Exploiting 2.5D/3D Heterogeneous Integration for AI Computing

(Invited Paper)

Zhenyu Wang*, Jingbo Sun*, Alper Goksoy†, Sumit K. Mandal‡, Yaotian Liu*,
Jae-sun Seo¶, Chaitali Chakrabarti*, Umit Y. Ogras†, Vidya Chhabria*, Jeff Zhang*, Yu Cao§
*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA
†Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI, USA
‡Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India
¶Department of Electrical and Computing Engineering, Cornell Tech, NY, USA
§Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities, MN, USA

2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC) | 979-8-3503-9354-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ASP-DAC58780.2024.10473875

Abstract— The evolution of AI algorithms has not only revolutionized many application domains, but also posed tremendous challenges on the hardware platform. Advanced packaging technology today, such as 2.5D and 3D interconnection, provides a promising solution to meet the ever-increasing demands of bandwidth, data movement, and system scale in AI computing. This work presents HISIM, a modeling and benchmarking tool for chiplet-based heterogeneous integration. HISIM emphasizes the hierarchical interconnection that connects various chiplets through network-on-package. It further integrates technology roadmap, power/latency prediction, and thermal analysis together to support electro-thermal co-design. Leveraging HISIM with in-memory computing chiplets, we explore the advantages and limitations of 2.5D and 3D heterogenous integration on representative AI algorithms, such as DNNs, transformers, and graph neural networks.

***Index Terms*—Heterogeneous Integration, 2.5D, 3D, Chiplet, ML accelerators, Performance Analysis**

I. Introduction

State-of-the-art artificial intelligence (AI) algorithms proficiently address practical problems across numerous domains. Existing monolithic Integrated circuits (ICs) offer many benefits for AI acceleration with high energy efficiency and small die size. However, emergent AI models with high connection density have a large number of connections between neurons or nodes which require excessive data communication. The AI models demand larger on-chip memory for the weight storage and have higher complexity. This trend amplified energy consumption, computation time, and memory access of monolithic AI acceleration chips. Consequently, monolithic chips will dramatically increase the chip area, fabrication cost, and data communication. Thus 2.5D/3D architectures have been proposed to address the power, performance, and area challenges for advanced technologies. This architecture leverages advanced packaging technology and integrate multiple monolithic chips on the same substrate. 2.5D heterogeneous chiplet-based architectures contain multiple chiplets connected with network-on-package (NoP) facilitated by *the active interposer* or *the embedded bridge*. This network-on-package enables inter-chiplet communication, ultimately delivering higher performance and reduced costs compared to traditional monolithic designs [1]–[3]. However, AI applications such as ChatGPT, require large amounts of data to be processed to achieve real-time responses as rapidly as possible. The 2.5D architecture with lateral interconnect is not enough to provide high bandwidth and fast signal speed between chiplet to chiplet. Hence, the 3D chiplet architecture as shown in Fig. 1 with through-silicon via (TSV) is proposed to offer significant performance enhancements [4]–[6].

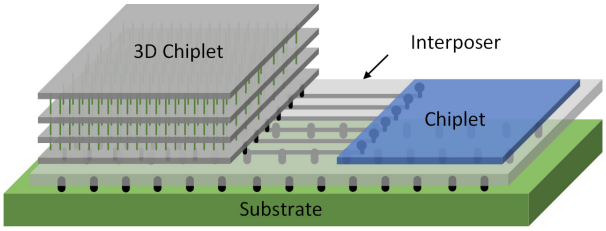

Fig. 1. Overview architecture of heterogeneous 2.5D/3D integration.

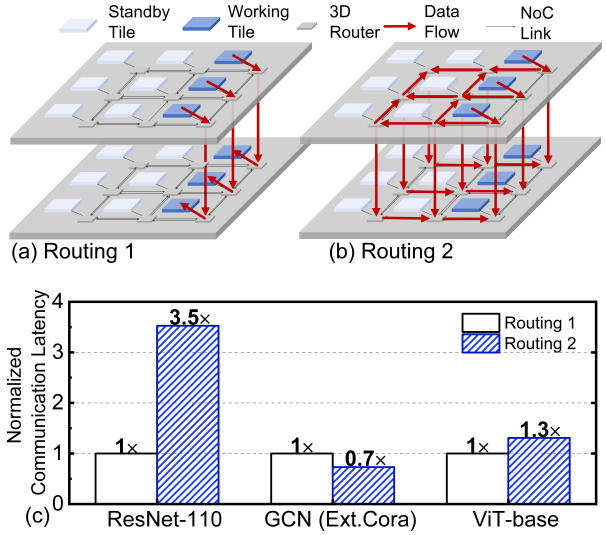

Fig. 2. The comparison of 2 routing methods of 3D chiplet-based architecture for different AI models. (a) Routing method 1 uses only local routers and TSVs to transport data to the tiles in the next tier. (b) In order to use maximum vertical links, routing method 2 uses all the routers and TSVs to transport data to the tiles in the next tier. (c) The comparison of the communication latency for different AI models mapped in 3D chiplet-based architecture with different routing methods.

Fig. 2 shows two data routing methods in 3D chiplet-based architecture for three different AI models. Routing method 1 uses only local routers and vertical TSVs to transport data to the tiers in the bottom tier. Routing method 2 uses maximum vertical TSVs from the entire tier to achieve the largest bandwidth for sending the data from the top tiles to the bottom tiles. It will scatter the data to all the routers through the 2D NoC which introduces extra intra-chiplet NoC latency.

Fig. 2(b), we compare the communication latency which includes the data communication on the 2D NoC and vertical TSVs for two routing methods on three AI models. The tradition convolutional neural network (CNN) model which has 3.52× higher latency with routing method 2 compared with routing method 1. The transformer model, ViT-based on ImageNet, has 1.3× higher latency with routing method 2 compared with routing method 1. However, the graph neural network, 2-layer GCN on Ext.Cora, has lower latency with routing method 2 compared with routing method 1. This is because the data volume of activation in GCN [7] is pretty large and the size of the weight is relatively smaller, and it requires the use of as many vertical interconnects as possible to speed up the communication even though it will add extra intra-chiplet NoC latency.

Several prior studies proposed 2.5D/3D chiplet-based architectures for accelerating AI models [1], [8]–[11]. The authors of [1] analyzed the effect of different generations of 2.5D interconnect on AI models. A chiplet-based benchmarking tool, SIAM, is proposed in [8]. The authors of [9] proposed the 2.5D architecture with 36 chiplets for DNN inference acceleration. The authors of [10] proposed 3D architecture with five-tiers AI accelerator designs. However, all these prior studies only handled one type of architecture – 2.5D or 3D – at a time. None of these works compared the performance of monolithic, 2.5D, and 3D architectures for different AI models, or optimize the configuration of the chiplet-based architecture for a specific AI model.

In this study, we introduce HISIM, a benchmark tool developed for chiplet-based heterogeneous integration (HI). HISIM comprises of various engines including partitioning, mapping and placement, computing unit/processing unit, heterogeneous interconnection, network/routing engine, and thermal analysis. Our proposed simulator can evaluate the performance of AI models under monolithic, 2.5D, and 3D architecture. In HISIM, we integrated the 2.5D/3D interconnection technology roadmap with the cycle-accurate simulation for technology-based design space exploration. For 3D interconnection, we employ a simplified TSV model for electrical parameter extraction. We align this interconnection roadmap with our custom heterogeneous cycle-accurate simulation to assess data communication performance. We demonstrate HISIM's capabilities by conducting experiments on state-of-the-art AI algorithms. Additionally, we analyze and compare the impacts of different placement and routing methods for AI models under 3D architectures. We examine 3D architecture configurations for AI models to facilitate design space exploration and pinpoint the optimal configuration to achieve enhanced performance, power, and area (PPA). The major contributions of this work are as follows:

- We generalize the technology roadmap and electro-modeling for 2.5D/3D interconnect, with the highlight that they are comparable to on-chip interconnect.
- We propose an end-to-end benchmarking simulator, HISIM, for chiplet-based heterogeneous integration (HI) with 2.5D/3D interconnect modeling. HISIM is the first simulator to provide support for the hardware performance evaluation of heterogeneous chiplet-based architectures.
- We use HISIM to explore the design space by varying technical parameters and heterogeneous architectures of the chiplet-based system. We exploit the trade-off of performance metrics under different methods of placement and routing with insight into AI models.

II. Background and Related Work

A. Chiplet-based Heterogeneous Integration

Chiplet-based heterogeneous integration has been proposed to address the manufacturing and cost challenges posed by large monolithic chips. This approach capitalizes on advanced packaging technology to integrate multiple chiplets into the system. Within a chiplet-based system, multiple chiplets are interconnected through a Network-on-Package (NoP) within a silicon interposer or an organic substrate. Various chiplet-based architectures catering to different applications have been proposed. For instance, a high-performance SoC for multichip architecture featuring 8 CPU cores is described in [12]. A silicon prototype system consisting of 36 chiplets with high energy-efficiency is proposed in [9] to accelerate the inference process of deep learning. In [13], a chiplet-based hybrid sparse-dense accelerator that includes CPU and FPGA is proposed to address both the memory throughput challenges and the compute limitations of AI models. As mentioned in [14], 3D IC packaging is realized by using wire bonding easily. However, this interconnection is fragile and introduces high latency. With the development of TSV [15], electrical connections through the silicon substrates can be used for building up the 3D IC integration. 3D chip stacking technologies for CMOS image sensors are proposed in [16]. In [17], the authors proposed the logic-on-logic die stacking, Foveros, which enables a robust face-to-face die 3D connection. These works show the available 3D chiplet-based architecture with high bandwidth and computation density due to the benefit of the development of 3D interconnection.

B. 2.5D/3D Interconnection

In order to connect the chiplets with each other in a 2.5D heterogeneous system, the advanced interconnection technology has been proposed to provide high density and bandwidth. Embedded Multi-Die Interconnect Bridge (EMIB) was proposed to connect two or more dies on the organic package substrate which provides new opportunities for heterogeneous integration [18]. The 2.5D chiplet-based design with Chip-on-Wafer-on-Substrate (CoWoS) technology communicates with each other through 0.5 mm interposer channels using a Low-voltage-In-Package-INterCONnect (LIPINCON) [19]. However, the number of interconnect links in a 2.5D system is constrained by the length of the shoreline or die edge. In contrast, the 3D chiplet-based architecture offers significantly higher bandwidth and a shorter signal distance between chiplets. For instance, Intel's Foveros 3D stacking technology, based on solder μbumps, can support area escape density as 1,600 IO/mm^2. In this work, we focus on Through-Silicon Vias (TSVs), a well-developed interconnection method in 3D chiplet-based integration. In the following section, we present the TSV electro-modeling for multiple generations and integrate it into HISIM. Compared with the traditional processing unit for the computing unit, in-memory computing (IMC) can

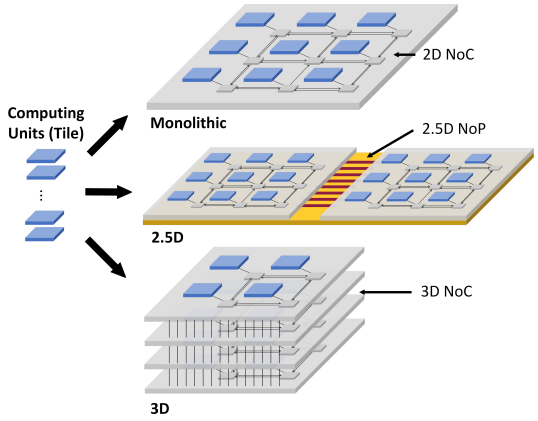

Fig. 3. Three packaging technologies utilized to fabricate computing units (tiles) into monolithic, 2.5D chiplet-based, and 3D chiplet-based chips. The tiles in the monolithic chip are connected via the NoC; The chiplets in 2.5D integration are connected via the NoP; The tiers in 3D integration are connected through the vertical links of 3D NoC.

support parallel computing and high memory density for AI acceleration [20]–[24]. With the increasing size of AI models and datasets, the monolithic IMC design will face the problem of larger chip area, high fabrication cost, and high volume of data communication. Hence, the 2.5D/3D chiplet-based integration is regarded as the effective solution for accelerating large-scale AI models. In this work, we integrate the IMC computing unit into HISIM to study the overall performance.

C. Benchmarking Tools

For monolithic or chiplet-based architecture, design space exploration necessitates comprehensive benchmarking tools to generate reliable performance results. For the DNN inference benchmarking simulator, the architecture includes the IMC crossbars connected by the point-to-point interconnection for the on-chip network communication in [25]. The authors in [8] proposed the simulator for the design space exploration of chiplet-based IMC architectures. This simulator combines the IMC circuit, NoC, NoP, and DRAM units together to provide the overall performance evaluation of 2.5D chiplet-based system. For the 3D chiplet-based simulation, the authors [10] explored the design configuration for multi-stacking IMC accelerator with TSVs and showed the tradeoff between different design schemes and associated thermal problems for different deep learning models. However, these works have limitations as these simulations support only one architecture—like monolithic, 2.5D, or 3D—for analysis. There is a dearth of robust comparisons between different architecture performances to find the optimized configuration. Furthermore, these works typically support only traditional CNN models as benchmarking examples. Our work, HISIM, supports simulation for chiplet-based heterogeneous integration (HI) with 2.5D/3D interconnect modeling. We meticulously model architectural components such as 2D/3D NoC and 2.5D NoP, providing a benchmarking tool that enables design space exploration across different heterogeneous architectures.

III. HISIM: Heterogeneous Integration Simulator

A. HISIM Overview

HISIM is the benchmark tool for the chiplet-based heterogeneous integration which evaluates the performance of monolithic, 2.5D and 3D architectures, as shown in Fig. 4. Based on the user inputs, the HISIM generates the specific

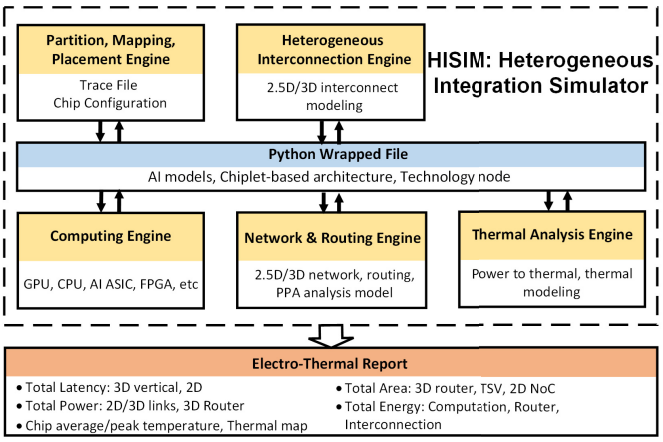

Fig. 4. Overview architecture of HISIM simulator. HISIM consists of four engines: the partition, mapping, and placement, the heterogeneous interconnection, the computing unit, network/circuit, and thermal analysis. The Python wrapped file works as the control tool and data communication among the engines.

TABLE I
DEFINITION OF THE USER INPUTS TO HISIM

User Input	Description
AI Algorithm	
Model Structure	Network Structure Information
Data Precision	Weights and activation precision
Sparsity	Layer-wise sparsity
Computing Engine	
Tech Node	Technology node for fabrication
Processing Unit	IMC, CPU, GPU, etc
Memory Cell	RRAM/SRAM/MRAM
Bits/Cell	Number of levels in memory cell
Memory array size	Memory bank size
Buffer Type	SRAM or Register File
ADC Resolution	Bit-precision of flash ADC
Computing Frequency	Frequency of processing unit
Partition, Mapping, Placement Engine	
Chip Type	Monolithic/2.5D/3D
Chiplet Size	Number of IMC tiles within each chiplet
Total Chiplet Number	Fixed or specific count
Network and Routing Engine	
2D NoC Topology	Mesh or Tree
Routing	XYZ routing, local routing, global routing
2D NoC Channel Width	Number of links per channel
2D NoC Frequency	Frequency of NoC
2.5D NoP/3D NoC topology	Mesh, Tree, 3D-Mesh, etc
2.5D NoP/3D NoC Frequency	Frequency of the NoP/NoC operation
2.5D NoP/3D NoC Channel Width	Number of links per channel
Heterogeneous Interconnection Engine	
Interconect	Technology of 2.5D/3D links
Tech Node	Technology node for Interconnection
Dimensions	Dimensions of the link
Thermal Analysis Engine	
Materials	Thermal conductivity, thickness

heterogeneous chiplet-based architecture and provides hardware performance analysis with respect to the chip area, power estimation, latency, energy efficiency, leakage energy, and processing unit utilization. HISIM focuses on the data communication in 2D/3D NoC and 2.5D NoP and supports the user in defining the chiplet block. In this tool, the user can choose the network topology and floorplan for network-on-chip and network-on-package. By the comprehensive approach, HISIM enables the rapid analysis of PPA and realizes the design-space exploration. The overall simulator is developed by Python and

TABLE II
DESCRIPTION OF THE PARAMETERS OF INTERCONNECTION

Architecture	Interconnect	Parameter	Description
Monolithic/2.5D	Wire (RDL)	l_{wire}	Length of wire
		w_{wire}	width of wire
		t_{wire}	thickness of wire
		p_{wire}	picth of wire
3D	TSV	d_{TSV}	TSV metal diameter
		h_{TSV}	TSV height
		p_{TSV}	TSV pitch
		t_{ox_TSV}	Thickness of TSV insulation layer

C++ programming languages. Table I shows the user inputs and descriptions of the HISIM.

B. HISIM Workflow

There are many components for different functions inside the HISIM. We use the top-level Python wrapper to combine different components with each other and generate the workloads for communication. Each engine works independently with the user inputs. To better comprehend the overall framework of HISIM, we illustrate the overview architecture of HISIM as shown in Fig. 4. From the aspect of technology, users need to select or configure the type of the chiplet. For example, the computing unit within the chiplet can be the IMC crossbar array, the GPU unit, the FPGA array, and any other computation unit. Additionally, the chip architecture and technology node need to be decided. In cases where users select the 3D chiplet-based architecture as their subject of study, interconnect parameters have to be specified as 3D interconnects, and the user must define the parameters for the next step of electro-modeling.

After the decision of technology configuration, the user also needs to decide on the algorithm model as the object of analysis. The precision of the activation and weights need to be decided. Then, HISIM will perform the layer partition, mapping, and placement onto the chiplets of the 2.5D or 3D architectures. HISIM then generates an architecture report detailing the overall chip structure, the required number of chiplets/tiers and processing unit tiles per layer, the placement of chiplets and tiles, and the volume of intra-chiplet and inter-chiplet data movement. Upon completion, HISIM selects the appropriate interconnection engine to calculate the performance of data communication. In the next section, we will detail the functionality of the interconnection simulation. After the simulation is done, the outputs of HISIM will include the overall performance metrics of intra-chiplet and the inter-chiplet for energy, area, and latency.

C. Heterogeneous Interconnection Engine

1) 2.5D/3D interconnection technology and dimensions: In this section, we will show the interconnection of monolithic, 2.5D, and 3D chiplet-based architecture. In traditional monolithic SoC, the segments will be connected by metal wires and routers. In the 2.5D chiplet-based system. the chiplets are connected by carefully designed multiple layers of metal wires. As shown in Table II, we listed out the key dimensions for monolithic and 2.5D interconnection: the metal wire. The size of l_{wire}, w_{wire}, t_{wire} and p_{wire} will define the structure of interconnection and be the key parameters for electro-modeling. As for the 3D interconnection, there are many technologies, like μbump, hybrid bonding and TSV, to realize the data communication between tiers vertically. In HISIM, we focus on building up the electro-modeling of the TSV with d_{TSV}, h_{TSV}, p_{TSV} and t_{ox_TSV}. We have generated the roadmap of TSV based on the lumped circuit elements values of TSV in [26]. As TSV technology advances, the scale of its electrical parameters is becoming increasingly comparable to, and in some cases even smaller than, on-chip interconnects. This technological advancement is crucial, as it opens up the possibility of data transmission via vertical links in the 3D chiplet-based architecture, rather than relying solely on on-chip interconnection. We choose the generation TSV when radius equals to 5 μm as the benchmarking model parameters in the next section simulation and experimental analysis.

TABLE III
DEFINITION OF THE USER INPUTS TO NETWORK AND ROUTING ENGINE

User Input	Description
Technology	
Tech Node	Technology Node for fabrication
On-chip link	2D on-chip NoC links parastics
TSV link	3D vertical links parastics
Frequency	Operation frequency
Network	
Topology	2d NoC, 3D NoC
Routing	3D Routing method: XYZ routing
TSV channel width	Channel number for vertical link
2D NoC channel width	Channel number for 2D NoC
Router port	Input and output port number
Num VCS	Number of virtual channels
VC Buf Size	Buffer size per virtual channels
Data	
Packet size	Number of flits per packet
Flit size	Number of bit of flit

D. Network and Routing Engine

1) Overview Architecture: For AI acceleration, data communication plays a crucial role in the overall hardware performance of the chip [27]. With large volumes of input and output activations being transferred between adjacent layers, an efficient interconnection simulation tool for DNN data communication is required to provide precise performance results and enable quick design space exploration. The traditional NoC simulators, like BookSim [28], can provide the cycle-accurate simulation and performance evaluation of 2D NoC architectures. Orion [29] can deliver power and area models for interconnection and network, while Nirgam [30] does not have the latency report for the 2D NoC simulation.

All these simulators lack adequate TSV modeling for 3D NoC simulation. The latest 3D NoC simulator, Ratatoskr [31], can provide accurate performance results for 3D routers and links through RTL simulation. However, the simulation time is longer if the user wants to do a large range of quick design space exploration. Furthermore, it does not offer trace-based simulation beneficial to layer-to-layer data communication in DNNs. In HISIM, we have customized and extended the traditional cycle-accurate NoC simulation to the 2.5D/3D heterogeneous NoC simulator. We integrated the power and area models from ORION 3.0 into our simulator. The cycle-accurate link delays are generated in our customized 2.5D/3D heterogeneous NoC simulator, and the routing network and models are modified to model the behavior of a 3D NoC network.

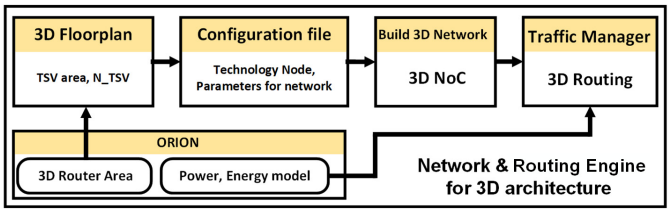

Fig. 5. The workflow and architecture of 3D chiplet-based network and circuit engine

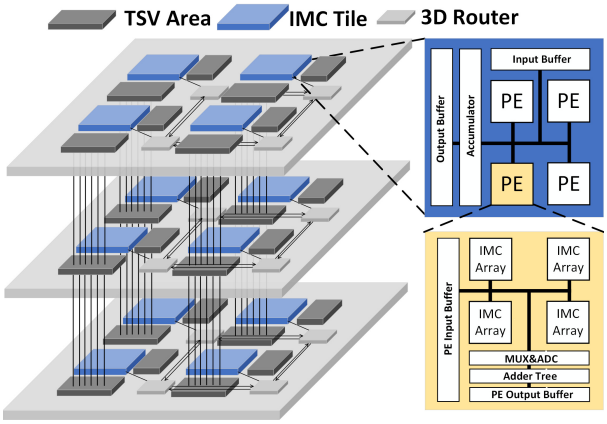

Fig. 6. The 3D IMC chiplet-based architecture

2) Workflow of 3D network and routing engine: The overall workflow of the 3D network and routing engine is depicted in Fig. 5. The trace files produced by the partition, mapping, and placement engine are transferred to the network engine. This engine will build up the floorplan of 3D architecture based on the information of chiplets and interconnections from other engines. Subsequently, the engine constructs the 3D network based on the trace files and configurations. The traffic engine runs the cycle-accurate simulation following the specific routing method to give the latency report. The 3D router is designed to realize the east, west, north, south, up, and down data communication. The latency results will be organized together with the area and power results generated by the ORION model.

E. Computing Engine

In order to build up the complete simulation for heterogeneous integration, the computing engine needs to choose the computation core for obtaining the performance results. Nowadays, the AI model inference can be realized by CPU, GPU or other ASIC design acceleration. The IMC acceleration [32] is one of design techniques for accelerating the inference or training process of AI models. In HISIM, we have integrated the IMC unit inside the computing engine as the choice which is shown in Fig. 6. The IMC tile includes the processing units, accumulator, and buffers. The user can define the different configurations of crossbar size, the number of IMC array, ADC precision and other parameters in the computing engine for further design space exploration.

F. Thermal Analysis Engine

To conduct the thermal analysis, we use the nodal analysis GT = P to predict the temperature (T) given the power consumption P and the conductance matrix (G). More specifically, we divide our design into small sub-nodes based on its functionality and material. The conductance (g) between two physically connected nodes is evaluated based on their contact area, dimensions, and the corresponding conductivity. We add the heat sink layer on top, substrate, and air boundary layer on the bottom to mimic the real working environment of the chip. The top face of the heat sink and the bottom face of the air boundary are set to constant temperature. We assume all other faces are fully thermal isolated thus the heat can only be conducted in the vertical direction. We also collected and calibrated the data points for different material thermal properties from the published results.

IV. Benchmark Study with HISIM

A. Experimental Setup

The primary target of HISIM is to provide the end-to-end simulator for chiplet-based heterogeneous integration (HI) incorporating varying interconnect modeling. This simulator must be sufficiently representative, ensuring extensive coverage of a wide range of heterogeneous architectures from monolithic to 3D structures. By altering the type of the computing unit, users can conveniently execute simulations for the design space exploration of the interconnection network and associated parameters. In our evaluations, we employed various AI models with HISIM, including ResNet-110 (1.7M) on CIFAR-100, DenseNet-121 (7.05M) on ImageNet, 2-layers GCN on Ext.Cora, and ViT-base (86M) on ImageNet. All experiments were performed on the Intel Xeon CPU platform. For the computing engine, we choosed the RRAM-based in-memory computing chiplet as the example to run the simulation which is the same as [25]. In IMC chiplet. the RRAM cell is one bit, the R_{off}/R_{on} ratio is 100, the ADC resolution is 4-bit with 8 columns multiplexed, the operation frequency is 1 GHz [20]. We quantized the weights and activations of AI models as 8-bit and the technology node for IMC core is 32nm CMOS technology node. For the partition and mapping on IMC crossbar array, we keep the same method as the prior works. From the 3D chiplet-based architectural aspect, we follow the same structure in [33].

The 3D NoC using TSVs in the Face-to-Back configuration to build up the 3D multi-chiplet systems. The maximum number of 3D chiplet/tier is not more than 7. There have been recent studies [34] of 12-/16-Hi stacks but that is not included in the scope of HISIM. The 3D 7×7 NoC router in each chiplet not only realizes the data communication inside the chiplet, but also the vertical communication through the TSVs. The operation frequency of 3D NoC is also 1 GHz. The number of virtual channels is 3 and the buffer size per virtual channel is 10. The packet size is 1 and the flit size depends on the channel width. For the TSV array, not only contains the TSVs for the 3D NoC signals but also additional power and ground TSVs for power delivery and heat dissipation. We follow the same percentage of signal in all TSVs [33] as 70%. We choose the parameters of TSV in which the diameter equals 10μm from the roadmap as the benchmarking modeling in the next section's simulations.

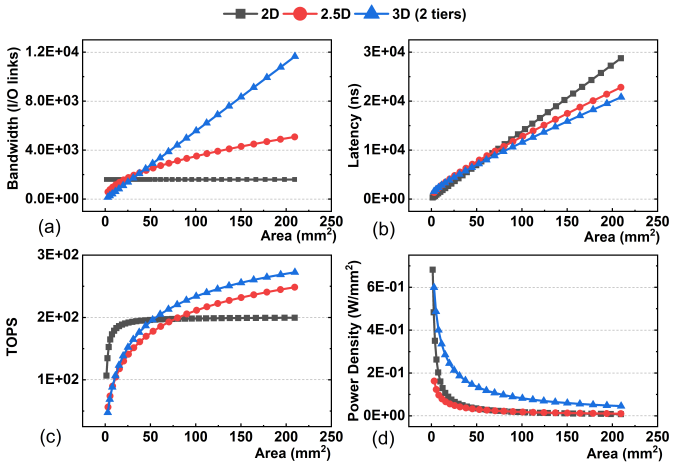

Fig. 7. The chip area scale study for AI accelerator with monolithic, 2.5D, and 3D chiplet-based architecture.

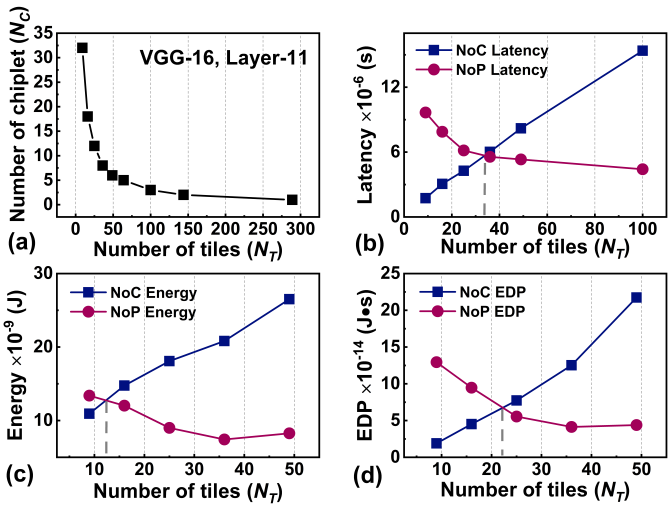

Fig. 8. The trade-off between 2D NoC and 2.5D NoP for different chiplet sizes; more number of tiles means larger size of chiplet; (a) Number of chiplets (N_C) for mapping VGG-16 Layer-11 for different number of tiles (N_T) per chiplet; (b) latency, (c) energy consumption, and (d) EDP of NoC and NoP for different number of tiles (N_T).

B. Scalability Study

Similar to the roofline model [35] of the AI accelerators to evaluate and compare these processors' performance, we scale the chip area and conduct the performance analysis for 2D, 2.5D, and 3D architecture. As shown in Fig. 7, the interconnect bandwidth of 2.5D and 3D increase with the enlargement of the chip area. We assume the number of interconnect links of the monolithic chip remains static, irrespective of changes in the chip area. As shown in Fig. 7(a), the bandwidth of 3D interconnection increases with the area and the bandwidth of 2.5D interconnection increases with the square root of the area. As the chip area increases, the monolithic system will face the problems of larger latency and smaller tera-operations per second (TOPS) compared to the other two systems. However, the 3D architecture composed of 2 tiers, the same as the structure shown in Fig. 2, has higher power density and energy consumption compared to the other two systems. This results in more power dissipation and thermal issues that other researchers are striving to resolve [36], [37].

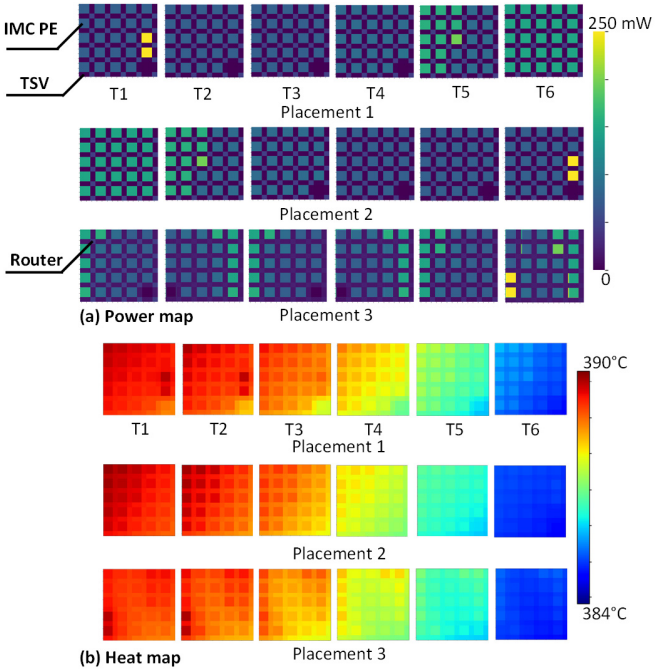

Fig. 9. Power distribution map and heat map of ResNet-110 on CIFAR-100 with three different placement methods in the 3D stacks; the placement 1 maps from top tier to bot tier for higher power consumption tiles close to heat sink; the placement 2 maps from bottom tier to top tier; the placement 3 makes the high-power tiles far with each other.

C. Tradeoff between 2.5D NoP and 2D NoC

In this section, we use a single layer simulation to illustrate why the optimized chiplet configuration achieves a balance between NoC and NoP. We use Layer-11 in VGG-16 as an example by setting the crossbar array size A_x as 64. Based on the layer structure, the total number of tiles required to map the weights of the layer is 288. As shown in Figure 8(a), change in the number of tiles in the chiplet affects the number chiplets that are needed to map this layer. If we change the number of tiles per chiplet from 289 to 9, more chiplets are required to map the layer, the latency, energy, and energy-delay product (EDP) of NoC will decrease as shown in Figure 8(b-d). This is due to the smaller NoC mesh size and the complexity of data communication on NoC. However, the mesh size of NoP will increase making the latency, energy, and EDP of NoP higher. This analysis explains the reason why we can find the optimized configuration of chiplets considering such trade-offs, and choose that as the standard type in the chiplet library which is discussed in the following section.

D. AI Placement Analysis

Thermal-aware task scheduling, placement and optimization are common considerations when designing the 3D system. Due to the heat sink being on the top of the 3D stacks, placement method 1 places the high power consumption tiles of AI models to the top tier. The placement method 2 maps the high power consumption tiles to the bottom tier first which is set up to compare with method 1. The placement method 3 places the high-power consumption tiles far from each other. As shown in Fig. 9, although the power map distribution is different from these three methods, the peak temperatures of these three methods are close to each other: 388 °C, 390

°C, and 385 °C. This similarity is due to the high thermal conductivity of the material and the limited thickness of the chip which is much smaller than the heat diffusion length.

V. Conclusion

In this work, we propose HISIM, a benchmarking tool for 2.5D/3D chiplet-based heterogeneous integration (HI).HISIM emphasizes the hierarchical interconnection and associated data movement in the HI system. It integrates the technology roadmap of 2.5D/3D wires, conducts electrical modeling and analysis, and performs cycle-accurate simulations. Combined with the performance model of various types of computing elements, HISIM provides a flexible and efficient platform for HI system mapping. HISIM also includes a thermal analysis module that enables electro-thermal co-design and design space exploration, which benefits early-stage 2.5D/3D heterogeneous integration design.

Acknowledgments

This work is supported by the Center for the Co-Design of Cognitive Systems (CoCoSys), one of seven centers in Joint University Microelectronics Program 2.0 (JUMP 2.0), a Semiconductor Research Corporation (SRC) program sponsored by the Defense Advanced Research Projects Agency (DARPA).

References

[1] Z. Wang *et al.*, "Ai computing in light of 2.5d interconnect roadmap: Big-little chiplets for in-memory acceleration," *IEEE IEDM*, 2022.

[2] R. Radojcic, *More-than-Moore 2.5 D and 3D SiP Integration*. Springer, 2017.

[3] T. Zhang, K. Kasichainula, D.-W. Jee, I. Yeo, Y. Zhuo, B. Li, J.-s. Seo, and Y. Cao, "Improving the efficiency of cmos image sensors through in-sensor selective attention," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023, pp. 1–4.

[4] E. Beyne, "Heterogeneous system partitioning and the 3d interconnect technology landscape," in *2020 Symposia on VLSI technology and Circuits*, 2020, pp. SC2–2.

[5] S. Sinha, S. Hung, D. Fisher, X. Xu, C. Chao, P. Chandupatla, F. Frederick, H. Perry, D. Smith, A. Cestero *et al.*, "A high-density logic-on-logic 3dic design using face-to-face hybrid wafer-bonding on 12nm finfet process," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 15–1.

[6] J. C. Lee *et al.*, "High bandwidth memory (hbm) with tsv technique," in *2016 International SoC Design Conference (ISOCC)*. IEEE, 2016, pp. 181–182.

[7] S. K. Mandal, G. Krishnan, A. A. Goksoy, G. R. Nair, Y. Cao, and U. Y. Ogras, "Coin: Communication-aware in-memory acceleration for graph convolutional networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 12, no. 2, pp. 472–485, 2022.

[8] G. Krishnan *et al.*, "Siam: Chiplet-based scalable in-memory acceleration with mesh for deep neural networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–24, 2021.

[9] Y. S. Shao *et al.*, "Simba: Scaling deep-learning inference with multi-chip-module-based architecture," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 14–27.

[10] X. Peng *et al.*, "Heterogeneous 3-d integration of multitier compute-in-memory accelerators: An electrical-thermal co-design," *IEEE Transactions on Electron Devices*, vol. 68, no. 11, pp. 5598–5605, 2021.

[11] H. Peng, S. Huang, T. Geng, A. Li, W. Jiang, H. Liu, S. Wang, and C. Ding, "Accelerating transformer-based deep learning models on fpgas using column balanced block pruning," in *2021 22nd International Symposium on Quality Electronic Design (ISQED)*, 2021, pp. 142–148.

[12] N. Beck *et al.*, "'zeppelin': An soc for multichip architectures," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2018, pp. 40–42.

[13] R. Hwang *et al.*, "Centaur: A chiplet-based, hybrid sparse-dense accelerator for personalized recommendations," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2020, pp. 968–981.

[14] L. Wu *et al.*, "The advent of 3-d package age," in *Twenty Sixth IEEE/CPMT International Electronics Manufacturing Technology Symposium (Cat. No. 00CH37146)*. IEEE, 2000, pp. 102–107.

[15] E. Beyne, "The 3-d interconnect technology landscape," *IEEE Design & Test*, vol. 33, no. 3, pp. 8–20, 2016.

[16] Y. Kagawa *et al.*, "3d integration technologies for the stacked cmos image sensors," in *2019 International 3D Systems Integration Conference (3DIC)*. IEEE, 2019, pp. 1–4.

[17] D. Ingerly *et al.*, "Foveros: 3d integration and the use of face-to-face chip stacking for logic devices," in *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019, pp. 19–6.

[18] R. Mahajan *et al.*, "Embedded multi-die interconnect bridge (emib)–a high density, high bandwidth packaging interconnect," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*. IEEE, 2016, pp. 557–565.

[19] M.-S. Lin *et al.*, "A 7-nm 4-ghz arm[1]-core-based cowos[1] chiplet design for high-performance computing," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 956–966, 2020.

[20] M. Imani *et al.*, "Floatpim: In-memory acceleration of deep neural network training with high precision," in *Proceedings of the 46th International Symposium on Computer Architecture*, 2019, pp. 802–815.

[21] G. Krishnan *et al.*, "Hybrid rram/sram in-memory computing for robust dnn acceleration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 11, pp. 4241–4252, 2022.

[22] A. Shafiee *et al.*, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 14–26, 2016.

[23] G. R. Nair, H.-S. Suh, M. Halappanavar, F. Liu, J.-s. Seo, and Y. Cao, "Fpga acceleration of gcn in light of the symmetry of graph adjacency matrix," in *2023 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2023, pp. 1–6.

[24] J. Meng, L. Yang, X. Peng, S. Yu, D. Fan, and J.-S. Seo, "Structured pruning of rram crossbars for efficient in-memory computing acceleration of deep neural networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 5, pp. 1576–1580, 2021.

[25] P.-Y. Chen *et al.*, "Neurosim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 12, pp. 3067–3080, 2018.

[26] J. Cho *et al.*, "Modeling and analysis of through-silicon via (tsv) noise coupling and suppression using a guard ring," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 1, no. 2, pp. 220–233, 2011.

[27] G. Krishnan *et al.*, "Big-little chiplets for in-memory acceleration of dnns: A scalable heterogeneous architecture," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.

[28] N. Jiang *et al.*, "Booksim 2.0 user's guide," *Standford University*, p. q1, 2010.

[29] A. B. Kahng *et al.*, "Orion3. 0: A comprehensive noc router estimation tool," *IEEE Embedded Systems Letters*, vol. 7, no. 2, pp. 41–45, 2015.

[30] L. Jain *et al.*, "Nirgam: a simulator for noc interconnect routing and application modeling," in *Design, automation and test in Europe conference*. IEEE, 2007, pp. 16–20.

[31] J. M. Joseph *et al.*, "Ratatoskr: An open-source framework for in-depth power, performance and area analysis in 3d nocs," *arXiv preprint arXiv:1912.05670*, 2019.

[32] F. Zhang and M. Hu, "Mitigate parasitic resistance in resistive crossbar-based convolutional neural networks," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 16, no. 3, pp. 1–20, 2020.

[33] P. Vivet *et al.*, "A $4 \times 4 \times 2$ homogeneous scalable 3d network-on-chip circuit with 326 mflit/s 0.66 pj/b robust and fault tolerant asynchronous 3d links," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 33–49, 2016.

[34] M. Chen *et al.*, "Low temperature soic bonding and stacking technology for 12-/16-hi high bandwidth memory (hbm)," *IEEE Transactions on Electron Devices*, vol. 67, no. 12, pp. 5343–5348, 2020.

[35] S. Williams, "Roofline: An insightful visual performance model for floating-point programs and multicore," *ACM Communications*, p. 16, 2009.

[36] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3d ics," in *IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004.* IEEE, 2004, pp. 306–313.

[37] T.-Y. Wang and C. C.-P. Chen, "3-d thermal-adi: A linear-time chip level transient thermal simulator," *IEEE Transactions on computer-aided design of integrated circuits and systems*, vol. 21, no. 12, pp. 1434–1445, 2002.