# Micro-bumping, Hybrid Bonding, or Monolithic? A PPA Study for Heterogeneous 3D IC Options

Jinwoo Kim, Lingjun Zhu, Hakki Mert Torun, Madhavan Swaminathan, and Sung Kyu Lim
School of ECE, Georgia Institute of Technology, Atlanta GA, USA
jinwookim@gatech.edu; limsk@ece.gatech.edu

***Abstract*—In this paper, we present three commercial-grade 3D IC designs based on state-of-the-art design technologies, specifically micro-bumping (3D die stacking), hybrid bonding (wafer-on-wafer bonding) and monolithic 3D IC (M3D). To highlight trade-offs present in these three designs, we perform analyses on power, performance, and area and the clock tree. We also model the tier-to-tier interconnection in each 3D IC methodology and analyze signal integrity to assess the reliability of each design. From our experiments, hybrid bonding design shows the best timing improvement of 81.4% when compared to its 2D counterpart, while micro-bumping shows the best reliability among 3D IC designs.**

## I. Introduction

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

Most recently, Intel has introduced the Foveros technology which enables 3D die stacking using micro-bump technology [2]. In micro-bumping 3D IC, two dies are stacked vertically with a dense array of micro-bumps in a F2F fashion, which provides high yield and reliability. Moreover, micro-bonding 3D IC enables heterogeneous 3D die stacking with a large flexibility in the technology selection and IP configurations.

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

M3D is an emerging technology which integrates device layers sequentially in the vertical direction [4]. Thanks to small monolithic inter-tier 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 top tier, which is processed at low-temperature to avoid the degradation of bottom tier.

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

1) This is the first work that compares the three key heterogeneous 3D integration approaches aforementioned. Our study is done using GDS layouts and sign-off quality simulations to convincingly quantify the power, performance, area (PPA) and signal integrity metrics.
2) We extend the state-of-the-art EDA tools built for 3D ICs to obtain competitive designs. In addition, we develop a new flow to handle 3D ICs that utilize the micro-bumping technology. We build our 2D IC baseline designs with a leading commercial vendor tool to substantiate our 2D vs. 3D IC comparisons.
3) Our study reveals useful PPA and signal integrity trade-offs among micro-bumping, hybrid bonding, and monolithic 3D 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.

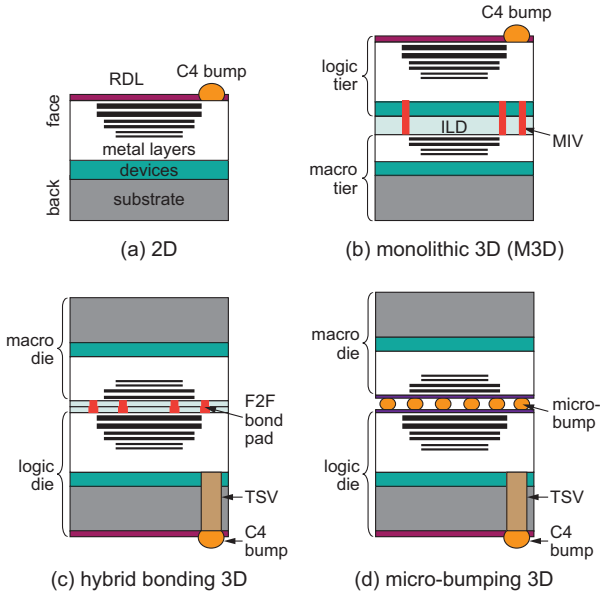



Fig. 1: The vertical stack-up of 2D and 3D integration options studied in this paper. M3D is face-to-back bonding, while hybrid bonding and micro-bumping are face-to-face.

## II. Experimental Setup

We choose a commercial $28nm$ technology node with high-$k$ metal-gates to perform the physical designs. Fig. 1 shows the vertical stack-ups of 3D designs. As we adopt logic-on-memory partitioning in our 3D design, the full metal stack is divided into logic and macro tiers/dies. In M3D IC and hybrid bonding 3D IC designs, we duplicate the 2D metal stack and

978-1-6654-3274-0/21/$31.00 ©2021 IEEE

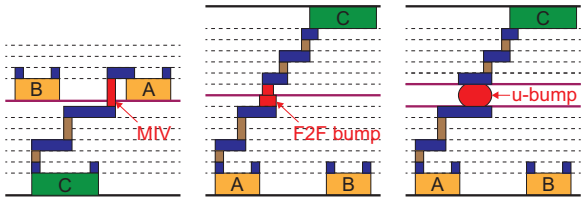



Fig. 2: Inter-tier/die interconnections of heterogeneous 3D integration options. Logic gates in logic tier/die is marked as yellow, and macro block as green.

TABLE I: Physical dimensions of inter-tier/die connections assumed in this paper.

| | 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 | $0.1\mu m$ | $0.17\mu m$ | $25.0\mu m$ |

form the doubled BEOL 3D metal stacks to generate our 2-tier designs as shown in Fig. 1(b) and (c). In case of micro-bumping 3D IC design shown in Fig. 1(d), we use a 2D metal stack for each die design, and integrate those 2D designs into a single 3D design with a micro-bump model.

Fig. 2 represents the vertical view of inter-tier/die connections between logic and macro tiers/dies. In M3D, MIVs provide connections from the pins of logic gates to the top metal layer of BEOL in macro tier. In hybrid bonding 3D, F2F bumps in logic and macro dies are bonded to provide the inter-die interconnections. Micro-bumping 3D uses micro-bumps for inter-die connections. Target nets are connected to the corresponding bump pads in each die while pads are bumped through micro-bumps.

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

We choose OpenPiton [5], 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 cache. 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 8kB of L1 instruction cache, 16kB of L1 data cache, 16kB of L2 cache, and 256kB of L3 cache.

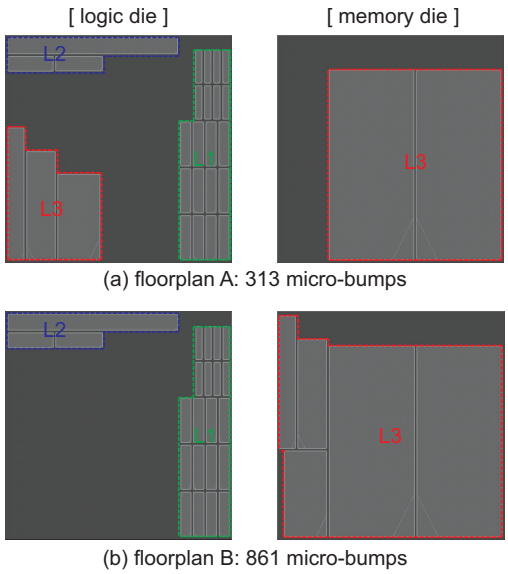



Fig. 3: Two partitioning and floorplanning options. We select Floorplan A in this paper due to a limit on the maximum micro-bump count.

## III. Heterogeneous 3D IC Design Flows

### A. Partitioning of Memory Modules

We adopt logic-on-memory partitioning in our study. We place the logic gates only in the logic tier while the memory tier includes macro blocks only such as memory modules. Therefore, the tier partitioning of macro blocks is important because the partitioning result affects the number of vertical interconnections.

As the size of micro-bump is larger than MIV and F2F bond pad, the micro-bump counts in micro-bonding design should be carefully considered to maintain small form factor of design. Fig. 3 shows the number of micro-bumps according to different floorplans of the OpenPiton architecture. Considering the footprint of 3D design as $0.88um \times 0.88um$, the maximum allowable number of micro-bumps is 400. Therefore, we decide to only assign L3 data cache in 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.

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

We design monolithic and hybrid bonding 3D designs with the flow shown in Fig. 4(a) [6]. While both designs use the full 3D metal stack, different 3D technology files are used. In M3D design, the 3D technology file includes MIV layer, while F2F bond pads are included as vias in hybrid bonding 3D design.

In the floorplanning stage, we generate 2D floorplans for logic and memory tiers separately. As discussed in Section III-A, we place L3 data cache blocks in memory tier and other memory parts in logic tier. We then project the floorplan of memory tier to the logic tier and generate a single floorplan.To avoid overlap issues between logic and memory tiers, we shrink memory blocks in the memory tier to the minimum

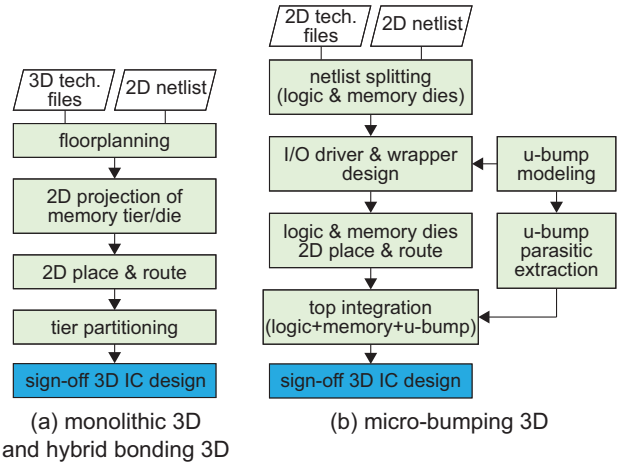



Fig. 4: Design flows used for monolithic [6], hybrid bonding [6], and micro-bonding 3D IC designs.

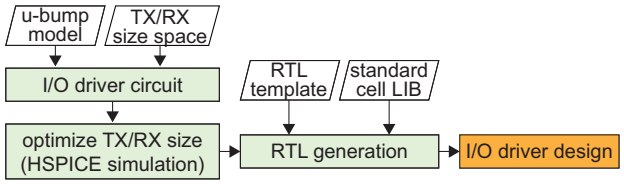



Fig. 5: I/O driver design and optimization flow used in our micro-bumping 3D IC design of Figure 4(c). We adopt Intel's AIB.

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

### C. Micro-bumping 3D Design Flow

Fig. 4(b) shows our design flow of micro-bumping 3D IC. Using ANSYS HFSS, we first perform the micro-bump modeling using physical dimensions presented in Table I. Then, we export the S-parameter of micro-bump 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 2D netlist considering the memory partitioning of Section III-A. We then design the I/O driver for inter-die connections which contains micro-bumps in the path. Unlike MIV or F2F bond pad, the size of micro-bump is significantly large, with a $25um$ diameter. Therefore, the I/O driver is necessary to transfer the signal properly through the micro-bump.

In micro-bumping design, we adopt Intel's AIB and select the proper size of transceiver with the micro-bump model shown in Fig. 5. For a wide range of TX/RX sizes and micro-bump 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 $\times 2$ and $\times 1$ respectively. The optimized I/O driver produces a $23.1\mu W$ of power and $20.2ps$ of propagation delay, which are within the design limit.

With pre-designed 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 2D P&R tool. In the P&R stage, we first place the micro-bump array and perform I/O assignment. By setting proper output loads and input delays for I/O micro-bumps, we perform P&R to obtain the final design of each die. I/O drivers are treated as macro blocks and placed automatically by the tool. As we design logic and memory dies separately, we finally integrate those designs with micro-bumps and perform sign-off analysis.

TABLE II: Power, performance, and area (PPA) comparisons between 3D IC designs. The percentage gain over 2D design is shown in (), where negative means gain.

| | 2D | monolithic | hybrid bonding | micro-bumping |
|---|---|---|---|---|
| Metal layer usage (logic+memory tiers/dies) | 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 wirelength ($mm$) | 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 frequency ($MHz$) | 700.0 | 700.0 | 700.0 | 700.0 |
| Worst negative slack ($ns$) | -0.40 | -0.10 (-74.7%) | -0.07 (-81.4%) | -0.41 (+3.8%) |
| Effective frequency ($MHz$) | 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%) |

## IV. Experiment Results

### A. Power, Performance and Area (PPA) Comparison

Fig. 6 and Fig. 7 show GDS layouts of our 2D and 3D IC designs targeting $700MHz$ operating frequency. Table II summarizes and compares the PPA results of different designs. In GDS layouts of micro-bumping design, micro-bumps are marked in blue and I/O drivers in red. In 3D IC designs, 6 metal layers are used in logic tier/die, and 4 metal layers in memory tier/die. As memory modules occupy only 4 metal layers, we can minimize the number of metal layers in memory tier/die. However, micro-bumping 3D design uses one additional layer in memory die due to micro-bump placement and routing.

In all 3D designs, the design areas have been reduced by -49.4% when compared to the 2D counterpart while the areas of logic gates remain similar. The M3D design achieves 18.8% total wirelength reduction while hybrid bonding design

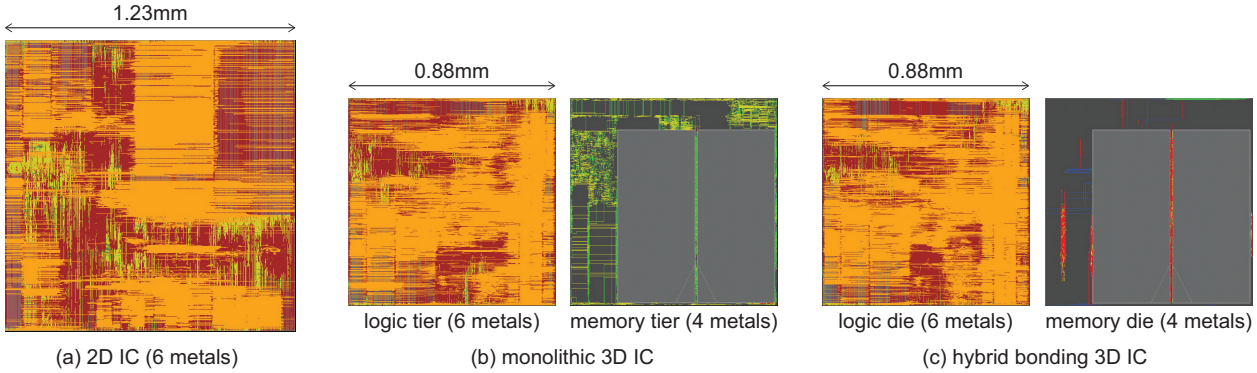



Fig. 6: GDS layouts of our 2D, monolithic 3D, and hybrid bonding 3D IC designs.

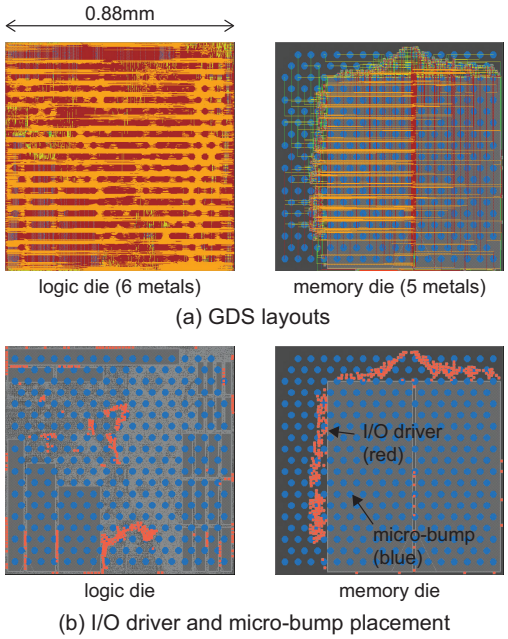



Fig. 7: Micro-bumping 3D design: GDS layout, I/O driver and micro-bump placement.

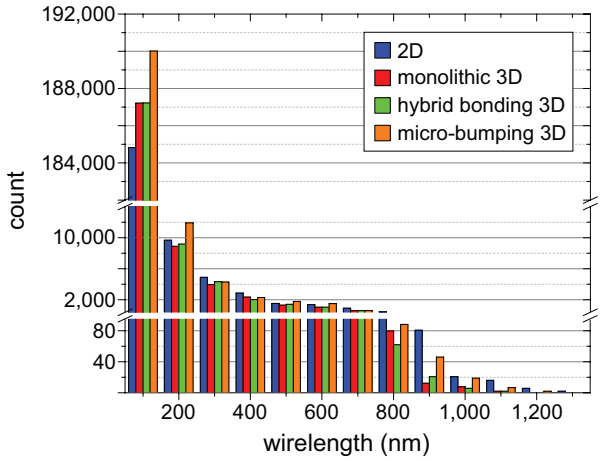



Fig. 8: Wirelength distribution in the 2D and 3D designs.

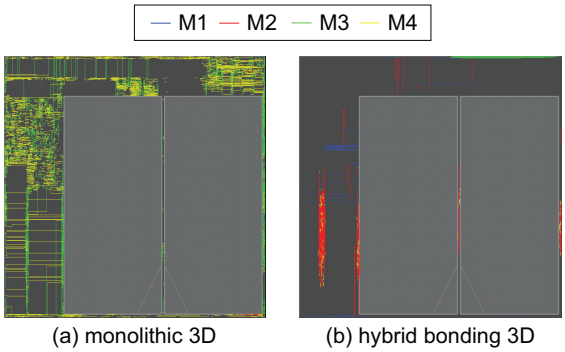



Fig. 9: Metal layer sharing in monolithic and hybrid bonding 3D designs. We highlight the logic nets in the memory tier/die.

exhibits 18.0% reduction. As shown in Fig. 8, micro-bumping design has longer wires than other 3D options because the locations of micro-bumps are fixed, whereas MIVs and F2F bond pads are not. Therefore, the overall wirelength in micro-bumping 3D is reduced by 6.9% compared to 2D design.

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

In terms of timing closure, hybrid bonding design shows 81.4% worst negative slack (WNS) improvement when compared to 2D design. As shown in Fig. 10 and Table III, the wirelength of critical path in hybrid bonding is 40.7% shorter than 2D, while monolithic shows 15.7% reduction. Moreover, hybrid bonding shows 14.7% shorter clock launch delay than monolithic, leading to 26.4% improvement in timing. In case of micro-bumping, WNS has been increased by 3.8% when compared to 2D design even though the wirelength of critical path is shorter than other integration options. Unlike other 3D designs, the clock launch path in micro-bumping design is formed across logic and memory dies with micro-bump.

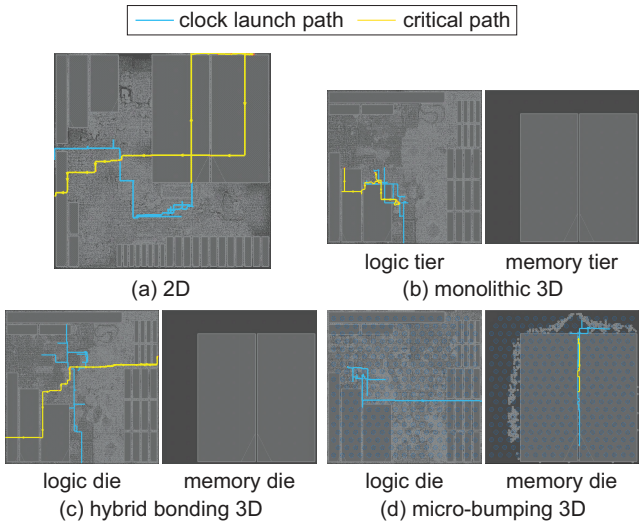



Fig. 10: Critical path and clock launch path in 2D and 3D designs.

TABLE III: Critical path analysis of 2D and 3D designs. Target clock period is $1.43ns$ (= $700MHz$).

| | 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 delay ($ns$) | 1.41 | 1.47 (+4.4%) | 1.21 (-13.7%) | 0.57 (-59.7%) |
| Clock launch delay ($ns$) | 0.42 | 0.34 (-20.1%) | 0.29 (-31.2%) | 1.27 (+200.9%) |

Therefore, $1.27ns$ of clock launch delay with $3.71mm$ of wirelengh, which is 60.5% longer than 2D design, has led the timing degradation in micro-bumping 3D design.

Monolithic and hybrid bonding 3D designs reduce power by 4.3% and 2.7%, while micro-bumping design consumes 3.5% more power, when compared to 2D design. Since the gate counts in designs are similar, the internal and leakage powers are similar as well. However, as the wirelength of monolithic and hybrid bonding designs have decreased, the switching powers have also reduced by 7.7% and 4.6%, respectively. In micro-bumping design, the switching power has increased by 9.7% due to micro-bump array between logic and memory dies. Even though micro-bump design has 6.9% shorter wirelength, the parasitic of micro-bump has mainly increased the switching power of the design.

### *B. Clock Tree Comparison*

In this section, we compare the clock tree metrics in heterogeneous 3D designs and propose guidelines for a robust clock tree design. Fig. 11 demonstrates the clock tree layouts in the various heterogeneous 3D designs and Table IV shows the comparison of clock tree metrics. As shown in Fig. 11(a), the 2D 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 2D clock nets are replaced with short 3D vertical connections in the 3D designs.

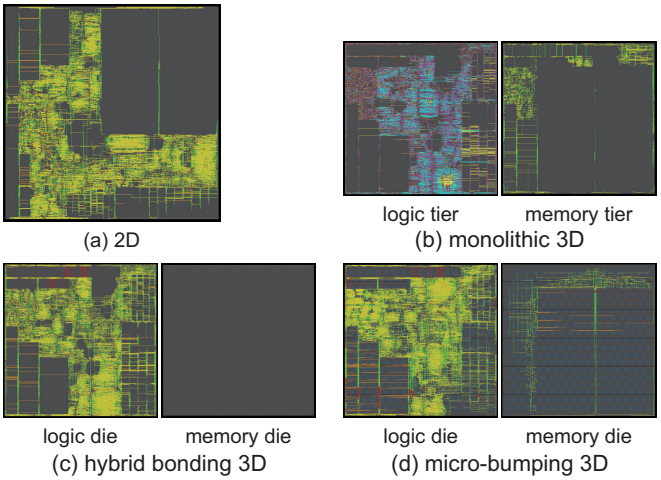



Fig. 11: Clock tree layouts in our 2D and heterogeneous 3D designs.

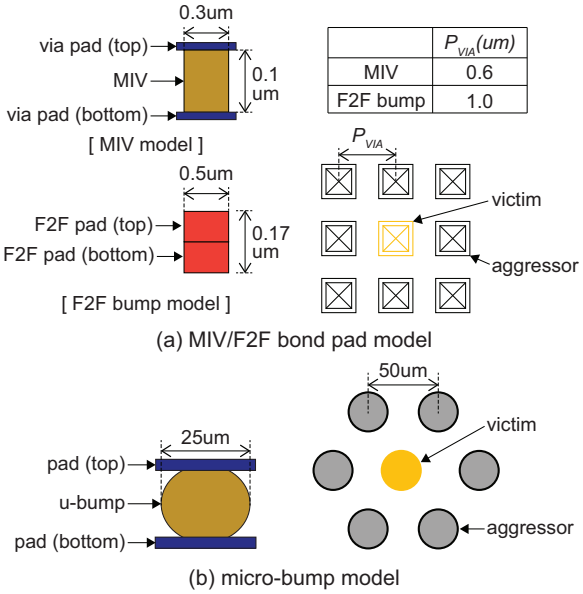
TABLE IV: Clock tree metrics in our 2D and heterogeneous 3D designs.

| | 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%) |

For hybrid bonding 3D design, the memory clock pins are all connected to F2F bumps directly and there is almost no clock net on memory die. On the other hand, in 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 3D designs provide a significant clock tree wirelength saving (>11%) compared to 2D and require -8% fewer buffers to drive clock nets, which helps reduces the clock latency and power. M3D clock tree has the lowest skew, which enables the high performance for heterogeneous 3D systems.

The clock tree in micro-bumping 3D shows inferior quality in clock wirelength and latency. One reason is that the large micro-bump pitch leads to longer routing wires between micro-bumps and clock pins, and micro-bumps themselves introduce non-negligible $RC$ delays. On the other hand, the clock tree of each die is implemented separately, which means the tool cannot optimize the 3D 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 micro-bumping 3D 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.




| | $P_{VIA}$(um) |
|---|---|
| MIV | 0.6 |
| F2F bump | 1.0 |

Fig. 12: MIV, F2F bond pad, and micro-bump models for HFSS signal integrity (SI) analysis. The width, height, and pitch used for MIV are $0.3um$, $0.1um$, and $0.6um$, respectively. The width, height, and pitch used for F2F bond pad are $0.5um$, $0.17um$, and $1.0um$, respectively.

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 3D design. The results show that the 3D designs provide considerable clock power savings (up to 9.3%) compared with 2D, and the best power reduction is provided by M3D design because of the optimized 3D clock tree.

### C. Signal Integrity Comparison

*1) Modeling Inter-tier/die Connections:* In this section, we model inter-tier/die connections of 3D integration options by using ANSYS HFSS to analyze SI. In cases of monolithic and hybrid bonding, as shown in Fig. 12(a), we model MIV and F2F bond pad as a single via and form 3×3 via array. To observe the cross-talk effect, we set the center via as a victim and other surrounding vias as aggressors. For micro-bumping design, we model a hexagonal micro-bump array and set the center bump as a victim and surroundings as aggressors. Each generated model is then converted to S-parameter and imported to Keysight ADS to perform SI analysis.

*2) Signal Integrity (SI) Results:* In Keysight ADS, we conduct eye diagram simulations at $0.7Gbps$ with the cross-talk model at each input of aggressors, the I/O driver impedance of $50\Omega$ as the ideal case on the transmitter side and $5pF$ for the parasitic on the receiver side as shown in Fig. 13(a). The simulation results of MIV, F2F bond pad and micro-bump are shown in Fig. 13(b)-(d), respectively.

From the results, the micro-bump model shows the best eye opening with $0.696V$ height and $1.34ns$ width because it has the smallest $R$ value among the interconnect models due to its large physical dimension. As MIV has the smallest

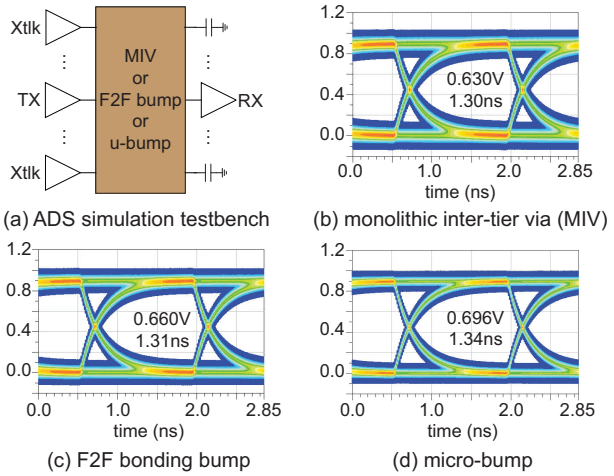



Fig. 13: Eye diagrams of inter-tier/die connections in our 3D IC designs.

via width and pitch, its eye diagram shows $0.630V$ of eye height and $1.30ns$ of eye width due to the stronger cross-talk effect. Micro-bumping 3D has shown the worst results in terms of PPA while the best in SI. Therefore, assuming the improvement of I/O drivers, micro-bumping can be the leading option in 3D IC design.

## V. Conclusion

In this paper, we present for the first time a comparative study between three key heterogeneous 3D integration options: using monolithic, hybrid bonding, and micro-bumping technologies. We conduct power, performance, and area comparison between 3D designs and their 2D counterpart. Moreover, we model the inter-tier/die connections of each topology and perform signal integrity analysis to assess reliability. From our experimental results, the hybrid bonding 3D design shows best timing performance whereas the micro-bumping design leads in reliability.


## Acknowledgements

This research is partially funded by the DARPA ERI 3DSOC Program under Award HR001118C0096, the Semiconductor Research Corporation under Task 2929, the National Research Foundation of Korea under NRF-2020M3F3A2A02082445, and Samsung Semiconductor, Inc.


## References


[1] E. Beyne, "The 3-D Interconnect Technology Landscape," *IEEE Design Test*, vol. 33, no. 3, pp. 8–20, 2016.

[2] D. B. Ingerly *et al.*, "Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 19.6.1–19.6.4.

[3] S. E. Kim and S. Kim, "Wafer Level Cu-Cu Direct Bonding for 3D Integration," *Microelectron. Eng.*, vol. 137, no. C, p. 158–163, Apr. 2015. [Online]. Available: https://doi.org/10.1016/j.mee.2014.12.012

[4] P. Batude *et al.*, "3D sequential integration opportunities and technology optimization," in *IEEE International Interconnect Technology Conference*, 2014, pp. 373–376.

[5] Http://parallel.princeton.edu/openpiton.

[6] 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)*, 2020, pp. 37–42.