Achieving Datacenter-scale Performance through Chiplet-based Manycore Architectures

Harsh Sharma, *Student Member, IEEE*, Sumit K. Mandal, *Member, IEEE,* Janardhan Rao Doppa, *Senior Member, IEEE,* Umit Ogras, *Senior Member, IEEE,* and Partha Pratim Pande, *Fellow, IEEE*

***Abstract*—Chiplet-based 2.5D systems that integrate multiple smaller chips on a single die are gaining popularity for executing both compute- and data-intensive applications. While smaller chips (chiplets) reduce fabrication costs, they also provide less functionality. Hence, manufacturing several smaller chiplets and combining them into a single system enables the functionality of a larger monolithic chip without prohibitive fabrication costs. The chiplets are connected through the network-on-interposer (NoP). Designing a high-performance and energy-efficient NoP architecture is essential as it enables large-scale chiplet integration. This paper highlights the challenges and existing solutions for designing suitable NoP architectures targeted for 2.5D systems catered to datacenter-scale applications. We also highlight the future research challenges stemming from the current state-of-the-art to make the NoP-based 2.5D systems widely applicable.**

***Keywords*— 2.5D, Chiplet, PIM, NoP**

I. Introduction

Chiplet-based architectures that integrate multiple small dies on an interposer are drawing the attention of leading silicon manufacturers due to their higher energy efficiency and lower fabrication cost [1]. ITRS 2.0 and IRDS roadmap highlight the unprecedented need for memory and processing over the next decade [2] [3] [4]. This need dictates large-scale chips with high memory and compute capabilities, offering a high degree of parallelism. Such large-scale chips include tens to hundreds of processing cores, significantly increasing the area of monolithic chips [2]. One of the major challenges in the silicon industry is the exploding fabrication cost as the monolithic chips approach the reticle limit. The chiplet-based design concept offers a promising solution for reducing the manufacturing cost of large monolithic chips [1]. Chiplet-based systems integrate multiple smaller chips (chiplets) on a single die. The chiplets are connected through the network-on-interposer (NoP). Since each chiplet consumes a smaller area than a monolithic chip, the overall fabrication cost of the overall 2.5D system is significantly lower than that of the monolithic counterpart [1]. Emerging 2.5D architectures are expected to enable datacenter-scale computing via handheld devices or embedded systems. However, the computing capabilities of current edge devices need to be enhanced at least by a factor of 30-50X to achieve a datacenter-scale performance [5]. To achieve this goal, leading foundries incorporate chiplet-based systems due to the yield and fabrication cost benefits over monolithic counterparts [6].

Manufacturing several smaller chiplets and combining them into a single system leads to the functionality of a larger chip while maintaining the cost advantages of the smaller chips. Moreover, integrating several chiplets in a single 2.5D system necessitates design and optimization of the NoP, which is the communication backbone of the chiplet-based system [7]. A given heterogeneous chiplet library can include manycore CPUs, GPUs, in-memory computing elements with resistive RAM (RRAM) and other types of accelerators, and memory (such as HBM-based 3D DRAM). Hence, the physical layout and NoP design play a crucial role in determining throughput, latency, and energy-efficiency, analogous to core placement and interconnection in intra-chip environments. This paper highlights the challenges and advantages of using NoP-based systems for achieving data-center scale performance.

The rest of the paper is organized as follows. Section II describes the overview of NoP architectures specifically considering a high number of chiplets. Section III presents the tool for reliable NoP performance evaluation and summarizes the underlying principles. Section IV presents a sample of performance evaluation results considering NoP architectures proposed so far. Finally, Section V highlights future research directions focused on designing more robust and innovative chiplet-based manycore systems.

II. Overview of NoP Architectures

Increasing fabrication costs can mask the performance improvement of large monolithic manycore architectures. Most chip vendors and foundries, including TSMC, NVIDIA, Intel, and AMD, are exploring non-monolithic alternatives such as 2.5D interposer-based systems to partition the on-chip resources into smaller discrete computing cores called chiplets. 2.5D-based manycore systems offer a promising alternative to monolithic chips [1] [8]. Novel 2.5D chiplet platforms provide a new avenue for compact scale-out implementations of various emerging compute- and data-intensive workloads. Integrating multiple small chiplets on a large interposer offers significant performance and manufacturing yield improvements compared to 2D ICs, reducing the fabrication cost [2]. Furthermore, it achieves higher thermal efficiency than 3D ICs and facilitates heterogeneous integration [9]. Hence, it has become possible to envision large-scale manycore systems on 2.5D platforms. However, scalable communication between chiplets is particularly challenging due to relatively large physical distances between chiplets, poor technology scaling of electrical wires, and shrinking power budgets. The aforementioned challenges make it difficult to design a viable NoP that can support ultra-high bandwidth, energy-efficient, and low-latency inter-chiplet data transfer without increasing fabrication costs. The demands on the NoP infrastructure will only be exacerbated as application complexity continues to scale. For example, the NoP area overhead alone can be up to 85% of the total system area [10].

Design of various general-purpose and application-specific NoP architectures has been explored so far. The first family of

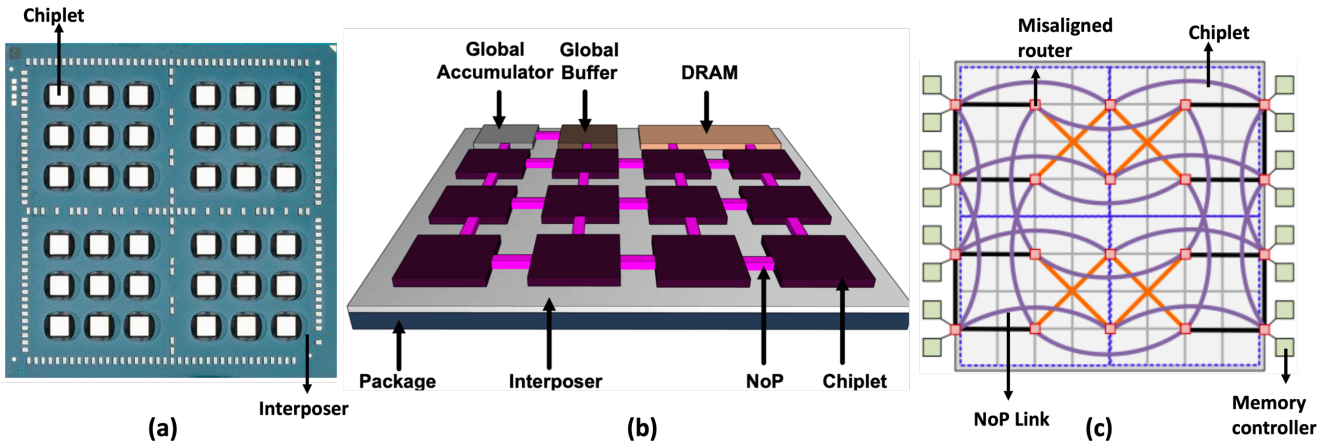

Fig. 1: NoP architectures designs based on multi-hop networks a) SIMBA, b) SIAM and c) Kite

NoP architectures are based on regular multi-hop networks. IntAct, for example, is a 2.5D prototype system with six chiplets stacked on an active interposer with a Mesh NoP [9]. In IntAct, the authors demonstrated the scalability of the 2.5D system with low-latency distributed interconnects. Simba is another 2.5D system with 36 chiplets specifically designed for deep neural network (DNN) inferencing [11]. It also uses a Mesh NoP. Simba employs tiling optimizations to limit the inter-chiplet traffic as shown in Figure 1(a). Recently, the Kite family of NoP topologies has been proposed for a 2.5D-based system considering synthetic traffic/workload as shown in Figure 1(c) [8]. NN-Baton is another recently proposed 2.5 D architecture that undertakes a design exploration considering several DNN applications [12]. The NoP topology adopted in NN-Baton is a ring architecture. Figure 1 shows the NoP architectures designed based on regular multi-hop networks.

We note that all the above-mentioned NoP architectures principally utilize multi-hop networks, which do not scale with higher number of chiplets. Moreover, these multi-hop NoP architectures create performance bottlenecks for datacenter scale applications. A high-performance and energy-efficient NoP architecture called SWAP has been recently proposed for designing chiplet-based systems for server-scale scenarios, running multiple deep learning (DL) workloads in parallel [13]. Figure 2 is an illustrative example of the SWAP architecture. SWAP is the first 2.5D accelerator with inter-chiplet communication-aware NoP to achieve high performance and energy efficiency with reduced fabrication cost with respect to state-of-the-art alternatives. SWAP leverages an efficient multi-objective optimization (MOO) mechanism to generate a NoP architecture with a smaller number of links and smaller routers than all the existing NoP counterparts mentioned above. The irregularity in the SWAP NoP improves the overall link utilization in the system. Moreover, it is scalable for a wide variety of DL workloads and number of chiplets in the system.

III. Software Tool for NoP Performance Evaluation

Chiplet-based architectures are proven to be more energy-efficient than their monolithic counterparts for various compute- and data- intensive applications (e.g., autonomous driving, machine vision, robotic medical diagnosis) necessitate high performance with small form-factor [14] [15]. These applications traditionally require datacenter scale computing infrastructures. However, various new data- and compute-intensive applications are emerging regularly. As an example, different neural network architectures including linear (e.g., VGG), residual (e.g., ResNet), and dense (e.g., DenseNet) connections are prevalent in widely used deep learning workloads. Even within the DL family, workloads vary widely. Chiplet-based systems can reduce the dependance on power-hungry datacenters if they are evaluated and fine-tuned for these emerging workloads. Therefore, there is a need for a full system performance evaluation framework for chiplet-based systems to enable fast design space exploration. There are two families of performance evaluation platforms targeting chiplet-based architectures. The first one is based on open-source traditional manycore simulators such as gem5, sniper, gpgpu-sim, etc. [16]. On the other hand, the recently proposed SIAM framework is a full system performance evaluation tool targeted specifically for 2.5D architectures consisting of processing-in-memory (PIM)-based chiplets [10].

Gem5-based HeteroGarnet is a recently proposed NoP performance evaluation tool [8]. HeteroGarnet is developed to characterize the performance of traditional von-Neuman based architectures consisting of CPU, GPU, and memory chiplets. Two types of interconnect architectures need to be considered in chiplet-based systems. The intra-chiplet network is principally a network-on-chip (NoC) and the inter-chiplet network is the NoP. Due to the size and locations, the physical interconnect materials (e.g., on-chip wires, TSVs, µbumps) and their individual widths vary across the whole system. Individual chiplets can operate at different voltages and frequencies, and they can be connected to form a larger system on the package.

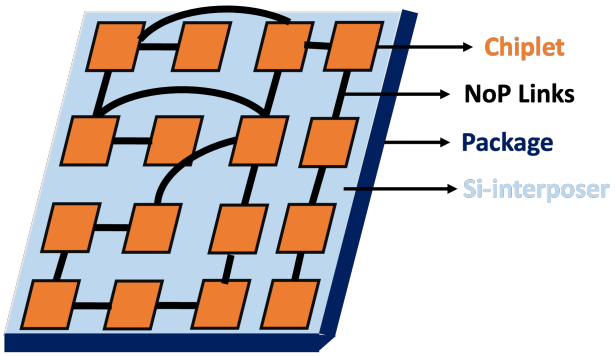

Fig. 2: Illustration of the SWAP architecture for a chiplet-based system with application-specific NoP links

Exploring the design space of such a chiplet-based architecture requires tools to model the heterogeneity. For example, in a hybrid GPU-CPU 2.5D architecture, both chiplet types need not be designed with similar technological and physical parameters, often requiring clock-domain crossings and serializer-deserializer to communicate, which is supported in HeteroGarnet.

PIM-based architectures for chiplet-based systems achieve higher performance and energy efficiency than traditional von-Neumann architectures, specifically for DL workloads. DL workloads, such as deep neural networks (DNNs), convolutional neural networks (CNNs), Graph Neural Networks (GNNs) and their variants, are employed in a large range of applications [14] [15]. To evaluate PIM-based systems, SIAM considers the properties of DL workload as inputs. The DL workload properties include the number of layers, input and output feature maps of each layer, kernel size of each layer, and activation function of each layer. The architecture specification includes the number of chiplets in the system, number of processing elements in each chiplet, device technological properties of the PIM device (ReRAM, SRAM, FeFET as few examples), and the properties of NoC as well as the NoP. The important properties of NoC and NoP to be considered include bus width, router port configuration, buffer sizes, NoP/NoC frequency, and link length. The performance metrics of interest for such a full system evaluation are latency, power, energy, and total area consumed.

SIAM's performance evaluation can be segregated into two components: circuit and network. There exists analytical model-based evaluation for the circuit component. Specifically, SIAM adopts the models of various basic circuit components (such as buffers, ADCs, decoders, switch matrix, etc.) from the well-known NeuroSim tool [17]. These models estimate the power consumption, latency, and area of those components. Then, the number of components in the design is computed through the architectural specifications and the activity of the components are estimated from the workload specifications. With this information, the overall performance of the circuit component of the system is evaluated.

The network component of SIAM evaluates the performance of the NoC and the NoP. SIAM incorporates BookSim to perform cycle-accurate simulations of the network [18]. The input to the cycle-accurate network simulation is a trace file. A trace file depicts the communication between multiple chiplets (inter-chiplet traffic) as well as between PIM elements within each chiplet (intra-chiplet traffic). BookSim injects packets into the network according to the trace file and evaluates the network performance along with energy and area numbers. SIAM supports architecture-level benchmarking with a focus on PIM architectures and helps determine area, energy, performance, and fabrication cost trade-off between design choices for an overall better architecture.

IV. NoP performance Evaluation

In this section, we present a comparative performance evaluation of various NoP architectures proposed so far in the literature. We evaluate the NoP architecture by considering a wide range of DNNs for inferencing. Table I shows different DNNs, corresponding datasets, and the number of parameters.

TABLE I: LIST OF DL INFERENCE WORKLOADS ALONG WITH THEIR CORRESPONDING NUMBER OF DNN PARAMETERS FOR 81 CHIPLET SYSTEM WITH IMAGENET

Network (ImageNet)	Workload	(in millions) # of parameters
VGG19,ResNet50	WL1	88M
ResNet101, ResNet50	WL2	136M
ResNet152	WL3	130M
ResNet101, ResNet34	WL4	114M
DenseNet169, ResNet50, ResNet18	WL5	944M

Each system can execute one large or more than one DL workloads simultaneously, representing a datacenter scenario. To represent a server-scale system, we consider a 2.5D architecture with 81 chiplets for this performance evaluation. We employ ReRAM-based chiplets as the enabling technology to accelerate DNN inference in this performance evaluation. It should be noted that all the architectures and associated design optimization methodologies are also applicable to other crossbar array (CBA)-based PIM chiplets. Beyond ReRAM, any other memory technologies such as SRAM, STT-MRAM, FeFETs, and any other types of chiplets can be adopted too. CBAs are by far the most popular representation for PIM. They are highly efficient for matrix-vector multiplication. Note that the DNNs considered in our evaluations consist of linear (VGG), residual (ResNet), as well as dense (DenseNet) connections. Moreover, all the DNNs consist of fully connected and convolution layers. Each layer of the DNN contains higher order of multi-bit weights (e.g., ResNet-101 on ImageNet with about 38M parameters, VGG16 on ImageNet with 93.4M parameters). In each considered scenario, multiple neural networks are running simultaneously (VGG19-ResNet50 on ImageNet dataset inferenced together as an example). SIMBA, IntAct, and SIAM principally are based on 2D Mesh NoP. We consider SIAM as the representative of this group. Kite is principally a Torus-based NoP that employs skip connections.

One of the main differences between SIAM, Kite, and SWAP is the router port configuration. Figure 3 shows the router port distribution of each NoP. Both Kite and SIAM have an average port count of around four, as shown in Figure 3. In the case of SWAP, the peak moves towards left with mean router port frequency being between two and three. SWAP mainly consists of routers with a lower number of ports due to the MOO mechanism to generate a NoP architecture with a smaller number of links and smaller routers than both Kite and SIAM. The irregularity in the SWAP NoP improves the overall link utilization in the system. It is scalable for a wide variety of DL workloads and the number of chiplets in the system. Smaller routers in SWAP helps in reducing NoP energy, area, and the fabrication cost. Next, we discuss the performance-energy-area-fabrication cost trade-offs associated with different NoP architectures.

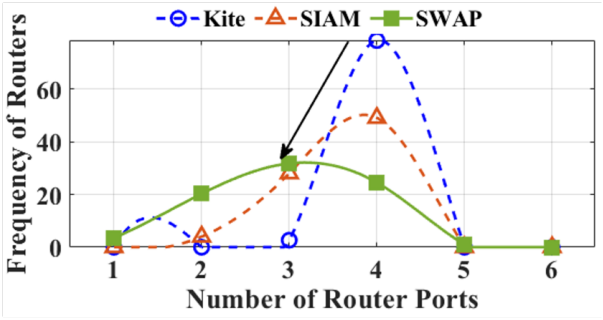

Fig. 3 Router port configuration for Kite, SIAM, and SWAP for a 2.5D system 81 chiplets. Peak of the plot is observed to move towards left.

Performance: Figure 4 presents the NoP latency for SWAP and the baseline designs (Kite and SIAM). Latency is normalized with respect to that of SIAM. We observe that SWAP outperforms both the baseline designs with up to 11% improvements in latency. Both Kite and SIAM incorporate regular NoP topologies and consist of several links which are not necessary for DL workloads. In contrast, SWAP consists of an optimized NoP that removes redundant links and places them appropriately based on inter-chiplet communication traffic. In summary, smaller routers and fewer appropriately placed links enable SWAP to achieve lower latency than SIAM and Kite.

Energy: By having smaller routers and hence reducing the unnecessary links, SWAP not only reduces the inference latency of DL workloads but also achieves significantly lower energy consumption. The energy consumption improvements compared to Kite and SIAM are shown in Figure 5. Energy consumption is normalized SIAM results. SWAP, for instance, achieves up to 47% lower energy than Kite for 81-chiplet-based system. On average we observe a 25% lower energy than SIAM for a system with 81 chiplets. The simultaneous energy and latency benefits result in significant EDP improvements over entire spectrum of considered datacenter scale scenarios. In summary, smaller routers and fewer appropriately placed links enable SWAP to achieve lower latency and energy consumption than both Kite and SIAM NoP architectures.

Cost: NoP consists of about 85% of the total 2.5D system area. Hence, the overall fabrication cost depends on the NoP. The normalized fabrication cost of an NoP is expressed as [10]:

$$C_{NoP} = \frac{L_{ref}}{L} \times e^{-D_0(A_{ref}-A_{NoP})} \quad (1)$$

where L_{ref} is the number of chiplets per wafer in the reference system and L is the number of chiplets per wafer for the system under consideration. The parameter D_0 represents the wafer defect density, and A_{ref} is the NoP area of the reference system. We consider a 2.5D system designed by AMD with 864 mm^2 interposer area and 64 chiplets as the reference in this work [1]. Using (1), we can compare the fabrication cost of two different NoP architectures. For an example, NoP fabrication cost for SWAP (C_{SWAP}) is:

$$C_{SWAP} = \frac{L_{ref}}{L} \times e^{-D_0(A_{ref}-A_{SWAP})} \quad (2)$$

Similarly, the fabrication cost of the mesh-based SIAM NoP is:

$$C_{SIAM} = \frac{L_{ref}}{L} \times e^{-D_0(A_{ref}-A_{SIAM})} \quad (3)$$

where A_{SWAP} and A_{SIAM} correspond to total NoP area of SWAP and SIAM respectively. Therefore, the fabrication cost of SWAP with respect to SIAM can be expressed as:

$$\frac{C_{SWAP}}{C_{SIAM}} = e^{-D_0(A_{SIAM}-A_{SWAP})} \quad (4)$$

The relative fabrication cost of SWAP and other architectures like SIAM principally boils down to the difference between the two NoP areas (4). Since the NoP area increases with the number of router ports and NoP links, the corresponding fabrication cost also increases. SWAP effectively reduces the number of NoP links and has smaller router ports. Hence SWAP reduces the area and the fabrication cost of the 2.5D system. As the scale of data-center applications is expected to reach an order of 100s of TOPS and equivalent to thousands of cores, the fabrication costs become an essential component for the affordability of such a system [5]. It is crucial to complement the low fabrication cost with performance and energy benefits. Figure 6 compares the trend in fabrication cost and EDP for SWAP vs Kite and SWAP vs SIAM for the 81-chiplet system. We observe that, for all the considered DL workloads, SWAP reduces both EDP and fabrication costs compared to Kite and SIAM. For instance, SWAP shows 57% improvement in EDP combined with a 13X reduction in fabrication cost with respect to Kite while executing ResNet101 and ResNet50 simultaneously, as shown in Figure 6(a). As shown in Figure 6(b), SWAP decreases the fabrication cost compared to SIAM by 17X with up to 63% EDP improvement. SWAP, having smaller router ports and fewer NoP links leads to high energy efficiency along with a significant reduction in fabrication cost

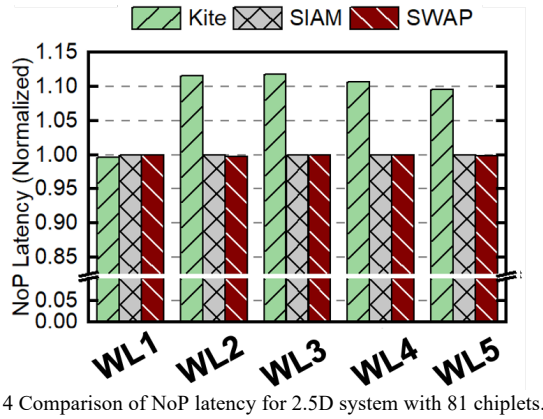

Fig. 4 Comparison of NoP latency for 2.5D system with 81 chiplets.

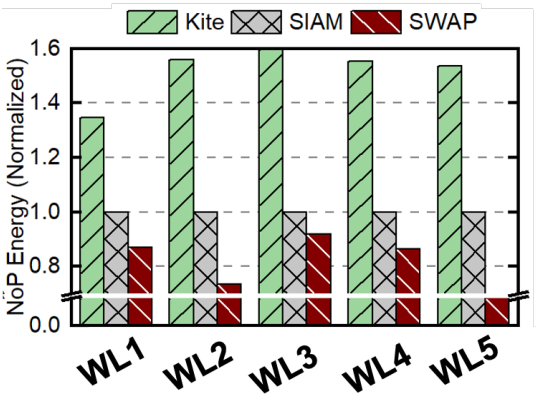

Fig. 5 Comparison of NoP energy for 2.5D system with 81 chiplets.

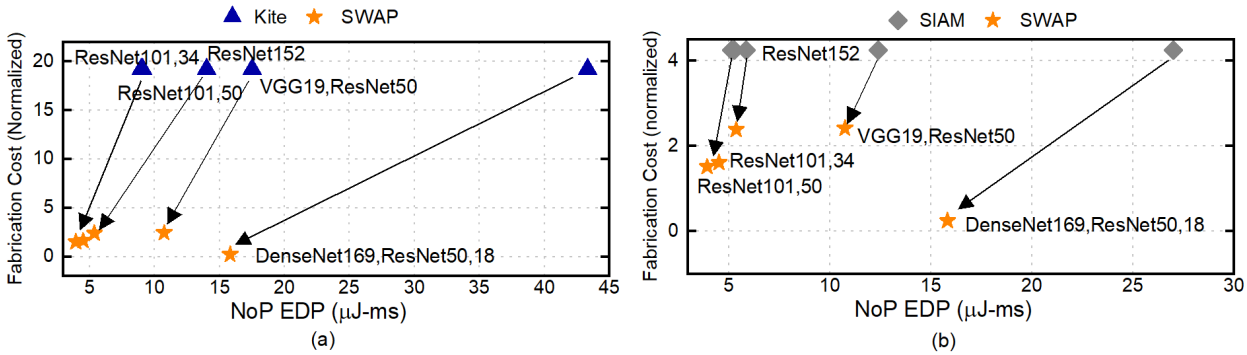

Fig. 6 Trend in fabrication cost and EDP for (a) Kite and SWAP; (b) SIAM and SWAP for a 2.5D system with 81 chiplets.

compared to state-of-the-art NoP architectures. This demonstrates scalability and affordability for achieving sustainable data-center scale of compute requirements.

V. Future Research Directions

This paper discusses how chiplet based systems should be designed to achieve datacenter scale performance. However, there are various future research directions stem from the current state-of-the-art.

Existing NoP architectures assume a single and typically fixed application workload executed one at a time. Therefore, the NoPs are optimized for a specific application, or a group of applications mapped onto the chiplet-based system. Offline NoP optimization is not practical for two main reasons. First, multiple workloads may need to be executed simultaneously in a real-world scenario. Second, various types of workloads may appear in a streamlined fashion. Specifically, the mapping of the neural layers onto the chiplets needs special attention for a stream of convolutional neural networks (CNNs) inference tasks appearing sequentially. Since each neural layer of a convolution neural network typically sends data to the subsequent layer, the consecutive neural layers must be mapped to neighboring chiplets to reduce latency. Most existing NoP architectures are primarily based on standard multi-hop regular topologies such as mesh, torus, etc. In these NoP architectures, it is not always possible to find contiguously placed chiplets available to map successive neural layers and hence they are suboptimal. Hence, design of a NoP architecture where the communicating neural layers can be executed on neighboring chiplets is of prime importance. This will also reduce the amount of long-range and multi-hop data exchanges significantly. Another application scenario that needs to be considered is natural language processing (NLP), which employs big transformer models with high memory footprint. For NLP workloads, the expected scale of parameters is in the order of hundreds of billions [19]. For instance, GPT-3 from OpenAI has over 175B parameters. Recent PALM design by Google contains 540B parameters [20]. This leads to much higher on-chip weight storage and access requirements than the existing chiplet-based systems. This in turn leads to thermal constraints. To address the thermal challenges, we may not be able to use the whole computing power of a chiplet-based architecture. Adopting the concept of dark silicon, where part of the chiplets is power gated to reduce the temperature, is a possible solution in this scenario [21] [22].

As the server-scale chips become mainstream, general CPU threads such as cache control, networking protocols, scheduling algorithms would have to run on such 2.5D based systems. A homogenous chiplet based system may not best serve all computation and algorithmic tasks requirements. Hence, heterogenous or hybrid systems (chiplets with different processing cores including CPU, GPU, or AI/ML accelerators) are to be considered for a pragmatic system design. 2.5D based systems, being modular, provides this freedom to connect multiple different chiplets through the NoP [8].

Chiplet-based systems can provide significant benefits in terms of fabrication cost. However, realistic design scenarios must consider the impact of silicon defects on the overall performance. Certain parts of each individual chiplet may not be fully functional due to intrinsic silicon defects [2]. However, none of the prior work takes silicon defects into consideration while designing a chiplet-based system. It is a common methodology that if a defect is present in a chiplet, the impacted segment is disabled and the chiplet is used with reduced functionality. Hence, we may need additional chiplet(s) to implement a particular computing kernel (e.g., mapping layer of a neural network to a chiplet). This would lead to an increase in the inter-chiplet data exchange and compromise the expected performance. Thus, any chiplet-based 2.5D system, which is designed without any provisioning for on-chip defects, cannot provide the expected performance. Therefore, it is critical to design the NoP by considering the impact of silicon defects in the chiplets. This tradeoff between reducing the fabrication cost against the performance penalty is a very important problem for more sustainable and cost effect datacenter scale system design.

Energy and policy considerations have been explored in the field of Green AI [23] [24]. Green AI refers to research that reveals novel insights without increasing computational cost (rather reducing the computational cost). Higher computational requirements lead to a larger carbon footprint for manufacturing and maintaining such systems. Green AI aims to explore the environmental effects regarding the capex (non-recurring) and opex (recurring) costs in semiconductor industry [25]. There is an overall increase of 300,000x in computing requirements in the last 10 years of deep learning, with training cost doubling every few months [24]. This necessitates larger monolithic chips that are costlier to manufacture (capex) and have higher energy requirements (opex) with respect to chiplet based 2.5D system. Reusing the defective chiplets instead of discarding them reduces the carbon footprint (capex). Hence, we should explore it to establish improved performance-sustainability

trade-offs and to pave the way towards more environmentally-friendly design paradigm, which is the need of the hour and is being pursued by many foundry and industries [26] [27].

VI. Conclusion

Datacenters require a high amount of compute and storage resources. Traditional monolithic IC-based manycore systems incur very high fabrication costs to achieve datacenter scale performance. Moreover, these monolithic chips have lower yields due to their large area. Chiplet-based 2.5D architectures are enablers to achieve datacenter-scale performance with lower fabrication costs than the monolithic counterpart. Network-on-interposer (NoP) is the communication infrastructure for the chiplet based architectures. Hence, it is a key component to achieve high performance and energy efficiency for 2.5D systems. In this paper, we discuss various design challenges associated with NoP-based 2.5D architectures. We also present a comparative performance evaluation considering various state-of-the-art NoP topologies. We also highlight important future research directions to make the NoP paradigm mainstream.

References

[1] A. Kannan, N. Jerger and G. Loh, "Enabling interposer-based disintegration of multi-core processors," in *Proceedings of the 48th International Symposium on Microarchitecture (MICRO-48)*, New York, 2015.

[2] D. Stow et al., Cost-Effective Design of Scalable High-Performance Systems Using Active and Passive Interposers, IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2017.

[3] J. A. Cunningham., "The use and evaluation of yield models in integrated circuit manufacturing.," *IEEE Trans. Semicond. Manuf.,* 1990.

[4] J. Carballo et al., "ITRS2.0:Towardare-framingofthesemiconductor technology roadmap.," in *IEEE 32nd Intl. Conf. Computer Design*, Oct 2014.

[5] B. Zimmer et al., "A 0.32–128 TOPS, Scalable Multi-Chip-Module-Based Deep Neural Network Inference Accelerator With Ground-Referenced Signaling in 16 nm," *IEEE Journal of Solid-State Circuits,* vol. 55, no. 4, 2020.

[6] https://www.amd.com/system/files/documents/amd-epyc-7003-pb-fsi-performance-comparison.pdf. [Accessed Jan 2023].

[7] N. Jerger, A. Kannan, Z. Li and G. Loh., "NoC Architectures for Silicon Interposer Systems: Why Pay for more Wires when you Can Get them (from your interposer) for Free?," in *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, USA, 2014.

[8] S. Bharadwaj, J. Yin, B. Beckmann and T. Krishna, "Kite: A Family of Heterogeneous Interposer Topologies Enabled via Accurate Interconnect Modeling," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020.

[9] P. Vivet et al., "IntAct: A 96-Core Processor With Six Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management," *IEEE Journal of Solid-State Circuits,* vol. 56, no. 1, 2021.

[10] G. Krishnan et al., "SIAM: Chiplet-based Scalable In-Memory Acceleration with Mesh for Deep Neural Networks," *ACM Trans. Embed. Comput. Syst,* vol. 20, no. 5, 2021.

[11] Y. Shao et al., "Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture," in *In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '52)*, New York, 2019.

[12] Z. Tan et al., "NN-Baton: DNN Workload Orchestration and Chiplet Granularity Exploration for Multichip Accelerators," *International Symposium on Computer Architecture (ISCA),* 2021.

[13] H. Sharma et al., "SWAP: A Server-Scale Communication-Aware Chiplet-Based Manycore PIM Accelerator," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 41, no. 11, pp. 4145-4156, 2022.

[14] Z. Wu et al., "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems,* vol. 32, no. 1, 2020.

[15] W. Liu et al., "A survey of deep neural network architectures and their applications," *Neurocomputing,* no. 234, 2017.

[16] N. Binkert et al., "The gem5 simulator," *SIGARCH Comput. Archit.,* vol. 39, 2011.

[17] X. Peng et al., "DNN+NeuroSim: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators with Versatile Device Technologies," in *International Electron Devices Meeting (IEDM)*, 2019.

[18] N. Jiang et al., "A Detailed and Flexible Cycle-Accurate Network-on-Chip Simulator," in *IEEE ISPASS*, 2013.

[19] A. Vaswani et al., "Attention Is All You Need," *Neurips,* 2017.

[20] J. Goldstein et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," *Computers and Society,* Jan 2023.

[21] A. Coskun et al., "Reclaiming Dark Silicon Using Thermally-Aware Chiplet Organization in 2.5D Integrated Systems," *Boston Area Architecture Workshop,* 2018.

[22] F. Eris et al., "Leveraging thermally-aware chiplet organization in 2.5D systems to reclaim dark silicon," *DATE ,* 2018.

[23] E. Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP," in *57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[24] R. Schwartz et al., "Green AI,", arxiv 2019.

[25] U. Gupta et al., "Chasing Carbon: The Elusive Environmental Footprint of Computing," in *HPCA*, 2020.

[26] https://www.apple.com/newsroom/2020/07/apple-commits-to-be-100-percent-carbon-neutral-for-its-supply-chain-and-products-by-2030/ [Accessed December 2022].

[27] https://esg.tsmc.com/download/file/2018_tsmc_csr_report_published_May_2019/english/pdf/e_all.pdf. [Accessed December 2022].