Design and Optimization of Heterogeneous Manycore Systems Enabled by Emerging Interconnect Technologies: Promises and Challenges

Biresh Kumar Joardar*, Ryan Gary Kim†, Janardhan Rao Doppa*, Partha Pratim Pande*

*School of EECS, Washington State University
Pullman, WA 99164, U.S.A.
{biresh.joardar, jana.doppa, pande}@wsu.edu

†Department of ECE, Colorado State University
Fort Collins, CO, 80523, U.S.A.
Ryan.G.Kim@colostate.edu

***Abstract*— Due to the growing needs of Big Data applications (*e.g.*, deep learning, graph analytics, and scientific computing) and the ending of Moore's law, there is a great need for low-cost, high-performance, energy-efficient, and small form-factor manycore systems. With more stringent design objectives, application specialization, and more cores on a single chip, design-time optimization becomes more complex. Moreover, with the advent of emerging interconnect technologies like 3D integration, the design optimization process has become more challenging. This increases the need for a holistically optimized design process that makes design decisions across multiple layers of the system, *e.g.*, memory, compute, interconnect technology, and network infrastructure. In this paper, we present various challenges of designing heterogeneous manycore architectures using 3D integration and viable optimization techniques to solve them.**

I. INTRODUCTION

Neural networks, graph analytics, and other big-data applications are now frequently used in diverse application domains. High-performance heterogeneous computing systems, *e.g.*, manycore platforms with CPUs, GPUs, and accelerators, are suitable for big data applications [1]. However, the use of off-chip interconnects (*e.g.*, PCIe) in existing discrete GPU systems gives rise to latency and energy overheads. Heterogeneous manycore architectures where the CPUs and GPUs are interconnected via an efficient network-on-chip (NoC) would avoid such expensive data transfers and lead to better performance and energy-efficiency.

To further reduce data transfer costs, three-dimensional (3D) integrated circuits (ICs) have made significant strides towards improving communication efficiency [2], [3]. By connecting planar dies stacked on top of each other with through-silicon vias (TSVs), the communication latency, throughput, and energy consumption can be further improved [2]. 3D ICs together with NoCs, enable the design of highly integrated heterogeneous (*e.g.*, CPUs, GPUs, and accelerators) manycore platforms for big-data applications. Specifically, recent works have addressed various issues associated with designing a 3D NoC architecture targeting deep learning, graph analytics, etc. [4]

However, the design of 3D NoC-based heterogeneous manycore architectures pose unique challenges. For example, in a heterogeneous CPU-GPU based system, the disparate natures of CPU and GPU architectures introduce significantly different and conflicting communication requirements. In addition, due to more dense circuit integration in 3D ICs, the power density is much higher than their 2D counterparts [4]. Reducing thermal hotspots resulting from the high-power density in 3D integration is one of the key challenges in 3D NoC-based designs. However, optimizing only for a single objective can have negative effects on the other objectives (e.g., optimizing for temperature can have ill effects on performance). Therefore, it is necessary to consider a joint multi-objective optimization (MOO) framework that simultaneously optimizes all relevant objectives.

As system size and the level of heterogeneity increases, conventional MOO algorithms take significantly more time to find near-optimal solutions due to their relatively unguided nature of design space exploration. To reduce the design time for finding near-optimal solutions, more efficient and scalable optimization techniques are required. Machine learning (ML) inspired techniques present a promising direction to explore in this regard [5]. In addition, efficient optimization methodologies should be complemented using other design innovations, e.g., emerging monolithic 3D (M3D) integration. Together, they enable the design of high-performance yet thermally viable 3D heterogeneous manycore systems.

In this paper, we first discuss the limitations of existing architectures to motivate the need for more efficient design techniques. Next, we elaborate on key objectives that should be considered to design a high-performance 3D heterogeneous (CPUs and GPUs) NoC targeted for big data workloads. In this regard, we also highlight the benefits of using ML-based optimization techniques. Subsequently, we focus on complementary ways to improve performance and temperature in 3D heterogeneous NoCs. More specifically, we discuss the necessity of MOO in design space exploration, the trade-offs and use of emerging interconnect technologies, and an ML-based technique that can be adopted to design efficient 3D heterogeneous architectures. This paper is a part of the DATE 2019 Special Session on "Smart Resource Management and Design Space Exploration for Heterogeneous Processors." The other two papers of this Special Session are: "Smart Thermal Management for Heterogeneous Multicores," [24] and "Power and Thermal Analysis of Commercial Mobile Platforms: Experiments and Case Studies," [25].

II. LIMITATIONS OF CONVENTIONAL NOC ARCHITECTURES

Due to its simplicity, mesh-based NoC architectures have been widely adopted [6,18]. Unfortunately, these types of networks are not well-equipped to handle the traffic seen in heterogeneous systems [7]. Fig. 1 shows the traffic distribution between cores (CPUs and GPUs), and between cores and last level cache (LLC) for several common heterogeneous system benchmarks from the Rodinia suite [19]: Backpropagation (BP), Breadth-first search (BFS), Gaussian elimination (GAU),

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.

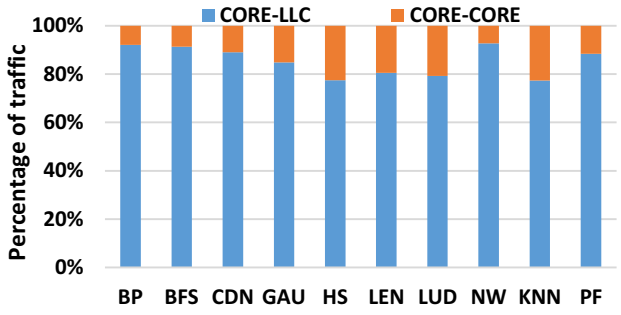

Fig. 1. Amount of traffic going to/from LLC (CORE-LLC: CPU-LLC, GPU-LLC, LLC-LLC and vice versa) and between cores (CORE-CORE: CPU-CPU, GPU-GPU) with different Rodinia [19] applications.

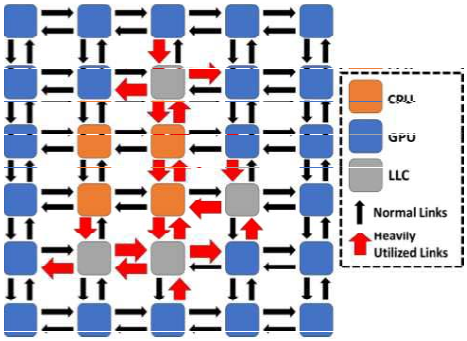

Fig. 2. Relative distribution of traffic among the links in a mesh NoC. Red arrows indicate heavily utilized links that carry more than twice the average number of messages [7]

Hotspot (HS), LU Decomposition (LUD), Needleman-Wunsch (NW), K-nearest neighbors (KNN), Pathfinder (PF), and two well-known deep learning frameworks: convolutional neural network (CNN) for CIFAR-10 (CDN) and CNN for MNIST (LEN). As a result of the data-intensive nature of heterogeneous system workloads, we see heavy communication coming from and going to the LLCs. Due to this *many-to-few communication*, these LLCs can potentially become traffic hotspots and bottleneck the system [7, 8].

Even for a mesh-based 3D NoC architecture with optimized CPU, GPU, and LLC placements, there still exists a few links (connected to the LLCs) that are heavily utilized when compared to the rest of the links present in the NoC [7] (see Fig. 2). During high traffic, such links become bandwidth bottlenecks, negatively affecting the overall performance. The presence of these bandwidth bottlenecks is due to the traffic hotspots, the inherent multi-hop nature of the mesh architecture, and the low path diversity introduced by simple routing mechanisms, which leads to high traffic aggregation at the intermediate routers and links associated with LLCs [7].

These issues necessitate the investigation of more complex NoC designs. In particular, we need to account for the specific traffic characteristics and requirements introduced by these heterogeneous systems. In this effort, irregular networks that allow more path diversity between the routers (e.g., small-world networks [5]) can alleviate some of these issues by better distributing the traffic among different links. However, as we'll see in the following sections, this requirement and other design objectives associated with heterogeneous systems significantly increases the complexity of the design-time optimization process and thereby require more efficient design techniques.

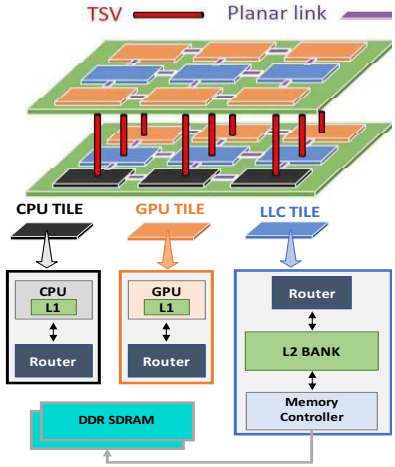

Fig. 3. Illustration of a 3D heterogeneous manycore system with CPUs, GPUs, and LLCs connected via TSVs [4].

III. Rationale for MOO in Design Space Exploration

In recent years, hardware for applications like machine learning and big-data analytics need to consider not only the application characteristics, but also where these applications are deployed. For example, hardware support for machine learning in the cloud or an automotive embedded system can look drastically different due to different design constraints, e.g., higher emphasis on reliability for the automobile deployment while prioritizing performance for the cloud. These demands for computing platforms with increasingly higher performance in highly constrained scenarios have led to more specialized systems. Naturally, this leads us to abandon design solutions that consider only a single aspect, e.g., performance, and pursue designs that jointly consider power, reliability, and the performance of individual components.

For example, CPU cores use instruction-level parallelism to achieve high performance on a limited number of threads. If any of these threads stall, CPUs incur a large latency penalty. On the other hand, GPUs achieve high throughput through massive data-parallelism. This parallelism, coupled with quick context-switching capabilities, allow GPUs to hide much of the memory access latency, thus relying more on higher throughput memory accesses. Therefore, a CPU-GPU based heterogeneous manycore system should provide low-latency CPU *and* high-throughput GPU communication.

In addition, one of the key challenges of 3D systems is the higher temperatures due to greater power density. High temperature not only affects performance but also the lifetime of the device. Therefore, thermal constraints need to be considered while designing an efficient CPU-GPU based 3D manycore heterogeneous system. However, optimizing only for temperature can move power hungry cores towards the sink and further away from cores they highly communicate with, negatively affecting the performance.

Due to these diverse and sometimes conflicting objectives, it is necessary to jointly consider all relevant objectives in order to design hardware systems that satisfy the requirements of applications for highly constrained scenarios. Suitably, we should formulate the design of 3D heterogeneous manycore

systems in a MOO framework. Instead of finding a single "best" solution, MOO frameworks attempt to find a set of solutions along the Pareto frontier. This set of solutions represent the optimal trade-offs a designer could make between all considered objectives. For example, let us consider a 3D heterogeneous manycore system with two layers consisting of CPUs, LLCs, and GPUs (see Fig. 3). When designing this 3D heterogeneous NoC, we should consider the CPU performance requirements, GPU performance requirements, and the thermal characteristics. We discuss these design objectives next. Note that designing an efficient NoC for CPU-GPU based manycore platform is not limited to these objectives only. Other design objectives can also be included as demanded by specific applications.

A. Design Objectives for 3D Heterogeneous NoCs

For the CPUs, their performance is highly sensitive to latency. Therefore, we should attempt to reduce the overall latency for CPU traffic or Lat:

$$Lat = \frac{1}{C \cdot M} \sum_{i=1}^{C} \sum_{j=1}^{M} (r \cdot h_{ij} + d_{ij}) f_{ij} \quad (1)$$

where C is the number of CPUs, M is the number of LLCs, r is the number of router stages, and h_{ij}, d_{ij}, f_{ij} are the number of hops, link length, and frequency of interaction between nodes i and j respectively. The frequency of interaction f_{ij} can be obtained using detailed full-system simulations.

For the GPUs, their performance relies much more on obtaining large blocks of data from memory. Therefore, we should attempt to increase the overall throughput of the network. Here, we can model throughput as a combination of minimizing the expected link utilization $\overline{U}$ and standard deviation of link utilization σ to ensure that the network load is well-balanced:

$$U_k = \sum_{i=1}^{R} \sum_{j=1}^{R} (f_{ij} \cdot p_{ijk}) \quad (2)$$

$$\overline{U} = \frac{1}{L} \sum_{k=1}^{L} U_k \quad (3)$$

$$\sigma = \sqrt{\frac{1}{L} \sum_{k=1}^{L} (U_k - \overline{U})^2} \quad (4)$$

where $p_{ijk} = 1$ if nodes i and j communicate via link k and $p_{ijk} = 0$ otherwise.

For temperature, we can employ quick approximation models [9] that examine the vertical and horizontal heat flows in the system and summarize the objective as T.

$$T_{n,k} = \sum_{i=1}^{k} \left(P_{n,i} \sum_{j=1}^{i} R_j \right) + R_b \sum_{i=1}^{k} P_{n,i} \quad (5)$$

$$\Delta T(k) = \max_{n} T_{n,k} - \min_{n} T_{n,k} \quad (6)$$

$$T = \left(\max_{n,k} T_{n,k} \right) \left(\max_{k} \Delta T(k) \right) \quad (7)$$

where $P_{n,i}$ is the power consumption of the core at layer i from the sink in a single-tile stack n, R_j is the thermal resistance in the vertical direction, and R_b is the thermal resistance of the base layer on which the dies are placed.

Using these models for performance and thermal objectives, we look to place nodes (CPUs, GPUs, and LLCs) and design the network (link placement between cores) to create 3D heterogeneous NoCs jointly optimized for both performance and thermal objectives. By using a MOO framework such as AMOSA [10], we can jointly optimize for Lat, $\overline{U}$, σ, and T. Fig. 4 shows the results of optimizing a 3D mesh (*3DMesh*) and a 3D NoC with optimized link placement (*3DHet*) with different objectives: (1) performance (Eqns. 1-4, $3DHet_{perf}$ and $3DMesh_{perf}$), and (2) joint performance-thermal (Eqns. 1-7, $3DHet_{therm}$). We observe that $3DHet_{perf}$ reduces the energy-delay product (EDP) by 35% and 29% over $3DMesh_{perf}$ executing two CNNs for CIFAR and LeNet respectively. $3DHet_{therm}$ shows a 25% and 18% improvement in EDP compared to $3DMesh_{perf}$ for CIFAR and LeNet respectively. The performance benefits of *3DHet* architectures with respect to its mesh-based counterparts are achieved by placing cores and links to bring highly communicating cores closer together with more path diversity between them. As a result, *3DHet* architectures achieve lower latency and energy consumption leading to better performance. Due to the joint optimization, $3DHet_{therm}$ reduces maximum core temperature by 24% over $3DMesh_{perf}$ by placing cores and links appropriately. However, due to the selection of a design with the best performance-thermal trade-off, $3DHet_{therm}$'s temperature improvement results in slight EDP degradation when compared to the best-case performance observed in $3DHet_{perf}$.

B. Machine Learning enabled MOO Framework

Unfortunately, the design space for joint core and link placement is enormous even for a modest system size of 64. For more intuition, in a 4x4x4 (64-tile) system with 144 links (96 planar + 48 vertical), the total number of possible tile placements is 64 factorials. Then, each of these tile placements has $C(C(16,2) * 4,96)$ different ways to place the planar links. Therefore, exploring the design space of heterogeneous NoCs is very challenging and computationally hard. Machine learning-based search techniques present a novel direction to intelligently explore the design space.

Fig. 5 shows a high-level overview of how conventional algorithms, e.g., Simulated Annealing (SA), and ML-inspired optimization techniques work. Here, we show one instance of how ML can be applied in design space exploration for 3D NoCs. It is well known that SA-based searches are sensitive to the initial starting state and hence, rely on multiple independent searches from random starting states to find global optima (Fig. 5(a)). However, choosing random starting states is equivalent to guesswork. As a result, SA requires many independent searches to ensure that enough states have been visited before declaring a good solution with high confidence. To avoid this,

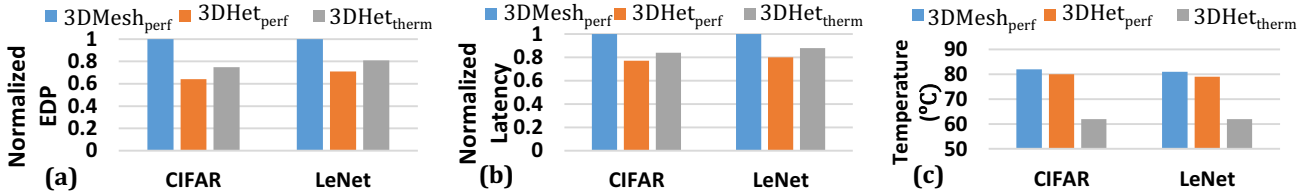

Fig. 4. (a) Network EDP, (b) latency, and c) max. temperature for $3DMesh_{perf}$, $3DHet_{perf}$, and $3DHet_{therm}$ for two deep learning benchmarks: CIFAR and LeNet. All values normalized with respect to $3DMesh_{perf}$. [4]

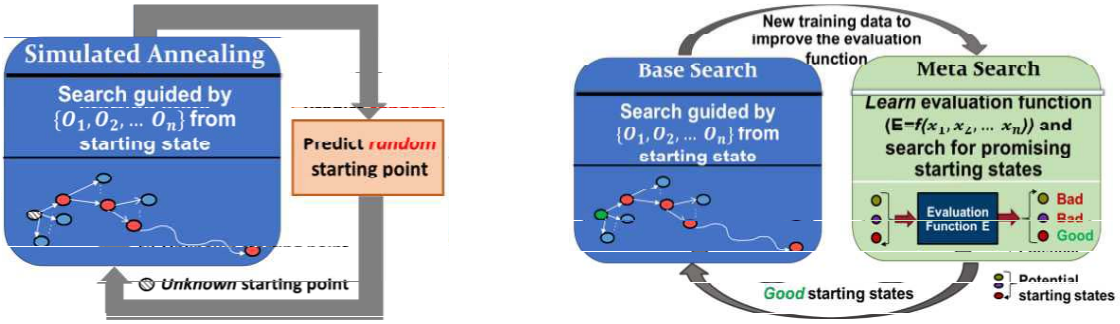

Fig. 5. Illustration of (a) conventional optimization algorithms, e.g., simulated annealing, and (b) ML-based optimization techniques.

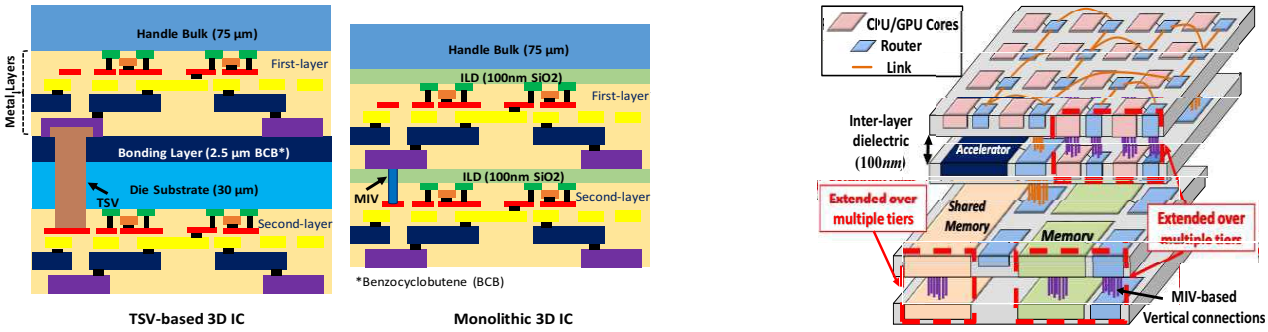

Fig 6: Comparison of (a) TSV-based and M3D-based 3D IC design, and (b) Example heterogeneous NoC architecture using M3D [15].

ML-based algorithms, e.g., STAGE [11], can learn the structure in the search space to predict the relationship between the final solution of local search and the starting state. In [5], authors have shown that STAGE can significantly speedup traditional optimization techniques, namely, SA and genetic algorithms for NoC design optimization with homogenous cores.

By incorporating a multi-objective function that can greedily choose neighboring designs and a solution archive, we can extend STAGE to a MOO setting (MOO-STAGE). Fig. 5(b) describes how MOO-STAGE can help improve design time and quality of the solution much faster than conventional algorithms. The first stage (*Local search*) performs a typical search from a given starting state guided by a cost function considering all objectives. Then, the search trajectories collected from the *Local search* are used for the next stage (*Meta search*) to learn an evaluation function. In other words, this evaluation function attempts to learn the potential of performing a *Local search* starting from a particular state (quantified using the cost function). This allows the algorithm to prune away bad starting states to reduce the number of *local search* calls needed to find (near-) optimal designs in the given design space. Unlike MOO-STAGE, other MOO algorithms based on random restarts, e.g., AMOSA [10], do not leverage any such knowledge and spend significant amount of time searching from states that would otherwise be rejected by MOO-STAGE. Therefore, MOO-STAGE's *guided exploration* progressively focusses the search in more promising areas of the search space to unearth better solutions much faster than conventional MOO algorithms. Here, we have discussed just one instance of how ML-based methodologies (MOO-STAGE) can benefit the design and optimization of heterogeneous NoCs. The diverse set of ML techniques available today can be used to develop appropriate algorithms similarly to improve different aspects of the hardware design.

IV. NoC Design with Emerging Integration Technologies

In this section, we discuss the role of emerging 3D technology that can be used in conjunction with the design optimization techniques to produce better NoC designs.

A. Limitations of TSV-based 3D NoCs

Although TSV technology (Fig. 6(a)) is predominantly employed in 3D integration, it has several limitations. The keep-out-zone (KOZ) required for TSVs and low die alignment precision impose limits on the achievable device integration density when using TSV-based 3D stacking. A minimum KOZ of 3um is required for ICs fabricated at the 20nm technology node [12], and the die alignment precision is currently limited to 0.5 um [13]. Therefore, the TSV pitch will remain in the micron range and will not reduce appreciably in the future due to bonder alignment limitations (0.5-1 um) and stacked silicon layer thickness (6-10 um). While TSV pitches in the micron range may provide enough vertical connections for stacking memory atop processors and memory-on-memory stacking, they are not enough to enable true 3D ICs.

Also, conventional TSV-based 3D architectures are susceptible to higher temperatures [14, 15]. As can be seen in Fig. 6(a), consecutive layers in TSV-based designs are attached using a bonding material, e.g., Benzocyclobutene (BCB) that exhibit very poor thermal conductivity [14, 15]. This impedes the seamless flow of heat across the layers resulting in considerable increase in temperature. Moreover, the relatively

thicker silicon substrate (several micrometers) in TSV-based architectures causes the heat to spread laterally within the substrate [16]. Together, they result in higher on-chip temperatures which is undesirable in a 3D architecture.

B. M3D integration: Opportunities in 3D NoC design

On the other hand, emerging integration technology like monolithic 3D (M3D) presents an attractive alternative to TSV-based designs. In M3D designs, multiple tiers are processed sequentially on the same wafer and are connected using monolithic inter-tier vias (MIVs) [16]. Since the physical dimensions of MIVs are comparable to standard copper vias, M3D designs offer improved integration density, total wirelength, and interconnect energy when compared to TSV-based designs [17]. Apart from these advantages, M3D also allows faster dissipation of heat than their TSV-based counterparts [14]. The absence of a bonding material and relatively smaller dimensions (nanometers as opposed to micrometers) leads to superior thermal characteristics than TSV-based architectures [14]. Therefore, we can design high-performance and thermally viable 3D architectures with multiple layers of logic using M3D integration. M3D opens up the possibility of designing true 3D systems by fully utilizing the vertical dimension. Circuit designers have designed high-performance and energy-efficient M3D circuits by utilizing logic blocks that span multiple tiers [16][17]. Similarly, 3D heterogeneous systems can take advantage of M3D to design NoC routers that span multiple tiers. As we can see in Fig. 6(b), we can create shortcuts and improve the energy and performance of 3D NoCs by including multi-tier routers in the network.

More specifically, the benefit of multi-tier M3D routers are three-fold:

(1) Lower intra-router delay and energy due to the shorter wire lengths enabled by MIVs;
(2) Lower inter-router hop-counts as routers extended over multiple tiers eliminate vertical hops; and
(3) Lower inter-tier communication energy as MIVs are significantly more energy-efficient than TSVs.

By using M3D as the enabler for 3D NoCs, we can improve the performance through better connectivity and faster routers, reduce the energy through lower wirelength and hop counts, and improve thermal characteristics by having all tiers of the 3D system closer to the thermal sink.

These M3D characteristics have allowed 28% higher energy efficiency for NoCs using M3D over TSVs [26]. In addition, these characteristics require us to revisit some of the assumptions made when designing TSV-based systems. For example, M3D significantly reduces the distance between layers and therefore, reduce the distance to the thermal sink. Hence, power management strategies that try to optimize both performance and thermal profile need to be adjusted for M3D NoCs. In the next section, we will focus on this aspect of M3D NoCs for heterogeneous systems.

C. Performance and thermal trade-offs in 3D NoCs

To further improve the energy-efficiency of a manycore system, voltage frequency island (VFI)-based power management has become a common practice [20]. In VFI-based designs, a group of cores and their associated network elements are clustered together based on their computation and communication patterns. Ideally, if the cores and routers have similar utilizations, they can share a single voltage/frequency (V/F) level. This offers the flexibility to scale the V/F of each VFI in order to minimize the overall energy consumption [20].

In order to optimize these VFI-enabled 3D NoCs, we must constrain the cores in the same VFI to be physically placed together. For traditional local search-based algorithms, this introduces additional constraints on the possible optimization policies for a particular design. This can potentially increase the number of steps needed to reach good designs, thus requiring more sophisticated techniques like MOO-STAGE.

When using TSVs, these VFI-enabled 3D systems need to place VFI clusters with high power consuming cores closer to the heat sink to remove maximum possible heat from the system and to ensure thermal efficiency. By incorporating power management in a M3D-based NoC in conjunction with a sophisticated ML-based design-time optimization algorithm like MOO-STAGE, we can create high-performance and thermally-efficient 3D heterogeneous systems. However, since M3D-based system significantly reduces the distances of layers to the heat sink, we can place high-power VFIs further away from the sink while maintaining thermal-efficiency.

In Fig. 7, we show TSV- and M3D-enabled designs jointly optimized for performance and thermal objectives. In Fig. 7(a), the high power cores are pushed towards the heat sink to maintain better temperature profile. However, in this design, some highly communicating cores are placed in the top and bottom layers, increasing their latency and reducing the performance of the system. On the other hand, the M3D design is able to take advantage of the short distance to the heat sink and optimize more for performance. Fig. 7(b) shows that high power cores can be placed elsewhere in the system and that the highly comunicating cores can all be placed within the middle layer. Hence, we can enhance performance while maintaining thermal limits. Therefore, M3D-enabled NoC design is beneficial in terms of both performance and temperature compared to its TSV-based counterpart.

D. Design optimization to address M3D integration

Unfortunately, M3D systems don't come without a cost. During M3D fabrication, each tier is fabricated sequentially, from the bottom to the top. One of the major challenges 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, rather than the standard thermal budget (≈1050 °C), to prevent any deterioration of the bottom transistors. Two techniques have

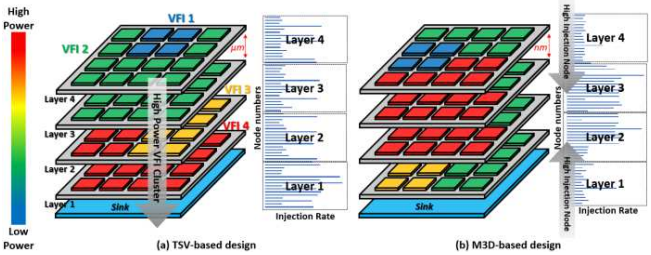

Fig 7: Comparison of the placement of (a) TSV-based and (b) M3D-based VFI-enabled 3D IC design. [15]

been proposed to fabricate the top tier at a lower temperature: solid phase epitaxy regrowth (SPER) [21] and laser annealing [22].

Although these techniques allow transistors to be built on top of each other (SPER = 500−600 °C [21] and laser annealing = 400−500 °C [22] thermal budgets), they each have their disadvantages. Transistors created using SPER show three times higher source-drain resistance when compared to conventional transistors [22]. On the other hand, laser-based annealing creates transistors with 15-28% lower on-current [23], increasing the delay for the top tier. As a result, both processes introduce performance degradation for the top-tier transistors. Additionally, tungsten is required for the back-end-of-line interconnect in the bottom tier since copper only supports temperatures up to the 400 °C range. This increases the resistivity of the interconnect in the bottom tier (copper = 1.68 μΩ·cm vs. tungsten = 5.28 μΩ·cm) and significantly impacts the performance of the design. This performance degradation of the top-tier transistors and increased delay in bottom-tier interconnects, must be considered during the design and optimization stages of any M3D NoC alongside the performance benefit due to routers utilizing multiple tiers. With these factors, future NoC designs will be far more complex with multiple objectives and constraints. Hence, sophisticated ML-based techniques, e.g., MOO-STAGE, are necessary to design efficient NoCs for heterogeneous systems.

V. CONCLUSION

Advanced computing systems have long been enablers for breakthroughs in science, engineering, and new technologies, and they have continued to play key roles in today's Big-Data era. However, with the slowing down of Moore's law and the relentless needs of Big-Data applications (e.g., deep learning, graph analytics, autonomous driving, personalized medicine etc.), we need innovative architectures and computationally efficient methods to design application-specific hardware systems to optimize performance, power, and thermal objectives. 3D integration is a catalyst for designing high-performance and energy- and thermally-efficient manycore architectures. However, 3D integration brings various unsolved challenges and it is crucial to establish suitable power-performance-thermal trade-offs. In this paper, we have highlighted various challenges associated with the design and optimization of heterogeneous manycore systems developed using emerging 3D integration technologies.

REFERENCES

[1] D. Strigl, K. Kofler, and S. Podlipnig, "Performance and Scalability of GPU-Based Convolutional Neural Networks," in *Proc. of the 18th Euromicro Conference on Parallel, Distributed and Network-based Processing*, Pisa, Italy, 2010.

[2] B.S. Feero and P.P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation," *IEEE Trans. on Computer*, vol. 53, no. 1, pp. 32-45, 2008.

[3] W.R. Davis et. al., "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *IEEE Design & Test*, vol. 22, no. 6, pp. 498-510, 2005.

[4] B.K. Joardar et al., "3D NoC-Enabled Heterogeneous Manycore Architectures for Accelerating CNN Training: Performance and Thermal Trade-offs," in *Proc. of the 11th Intl. Symp. on NOCS*, Seoul, Korea, 2017.

[5] 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 Trans. on Computer-Aided Design of Integrated Circuits and Syst.*, vol. 36, no. 5, pp. 719-732, 2017.

[6] V. Dmitri and R. Ginosar. "Network-on-chip architectures for neural networks.". in *Proc. of the 4th Intl. Symp. on NOCS*, Grenoble, France. 2010.

[7] W. Choi et al., "On-Chip Communication Network for Efficient Training of Deep Convolutional Networks on Heterogeneous Manycore Systems," *IEEE Trans. on Computers*, vol. 67, no. 5, pp. 672-686, 2018.

[8] H. Jang et al., "Bandwidth-efficient on-chip interconnect designs for GPGPUs," in *Proc. of the 52nd DAC*, San Francisco, CA, 2015.

[9] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. of the 2004 ICCAD,* San Jose, CA, 2004.

[10] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multi-objective optimization algorithm: AMOSA," *IEEE Trans. on Evolutionary Computation*, vol. 12, no. 3, pp. 269-283, 2008.

[11] J.A. Boyan and A.W. Moore, "Learning Evaluation Functions to Improve Optimization by Local Search," *Journal of Machine Learning Research*, pp. 77-112, 2000.

[12] S. Kannan, R. Agarwal, A. Bousquet, G. Aluri, and H.S. Chang, "Device performance analysis on 20nm technology thin wafers in a 3D package," in *Proc. of the IEEE International Reliability Physics Symposium*, Monterey, CA, 2015.

[13] [online] Available: http://www.itrs2.net/2013-itrs.html

[14] S.K. Samal et al., "Fast and accurate thermal modeling and optimization for monolithic 3D ICs," in *Proc. of the 51st DAC*, San Francisco, CA, 2014.

[15] 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," *ACM Trans. on Design Automation of Electronic Systems*, vol. 23, no. 5, 2018

[16] Y.J. Lee, P. Morrow, and S.K. Lim, "Ultra high density logic designs using transistor-level monolithic 3D integration," in *Proc. of the 2012 ICCAD*, San Jose, CA, 2012.

[17] 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 *Proc. of the 2016 S3S*, Burlingame, CA, 2016.

[18] A. Firuzan, M. Modarressi, and M. Daneshtalab, "Reconfigurable communication fabric for efficient implementation of neural networks," in *Proc. of the 10th ReCoSoC*, Bremen, Germany, 2015.

[19] S. Che et al., "Rodinia: A benchmark suite for heterogeneous computing," in *Proc. of the 2009 IISWC*, Austin, TX, 2009.

[20] U.Y. Ogras, R. Marculescu, D. Marculescu, and E.G. Jung, "Design and management of voltage-frequency island partitioned networks-on-chip," *IEEE Trans. Very Large Scale Integr. Syst*., vol. 17, no. 3, pp. 330-341, 2009.

[21] L. Pasini et al., "nFET FDSOI activated by low temperature solid phase epitaxial regrowth: Optimization guidelines," in *Proc. of the 2014 S3S*, Millbrae, CA, 2014.

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

[23] B. Rajendran et al., "Low Thermal Budget Processing for Sequential 3-D IC Fabrication," *IEEE Trans. on Electron Devices*, vol. 54, no. 4, pp. 707-714, 2007.

[24] J. Henkel, H. Khdr, and M. Rapp, "Smart Thermal Mangement for Heterogeneous Multicores," in *Proc. of the 2019 DATE*, Florence, Italy, 2019.

[25] G.M. Bhat, S. Gumussoy, and U.Y. Ogras, "Power and Thermal Analysis of Commercial Mobile Platforms: Experiments and Case Studies," in *Proc. of the 2019 DATE*, Florence, Italy, 2019.

[26] S. Das, J.R. Doppa, P.P. Pande, and K. Chakrabarty, "Monolithic 3D-Enabled High Performance and Energy Efficient Network-on-Chip," In *Proc. of the 2017 ICCD*, Boston, MA, 2017.