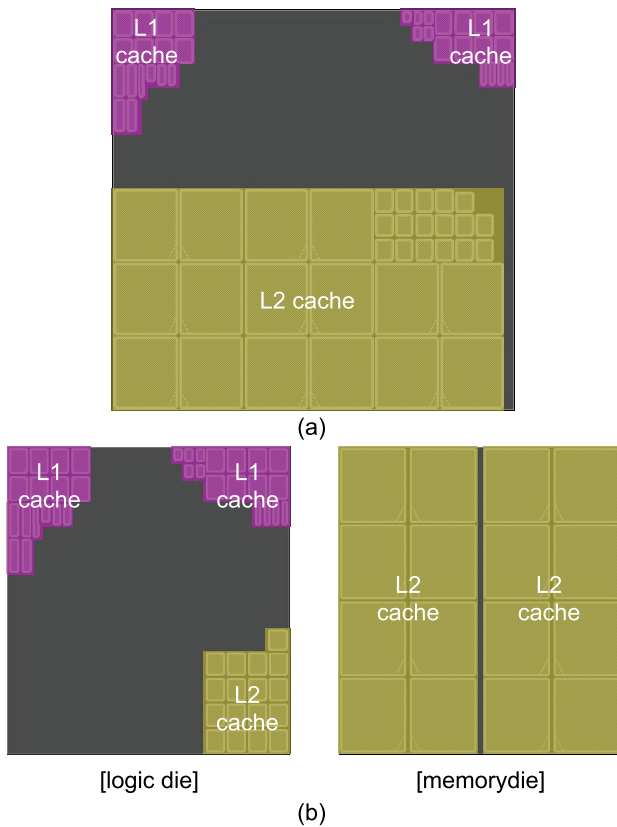
Fig. 15. Floorplans of Arm Cortex-A53 2-D and 3-D designs. (a) 2-D floorplan. (b) 3-D floorplan.

Fig. 16. Conceptual view of the microbump placement/assignment and the result. (a) Initial bump placement. (b) Final bump placement. (c) μ-bump placement/assignment result: the assigned bumps are highlighted in blue, and the unassigned bumps in red.

we have manually assigned the microbumps. However, in the microbumping 3-D design of Cortex-A53, there are 1173 die-to-die interconnections and 3416 available microbumps as shown in Fig. 16(c), which make the manual assignment no longer available.

Fig. 16 shows how we perform the microbump assignment in our microbumping 3-D design. As we adopt the MoL configuration, it is obvious that all 3-D nets, which go through microbumps, are directly connected to the I/O of memory modules in the memory die. Therefore, we first create the microbump array and place the primary I/O pins at the coordinates of the corresponding memory I/O pins as shown in Fig. 16(a).

Then, we generate the search boundary of microbumps whose center is the coordinates of the I/O pin and whose radius (R) is 1.5× of the microbump pitch to find the available microbumps. If there is no microbump available in the current boundary, we expand R to 3.0× of the bump pitch and perform the search again. Finally, we choose the nearest target from the initial coordinate as the proper location of the microbump to achieve the shortest routing lengths of 3-D nets as shown in Fig. 16(b).

B. Power, Performance, and Area: Max. Performance Comparison

Fig. 17 shows the GDS layouts of Arm Cortex-A53 2-D and 3-D IC designs. In the GDS layouts of the microbumping design, the microbumps are marked in blue and the I/O drivers
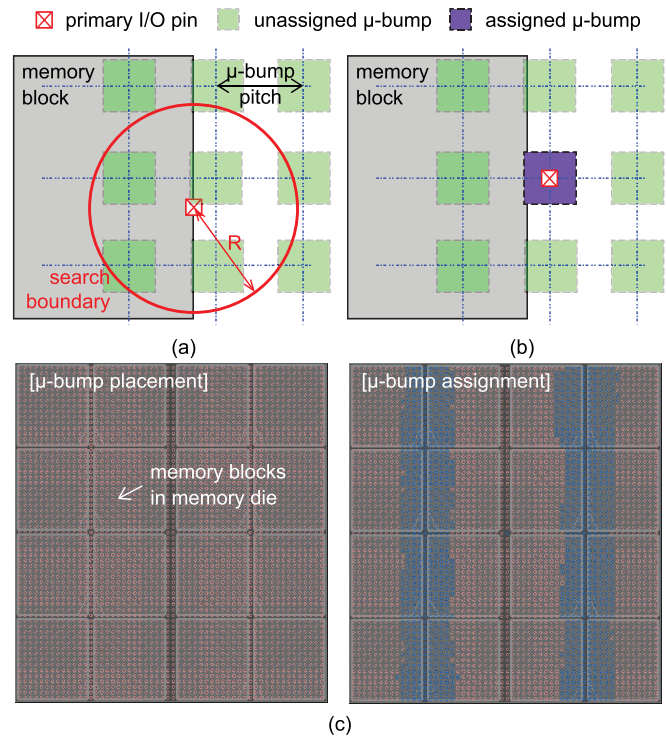
in red. In 3-D IC designs, six metal layers are used in the logic tier/die, and four metal layers in the memory tier/die same as in previous 3-D designs. However, the microbumping 3-D design takes six metal layers in its memory die considering the routability of the microbumps.

Table VIII summarizes and compares the PPA results of different designs at their maximum frequencies. As we expected, the areas of 3-D designs have been reduced by 50.0% when compared to the 2-D counterpart, while the areas of logic gates remain similar. In terms of the total wirelength, the savings in 3-D designs are negligible. The M3D design and the hybrid bonding design have achieved 5.3% and 0.9% wirelength reductions, respectively. The microbumping design has 1.6% longer wires than other 3-D options because of the fixed microbumps, whereas MIVs and F2F bond pads are not. Therefore, the overall wirelength in the microbumping 3-D is increased by 1.6% compared to the 2-D design.

When comparing the number of 3-D intertier/interdie connections, the M3D design also has 13.29× more vertical connections than hybrid bonding. This is due to the *metal layer sharing* as we have discussed in Section V-A. As shown in Fig. 18, the *metal layer sharing* is favored in the M3D design as the number of intertier/interdie 3-D nets is 7.96× higher than hybrid bonding in Cortex 3-D designs as well. The number of bumps in the microbumping design is fixed at 1173 according to the memory partitioning.

In terms of the maximum frequency, the hybrid bonding 3-D shows a 34.7% improvement when compared to the 2-D design. As shown in Fig. 19 and Table IX, the critical path delay in the hybrid bonding design is 27.0% shorter than 2-D,
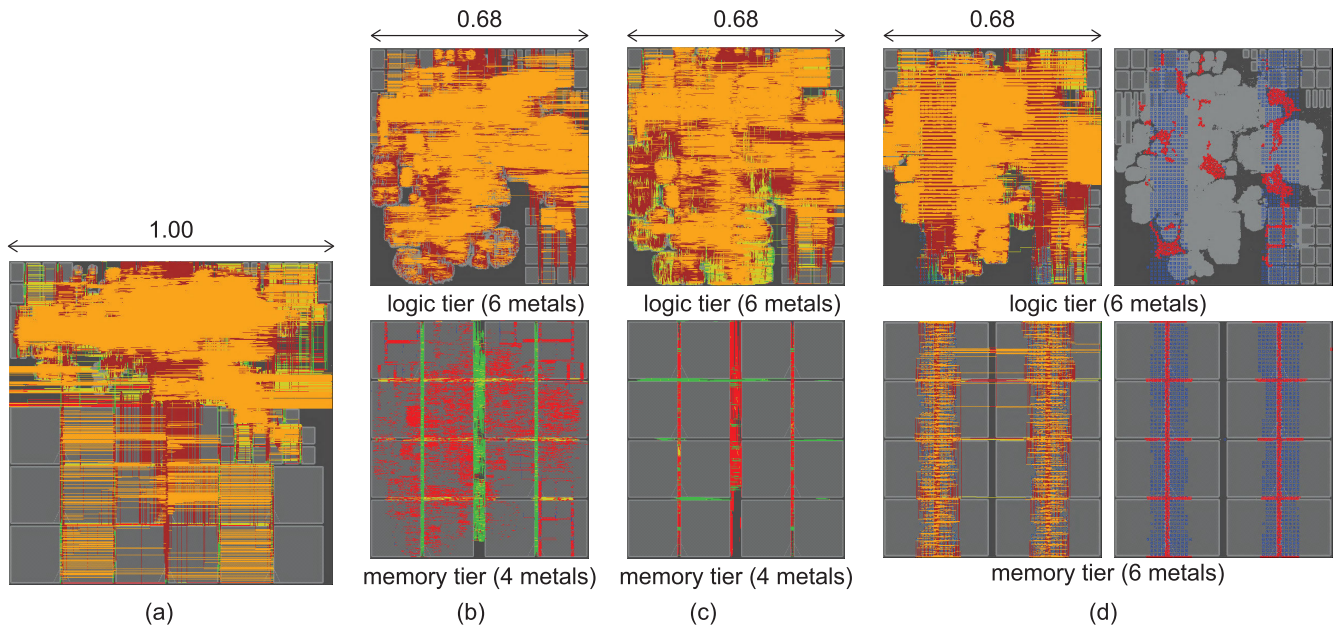
Fig. 17. GDS layouts of Cortex-A53 2-D and 3-D designs. In the microbumping 3-D design, microbumps are highlighted in blue and I/O drivers in red. (a) 2-D IC (six metals). (b) M3D IC. (c) Hybrid bonding 3-D IC. (d) Microbumping 3-D IC.

TABLE VIII

MAX. PERFORMANCE COMPARISONS BETWEEN 3-D IC DESIGNS. THE PERCENTAGE GAIN OVER THE 2-D DESIGN IS SHOWN IN (), WHERE NEGATIVE MEANS GAIN. ALL NUMBERS ARE NORMALIZED DUE TO NDA

	2D	monolithic	hybrid bonding	micro-bumping
Metal layer usage (logic+memory)	6	6+4	6+4	6+6
Area	1.00	0.50	0.50	0.50
Logic area	1.00	1.00	1.00	0.89
Total WL	1.00	0.95	0.99	1.02
MIV/bump count (#)	-	43,652	3,285	1,173
3D net count (#)	-	17,864	2,244	1,173
Max. freq.	1.00	1.33	1.35	0.89
Total power	1.00	1.30	1.35	0.84
Energy-delay-product	1.00	0.73	0.74	1.06

while M3D shows 24.8% reduction. Moreover, both M3D and hybrid bonding 3-D designs show 21.0% and 22.5% shorter clock launch delays when compared to the 2-D design. In the case of microbumping 3-D, the critical path delay has been reduced by 69.5% when compared to the 2-D design. However, the clock launch delay has significantly increased by $2.58\times$. Unlike other 3-D designs, the clock launch path in the microbumping design is formed across logic and memory dies with the microbump. Therefore, a huge clock launch delay has led to the timing degradation in the microbumping 3-D design.

For the cross-comparison of 2-D and 3-D designs at their own maximum frequencies, we have also tabulated the normalized EDP metric in Table VIII. We first observe that M3D and the hybrid bonding 3-D designs achieve 33.1% and 34.7% faster clock frequency compared to the 2-D counterparts, respectively. However, the microbumping 3-D design shows an 11.3% slower frequency than the 2-D design. As we have
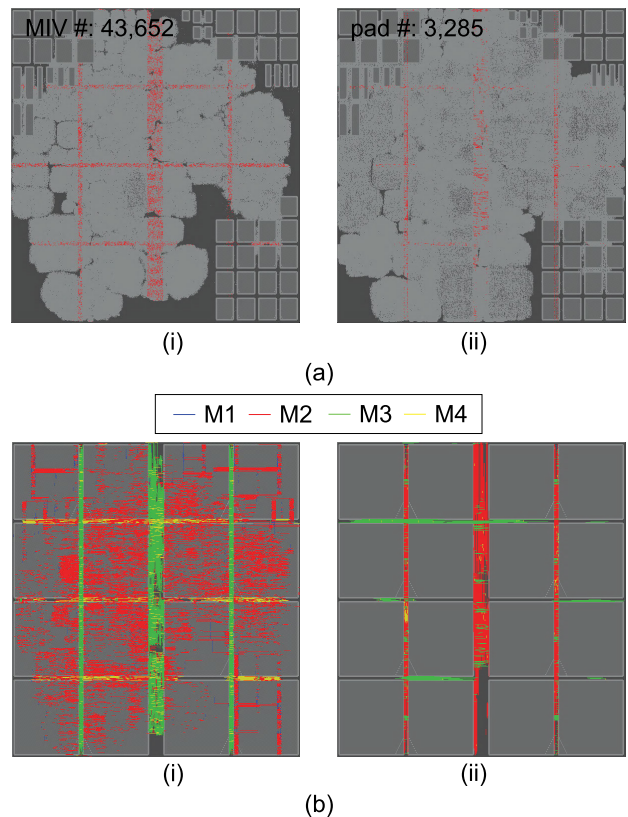

Fig. 18. Metal layer sharing effects in monolithic and hybrid bonding 3-D designs. (a) MIV and hybrid bond pad placements. (b) Metal layer sharing. (i) M3D. (ii) Hybrid bonding 3-D.

discussed earlier, the clock launch path in the microbumping 3-D has formed across the microbump. Therefore, the clock propagation delay through the microbump limits its frequency improvement.
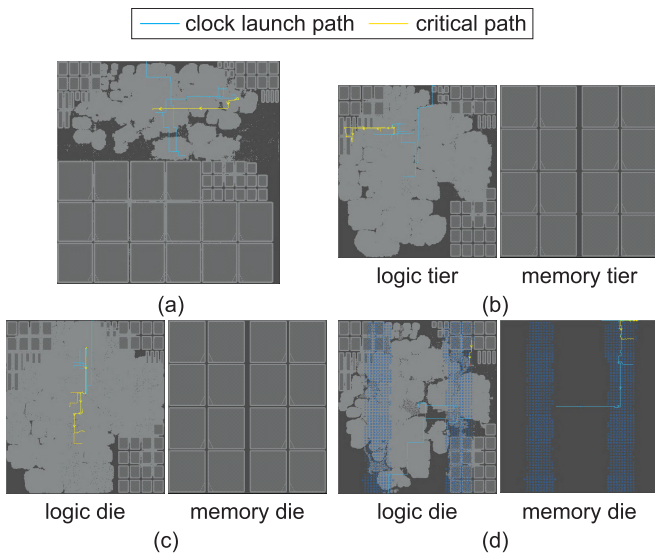
Fig. 19. Critical path and clock launch path in Cortex-A53 designs. (a) 2-D. (b) M3D. (c) Hybrid bonding 3-D. (d) Microbumping 3-D.

TABLE IX
NORMALIZED CRITICAL PATH ANALYSIS OF 2-D AND 3-D DESIGNS

	2D	monolithic	hybrid bonding	micro-bumping
Critical path delay	1.00	0.75	0.73	0.31
Clock launch delay	1.00	0.79	0.78	2.58

TABLE X
PARAMETERS FOR THE 3-D INTERTIER/INTERDIE CONNECTION MODELING

3D interconnect	MIV	F2F bond	micro-bump
Set A: OpenPiton			
$P_{VIA/BUMP}$	$0.6\mu m$	$1.0\mu m$	$50.0\mu m$
Via/bump height	$0.1\mu m$	$0.17\mu m$	$25.0\mu m$
Set B: Arm Cortex-A53			
$P_{VIA/BUMP}$	$0.5\mu m$	$5.0\mu m$	$25.0\mu m$
Via/bump height	$0.1\mu m$	$0.17\mu m$	$12.5\mu m$

TABLE XI
SNR OF 3-D INTERCONNECT MODELS

	MIV	F2F bond	micro-bump
Set A	$20.0dB$	$21.0dB$	$22.4dB$
Set B	$19.3dB$	$26.2dB$	$21.9dB$

The total power and energy values of all designs are calculated at their maximum frequencies. Finally, we observe that the M3D and hybrid bonding 3-D designs improve EDP by 26.7% and 25.8% compared to the 2-D design, respectively. However, the microbumping 3-D shows a 6.3% reduction due to its lower frequency. This degradation indicates the need for a 3-D clock synthesis scheme for the microbumping 3-D design. These EDP improvements represent the benefit of 3-D integration options.
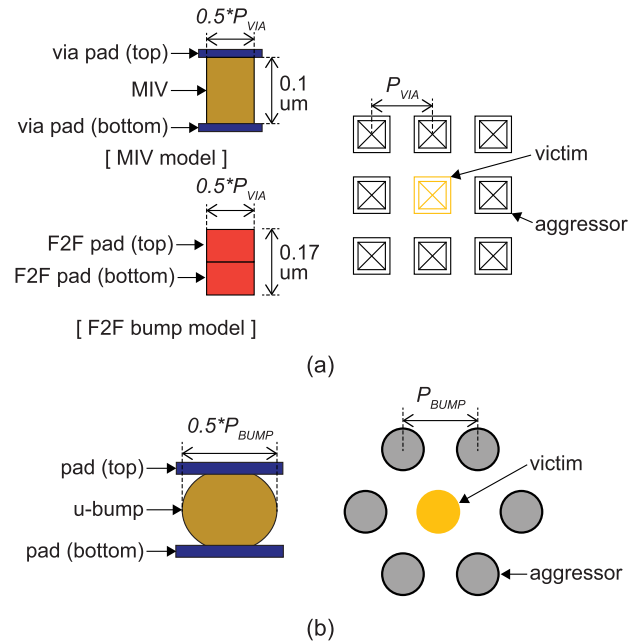

Fig. 20. MIV and F2F bond pads and microbump models for HFSS SI analysis. (a) MIV/F2F bond pad model. (b) Microbump model.

VIII. SIGNAL INTEGRITY ANALYSES ON 3-D INTERTIER/INTERDIE CONNECTIONS

A. Modeling Intertier/Interdie Connections

In this section, we model intertier/interdie connections of 3-D integration options by using ANSYS HFSS to analyze SI. In cases of M3D and hybrid bonding 3-D, as shown in Fig. 20(a), we model MIV and F2F bond pads as a single via and form the 3 × 3 via array according to the design rules. To observe the crosstalk effect, we set the center via as a victim and other surrounding vias as aggressors.

Fig. 20(b) shows the microbump modeling for the microbumping design. We model a hexagonal microbump array based on its physical dimensions and set the center bump as a victim and the surroundings as aggressors. Each generated model is then converted to an S-parameter and imported to Keysight ADS to perform SI analysis. Table X tabulates the physical dimensions of 3-D intertier/interdie connections, which we have used in OpenPiton and Cortex-A53 designs.

B. Signal Integrity Results

In Keysight ADS, we conduct eye diagram simulations at 0.7 Gb/s with the crosstalk model at each input of aggressors, the I/O driver impedance of 50 Ω as the ideal case on the transmitter side, and 5 pF for the parasitic on the receiver side as shown in Fig. 21(a). The simulation results of MIV and F2F bond pads and microbump in Set A are shown in Fig. 21(b)–(d), respectively.

From the results in Fig. 21(b)–(d), the microbump model shows the best eye-opening with 0.696-V height and 1.34-ns width because it has the smallest R value among the interconnect models due to its large physical dimension. As MIV has the smallest via width and pitch, its eye diagram shows
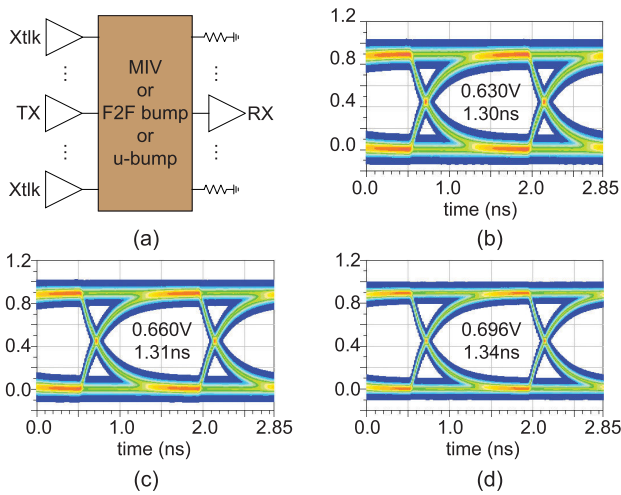
Fig. 21. Eye diagrams of intertier/interdie connections in our 3-D OpenPiton designs (Set A). (a) ADS simulation testbench. (b) MIV. (c) F2F bonding bump. (d) Microbump.
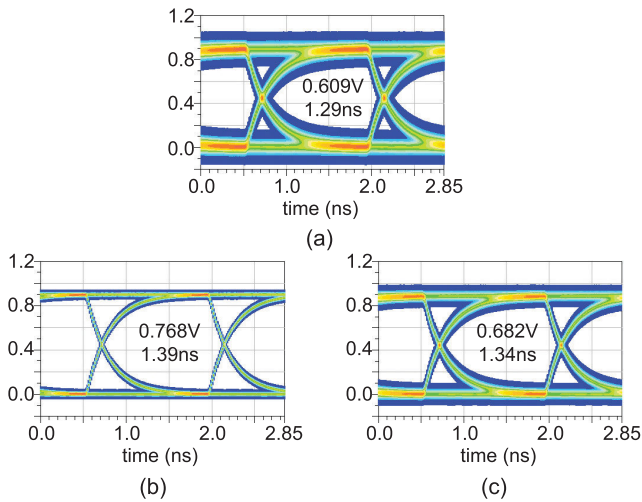

Fig. 22. SI analyses results of 3-D Cortex-A53 designs (Set B). (a) MIV. (b) F2F bonding bump. (c) Microbump.

0.630 V of eye height and 1.30 ns of eye width due to the stronger crosstalk effect.

Fig. 22 shows the eye diagrams of 3-D intertier/interdie connections, which are used in the second experiment: Arm Cortex-A53. Unlike the first SI analyses, the F2F bonding model shows the best eye-opening result with 0.768 V of eye height and 1.39 ns of eye width when compared to other 3-D interconnect models. Since the pitch of the F2F bonding pad has increased from 1 to 5 μm, the parasitic R has been significantly reduced when compared to the previous model. Moreover, the crosstalk effect has reduced with the larger pitch, and, therefore, the eye parameters have been improved significantly. In cases of MIV and microbump, the eye height and width remained similar as before.

Table XI summarizes the signal-to-noise ratios (SNRs) that are calculated from Figs. 21 and 22. When comparing each 3-D interconnect option, the finer pitch leads to the worse SNR due to the crosstalk from the adjacent vias or bumps. The worst SNR is shown in the 0.5-μm pitch MIV as 19.3 dB and the best SNR in the 5.0-μm pitch F2F bond pad as 26.2 dB.

Since all models show approximately over 20 dB, these 3-D interconnections are acceptable.

IX. CONCLUSION

In this article, we have presented a comparative study between three key heterogeneous 3-D integration options: monolithic, hybrid bonding, and microbumping technologies. We have conducted a PPA comparison between 3-D designs and their 2-D counterpart with the benchmark design of OpenPiton. Moreover, we have modeled the intertier/interdie connections of each topology and performed SI analysis to assess reliability. To observe the impact of intertier/interdie partitioning, we explored a new partitioning for M3D and hybrid bonding 3-D and performed a comparative analysis. In addition, we have expanded our analyses to the commercial processor design, which is Arm Cortex-A53, to make our study solid. In the additional experiment, we have proposed the microbump assignment methodology automatically to handle a number of microbumps in the 3-D design. Moreover, we have performed SI analysis on the new set of 3-D intertier/interdie connections. From our experimental results, the hybrid bonding 3-D design shows the best timing performance in all benchmark designs. In terms of SI, the microbumping 3-D has led to reliability in the OpenPiton benchmark, however, the hybrid-bonding 3-D in Arm Cortex-A53. This indicates that the physical dimension of 3-D intertier/interdie connections should be determined thoroughly since it affects both performance and reliability.

REFERENCES

[1] E. Beyne, "The 3-D interconnect technology landscape," *IEEE Des. Test*, vol. 33, no. 3, pp. 8–20, Jun. 2016.
[2] D. B. Ingerly et al., "Foveros: 3D integration and the use of face-to-face chip stacking for logic devices," in *IEDM Tech. Dig.*, Dec. 2019, p. 19.
[3] S. E. Kim and S. Kim, "Wafer level Cu–Cu direct bonding for 3D integration," *Microelectronic Eng.*, vol. 137, pp. 158–163, Apr. 2015, doi: 10.1016/j.mee.2014.12.012.
[4] P. Batude et al., "3D sequential integration opportunities and technology optimization," in *Proc. IEEE Int. Interconnect Technol. Conf.*, May 2014, pp. 373–376.
[5] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Shrunk-2-D: A physical design methodology to build commercial-quality monolithic 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1716–1724, Oct. 2017.
[6] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A physical design methodology to build two-tier gate-level 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 6, pp. 1151–1164, Jun. 2020.
[7] L. Bamberg, A. García-Ortiz, L. Zhu, S. Pentapati, D. E. Shim, and S. K. Lim, "Macro-3D: A physical design methodology for face-to-face-stacked heterogeneous 3D ICs," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2020, pp. 37–42.
[8] Y.-C. Lu, S. Pentapati, L. Zhu, G. Murali, K. Samadi, and S. K. Lim, "A machine learning powered tier partitioning methodology for monolithic 3D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 11, pp. 4575–4586, 2022.
[9] S. S. K. Pentapati, D. E. Shim, and S. K. Lim, "Logic monolithic 3D ICs: PPA benefits and EDA tools necessary," in *Proc. Great Lakes Symp. VLSI*. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 445–450, doi: 10.1145/3299874.3319486.
[10] L. Zhu et al., "High-performance logic-on-memory monolithic 3-D IC designs for arm Cortex-A processors," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 6, pp. 1152–1163, Jun. 2021.
[11] J. Balkind et al., "OpenPiton: An open source manycore research framework," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 2, pp. 217–232, 2016.