3D++: Unlocking the Next Generation of High-Performance and Energy-Efficient Architectures using M3D Integration

Biresh Kumar Joardar[†], Aqeeb Iqbal Arka[*], Janardhan Rao Doppa[*], Partha Pratim Pande[*]

[†]Department of ECE, Duke University
Durham, NC 27708, USA.
bireshkumar.joardar@duke.edu

[*]School of EECS, Washington State University
Pullman, WA 99164, U.S.A.
{aqeebiqbal.arka, jana.doppa, pande}@wsu.edu

***Abstract*—Three-dimensional (3D) integration has frequently been described as a means to overcome scaling bottlenecks, and advance both "More Moore" and "More Than Moore" through the use of vertical interconnects and die/wafer stacking. Recent industry trends show the viability of 3D integration in real products. Flash memory producers have also demonstrated multiple layers of memory on top of each other. However, conventional TSV-based 3D designs cannot achieve the full-potential of vertical integration and perform sub-optimally. Monolithic 3D (M3D) is an emerging vertical integration technology that promises significant power-performance-area benefits compared to TSVs. Hence, it is important to understand the necessary design trade-offs and challenges associated with this new paradigm. In this paper, we present both the advantages and the various design challenges in M3D-enabled system design considering Processing-in-Memory (PIM) and manycore systems as suitable case-studies.**

***Keywords*—M3D, Manycore, PIM, Thermal, Power**

I. Introduction

The unprecedented growth in big-data computing has contributed to the boom in the volume of information to be analyzed, classified, and stored. Gene sequencing, machine learning and social networking are some of the most popular big-data applications. However, these applications are both compute- and data-intensive in nature, which necessitates the design of suitable hardware platforms. The traditional approach of employing massive data centers for big data analytics has serious limitations in terms of scalability and power consumption to handle the expected growth in data volume. The volume of data is only bound to grow further in the upcoming years. This has led to the search for a suitable miniaturized computing system that can efficiently utilize the tremendous amount of data parallelism associated with these applications. Three-dimensional (3D) integrated circuit (IC) has emerged as a promising solution in this regard. 3D ICs enable new generation of architectures like Processing-in-Memory (PIM), which can significantly improve performance of big-data applications [1]. In addition, 3D ICs enable the integration of heterogeneous components such as CPUs, memories, analog circuits, etc. in the same die, which may be impossible otherwise.

Traditionally, 3D ICs are enabled by bonding two (or more) prefabricated dies using Through-Silicon Vias (TSVs). TSV is the most mature 3D integration technology. This vertical connectivity enables design of high performance and energy-efficient circuits and systems. Despite these advantages, the relatively large dimensions of TSVs (~um) present some fundamental limitations in high-performance, low-power system design: (a) fine-grained partitioning of logic blocks across multiple tiers is not possible [2], forcing only planar implementations of Processing Elements (PEs) and their associated logic blocks; (b) Thick bonding material introduces heat dissipation challenges [3]; and (c) TSVs add non-negligible area and power overheads. Overall, TSV-based 3D designs cannot achieve the full-potential of vertical integration. This can lead to relatively poor power-performance-area trade-offs in 3D architectures.

Meanwhile, monolithic 3D (M3D) has emerged as a technology that uses Monolithic Inter-tier Vias (MIVs) for fine-grained vertical integration. In M3D, two or more tiers of devices are fabricated sequentially, one on top of another. This eliminates the need for any die alignment, which enables considerably smaller via sizes. An MIV is essentially of the same size as a regular local via (diameter of 100 nm) [2]. This enables an integration density that is a few orders of magnitude higher than TSV-based 3D ICs. By utilizing these extremely small MIVs, M3D architectures have the potential to address the limitations of their TSV-based counterparts. The ultra-high density in M3D enables the design of hardware logic spanning multiple tiers. This leads to significant performance gains and better energy efficiency. For instance, an M3D-enabled adder spanning two tiers outperforms conventional designs by 33% [4]. These unique properties of M3D integration can be used to enhance existing 3D architectures and should be explored further.

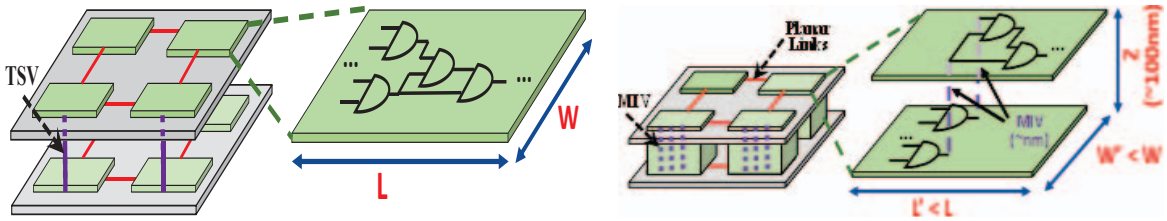
In this paper, we first discuss the limitations of existing TSV-based 3D architectures to motivate the need for M3D integration. Next, we highlight the advantages of M3D that can be used to design the next generation of high-performance and energy-efficient architectures. As an example, we first examine how M3D can be used to design more powerful PIM architectures *without creating thermal hotspots*. Next, we present the benefits of using M3D for manycore chip design. We show that by vertically partitioning core (e.g. CPU/GPU) and uncore (e.g. NoC, cache) elements using M3D, we can achieve better performance and lower on-chip temperature simultaneously without any thermal-aware optimizations.

II. TSV vs M3D: The Battle of 3D

A TSV-based 3D architecture is realized by physically stacking two prefabricated planar dies on top of each other. The adjacent layers are physically attached using a bonding material e.g., Benzocyclobutene [3] and are connected using TSVs. Fig. 1(a) shows an example TSV-based architecture. As discussed earlier, the logic blocks in each tier (the green blocks) are planar as shown in Fig. 1(b) and are physically stacked on top of each other to create the 3D architecture.

On the other hand, M3D integration is enabled by fabricating two or more silicon layers sequentially on the same

This work was supported, in part by the US National Science Foundation (NSF) grants CNS-1955353, CNS-1564014, and USA Army Research Office grant W911NF-17-1-0485.



(a) (b) (c) (d)

Figure 1. Illustration of (a) TSV-based 3D architecture and (b) one of its planar logic blocks; (c) M3D-based architecture and (d) one of its logic blocks partitioned across two-tiers used in M3D-based designs. The width (W) and length (L) of an equivalent logic block is substantially smaller in M3D. This figure is for illustration purpose only; it does not implement any specific logic.

substrate and interconnecting the layers using small MIVs [5]. This is fundamentally different from 3D integration using TSVs to interconnect separately fabricated dies. Fig. 1(c) and Fig. 1(d) illustrate an M3D-enabled architecture spanning two tiers. The gates spread across different tiers are connected using MIVs. By distributing the gates across multiple tiers, the physical dimensions of M3D-based designs are considerably smaller than those of its planar counterparts used in TSV-based architecture. Next, we present the power-performance-area trade-offs associated with TSV and M3D integration.

Performance: As shown in Fig. 1(b), the individual logic blocks in TSV-based 3D systems are principally planar. Hence, in TSV-based designs, the main performance benefits arise from physical nearness in the vertical direction (few µm) and not from any performance gain of individual logic blocks. On the other hand, M3D enables the design of logic blocks spanning multiple tiers (as shown in Fig. 1(d)). Hence, the logic blocks effectively become smaller, resulting in significantly smaller wires; Note that the vertical dimensions is few nm only. Due to this wirelength reduction, critical path delay is improved. Hence, multi-tier logic blocks can operate at higher frequencies compared to its TSV-based counterparts, leading to better performance. For instance, an M3D-enabled adder spanning two tiers achieves 33% lower execution time than conventional planar designs [4]. Many such multi-tier logic blocks can then be stacked on top of each other (as in TSV-based designs) for further performance benefits.

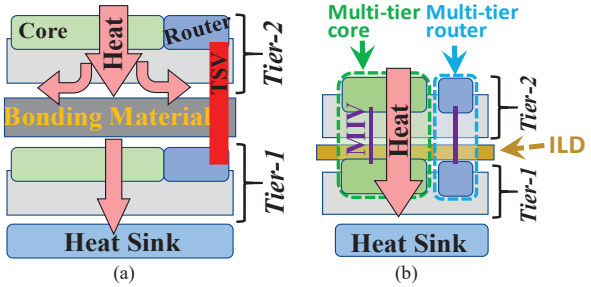
Power and Thermal: It is well known that 3D integration has inherent thermal issues [1]. This happens as heat dissipation is challenging, particularly for layers that are farther away from the sink. Hence, it is important to consider the power dissipation and the related on-chip temperature in any 3D architecture. As discussed earlier, multi-tier logic design enabled by M3D reduces wirelength, which in turn necessitates fewer repeaters. As a result, M3D-based architectures inherently consume less power than their TSV counterparts. For instance, an M3D enabled CPU is 39% more energy-efficient than its TSV counterpart without compromising performance [15].

Next, the different physical structures of TSV and M3D-based designs (as shown in Fig. 2) also influence the on-chip heat flow (and the thermal profile). The layer of bonding material between adjacent silicon tiers in a TSV-based 3D system has very poor thermal conductivity (Fig. 2(a)) [3]. This prevents heat from easily flowing towards the heat sink. In addition, TSVs are much larger than MIVs, resulting in thicker silicon layers and a longer path for heat flowing towards the sink. Due to these reasons, a major portion of the heat spreads laterally rather than flowing vertically towards the sink (as shown in Fig. 2(a)). As a result of this gradual heat accumulation among the layers, the overall temperature of the chip increases, which can also negatively affect the performance and lifetime of the system. Unlike their TSV counterparts, M3D integration (shown in Fig. 2(b)) inherently exhibits better thermal properties due to thinner tiers and the absence of any bonding material [3]. The inter-layer dielectric (ILD) in M3D is significantly thinner and has better thermal characteristics than the equivalent "Bonding Layer" of TSV. As a result, M3D-based architectures tend to have lower *maximum* on-chip temperature than their TSV counterparts. This property can be utilized to design better Processing-in-Memory architectures as we discuss later.

Area: By placing different logic gates across multiple tiers, i.e., in M3D, the overall planar dimensions is reduced significantly (L'<L and W'<W as shown in Fig. 1(d)) while the vertical dimension (H') is relatively small (~100 nm). As a result, M3D-based architectures are more area efficient than their TSV counterpart. For instance, a M3D-enabled L1 cache requires 49% less area than conventional designs [18]. The physical nearness can be utilized to design high-performance manycore chips as we elaborate in this work.

III. Deisgn Examples Using M3D

Depending on the granularity with which devices are partitioned across multiple tiers, M3D-based architectures can be grouped into three main categories: (a) transistor-level or N/P partitioning: the nFETs and pFETs of a gate are placed on two separate tiers and connected via MIVs [6]; (b) gate-level partitioning: planar gates are placed in different tiers and connected using MIVs [7]; and (c) block-level partitioning: intellectual property (IP), functional, and memory blocks are placed in different tiers and connected using MIVs [8]. Among these different partitioning techniques, gate-level partitioning has been shown to achieve the highest footprint reduction and performance improvement [7]. By placing different logic gates across multiple tiers, i.e., in 3D, the overall wirelength is reduced significantly. This leads to higher clock frequencies due to lower latency along the critical paths and a simplified and more energy-efficient clock tree and power delivery network. Next, we discuss few examples of how M3D can be used to design the next generation of high-performance hardware architectures.



(a) (b)

Fig. 2. Illustration of physical structure and heat flow in (a) TSV two-tier cross-section, (b) M3D two-tier cross-section. [16]

A. M3D-enabled PIM Architectures

Processing-in-Memory (PIM) involves moving the computational units closer to memory. This allows efficient data transfer, which is necessary for several memory-intensive big-data applications like neural networks, graph analytics and bioinformatics. Prior works have shown that PIM achieves significant speed-up and energy savings for these applications compared to conventional architectures [9][10]. The faster memory access enabled by PIM is crucial for these applications as a large fraction of time (>85% [9]) can be spent in fetching/storing data to/from memory resulting in poor hardware utilization and performance.

However, temperature presents an important limitation in conventional PIM architectures. DRAM retention capability is lowered beyond 85°C. Once temperature exceeds this threshold, refresh rate must be doubled for every ~10°C increase. Higher refresh rates consume more power and, results in lower memory performance [10]. Also, traditional power management techniques are often not tailored for memory. Therefore, placing memory directly on top of (or nearby) the processing elements (PEs), without addressing thermal issues, can be detrimental to the achievable performance. To avoid this problem, Non-Volatile Memories like Resistive Random-Access Memory (ReRAM) can be used as an alternative to DRAM for the PIM architecture. ReRAMs store bits as conductance levels and do not need to be refreshed [11]. However, ReRAMs are affected by thermal noise which can cause outputs to be misinterpreted at higher temperatures [11]. Moreover, higher temperatures are not desirable in general as it can affect the performance and lifetime of the device.

In [12], the authors found that 2.5D PIM architectures are prone to lateral heat flow from PEs even when placed 10mm farther from the memory. Placing memory farther away to reduce temperature, also defeats the main purpose of PIM, which is to bring computation closer to memory. 3D PIM architectures where PEs are in the same vertical stack as memory, are even more sensitive. Therefore, conventional PIM architectures (both 2.5D and 3D) typically use either (a) PEs with simpler architectures (as complex architectures e.g. Out-of-Order (OoO) CPUs tend to consume more power [10]), (b) fewer number of PEs, or (c) minimal computing power [12], (or all of the above) to remain within the temperature threshold. Due to these restrictions, conventional PIM architectures have lower computation capability that affects performance and are not scalable with increasing system size.

Moreover, traditional PIM architectures are restricted to single logic layer and multiple memory layers, as logic (PEs) dissipates more heat than memory [10]. It is well known that 2D logic provides limited floor-planning choices and require more die area than an equivalent 3D counterpart. However, multiple logic layers stacked vertically in 3D ICs are prone to higher temperatures as PEs farther away from the sink cannot dissipate heat easily, resulting in worse temperature [1]. As PEs consume more power than memory, use of multiple layers of logic in PIMs is typically avoided. As a result, only a few PEs can be integrated given a fixed area constraint. Overall, our objective for a "suitable PIM architecture" should be one that: (a) allows larger amount of logic to be integrated without incurring extra area and thermal overheads; and (b) enable efficient data exchange between PEs and memory.

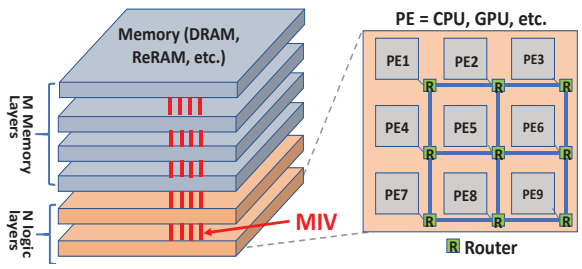


Fig. 3: Envisioned PIM architecture with M memory and N logic layers (where M, N >1) enabled via M3D integration

Fig. 3 shows the envisioned PIM architecture with multiple logic layers (similar to 3D ICs [1]) and multiple memory layers. Each logic layer consists of multiple PEs, while the memory layers can be based on DRAM, ReRAM or any other memory technology. For the sake of discussion, we consider DRAM as an example here. The use of multiple logic layers enables large number of PEs to be integrated compared to traditional PIM (single logic layer) under an "iso-area" setting. As discussed earlier, cconventional TSV-based 3D architectures are susceptible to higher temperatures and hence cannot be used to design the envisioned architecture (Fig. 3) [3]. On the other hand, emerging M3D integration allows faster dissipation of heat than its TSV-based counterparts [9]. Hence, we argue that we should design high-performance yet thermally viable PIM architectures with multiple logic (and memory) layers as shown in Fig. 3 using M3D. Similar architectures with multiple logic and memory layers in M3D have also been proposed in [13] and [14].

To evaluate the characteristics of the 3D-PIM architecture, a memory intensive bioinformatics application: k-mer counting was chosen. The task of *k-mer* counting involves creating a histogram of all k-length substrings in a DNA sequence [9]. Fig. 4(a) shows the variation of maximum on-chip temperature as more logic layers are added. We fix the number of memory layers while varying the number of logic layers beneath it. Here, we assume a DRAM-based memory as an example. However, similar observations are made for other memory technologies like ReRAM [11]. For all experiments, an ambient temperature of 45°C and an inexpensive cooling (convection resistance = 2°C/W [10]) is used. Fig. 4(a) indicates that even with a simple cooling solution, up to four layers of logic can be easily integrated in M3D-based 3D-PIM without reaching the temperature

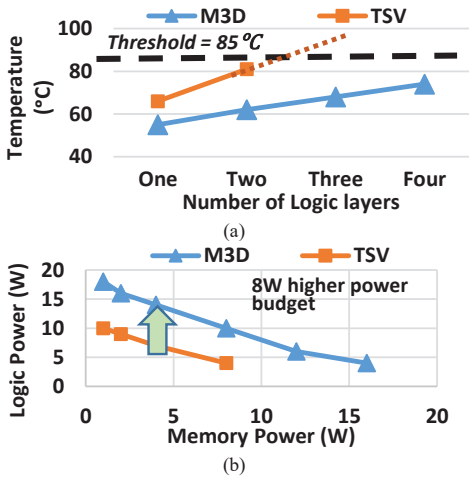


Fig 4: (a) Max. on-chip temperatures with varying number of logic layers, and (b) Power budget study, in TSV and M3D-based PIM architectures [9]

threshold of 85°C. On the other hand, TSV-based PIM only allows a maximum of two logic layers. Beyond two layers, TSV-based PIM architectures necessitate higher refresh rates and more expensive cooling solutions to be viable. Note that adding the second layer of logic results in temperature close to the threshold (81°C) which may still necessitate precautions for safe operation. However, even with four logic layers, M3D-based PIM architecture exhibits maximum temperature of 74°C only. Therefore, contrary to conventional PIM architectures, it is possible to have multiple logic layers in an M3D-enabled PIM without violating thermal constraints.

Also, since PE and memory power depend on several factors e.g., voltage-frequency settings, technology node etc., it is important to study the power budget available for both logic and memory (without exceeding 85°C) for a complete analysis. Fig. 4(b) shows the amount of power budget available in both PIM architectures when logic and memory power are varied simultaneously to study the maximum on-chip temperature. From Fig. 4(b) we note that M3D-based PIM provides a much higher power budget (up to 8W more) than their TSV-based counterpart under similar settings. The higher power budget is achieved as M3D-based architectures do not have layers with poor thermal conductivity and have relatively smaller dimensions (discussed in Sec. 4) which aide in quick dissipation of heat. As a result, the temperature increase is significantly contained allowing more power budget (and multiple layers of logic) in an M3D-enabled PIM without exceeding 85°C.

The better heat dissipation profile in M3D is also useful for ReRAM-based PIM architectures as shown in [11]. In [11], the authors presented a 3D PIM architecture (similar to Fig. 3) with multiple GPU and ReRAM layers for training Convolutional Neural Network (CNNs). Here, ReRAMs act as both the memory and compute layer due to their ability to support in-situ Multiply-and-Accumulate (MAC) operations. However, ReAMs are sensitive to temperature, particularly thermal noise. As shown in [11], thermal noise can affect the accuracy of the trained CNN models, particularly at higher frequencies. The use of M3D not only allows multiple layers of ReRAMs and GPUs to be stacked, but also reduces temperature. An M3D PIM, complemented with a thermal-aware optimization can make the architecture agnostic to thermal noise even at extremely high frequencies.

B. Manycore design using M3D

In this sub-section, we first discuss how M3D can benefit different core (CPU, GPU) and uncore (like Network-on-chip (NoC) routers and cache) components. Here, the term "core" represents the processing elements (a.k.a. PEs), and generally refers to CPU and GPU cores.

1) Core Design using M3D: M3D integration has been utilized to design high-performance yet energy-efficient 3D CPU cores [15]. The M3D CPU design in [15] is based on a typical pipelined, x86 architecture. By considering different stages of the CPU execution pipeline, we can identify critical paths and explore different design strategies to improve them. The various stages of the pipeline are then vertically partitioned across two tiers for best performance. Overall, compared to a conventional planar CPU, the M3D CPU improves critical path delay by 14%, which results in an average performance improvement of 14% and 26% in a single- and multi-core (4-cores) setting, respectively.

Similar methodology can be adopted to design an M3D GPU core. To the best of our knowledge, [16]is the only work

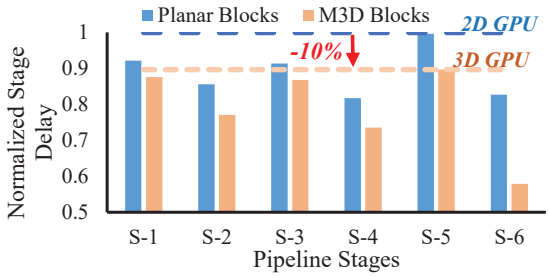


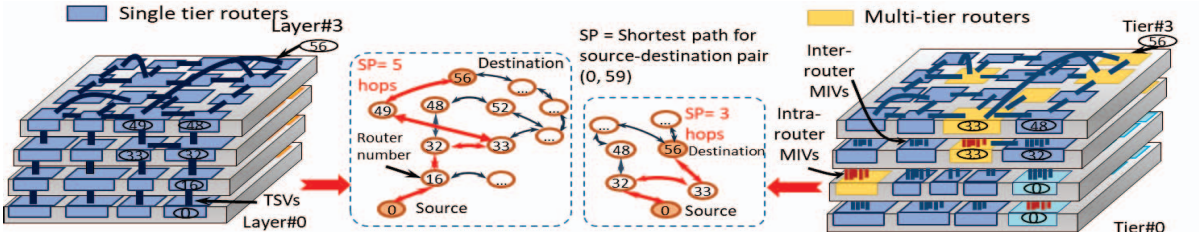
Fig. 5. End-to-end latency of each stage of the GPU execution pipeline for a planar and an M3D-enabled GPU core [16]

that discusses GPU design using M3D. A GPU core is analogous to a streaming multiprocessor (SM) in Nvidia terminology or a Compute Unit (CU) in AMD architecture. For the purpose of demonstration, we use the open-source MIAOW GPU [17] for designing the M3D GPU. The GPU core is made up of several blocks that operate in a pipelined fashion. Fig. 5 shows the timing characteristics of each pipeline stage normalized with respect to the clock period of the planar GPU core. Here, "S-i" denotes the "i^{th}" stage in the GPU execution pipeline.

In a pipelined architecture, the overall delay is bottlenecked by the slowest stage. From Fig. 5, we can see that the pipeline stage delay (hence, the operating frequency) is limited by the S-5 stage (that includes scalar ALU, vector ALU (Single Instruction Multiple Data aka. SIMD and Single Instruction Multiple Floating-Point aka. SIMF), and load-store units) for the planar GPU core. As shown in Fig. 5, M3D improves the timing characteristics of all the pipeline stags in the 3D GPU by 8-14%. However, as mentioned earlier, the slowest stage determines the clock frequency of the pipelined GPU. From Fig. 5 we note that the S-5 stage has the highest delay in the M3D GPU design. However, compared to the planar design, the M3D S-5 stage has 10% lower delay. Hence, *we can operate the M3D GPU at 10% higher frequency compared to its planar counterpart* (baseline GPU core) without violating any timing constraints, leading to better performance. In addition, the authors in [16] reported 21% lower energy consumption in the M3D GPU compared to its planar counterpart due to the use of MIVs and a smaller number of buffers in the overall design.

2) Uncore design using M3D: In addition to the cores, M3D can also benefit the uncore components like cache and NoC. On-chip cache (like L1 and L2) are repeatedly accessed by the cores during execution. A slow cache response can lead to delay in execution and memory access, creating performance bottlenecks. CPU performance can be particularly affected as they are latency sensitivity. Hence, faster caches are desirable, especially in manycore architectures that include CPUs. The dense M3D integration allows high-performance cache designs as shown in [18]. By investigating different types of partitioning for caches, such as bank stacking, bit line partitioning and word line partitioning in two tiers, the cache architecture proposed in [18] achieves up to 23.3% reduction in access latency. For a single core system, the faster M3D cache is able to improve the overall performance by 9.9%.

Similarly, M3D can be used to design better NoCs. The NoC paradigm has emerged as a revolutionary methodology for integrating many embedded cores in a single die. Prior work has shown the effectiveness of 3D NoCs compared to existing planar counterparts for several applications including



(a) (b)

Fig. 6. Illustration of (a) TSV- and (b) an M3D-based 3D Small-World NoC; both showing the shortest paths from router-0 (source) to router-56 (destination). M3D reduces the hop-count significantly compared to TSVs, leading to faster communication.

Neural Networks, Graph-analytics, etc. [1]. Traditional TSV-based 3D NoCs consist of planar routers, which are connected using planar wires for horizontal connectivity. The vertical connectivity is established using TSVs. By connecting cores in 3D, the physical and logical separation between cores is reduced, leading to low-latency, high-throughput, and energy-efficient communication. The vertical links act as long-range shortcuts and increase path-diversity between communicating cores. This leads to lower NoC congestion and better performance. On the other hand, M3D integration enable the design of routers across multiple tiers which leads to even better performance.

Fig. 6(a) and Fig. 6(b) show a conceptual view of TSV and M3D-based 3D Small World NoCs (SW-NoCs) respectively. By adding a handful of application specific long-range shortcuts, SW-NoCs can improve the communication latency significantly [19]. As seen from Fig. 6(b), in the M3D SW-NoC, in addition to the single-layer routers (blue), some routers are extended over multiple layers (yellow). In contrast, for a TSV-based design (Fig. 6(a)) all the routers are designed only over a single-layer (blue). To explain how this helps in improving the NoC performance, we consider the inter-node communications for the source-destination pair (0, 56) as an example. The associated SW-graph connectivity is shown in the middle part of Fig. 6. The path highlighted with red indicates the shortest available path for this node pair. The shortest-path SP lengths for M3D- and TSV-based SW-NoCs are calculated to be 3 and 5 hops respectively, which shows significant reduction in hop count for the M3D configuration. Additionally, we get improvements in energy as MIVs are more energy-efficient compared to TSVs. Overall, M3D NoCs outperform their TSV-based counterparts in terms of both performance and energy efficiency.

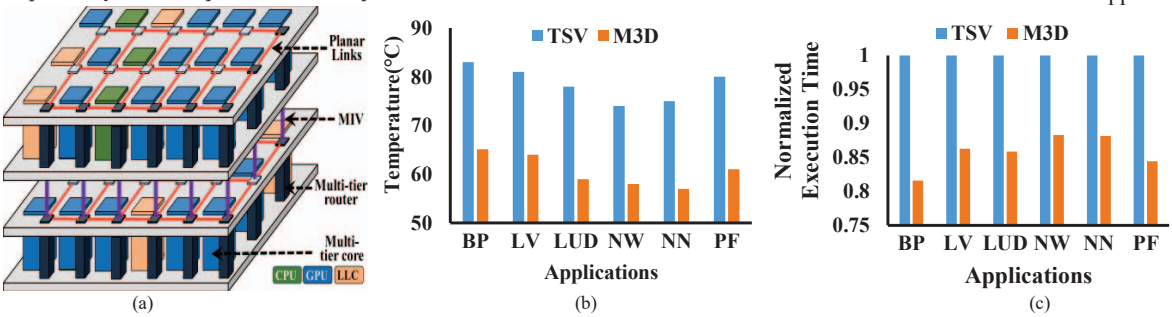
3) Heterogeneous Manycore Design using M3D: Fig. 7(a) shows an example heterogeneous architecture designed using M3D core and uncore components. Fig. 7(b) and Fig. 7(c) show the thermal (maximum on-chip temperatures) and performance characteristics of an M3D-based architecture respectively, and compare it with an equivalent TSV baseline. As shown in Fig. 7(b), the M3D-based architecture has 18°C lower temperature than the TSV baseline on an average for six different Rodinia benchmarks [21]: Backprop (BP), Needle (ND), Lava (LV), Lud (LUD), KNN, and Pathfinder (PF). The lower temperature of M3D compared to TSV is due to lower ILD thickness and better thermal conductivity (shown in Fig 2). Since the ILD layers are extremely thin, virtually all the cores are near the sink. Hence, the M3D based architecture achieves lower temperature than their TSV counterpart, without the use of any specialized cooling techniques. This lower temperature will make the M3D-based architecture more sustainable than its TSV-based counterpart.

Fig. 7(c) demonstrates the normalized execution time for the two different architectures. From Fig. 7(c), we note that the M3D-based architecture outperforms its TSV-based counterpart by 14.2% on average. This happens as: (a) M3D cores operate at higher frequencies than planar cores (M3D CPU and GPU operate at 14% and 10% higher frequencies than their corresponding baseline planar counterparts; (b) the M3D-enabled cache provide 23.3% faster cache accesses; and (c) the M3D routers and lower physical distance between adjacent cores enables high performance NoC designs. Overall, a M3D-based manycore architecture can achieve better performance and lower temperature simultaneously.

IV. Future Research Directions

Despite the advantages, M3D-based systems still have various unsolved challenges. In this section, we outline some of these open challenges in M3D-based hardware design that need to be addressed.

Process-Variation: During M3D fabrication, each tier is fabricated sequentially, from the bottom to the top. One of the major challenges here is to create the top-tier transistors without impacting the already-processed bottom-tier transistors and interconnects. This requires a low temperature process for the top tier, to prevent any deterioration of the bottom transistors. However, this leads to slower upper-tier transistors. Additionally, tungsten is required for the back-end-of-line interconnect in the bottom tier since copper only



(a) (b) (c)

Fig. 7. (a) Illustration of an M3D-enabled heterogeneous architecture, (b) Maximum on-chip temperature, and (c) Full system execution time (normalized), comparison of the M3D-based heterogeneous architecture and its TSV equivalent. [16]

supports temperatures up to the 400ºC range. This increases the resistivity of the interconnect in the bottom tier. The performance degradation of the top-tier transistors and increased delay in bottom-tier interconnects affect the achievable performance. In [20], the authors show that performance of a manycore system is overestimated by 50.8% if process-variation is not considered during design process. A process-variation aware design can address this problem by suitably partitioning logic blocks/gates across different tiers. Hence, new hardware design techniques for M3D should consider process-variation for best achievable performance.

Electrostatic coupling: Electrostatic coupling occurs when the separation between two neighboring circuits is reduced to the order of tens of nanometer as in M3D. In this situation, the threshold- and switching-voltages of any victim transistor can be affected by an aggressor. This, in turn, leads to variation in the transistor current and overall delay of the corresponding critical paths. For example, the threshold voltage, V_{th}, of a PMOS transistor increases by 130 mV when the gate voltage of a neighboring NMOS, located 10nm from the PMOS, is switched from 0 to 1 V in 45nm FDSOI technology [22]. This coupling phenomenon can significantly alter the delay of the victim circuit. A guard-band-based design can compensate such timing penalties to some extent. However, if the delay is more than the guard-band (in general 10% positive slack is employed), the design can violate the timing specification of the system, leading to significant performance degradation. In [22], the authors demonstrate that electrostatic coupling can degrade the performance of an M3D NoC by 18.1% on average. Hence, new electrostatic-coupling aware design methodologies need to be developed.

Optimization: The unique attributes of M3D like multi-tier logic design and process variation also makes the optimization of new architectures challenging. The number of design choices to consider during the optimization process are significantly higher [16]. Existing optimization algorithms like Simulated Annealing (SA) do not scale with the size of design space and require long time to find a good solution. Machine Learning (ML) based optimization algorithms present a promising direction in this regard [16][23]. By learning the search space, ML-based optimization algorithms can significantly reduce the time to find a good solution. For instance, in [16], the authors use an ML-based algorithm: MOO-STAGE that is able to find a good M3D architecture 7.38 times faster than SA-based methods on average. This will reduce design time of new M3D architectures significantly.

V. Conclusion

Advanced computing systems have enabled breakthroughs in science, engineering, and have continued to play key roles in today's Big-Data era. However, with the slowing down of Moore's law, new architectural innovations are necessary to meet the relentless needs of Big-Data applications (e.g., deep learning, graph analytics, autonomous driving, personalized medicine etc.). 3D integration enables the design of high-performance and energy-efficient architectures. However, conventional TSV-based 3D integration has several limitations, which can limit the achievable performance and energy-efficiency. On the other hand, M3D enables the design of true 3D circuits and systems that mitigates many of the issues common in TSV-based counterparts. In this paper, we have highlighted some of the challenges associated with TSV-based 3D systems. We discuss that M3D can alleviate some of these problems leading to better performance-thermal-area trade-offs. Hence, M3D should be the preferred technology to design the next generation of high-performance and energy-efficient circuits and systems.

References

[1] B. K. Joardar et. al., "Learning-Based Application-Agnostic 3D NoC Design for Heterogeneous Manycore Systems," in IEEE TC 68, pp 852-866, 2019.

[2] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna and S. K. Lim, "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology," *in IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)* 1-2, 2016

[3] S. K. Samal et. al., "Fast and accurate thermal modeling and optimization for monolithic 3D ICs," in *DAC,* pp 1-6, 2014

[4] S. Panth, K. Samadi, Y. Du and S. K. Lim, "High-density integration of functional modules using monolithic 3D-IC technology," in A*SP-DAC,* pp 681-686, 2013

[5] Y. Lee and S. K. Lim, "Ultrahigh Density Logic Designs Using Monolithic 3-D Integration," *IEEE TCAD,* 32, pp 1892-1905, 2013

[6] J. Shi et al., "A 14nm FinFET transistor-level 3D partitioning design to enable high-performance and low-cost monolithic 3D IC," in *IEDM,* 2.5.1-2.5.4, 2016

[7] C. Liu and S. K. Lim, "A design tradeoff study with monolithic 3D integration," in *ISQED,* 529-536, 2012

[8] S. Panth, K. Samadi, Y. Du and S. K. Lim, "Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations," in *DAC, pp* 1-6, 2014

[9] B. K. Joardar et. al., "NoC-enabled software/hardware co-design framework for accelerating k-mer counting," in NOCS, Article 4, pp. 1–8, 2019

[10] Y. Eckert, N. Jayasena, and G.H. Loh, 2014. Thermal Feasibility of Die-Stacked Processing in Memory. *Workshop on Near-Data Processing*

[11] B. K. Joardar, J. R. Doppa, P. P. Pande, H. Li and K. Chakrabarty, "AccuReD: High Accuracy Training of CNNs on ReRAM/GPU Heterogeneous 3D Architecture," in IEEE *TCAD*, 2020

[12] Y. Zhu, B. Wang, D. Li, and J. Zhao, "Integrated Thermal Analysis for Processing in Die-Stacking Memory," in *MEMSYS*. Alexandria, pp 402-414, 2016

[13] M. Valad Beigi and G. Memik, "THOR: THermal-aware Optimizations for extending ReRAM Lifetime," in IPDPS, 2018, pp. 670-679

[14] Mohamed M. Sabry Aly., "N3XT Monolithic 3D Energy-Efficient Computing Systems," In GLSVLSI '19, Association for Computing Machinery, New York, NY, USA, 463, 2019

[15] B. Gopireddy and J. Torrellas, "Designing vertical processors in monolithic 3D," In ISCA, *New York, NY, USA, 643–656,* 2019

[16] A. I. Arka et. al., "HeM3D: Heterogeneous Manycore Architecture Based on Monolithic 3D Vertical Integration," in arXiv preprint, arXiv:2012.00102v1, 2020

[17] R. Balasubramanian et al., "MIAOW - An open source RTL implementation of a GPGPU," in *IEEE COOL CHIPS XVIII,* 1-3, 2015

[18] Y. Gong, J. Kong and S. W. Chung, "Quantifying the Impact of Monolithic 3D (M3D) Integration on L1 Caches," in *IEEE TETC, 2019*

[19] D. Lee et. al., "Performance and Thermal Tradeoffs for Energy-Efficient Monolithic 3D Network-on-Chip," in ACM TODAES. 23, 5, Article 60, 2018

[20] S. Musavvir et al., "Inter-Tier Process Variation-Aware Monolithic 3D NoC Architectures," ArXiv, 2019. arXiv:1906.04293v1.

[21] S. Che et al., "Rodinia: A benchmark suite for heterogeneous computing.," in *IISWC,* 44-54, 2009

[22] D. Lee, S. Das, J. R. Doppa, P. P. Pande, and K. Chakrabarty, "Impact of Electrostatic Coupling on Monolithic 3D-enabled Network on Chip," in ACM TODAES, 24, 6, Article 62, 2019

[23] A. Deshwal, N. K. Jayakodi, B. K. Joardar, J. R. Doppa, and P. P. Pande, "MOOS: A Multi-Objective Design Space Exploration and Optimization Framework for NoC Enabled Manycore Systems," in ACM TECS. 18, 5s, Article 77 (October 2019), 23 pages