Power, Performance, and Thermal Trade-offs in M3D-enabled Manycore Chips

Shouvik Musavvir*, Anwesha Chatterjee*, Ryan Gary Kim†, Dae Hyun Kim*, Janardhan Rao Doppa*, Partha Pratim Pande*
*School of EECS, Washington State University
Pullman, WA, 99164, U.S.A.
{shouvik.musavvir,anwesha.chatterjee,daehyun.kim,
jana.doppa,pande}@wsu.edu
†Department of Electrical and Computer Engineering,
Colorado State University
Fort Collins, CO, 80524, USA
Ryan.G.Kim@colostate.edu

***Abstract*— Monolithic 3D (M3D) technology enables unprecedented degrees of integration on a single chip. The miniscule monolithic inter-tier vias (MIVs) in M3D are the key behind higher transistor density and more flexibility in designing circuits compared to conventional through silicon via (TSV)-based architectures. This results in significant performance and energy-efficiency improvements in M3D-based systems. Moreover, the thin inter-layer dielectric (ILD) used in M3D provides better thermal conductivity compared to TSV-based solutions and eliminates the possibility of thermal hotspots. However, the fabrication of M3D circuits still suffers from several non-ideal effects. The thin ILD layer may cause electrostatic coupling between tiers. Furthermore, the low-temperature annealing degrades the top-tier transistors and bottom-tier interconnects. An NoC-based manycore design needs to consider all these M3D-process related non-idealities. In this paper, we discuss various design challenges for an M3D-enabled manycore chip. We present the power-performance-thermal trade-offs associated with these emerging manycore architectures.**

Keywords*— Monolithic 3D, Manycore system, NoC, Electrostatic Coupling, Process variation, Thermal hotspots.

I. Introduction

The emergence of three-dimensional (3D) integration has revolutionized the design of high-performance and energy-efficient manycore chips. Moreover, recent industry trends show the viability of 3D integration in commercial products (*e.g.*, AMD's Radeon R9 Fury X graphics card and Xilinx's Virtex-7 2000T/H580T and Ultra-scale FPGAs). However, the achievable performance of conventional through-silicon-via (TSV)-based 3D manycore chips is ultimately bottlenecked by the planar interconnects (wires in each planar die).

Monolithic 3D (M3D) integration, a breakthrough technology to achieve "More Moore and More Than Moore," opens up the possibility of designing cores and their associated network routers and links using multiple tiers. Compared to TSV-based 3D ICs, M3D offers the "true" benefits of 3D circuits for system integration: the size of a monolithic inter-tier via (MIV) used in M3D is over 100x smaller than a TSV [1]. This dramatic reduction in via size and the resulting increase in density opens up numerous opportunities for design optimizations in 3D manycore chips: designers can use millions of MIVs for ultra-fine-grained 3D optimization, where individual cores and routers can be spread across multiple tiers for extreme power and performance optimization [2]. *Existing TSV-based 3D interconnects are not adequate as the interconnection fabric for communication-bandwidth-hungry manycore processors. These architectures are simple extensions of regular 2D architectures and they cannot exploit the advantages provided by dense M3D integration.*

In addition to the ability to create true 3D circuits, another advantage of M3D integration is that in an M3D-enabled architecture, the inter-layer dielectric (ILD) used between planar tiers is very thin. This, combined with the lack of a bonding layer (which has poor thermal conductivity and is typically used in TSV-based systems) between adjacent tiers, facilitates better cooling across the chip [3]. In contrast, in a TSV-based design, the thick silicon substrate and the presence of the bonding layer obstruct the heat flow from the source of power dissipation to the heat sink, aggravating temperature issues in an already power-dense 3D design [3]. These two qualities of M3D systems: the ability to create true 3D cores and routers, and a compact third dimension due to the presence of thin ILD structure between the tiers; make M3D technology a desirable candidate to provide high-performance and thermally viable 3D manycore chips.

Although M3D designs offer significant performance improvements over TSV-based solutions, its fabrication process is still not mature. The proximity of the tiers poses several fabrication challenges. First, if the ILD is very thin (<100nm), components (transistors and interconnects) in adjacent tiers can be close enough to experience electrostatic coupling [4]. This phenomenon changes the timing characteristics and reduces the signal integrity of the M3D-based circuits and systems. Secondly, the high annealing temperature of upper-tier fabrication can damage the lower tier [5]. To solve this problem, researchers have adopted low-temperature annealing [6] that leads to the inferior performance of top-tier transistors. Moreover, tungsten interconnects are also used in the bottom tier, resulting in lower conductance at the bottom tier compared to traditional copper interconnects [5]. Hence, the M3D architecture suffers from the inter-tier process variation due to top-tier transistor and bottom-tier interconnect degradations. Both of these non-ideal effects (electrostatic coupling and inter-tier process variation) can become so severe that they can offset the performance benefits of M3D integration in a manycore design. Hence, we need to take these effects into account while designing an M3D-enabled manycore chip. Otherwise, we will significantly overestimate the timing and energy gains enabled by M3D.

On the other hand, power management is inherent in any manycore chip to achieve suitable trade-offs between energy-efficiency and performance. This is no exception in an M3D-based manycore design [7]. The heart of any power management system is assigning voltage/frequency (V/F) levels to cores and

This work was supported, in part by the US National Science Foundation (NSF) grants CNS-1564014, CCF 1514269 and USA Army Research Office grant W911NF-17-1-0485.

uncore (network routers, memory controllers, etc.) elements in a manycore chip without sacrificing significant performance. V/F scaling in the presence of the non-ideal effects of M3D integration incurs additional performance penalties and energy degradation. Hence, the non-ideal effects of M3D need to be considered while designing the power management system.

In this paper, we first discuss manycore system design challenges in M3D and efficient optimization methods to resolve them. Next, we discuss the benefits of M3D networks-on-chip (NoCs) as the communication backbone of manycore systems. Subsequently, we elaborate on the effects of the fabrication and process-variation-related pitfalls on M3D NoC performance. We explore how to incorporate the effects of both electrostatic coupling and process variation in the M3D NoC design optimization flow. Finally, we focus on the role of power management for M3D-enabled manycore systems and its implication on the temperature.

II. Manycore System Design Challenges Using M3D

Although M3D provides us the flexibility and circuit benefits of having cores and routers spread over multiple tiers, this additional freedom tremendously increases the size of the design space. In this 3D system, the physical coordinates associated with the cores, routers, horizontal links, and vertical links significantly increase the size of the overall design space. Hence, we need to develop and apply efficient and scalable design optimization algorithms for M3D systems.

Researchers have been using traditional optimization methods like simulated annealing (SA), genetic algorithm (GA), etc. for decades to handle these combinatorial optimization problems with non-linear constraints. These algorithms typically use a local search procedure with random restarts to traverse the design space [8]. However, they waste a lot of time restarting the search in the hope of finding good locations within the design space, especially when they contain many local optima. Hence, the runtime of these algorithms increases rapidly for bigger system sizes. Moreover, both SA and GA are sensitive to the initial state [8].

Machine learning (ML) based methods can overcome these shortcomings in traditional algorithms to quickly uncover near-optimal solutions for ever-increasing design space. For example, ML based algorithm STAGE [9] is applied for minimizing the latency of M3D NoCs [8]. STAGE is highly scalable and uses the knowledge of past searches to explore the design space efficiently. STAGE iterates over two steps. In the first step, it applies hill climbing to optimize the primary design objective. STAGE uses the search trajectory obtained in the first step to learn an evaluation function. The evaluation function attempts to predict the best achievable value of the hill climbing search for any particular starting point. In the second step, the algorithm performs another local search to optimize the evaluation function and generates a suitable starting point for the first step. As time progresses, the evaluation function becomes more accurate and the algorithm can avoid searching through non-promising design points and hence, reduces the overall runtime.

Past work in the search community concluded that many practical optimization problems exhibit a "globally convex" or "big valley" structure [9], where the set of local optima appear to be convex with one global optimum in the center. The main advantage of STAGE over popular algorithms such as SA and integer linear programming (ILP) is that it tries to learn the solution space structure and uses this information in a clever way to improve the convergence time [10]. This aspect of STAGE is very advantageous for large system sizes to improve the design-validate cycle before mass manufacturing and for dynamically adapting the designs to new application workloads. Fig. 1 shows the runtime comparison of STAGE and SA for M3D NoC [8]. As the system size increases, SA runtime increases dramatically compared to the runtime of STAGE. Fig. 2 shows the difference in runtime improvement (STAGE vs. SA) for an M3D-enabled NoC design optimization over the same runtime improvement for a TSV-enabled NoC design optimization considering different SPLASH-2 and PARSEC benchmarks. On an average, the runtime benefit is 30% more for M3D-based designs compared to their TSV-based counterpart [8]. This difference in runtime improvements is due to the much larger design space of M3D systems, allowing the M3D systems to benefit more from the ML-based optimization.

III. M3D NoC Architectures

In manycore systems, NoC is the de-facto communication backbone. 3D NoC architectures combine the benefits of two paradigms (3D IC and NoC) to offer an unprecedented performance gain even beyond the Moore's law regime. Existing 3D NoC architectures mainly follow a simple extension of regular 2D NoCs. However, this approach does not fully exploit the advantages provided by M3D integration. In this context, small-world network-based NoC (SWNoC) architectures [11] are a notable example. It has been shown that either by inserting long-range shortcuts in a regular mesh to induce small-world effects or by adopting power-law based small-world connectivity, we can achieve significant performance gains and lower energy dissipation compared to traditional multi-hop mesh NoCs [11]. We advocate that this concept of small-worldness should be adopted in M3D NoCs. The vertical links in M3D NoCs should enable the design of

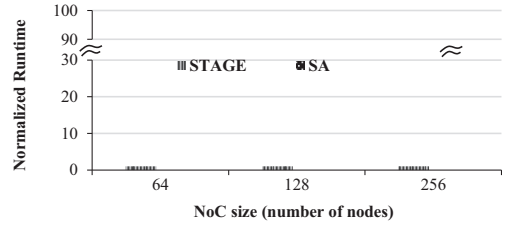


Fig. 1. Normalized runtime of STAGE and SA algorithm for different M3D NoC system sizes with respect to the STAGE runtime. [8]

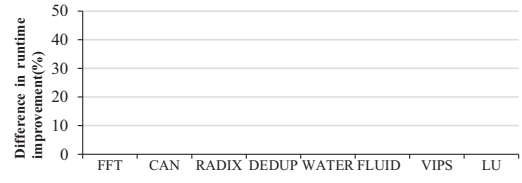


Fig. 2. Difference in runtime improvement (STAGE vs. SA) between M3D- and TSV-enabled NoC design optimization [8].

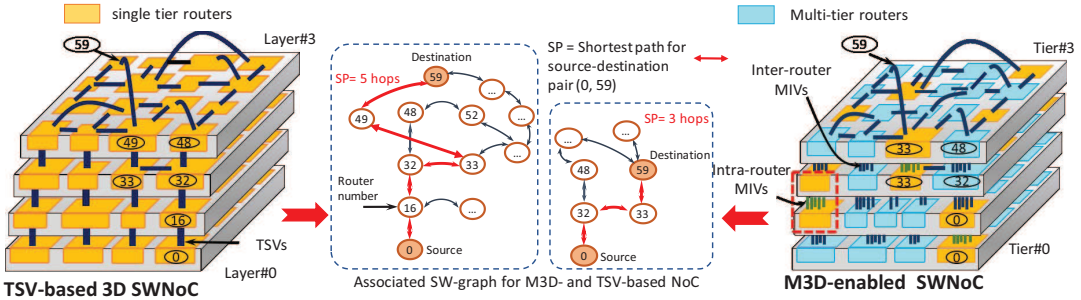


Fig. 3. Illustration of four-tier 64-node TSV- and M3D-enabled SWNoCs. In both cases, we show the shortest path to transfer data from router 0 to router 59. In M3D SWNoC, the shortest path (0->32->33->59) needs only three hops to reach router 59. However, in TSV based SWNoC, the shortest path (0->16->32->33->49->59) contains five hops [8].

long-range shortcuts necessary for small-world networks. By exploiting the MIV-based vertical connections in M3D, the multi-hop long-range planar links can be placed along the shorter z-dimension, and hence, overall system performance can be significantly improved. MIV-based vertical links are faster and consume less energy compared to TSV-based counterparts [8]. Moreover, we can partition routers in multiple tiers using MIVs. Multitier routers can transfer data vertically through intra-router MIVs and communicate with routers in other tiers. Hence, the hop count reduces in M3D NoCs compared to their TSV-based counterpart. All these advantages lead to higher performance and energy-efficiency in M3D NoCs.

In Fig. 3, we show how the multitier router can eliminate vertical hops. In a TSV-based design, all the routers are planar. However, in the M3D SWNoC, in addition to the single-tier routers, some routers are extended over multiple tiers. To see how this makes a difference in the data exchange, we consider the communication from node 0 to node 59 as an example. The associated connectivity is shown in the middle part of Fig. 3. The path highlighted with red indicates the shortest available path between these routers. Here, we can see that the shortest path for the M3D SWNoC contains three hops. By using two multitier routers, 0 and 33, we can reduce the amount of inter-router vertical communication. On the other hand, the TSV-based design already requires three hops to traverse the vertical dimension alone. Overall, the TSV-based design requires more hops to communicate between nodes 0 and 59. The reduced hop count and energy efficiency in vertical links lead to 28% and 30% savings in energy and EDP, respectively [8]. However, all the M3D NoC-related works so far principally considered ideal M3D process characteristics and have not considered the effects of transistor and interconnect degradation. In the next section, we illustrate how the NoC design methodology should incorporate the effects of process variation in the overall design flow.

IV. Reliable and Robust M3D NoC Design Challenges

As we mentioned earlier, the M3D NoC design methodology should consider various M3D process- and fabrication-related challenges. Here, we discuss their effects and show how to incorporate them in the M3D NoC design flow to minimize the performance penalties.

A. Electrostatic Coupling

In M3D designs, the thin ILD can cause coupling between circuit components in adjacent tiers [4]. The coupling effect from the bottom-tier interconnects to the top-tier transistors can be avoided by limiting their usage [12]. However, the electrostatic coupling between transistors in adjacent tiers is inevitable. When the gate-source voltage of a transistor changes (ΔV_{gs}), the threshold voltage of the affected transistor in the other tier changes as follows [13]:

$$\Delta V_{th} \approx \frac{C_{ILD}}{C_{ox}} \cdot \Delta V_{gs} \quad (1)$$

where C_{ox} and C_{ILD} are the capacitance of the gate-oxide of the transistor and ILD, respectively. Since the capacitance of the ILD is inversely proportional to its thickness, ΔV_{th} increases when ILD thickness (T_{ILD}) is decreased. This causes the transistor switching delay to increase since it is inversely proportional to the square of overdrive voltage (V_{gs}-V_{th}) [13]. In Fig. 4, we show the cross-section of a two-tier M3D circuit with transistor-level partitioning. Here, if the voltage changes in the top-tier NMOS, the switching delay of the PMOS increases and vice-versa.

Electrostatic coupling induces significant delay and energy overheads for multi-tier NoC routers. This, in turn, results in considerable performance degradation if the NoC design methodology does not incorporate the effects of electrostatic coupling. Hence, it is advisable that the NoC design and optimization methodology consider the effects of electrostatic coupling. It has been shown that an electrostatic coupling-aware design approach reduces the number of allowable multi-tier routers [12] to achieve a proper balance between the

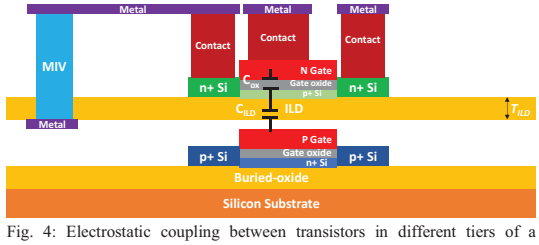


Fig. 4: Electrostatic coupling between transistors in different tiers of a monolithic 3D IC [12].

performance gains due to multitier routers and the negative effects of electrostatic coupling.

B. High Temperature Annealing Effects on Transistors and Inteconnects

In M3D fabrication, the tiers are placed within tens of nanometers of each other. Such close proximity poses a big challenge for the high temperature annealing of the top-tier transistors. In a traditional 2D fabrication process, the annealing is performed at 1050°C [5]. Such high temperatures will damage the copper interconnects and transistors in the bottom tier. Researchers have managed to bring down the temperature of annealing to 500-600°C [6] [14]. However, at these temperatures, we still cannot use copper interconnects in the bottom tier since copper can only endure temperatures up to 400 °C. Hence, tungsten has been proposed as a suitable bottom-tier interconnect material as it can withstand such high temperatures. However, tungsten interconnects exhibit lower conductivity compared to copper. This leads to the performance degradation in bottom-tier interconnects.

From an M3D NoC perspective, the bottom-tier inter-router links will experience increases in delay and energy consumption. Moreover, the top-tier transistor performance degrades due to the low-temperature annealing process [15] that results in slower intra-router logic gates in the top tier. Both transistor degradation in the top tier and interconnect degradation in the bottom tier will incur a performance penalty in the M3D NoC. If we design the M3D NoC assuming nominal transistor and interconnect characteristics and ignore inter-tier process variations, we will overestimate the NoC performance and possibly make sub-optimal design decisions. Hence, we should take these effects into account while designing the M3D NoC.

Unfortunately, these design decisions are not straightforward. As discussed earlier, the M3D process affects different aspects of the circuit in different tiers (logic degradation in the top tier vs. interconnect degradation in the bottom tier). This forces us to examine the circuit characteristics (logic heavy vs. interconnect heavy in particular) during the optimization. Since each router stage (in a standard pipelined router) may have different characteristics, *e.g.*, crossbar stage of a router is mostly dominated by interconnects [16], this naturally leads us to consider tier-wise placement of the intra-router logic stages. Hence, we consider three different tier placements: top tier only (TT), bottom tier only (BT), and multitier (MT).

Any process-aware NoC design optimization should then distribute the intra-router stages and inter-router links suitably among the M3D tiers for achieving the best performance while minimizing the process variation related penalties. For example, the crossbar stages should be MT since the benefit from interconnect capacitance reduction offsets the transistor degradation effect. Intra-router stages dominated by logic can become BT or MT depending upon the trade-off between the tier partitioning benefits and transistor degradation. For example, the switch allocator (SWA) and virtual channel allocator (VCA) stages are logic dominated. For these stages, for higher levels of transistor degradation, more of these stages should use BT placement. On the other hand, the placement of the links depends on the magnitude of interconnect degradation and link length. The long-range links (links between non-adjacent routers) cause more delay and energy consumption compared to the short-range (links between adjacent routers) links. Hence, the long-range links are placed mostly at top tier to reduce the interconnect performance penalty.

The energy-delay product (EDP) savings in the process-aware design (with respect to its process-oblivious counterpart) increases with the severity of process variation. We show the EDP comparison of process-aware and process-oblivious (all components are made MT) design in Fig. 5. EDP is normalized with respect to the EDP of an ideal M3D design with no inter-tier process variation. Considering all feasible values of process variation parameters, the process-aware design outperforms the process-oblivious counterpart by 27.4% on average. Even in the presence of the worst-case M3D process variation, the process-aware M3D NoC is still better than its TSV-based counterpart, improving the EDP by 11.7% on average across all benchmarks as shown in Fig. 6. Although the natural impulse is to make the entire system multitier to take advantage of M3D, these results demonstrate that all routers should not be made multitier. Depending on the process variation parameters and router microarchitecture, various parts of the NoC routers need to be placed in different tiers.

V. Energy and Thermal Impacts in M3D Manycore Systems

Power management strategies improve the power and thermal profiles of a manycore chip without sacrificing the overall achievable performance. 3D manycore chips are no exception in this regard. However, the inherent structure of a 3D chip plays an important role. As we discussed in Section I, TSV-based design shows major thermal hotspot regions in the chip. In TSV-based 3D dies, the die thickness and the distance between the active devices in adjacent layers range from 20 μm

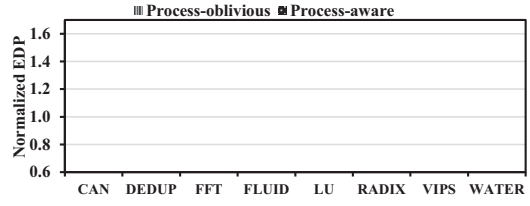


Fig. 5. EDP (averaged over all process variation parameter values) for process-oblivious and process-aware M3D NoCs considering different PARSEC and SPLASH-2 benchmarks. EDP is normalized with respect to the process-oblivious design under ideal conditions (no inter-tier process variation).

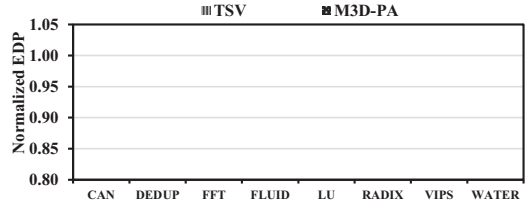
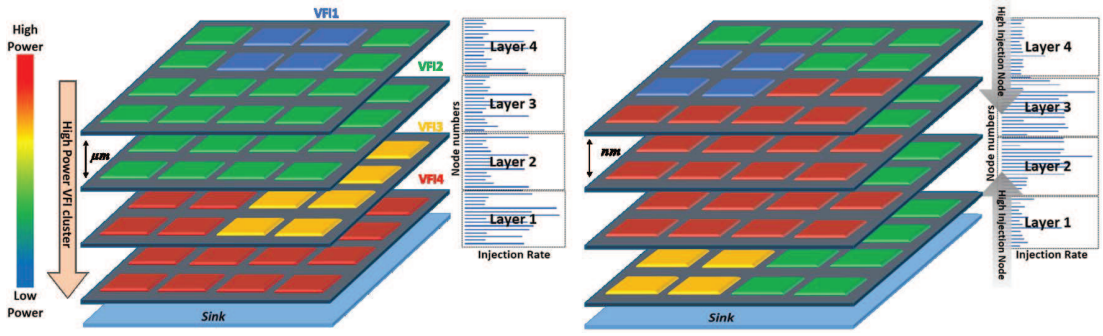


Fig. 6. EDP of TSV- and process-aware M3D-enabled NoCs considering only the worst-case process variation for different PARSEC and SPLASH-2 benchmarks. EDP is normalized with respect to the TSV-based design.



(a) (b)

Fig 7. (a) For the TSV-based design, high-power consuming cores are placed near the heat sink while highly communicating nodes are not located adjacent to each other. (b) For the M3D-based design, highly communicating cores are placed together in the middle two tiers regardless of their power consumption profile [7].

to 80 μm. Additionally, the thermal conductivity of the bonding layer (Benzocyclobutene (BCB)) in between the dies of a TSV integrated system is quite poor [3]. This hinders the heat flow in the vertical direction. On the other hand, a thin layer of SiO_2 with its higher thermal conductivity (four times more than BCB) facilitates better thermal flow for M3D designs [3]. Therefore, the M3D design leads to lower maximum temperature and fewer thermal hotspots than TSV-based designs.

Voltage-frequency island (VFI)-based power management is a well-known mechanism to lower the overall energy consumption of a manycore chip within a given performance constraint [7] [17]. Hence, it has been employed to handle the thermal hotspots of conventional 2D as well as 3D manycore chips. For a VFI-based system, a group of cores with similar communication and computation characteristics are clustered together and assigned suitable voltage/frequency (V/F) pairs. However, the cores and associated network elements within a cluster needs to be physically close. Hence, for VFI-based power management in a 3D manycore system, the placement of the VFI clusters introduces additional constraints for optimizing the thermal profile. The position of the VFI clusters affects both performance and thermal profiles of the 3D NoC.

Fig. 7 illustrates the VFI configuration of a four-layer 3D-NoC architecture with TSV and M3D integration. The network injection rates of each node, shown along to the right of each 3D system representation refer to the inter-node communication patterns. Thermal-aware optimization in TSV-based designs place the high-power cluster near the heat sink. However, this may place highly communicating cores farther apart. This is reflected in the core placement and respective traffic injection rates in Fig 7(a). Hence, energy- and thermal-efficiency is achieved at the cost of performance. The situation with M3D is different. As M3D has better thermal conductivity and much lower ILD thickness than TSVs there is no need to place high-power consuming cores near the sink. Virtually, every core may be considered to be "near the sink". Hence, we can avoid thermal hotspot without sacrificing performance. Consequently, the optimized M3D-based design in Fig 7(b) shows that the high-power cluster is not placed near the heat sink. On the other hand, the clusters with highly active and frequently communicating cores are placed in the middle layers so that they can communicate among themselves with reduced hop count. Thus, we see improved performance in M3D-based designs as these configurations need not follow strict placement constraints for thermal efficiency like the TSV-based counterparts [7].

For a 3D system, increasing the number of vertical layers leads to improved performance of the system due to smaller hop counts. However, the temperature rises due to the increased power density. Fig. 8 captures the relationship between the number of layers and EDP (Fig. 8(a)) and the maximum temperature (Fig. 8(b)) reached for different TSV- and M3D-based designs. For all systems, increasing the number of layers leads to large EDP improvements. However, the temperature of the system rises significantly due to poor thermal conductivity in TSV-based designs. This can be observed in Fig. 8(b) for the performance-only optimized TSV-based design without any VFI-based power management (NVFI). Although there is an improvement in the average EDP of the system, when the number of layers is increased up to four, the maximum temperature of the system rises up to 90°C. A VFI-enabled TSV-based manycore design achieves better EDP as well as lower temperature compared to its NVFI counterpart. However, the performance-only optimized VFI system reaches a maximum temperature of 83°C.

Thermal-aware VFI-enabled TSV-based architectures improves the thermal profile without large degradation in EDP. As we can observe in Fig. 8, the performance-thermal joint optimization TSV-based VFI reduces the maximum temperature to 70°C at the cost of 6.8% EDP degradation compared to performance only optimized 4-layer system. However, in a M3D-based VFI design, the performance profile improves significantly with increasing number of tiers with much lower maximum temperature compared to the TSV-based designs. Therefore, it is evident that the superior thermal profile of VFI-enabled M3D-based design allows scalability in the number of vertical layers in a 3D system to achieve much better performance efficiency than any TSV-based design.

VI. Conclusion

The demand for high-performance and energy-efficient computing is ever increasing. 3D manycore systems are a promising solution in this direction. However, current TSV-

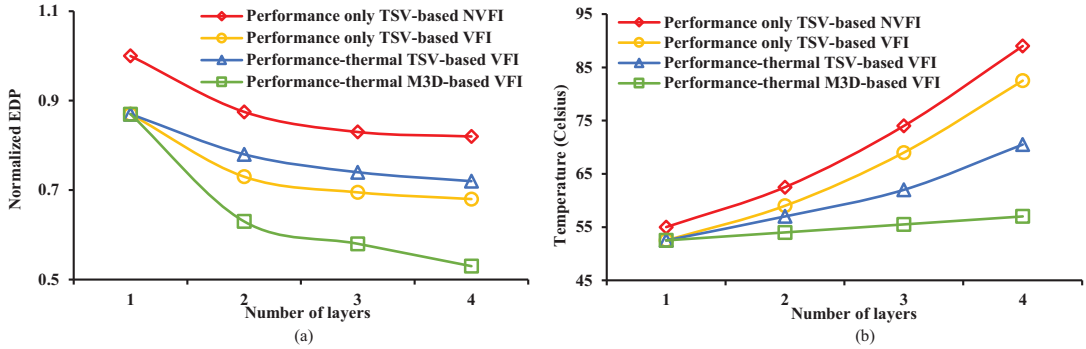


Fig 8. (a) Average EDP and (b) maximum temperature profiles of different TSV-based and M3D-based systems with varying number of layers [7].

based 3D architectures are not adequate. It suffers from limited integration capability, area overhead and thermal hotspots. M3D-based architectures enable true 3D circuits and hence, achieve better performance and energy-efficiency compared to the TSV-based counterparts. Moreover, M3D improves the thermal profile too. However, there are several challenges in designing M3D-based manycore system. We need to consider the effects of various types of process-related issues. We discuss that without considering these process-related issues, we lose significant achievable performance. Significantly, even after considering the impacts of process variation, M3D-based designs can achieve better performance and lower energy and temperature compared to TSV-based solutions. Hence, we conjecture that M3D will be the 3D technology of choice in the near future for designing manycore architectures.

REFERENCES

[1] 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)*, Burlingame, CA, 2016, pp. 1-2.

[2] M. M. Shulaker et al., "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in IEEE International Electron Devices Meeting, pp. 27.4.1-27.4.4,2014.

[3] S. K. Samal, S. Panth, K. Samadi, M. Saeidi, Y. Du and S. K. Lim, "Adaptive Regression-Based Thermal Modeling and Optimization for Monolithic 3-D ICs," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 35, no. 10, pp. 1707-1720, Oct. 2016.

[4] A. Koneru, S. Kannan and K. Chakrabarty, "Impact of Electrostatic Coupling and Wafer-Bonding Defects on Delay Testing of Monolithic 3D Integrated Circuits," ACM Journal on Emerging Technologies in Computing Systems, vol. 13, no. 4, pp. 54:1-54:23, July 2017.

[5] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain and P. E. Gaillardon, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 2, no. 4, pp. 714-722, Dec. 2012.

[6] C. Fenouillet-Beranger et al., "New insights on bottom layer thermal stability and laser annealing promises for high performance 3D VLSI," in IEEE International Electron Devices Meeting, San Francisco, CA, 2014, pp. 27.5.1-27.5.4.

[7] D. Lee, S. Das, J. R. Doppa, P. P. Pande and K. Chakrabarty, "Performance and Thermal Tradeoffs for Energy-Efficient Monolithic 3D Network-on-chip," Transactions on Design Automation of Electronic Systems, vol. 23, no. 5, pp. 60:1-60:25, Oct. 2018.

[8] S. Das, J. R. Doppa, P. P. Pande and K. Chakrabarty, "Monolithic 3D-Enabled High Performance and Energy Efficient Network-on-Chip," in IEEE International Conference on Computer Design (ICCD), Boston, MA, 2017, pp. 233-240.

[9] J. A. Boyan and A. W. Moore, "Learning evaluation functions to improve optimization by local search," The Journal of Machine Learning Research, vol. 1, p. 77–112, Nov. 2000.

[10] S. Das, J. R. Doppa, P. P. Pande and K. Chakrabarty, "Design-Space Exploration and Optimization of an Energy-Efficient and Reliable 3-D Small-World Network-on-Chip," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 5, pp. 719-732, May 2017.

[11] S. Das, D. Lee, D. H. Kim and P. P. Pande, "Small-world network enabled TSV enabled energy efficient and robust 3D NoC architectures," in Proceedings of the 25th edition on Great Lakes Symposium on VLSI, Pittsburgh, PA, 2015.

[12] D. Lee, S. Das, J. R. Doppa, P. P. Pande and K. Chakrabarty, "Impact of Electrostatic Coupling on Monolithic 3D-enabled Network on Chip," ACM Transactions on Design Automation of Electronic Systems, vol. 24, no. 6, pp. 62:1-62:22, Nov. 2019.

[13] Y. S. Yu, S. Panth and S. K. Lim, "Electrical Coupling of Monolithic 3-D Inverters," IEEE Transactions on Electron Devices, vol. 63, no. 8, pp. 3346-3349, Aug. 2016.

[14] L. Pasini et al., "nFET FDSOI activated by low temperature solid phase epitaxial regrowth: Optimization guidelines," in SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Millbrae, CA, 2014, pp. 1-2.

[15] S. Panth, S. K. Samal, K. Samadi, Y. Du and S. K. Lim, "Tier Degradation of Monolithic 3-D ICs: A Power Performance Study at Different Technology Nodes," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 8, pp. 1265-1273, Aug. 2017.

[16] L. Peh and W. J. Dally, "A delay model for router microarchitectures," IEEE Micro, vol. 21, no. 1, pp. 26-34, Jan.-Feb. 2001.

[17] U. Y. Ogras, R. Marculescu, D. Marculescu and E. G. Jung, "Design and Management of Voltage-Frequency Island Partitioned Networks-on-Chip," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 17, no. 3, pp. 330-341, March 2009.