Theme Article: Monolithic 3-D Integration

A Logic-on-Memory Processor-System Design With Monolithic 3-D Technology

Sai Pentapati
Georgia Institute of Technology

Lingjun Zhu
Georgia Institute of Technology

Lennart Bamberg
University of Bremen

Da Eun Shim
Georgia Institute of Technology

Alberto García-Ortiz
University of Bremen

Sung Kyu Lim
Georgia Institute of Technology

***Abstract*—In recent years, the size of transistors has been scaled down to a few nanometers and further shrinking will eventually reach the atomic scale. Monolithic three-dimensional (M3D) ICs use the third dimension for placement and routing, which helps reduce footprint and improve power and performance of circuits without relying on technology shrinking. This article explores the benefits of M3D ICs using OpenPiton, a scalable open-source Reduced Instruction Set Computer (RISC)-V-based multicore SoC. With a logic-on-memory 3-D integration scheme, we analyze the power and performance benefits of two OpenPiton single-tile systems with smaller and larger memory architectures. The logic-on-memory M3D design shows 36.8% performance improvement compared to the corresponding tile design in 2-D. In addition, at isoperformance, M3D shows 13.5% total power saving.**

Monolithic three-dimensional (M3D) integration is a growing trend in the semiconductor industry due to the performance and power benefits possible with its smaller footprint and shorter interconnects. Technology scaling predicted by Moore's law is gradually slowing down and new alternatives to silicon-based transistors are being explored. Some of the most promising solutions make use of materials such as carbon nanotubes[1]

Digital Object Identifier 10.1109/MM.2019.2944330
Date of publication 27 September 2019; date of current version 8 November 2019.

Published by the IEEE Computer Society

or ferroelectrics with negative capacitance effects.[2] Although such materials may be promising for performance and power benefits, many manufacturing related hurdles need to be cleared for their use in processors and other circuits in the near future.

Three-dimensional integration uses existing silicon manufacturing capabilities with few modifications and is a more practical option for the near future. Three-dimensional integration improves power, performance, and area (PPA) by stacking multiple smaller 2-D dies vertically instead of using a single 2-D die with a larger footprint. This leads to shorter interconnects and adds an extra degree of placement freedom in the z-direction along with the traditional x,y-directions.

M3D provides more flexible 3-D interconnections and potential benefits on routing and clock tree optimization. In this article, we will explore the impacts of logic-on-memory 3-D integration on the performance of an RISC-V-based processor system.

In this article, we use two different configurations of the RISC-V-based OpenPiton tile,[3] as our benchmark architecture: a memory-heavy case and a second case with smaller memory capacity. We present logic-on-memory partitioning for M3D ICs for the two different memory architectures, and show drastic PPA improvement of M3D over the respective 2-D designs for a commercial 28-nm process design kit (PDK). Furthermore, a detailed analysis outlines the cause of performance, power, and routing improvement of the logic-on-memory designs.

MONOLITHIC 3-D INTEGRATION

Background to Monolithic 3-D ICs

M3D is a 3-D integration methodology where the dies are fabricated sequentially on top of each other. The connections between the dies are realized by monolithic intertier vias (MIVs). The size and parasitics of MIVs are comparable to normal routing vias (see the "Technology Setup" section) and allow for a high 3-D connection density.

M3D-ICs are manufactured using low-temperature processes to enable sequential integration. Brunet *et al.*[4] successfully taped out and tested their low-temperature process technology that supports two device layers along with metal layers for each device layer.

Logic-on-Memory Monolithic 3-D Partitioning

The high demand for processor-to-memory bandwidth has become a challenge for modern computer systems. Three-dimensional integration of processor and memory is a promising solution to improve the memory bandwidth from the physical perspective because it reduces the wire delay between the processor and the memory by replacing long 2-D interconnections with shorter 3-D interconnections. Logic-on-memory is a special structure of 3-D integration as it separates the logic gates and memory blocks into different dies, which allows them to be fabricated with different technologies.

There have been a lot of studies on 3-D memory stacking with various 3-D integration technologies.[1,5,6] Major silicon companies such as AMD and Xilinx are also using 3-D integration techniques to improve the performance of memory in their next-generation products.[7] However, most of the studies use Through Silicon Via (TSV)-based or face-to-face bonded 3-D integration technology, which provide limited and predefined 3-D interconnections. On the other hand, M3D provides more flexible 3-D interconnections and potential benefits on routing and clock tree optimization. In this article, we will explore the impacts of logic-on-memory 3-D integration on the performance of an RISC-V-based processor system.

RTL-to-Graphic Database System (GDS) Tool Flow for Monolithic 3-D ICs

One of the challenges faced by M3D ICs is the lack of commercial tools to perform placement and routing (P&R) in the 3-D space. Currently available commercial tools only support placement in a single 2-D plane restricting their use in designing 3-D ICs. A 3-D placement should use the silicon area on both the top- and bottom-dies to optimize the design. A variety of flows have been developed that make use of 2-D commercial tools along with various heuristics to achieve a 3-D placement.[8–10]

Shrunk-2-D[8] is the first RTL-to-GDS flow developed to design commercial quality 3-D ICs from RTL using the design optimization capabilities of 2-

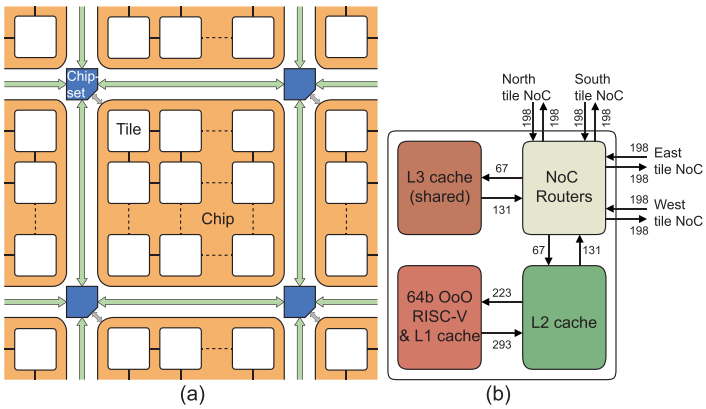

Figure 1. OpenPiton architecture. (a) Full system (adopted from the article by Balkind *et al.*[3]). (b) Single tile with data-flow width.

D P&R tools. Compact-2-D[9] flow has an added ability to perform complete routing and timing optimization in 3-D. Cascade-2-D[10] performs architecture-based 3-D placement. It also supports the complete routing and timing optimization for 3-D.

A poor partitioning choice undermines the benefits of 3-D ICs and architecture-based or heuristic-based placement should be done carefully. In this article, we make use of an extended version of the Shrunk-2-D flow tailored for P&R in logic-on-memory design. However, details on the EDA flow are outside the scope of the present publication.

EXPERIMENTAL SETUP

Benchmark Architecture

We use OpenPiton,[3] an open-source multicore processor system and framework, as the benchmark architecture. It is highly configurable, which makes it possible to change the core count, cache sizes, etc. The OpenPiton many-core system is shown in Figure 1(a). A full system consists of one or more chips and corresponding chipsets, while chips are made up of multiple tiles. Thus, a tile is an atomic piece out of which systems of arbitrary size are constructed. Hence, we only analyze the design of the tile while ensuring a correct functionality/timing when multiple tiles are later instantiated to create larger systems (more details on the resulting constraints in the "Design Setup" section). Thereby, we report results valid for systems with arbitrary tile counts.

The tile structure along with the bit-widths for data-flow is illustrated in Figure 1(b). It consists of a 64-bit out-of-order (OoO) RISC-V Ariane core and three levels of cache (L1–L3). The first two levels, L1 and L2, are private to the individual cores, while the third level is shared cache-coherent among all cores of the system. Thus, the physical memory of the shared L3 cache is distributed evenly among the tiles of a system. Network-on-chip (NoC) routers are used for the communication to provide good scaleability up to hundreds of cores/tiles.

Two variants of the tile are analyzed, which differ in their memory capacities. Case-I is a memory-heavy case with 16 kB of L1 instruction cache, 16 kB of L1 data cache, 128 kB of L2 cache, and 1 MB of L3 per tile. In Case-II, smaller memories are used with 8 kB of L1 instruction cache, 16 kB of L1 data cache, 16 kB of L2 cache, and 256 kB of L3 cache per tile. The memory macros occupy way more than 50% of the area in both cases, showing the suitability of logic-on-memory integration also for memory-light systems. The memory-macro floorplans for the 2-D and M3D designs are shown in Figure 2(a)–(f).

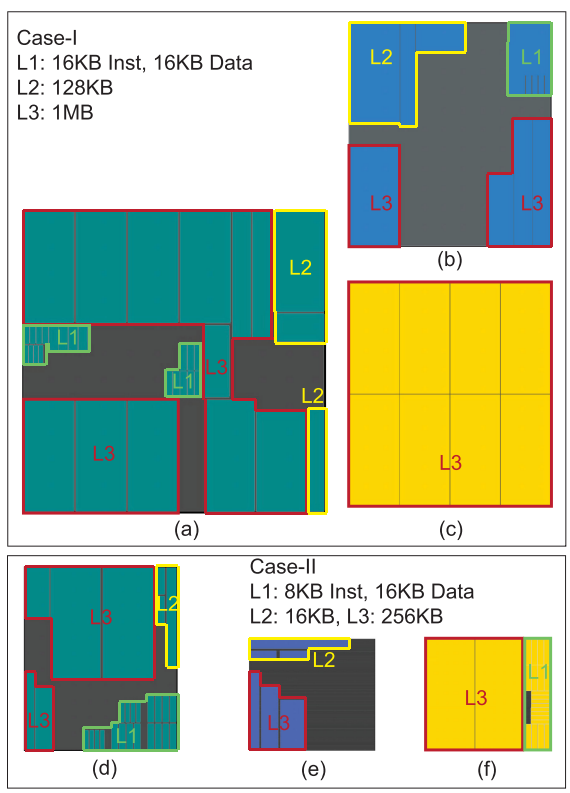

Figure 2. Physical layout of the memory modules. Case-I designs: (a) 2-D; (b) M3D top-die; and (c) M3D bottom-die. Case-II designs: (d) 2-D; (e) M3D top-die; and (f) M3D bottom-die.

Design Setup

In the tile design, the complete intertile NoC interconnection must be captured through constraints as corresponding paths start in one tile and end in another. For example, consider an NoC path starting in one tile-instance and ending in the north adjacent tile-instance. This path is represented in the tile-design by a path starting at an NoC-output-register and ending at a north-output pin, combined with a path starting at a south-input pin ending at an NoC-input register. Thus, both paths together have to finish in one clock-cycle and the north-output and the according south-input pin locations have to be aligned in a way that multiple instances of the tile can be connected without additional routing. Thus, in our tile-design input-to-NoC and NoC-to-output paths are constrained with a half-clock-cycle delay, all north–south output–input pin-pairs are horizontally aligned, all east–west output–input pin-pairs are horizontally aligned, and all pins are located in metal 3. This ensures timing closure for full systems with arbitrary tile counts. The frequency achieved per single-tile is used as the performance metric for the OpenPiton system in this article.

The power is calculated statistically using an input toggle rate of 0.1 and a flip-flop toggle rate of 0.2. The power–delay product (PDP) is the energy consumption of the design per clock-cycle. The energy–delay product (EDP) metric has a quadratic dependence on clock-period as it is the product of clock-period and the PDP.

In all the tile designs, a worst negative slack (WNS) below 3%–4% of the clock-period time is considered to be within the noise limit and is used as the timing-met condition.

Technology Setup

Using a commercial 28-nm PDK, the timing closure and analysis are done at the typical PVT corner at 0.9 V and 25°C. Six metal layers are used for routing in the 2-D designs. For M3D designs, we use six metal layers per die. Since each metal layer in M3D has half the XY-area of a 2-D metal layer, the overall available metal-layer area is equal for both 2-D and M3D designs.

For M3D, routing-technology files are created for the M3D metal-stack used (six metal layers per die). The metal-layer stack of each die resembles the 2-D metal-stack. A new cut-layer is added for the MIVs between the top metal of bottom-die and bottom metal of top-die. The MIV cut-layer passes through the transistor layer of top-die. This creates placement obstructions between MIVs and cells on the top-die. A width and spacing of 70 nm each is used for the MIVs. For comparison, this MIV is ~430 times smaller in area than the smallest inverter in this technology. Each MIV has R = 2 Ω and C = 0.02 fF. In comparison, a normal routing via connecting the metal layers M1 and M2 has a footprint of 50 nm × 110 nm with R = 8 Ω and C = 0.02 fF. Possible performance degradations due to the low-temperature M3D manufacturing are neglected as the transistors and metal layers on both tiers are assumed to be identical.

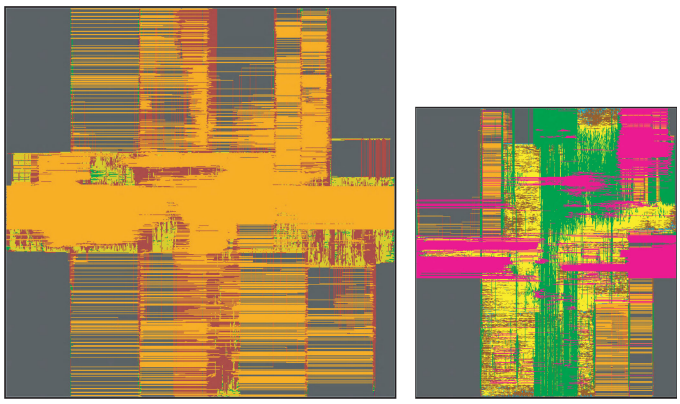

Figure 3. GDS layouts of single-tile OpenPiton Case-I (= large) memory architecture. (a) 2-D and (b) M3D.

DESIGN AND SIMULATION RESULTS

GDS Layouts

The memory layouts presented in Figure 2 are chosen as they minimize the distance between closely connected blocks and allow good memory-pin access for standard cells. In logic-on-memory M3D design, standard cells are placed only on the top-die (logic-die), leaving the bottom-die (memory-die) for macros. This arrangement is critical for the M3D design as huge macrocells on the top-die create huge obstructions for MIV insertion and restrict the MIV placement to the small channels between the memory macros. This setup would create routing congestion and increases the wirelength (WL) of the design. With the standard cells on the top-die, the MIVs can pass through the spacing between standard cells present

Table 1. Max-performance and isoperformance comparisons. (a) Max-performance comparisons of the 2-D and M3D designs of OpenPiton. (b) Iso-performance comparison of the Case-II (small memory architecture) 2-D and M3D designs of single-tile OpenPiton.

(a)

	Case-I: Large Memory			Case-II: Small Memory		
	2D	M3D	Δ_{M3D}	2D	M3D	Δ_{M3D}
Full-Chip Stats						
Frequency (MHz)	475	650	36.8	500	675	35.0
Width (mm)	1.97	1.32	-33.0	1.00	0.82	-18.0
Height (mm)	1.97	1.46	-25.9	1.20	0.73	-39.2
Silicon Area (mm^2)	3.880	3.854	-0.7	1.200	1.201	0.04
Cell Area (mm^2)	0.481	0.502	4.2	0.311	0.310	-0.5
Memory Area (mm^2)	2.856	2.856	0.0	0.775	0.775	0.0
Total WL (m)	12.09	10.96	-9.3	7.38	5.34	-27.6
Avg. WL/net (μm)	38.30	34.56	-9.8	35.05	26.06	-25.6
MIV Count	–	252,075	n/a	–	136,295	n/a
Total Power (mW)	292.69	399.27	36.4	146.21	178.04	21.8
PDP (mW*ns)	616.19	614.26	-0.3	292.42	254.34	-13.0
EDP ($mW*ns^2$)	1297.24	945.02	-27.1	584.84	363.35	-37.9
Critical Path Stats						
Inter-tile path	Yes	Yes	n/a	Yes	No	n/a
Path WL (μm)	2319	1440	-37.9	1824	1587	-13.0
Longest WL (μm)	772.2	385.6	-50.1	658.2	258.9	-60.7
Clock Period (ns)	2.1053	1.5385	-26.9	2.0000	1.4815	-16.7
Launch Latency (ns)	0.4028	0.3291	-18.3	0.3987	0.1501	-62.4
Cell Delay (ns)	0.3173	0.3306	4.2	0.4019	1.1846	195
Wire Delay (ns)	0.3422	0.1318	-61.5	0.2625	0.3611	37.6
Pin Delay (ns)	1.0526	0.7692	-26.9	1.0000	–	n/a
Setup Time (ns)	–	–	n/a	–	0.0040	n/a
Capture Latency (ns)	–	–	n/a	–	0.2557	n/a
Slack (ns)	-0.010	-0.022	120	-0.0631	0.0374	-159

(b)

	2D	M3D	Δ_{M3D}
Frequency (MHz)	500	500	0.0
Silicon Area (mm^2)	1.200	1.201	0.04
Cell Area (mm^2)	0.311	0.301	-3.3
Memory Area (mm^2)	0.775	0.775	0.0
Total WL (m)	7.38	5.33	-27.8
MIV Count	–	141,156	n/a
Total Power (mW)	146.21	126.53	-13.5
WNS (ns)	-0.0631	0.0626	n/a
Total Power distribution by power type			
Internal (mW)	73.12	70.02	-4.2
Switching (mW)	70.39	54.29	-22.9
Leakage (mW)	2.70	2.22	-17.8
Total Power distribution by cell type			
Sequential (mW)	38.01	35.42	-6.8
Combinational (mW)	71.61	55.24	-22.9
Macro (mW)	23.72	23.63	-0.4
Clock (mW)	12.86	12.23	-4.9
Clock Network Stats			
Clock Period (ns)	2.0000	2.0000	0.0
Clock WL (μm)	400,115	379,032	-5.3
Max Latency (ns)	0.4330	0.3059	-29.4
Max Skew (ns)	0.1464	0.1226	-16.3

throughout the top-die, allowing the router to place the MIVs throughout the die with only small obstructions of the standard cells.

Figure 3(a) and (b) shows the routing results on 2-D and M3D designs of Case-I OpenPiton tile. Thereby, routes of different metal layer are represented by different colors.

Analysis

Max-Performance Analysis The max-performance results of all the 2-D and M3D designs of the OpenPiton tile are presented in Table 1(a). In Table 1, $\Delta_{M3D} = (M3D -_2 D/_2 D) * 100\%$. Compared to the 2-D design, Case-I OpenPiton tile achieves 36.8% higher performance with logic-on-memory M3D integration [Table 1(a)]. A closer look at the delay numbers on the critical paths in 2-D and M3D helps understand the difference between these max-performance designs. The total delay of the critical paths are consisted of cell delay, wire delay, pin delay (added to capture intertile communication), and launch latency of the clock. As mentioned before, paths that end/start in an adjacent tile are assigned a delay equal to half the clock-period. Thus, only half clock-period is available for a flop-to-pin path and the other half clock-period is left for the pin-to-flop counterpart. While the 18.3% latency decrease is on-par with the clock-period reduction of 26.3% between 2-D and M3D, the main difference is the drastic reduction in wire delay portion of the critical path.

Although using NoCs reduces the interconnect bottleneck in many-core systems, the longer global wires for the NoC-based tile communication are still found out to be the system's performance bottleneck as they heavily contribute to the critical path in 2-D. M3D helps to overcome this bottleneck efficiently as long global wires are shortened in M3D. We observe that the WL of the critical path in M3D is 37.9% less than the critical path in 2-D, giving rise to a 61.5% lower wire delay in M3D compared to 2-D. In the 2-D design, wire delay makes up 32.5% of the half clock-period available, whereas in M3D the wire delay is only 16.7% of the half clock-period. Further timing analysis is done in the "Timing Path Analysis" section. The energy of the M3D system is nearly equal to the 2-D energy. As the average WL per net is only 9.8% smaller in the logic-on-memory M3D design, standard cells drive similar amount of wire-load in both 2-D and M3D. So a larger cell area is needed to meet timing with faster clock. This results in the power increase in M3D. Thus, the drastic performance improvement in M3D is obtained at a cost of power increase, nullifying the effect of M3D on the PDP.

In the Case-II OpenPiton design with smaller memories, we see a performance improvement of 35.0%, which is on-par with the Case-I design. The footprint in Case-II 2-D design is nearly one-third of the Case-I 2-D design and is also smaller than M3D design in Case-I. The critical path in Case-I 2-D is still between tiles, but the maximum WL and the wire delay portion of the total delay are smaller than 2-D Case-I due to the smaller footprint of the tile. In M3D Case-II, we see that the critical path is no longer a tile-communication path, but a memory-to-memory path. This is because the Case-II M3D footprint is small enough that global wires are no longer the performance bottleneck. In this Case-II M3D, memory latency is the performance bottleneck as it contributes to the 46.35% of the 1.1846 ns of cell delay on the critical path.

Another interesting aspect in the Case-II designs is that, even at a higher frequency, the standard cell area in the M3D design is still smaller than the cell area of the 2-D design. This is an indication that M3D is able to meet timing much better than 2-D. Because the average WL per net is 25.6% smaller in this M3D design, the standard cells meet timing easier, and removing the need to upsize the cells significantly. Therefore, the power increase is smaller than the frequency increase leading to an overall PDP benefit of 13.0% in M3D.

Isoperformance Analysis Table 1(b) compares the Case-II 2-D and M3D at an isofrequency of 500 MHz. In this comparison, we see a huge WL reduction of 27.8% in M3D. This is because, in Case-II M3D, the bottom-die contains both the shared L3 data cache and the L1 cache that is part of the RISC-V core. The standard cell placement by the commercial tool in the top-die is efficiently guided by both L1 and L3 blocks on the bottom-die. By placing logic cells on top of the L1 cache helps reducing the WL inside the core. This is the reason we see a much better WL savings in the Case-II designs using logic-on-memory M3D. The WL reduction leads to the 22.9% switching power savings in the M3D design. With the high switching power reduction, the total power is reduced by 13.5% in isoperformance Case-II M3D design compared to its 2-D counterpart.

MIV Count Analysis As seen in Table 1(a) and (b), the MIV count of the Case-I M3D designs

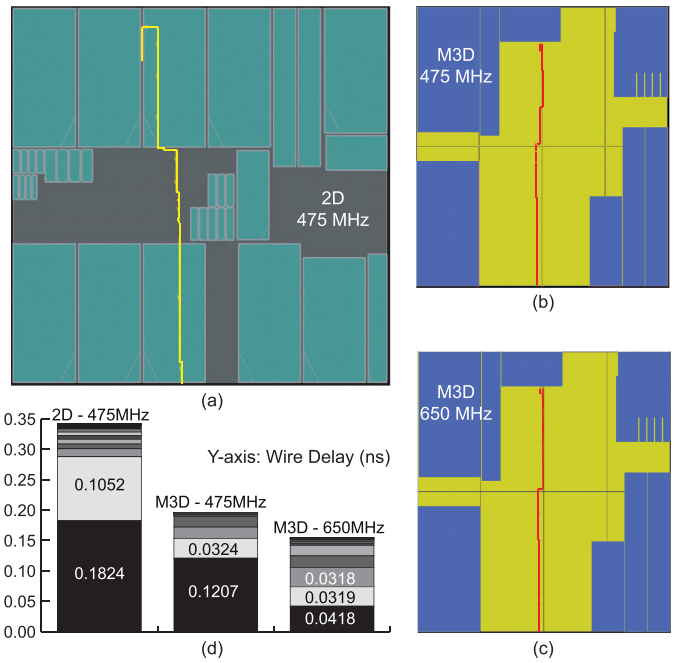

Figure 4. Timing critical path of Case-I 2-D memory architecture design in (a) 2-D at 475 MHz; (b) M3D at 475 MHz (iso-performance); and (c) M3D at 650 MHz (max-performance). (d) Detailed delay breakdown of the path in the designs.

is ~250 000 and for the smaller Case-II M3D designs, it is ~141 000. The standard cells are located on the top-die and have easier access to the top metal layers of the bottom-die. The bottom-die consists of memory macros, which blocks metal layers 1–4 for internal routing and have a sparse interblock routing. So, the routing resources on the top metal layers are underutilized by bottom-die macros. Commercial tools therefore access the bottom-die metals through MIVs to route the top-die interconnects. This leads to the high MIV counts observed and mitigates the routing congestion of the design.

Timing Path Analysis To better understand the impact of the logic-on-memory floorplan on the performance benefit of the OpenPiton processor system, we analyze the Case-I 2-D critical path in 2-D and M3D designs in Figure 4(a)–(c), highlighting the path in the respective layouts. By comparing a fixed path in all the designs, we can understand the effects of footprint reduction. A detailed wire delay breakdown of the path in these designs is shown in Figure 4(d). Here, the total wire delay is broken into delays of individual wires. Each block in the stacked-

column chart represents the delay of a wire between the output-pin of one standard cell to the input-pin of the next. Some of these wires have insignificant delays and cannot be seen in the stacked-column chart.

The 2-D critical path here is part of the tile communication and span a major portion of the width/ height of the floorplan. The tile-communication path is constrained and only half clock-period is available for the delay optimization. Out of the half clock-period, the wire delay makes up 32.5% of total delay in 2-D, 18.6% in 475-MHz M3D, and 17.2% in 650-MHz M3D design. The majority of the wire delay (0.2876 ns out of 0.3422 ns) in 2-D is caused due to two wires on the path, as seen in Figure 4(d). These wires are routed over the memory modules in 2-D and buffers cannot be placed to break down the long nets leading to large and widespread wire delays in 2-D. These paths are benefited by two main characteristics of the logic-on-memory M3D design: the first is due to the small footprint of M3D design reducing tile dimensions.

The second is an integral part of the logic-on-memory placement. In this placement, as the top-die is free of the huge memory blocks, it is easier for placing the buffers to split long wires if necessary. This is the reason the logic-on-memory-based M3D design shows high performance improvement in multicore processor systems. Thus, in the isoperformance M3D design at 475 MHz, the same timing path has a smaller wire delay of 0.1959 ns. Comparing the delay breakdown in the timing path in M3D at 475 and 650 MHz, the deviation in wire delay is also not as wide as that in 2-D. This is again due to the absence of the memory macros that obstruct placement and routing.

This discussion also explains the presence of the long wire in the Case-II 2-D design in Table 1 (a) even when the footprint is substantially smaller than those of the Case-I 2-D and M3D designs (refer to Figure 2 comparing the footprints head-to-head). Because the top and bottom portion of the 2-D-die is mainly occupied by macros, the vertical tile-communication paths need to bypass over the memories with a height of over 730 μm. This leads to the large maximum WL that is similar in Case-I and Case-II 2-D designs.

CONCLUSION

In this article, we benchmarked an RISC-V single-core system with a logic-on-memory M3D-integration scheme. We demonstrated a 37% improvement in maximum performance with M3D due to the critical paths being wire delay dominated. This shows that M3D alongside a good memory-macro floorplan is a very promising method for improving the performance of common NoC-based processor systems whose critical paths are still dominated by global wires. Using a smaller memory size for the tiles, we observe a 13.5% power savings, demonstrating the usability of logic-on-memory designs also for low-power designs.

The cost-impact of having huge MIV counts in the design is not considered here. A high MIV count can increase the cost of M3D ICs until the technology matures. Thus, an analysis of the max-performance as a function of MIV count is left for future work.

REFERENCES

1. M. M. Shulaker *et al.*, "Monolithic 3-D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *Proc. IEEE Int. Electron. Devices Meeting*, 2014, pp. 27.4.1–27.4.4.
2. Yadav *et al.*, "Spatially resolved steady-state negative capacitance," *Nature*, vol. 565, pp. 468–471, 2019.
3. J. Balkind *et al.*, "OpenPiton: An open source manycore research framework," in *Proc. Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2016, pp. 217–232.
4. L. Brunet *et al.*, "First demonstration of a CMOS over CMOS 3-D VLSI CoolCube^{TM} integration on 300 mm wafers," in *Proc. IEEE Symp. VLSI Technol.*, 2016, pp. 1–2.
5. D. H. Woo, N. H. Seong, D. L. Lewis, and H. S. Lee, "An optimized 3-D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth," in *Proc. Int. Symp. High-Performance Comput. Archit.*, 2010, pp. 1–12.
6. S. K. Lim, "3-D-MAPS: 3-D massively parallel processor with stacked memory," in *Design for High Performance, Low Power, and Reliable 3-D Integrated Circuits*, Berlin, Germany: Springer, 2013.
7. AMD, "High bandwidth memory," [Online]. Available: https://www.amd.com/en/technologies/hbm. Accessed: Sep. 10, 2019.

8. 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.
9. B. W. Ku, K. Chang, and S. K. Lim, "Compact-2-D: A physical design methodology to build commercial-quality face-to-face-bonded 3-D ICs," in *Proc. Int. Symp. Phys. Design*, 2018, pp. 90–97.
10. K. Chang *et al.*, "Cascade2-D: A design-aware partitioning approach to monolithic 3-D IC with 2-D commercial tools," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, 2016, pp. 1–8.

Sai Pentapati is currently working toward the PhD degree at the GTCAD Lab, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. He received the BTech degree in electronics and electrical communication engineering from IIT Kharagpur, Kharagpur, India, in 2017. His current research interests include physical design methodologies for monolithic stacked 3-D ICs. Contact him at: sai.pentapati@gatech.edu.

Lingjun Zhu is currently working toward the PhD degree at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. He received the BS degree in microelectronics from Fudan University, Shanghai, China, in 2018. His current research interests include wafer-on-wafer stacking technology, physical design methodologies, and high-performance and low-power 3-D IC designs. Contact him at: lingjun@gatech.edu.

Lennart Bamberg is currently working toward the Ph.D. degree at the University of Bremen, Germany, where he is employed since 2016 as a Teaching and Research Associate. In 2019, he was a Visiting Research Scholar at Georgia Institute of Technology, Atlanta, GA, USA. He received the BSc and MSc degrees with distinction in electrical and information engineering from the University of Bremen, Bremen, Germany, in 2014 and 2016, respectively. He received the Best Paper Award at PATMOS 2017 and PATMOS 2018. His research interests include low-power design and estimation, communication-centric design and 3-D SoC integration. He is a student member of the IEEE. Contact him at: bamberg@item.uni-bremen.de.

Da Eun Shim is currently working toward the PhD degree at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. She received the BS degree in general engineering from Harvey Mudd College, Claremont, CA, USA, in 2016. Her current research activities include exploration of monolithic 3-D IC design limits and PPA, and 3-D memory cubes. She is a student member of the IEEE. Contact her at: daeun@gatech.edu.

Alberto García-Ortiz is currently a Full Professor with the Chair of Integrated Digital Systems, University of Bremen, Bremen, Germany. He received the PhD degree with *summa cum laude* in 2003. His research interests include low-power design and estimation, communication-centric design, SoC integration, and variations-aware design. He received the "Outstanding Dissertation Award" in 2004 from the European Design and Automation Association. In 2005, he received from IBM an innovation award for contributions to leakage estimation. He is an editor and reviewer for several conferences, journals, and European projects. He is a senior member of the IEEE. Contact him at: agarcia@item.uni-bremen.de.

Sung Kyu Lim is currently the Dan Fielder Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. He received the PhD degree from the Computer Science Department, University of California, Los Angeles, Los Angeles, CA, USA, in 2000. He joined the School in 2001 and is currently a Full Professor. His research focus is on the architecture, design, test, and EDA solutions for 2.5-D and 3-D ICs. He has authored or co-authored more than 300 papers on 2.5-D and 3-D ICs. His research is featured as Research Highlight in the Communication of the ACM in January, 2014. He is the author of *Practical Problems in VLSI Physical Design Automation* (Springer, 2008) and *Design for High Performance, Low Power, and Reliable 3-D Integrated Circuits* (Springer, 2013). He received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006, the ACM SIGDA Distinguished Service Award in 2008. He was an Associate Editor of the IEEE Transactions on Very Large Scale Integration (VLSI) Systems during 2007–2009 and IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems during 2013–2018. He received the Best Paper Award from ATS'12, IITC'14, and EDAPS'17. His works have been nominated for the Best Paper Award at several top venues in EDA and circuit/packaging design. He is a senior member of the IEEE. Contact him at: limsk@ece.gatech.edu.