This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Thermal-Aware Design Space Exploration of 3D Systolic ML Accelerators

Rahul Mathur, *Senior Member, IEEE,* Ajay Krishna Ananda Kumar, *Member, IEEE,* Lizy John, *Fellow, IEEE,* and Jaydeep P. Kulkarni, *Senior Member, IEEE*

***Abstract*—ML accelerators have a broad spectrum of use-cases that pose different requirements on accelerator design for latency, energy, and area. In the case of systolic array-based ML accelerators, this puts different constraints on Processing Element (PE) array dimensions and SRAM buffer sizes. 3D integration packs more compute or memory in the same 2D footprint, which can be utilized to build more powerful or energy-efficient accelerators. However, 3D also expands the design space of ML accelerators by additionally including different possible ways of partitioning the PE array and SRAM buffers among the vertical tiers. Moreover, the partitioning approach may also have different thermal implications. This work provides a systematic framework for performing system-level design space exploration of 3D systolic accelerators. Using this framework, different 3D-partitioned accelerator configurations are proposed and evaluated. The 3D-stacked accelerator designs are modeled using hybrid wafer bonding technique with a 1.44 µm pitch of 3D connection. Results show that different partitioning of the systolic array and SRAM buffers in a 4-tier 3D configuration can lead to either 1.1-3.9X latency reduction or 1-3X energy reduction compared to the baseline design of the same 2D area footprint. It is also shown that by carefully organizing the systolic array and SRAM tiers using logic over memory, the temperature rise with 3D across benchmarks can be limited to 6°C.**

***Index Terms*—3D integration, energy-efficient, systolic accelerators, thermal**

I. INTRODUCTION

MACHINE Learning (ML) algorithms are composed of both computationally and memory-intensive matrix multiplication operations. Systolic array architectures [1] achieve high throughput with modest bandwidth for matrix multiplication operations and hence make a good choice for ML acceleration. Systolic array-based ML accelerators have seen deployment in data-centers [2] [3] as well as in mobile platforms [4] [5]. As the ML application space continues to expand with big data and as the Neural Network (NN) models continue to grow bigger to achieve higher accuracy, the accelerators must scale to meet the increasing demands of computation and energy-efficiency.

At the same time, the typical gains in energy-efficiency that dimensional scaling has brought over the past several decades are slowing down [6] [7] [8]. 2D enhanced architectures [9] place dies side-by-side and interconnect them through mediums such as a silicon interposer [10], or embedded bridge [11] [12] to achieve higher interconnect densities compared to mainstream packages. 3D architectures like hybrid wafer bonding [13] [14] directly stack two or more dies on top of each other without using the agency of the package, further reducing distances and increasing interconnect densities between dies. 3D architectures may offer complementary gains to traditional dimensional scaling for achieving high performance, low power, high bandwidth, fast time-to-market, all in a small footprint. Larger 2D dies can be replaced by a few smaller ones with potentially higher manufacturing yields [15] [16]. Besides, 3D allows heterogeneous integration of parts from different technologies instead of having to redesign every component for a specific process [17]. As 3D technologies evolve, increasingly finer pitches of 3D connections become viable [18] [19]. This opens interesting possibilities for designers to partition and fold designs onto multiple tiers [20] [21]. Deep Neural Network (DNN) processing is heavy in computation and data movement [22]. 3D makes it possible to pack more compute or memory in the same 2D footprint while reducing interconnect delay and power by bringing the blocks closer. Hence, 3D provides an opportunity to build powerful and energy-efficient accelerators.

Traditional 2D systolic array design involves careful partitioning of the silicon real estate between the PE units and SRAM buffers to balance the throughput and external memory transfer bandwidth. 3D accelerator design additionally involves the optimal distribution of the increased silicon real estate available in the same 2D footprint between the PE units and memory. Further, the power density of systolic accelerators is high due to their desired high computing capability and closely packed PEs. This is exacerbated in 3D due to higher logic integration density which may lead to worse thermal characteristics [23]. Hence, the designer must take into account the thermal implications when partitioning the accelerator components among 3D tiers. A systematic methodology for navigating the 3D systolic accelerator design space accounting for the thermal issues is necessary. This paper makes the following contributions to address this issue:

- Provide a systematic framework to navigate the design space of 3D systolic array-based ML accelerators under different workload conditions.
- Perform system-level analysis to evaluate and compare different 3D-partitioned accelerator approaches for performance, power, and thermal characteristics.
- Provide insights and takeaways for system designers to perform thermal-aware design of such 3D accelerators.

The remainder of the paper is organized as follows: Section II provides background and prior work on 3D integration technologies and systolic architectures. Section III describes the 2D baseline design and different 3D partitioned configurations. Section IV delineates the simulation framework used to perform performance, power, and thermal analysis. Section

This manuscript was submitted for review on 04/28/2021.

R. Mathur, A. Kumar, L. John and J. P. Kulkarni are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, TX, 78712 USA (e-mail: rahul.mathur@utexas.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

V describes the experimental setup. Section VI presents results from a comparative analysis of different 3D accelerator configurations. Section VII provides concluding remarks.

II. BACKGROUND AND PRIOR WORK

This section provides a brief overview on various 3D IC technologies. A refresher is also provided on the basic principles of a systolic array-based DNN accelerators.

A. Overview of 3D Integration Technologies

Traditionally, two or more dies are flip-chip attached to an organic package substrate and interconnected with the agency of the package. Certain 2D enhanced (also referred to as 2.5D integration) utilize an interposer made of silicon, glass or, ceramic for high-density communication between separate dies mounted side by side (figure 1a). The interposer may contain Through Silicon Vias (TSVs) [24] which are essentially holes etched out in the silicon wafer and then filled with a conductive metal like copper.

3D stacked ICs involve a die containing TSVs attached to the package substrate using conventional flip-chip technology and a second die, fabricated separately and bonded to the first die using micro bumps [25] or hybrid wafer bonds [13]. This leads to a back-to-face (B2F) configuration, as the back of the first die is bonded to the face of the second die (figure 1b). Similarly, other configurations like back-to-back (B2B) and face-to-face (F2F) are possible, especially when multiple dies are stacked in this manner. Compared to 2.5D (lateral) integration, 3D stacking worsen thermals due to increased power density with die overlap, and heat dissipation from tiers away from the heat sink is a challenge [26].

Monolithic 3D ICs consist of multiple device layers fabricated sequentially on the same die and connecting using Monolithic Inter-tier Vias (MIVs) which are essentially the same size as intra-tier vias [27]. MIVs offer better parasitics and a higher integration density compared to TSVs due to their smaller size [28]. Since monolithic 3D enables the finest pitch of 3D connection, it holds the most promise. However, more breakthroughs in low-temperature processing to fabricate transistors in the upper layers while preserving the transistors and Back end of line (BEOL) of the lower layer are desired [29]. Monolithic 3D suffers from limited lateral thermal conductivity due to the absence of substrate on upper layers. Besides, high device integration density and thin layers lead to strong tier-to-tier thermal coupling [30].

This work uses 3D stacked ICs using hybrid wafer bonding technology to model the design of 3D ML accelerators.

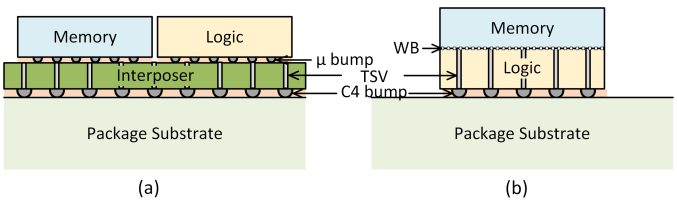


Fig. 1: (a) 2D enhanced: Side-by-side die stacked over interposer (2.5D) and (b) 3D: Memory die stacked directly on the logic die using hybrid wafer bond (WB) technology.

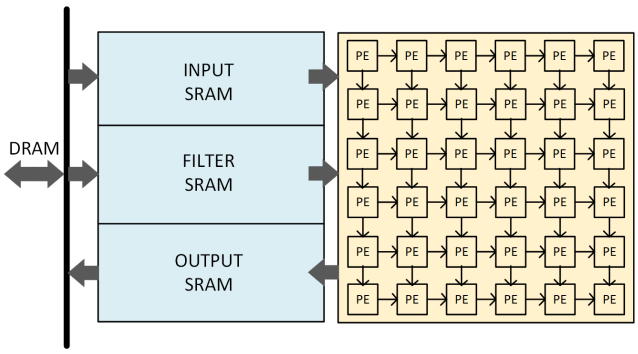


Fig. 2: A typical systolic array-based accelerator system.

Nonetheless, some of the ideas discussed in this paper around efficient partitioning of a ML accelerator design into 3D tiers can be helpful to design for other 3D IC technologies as well. Next, we will discuss the basic principles of operation of a systolic array-based ML accelerator system.

B. Systolic ML Accelerators

A systolic array consists of a simple and regular grid of PEs wired together using the nearest-neighbor interconnect [31] [32]. Data from banked scratchpad memory made of SRAMs is injected from the edges of the array in a rhythmic pipelined manner (similar to a systolic beat). The PEs perform the same operation on their inputs, typically multiply and accumulate, and pass the intermediate results or the original inputs to adjacent PEs. The key idea is to exploit data re-use so that fewer data transfers from memory are needed. Further, purely local data movement (neighbor to neighbor) means simpler interconnect and control. PEs operating in parallel achieve high computational concurrency. Moreover, systolic architectures are modular making them easy to floorplan and scale. Figure 2 shows a high-level diagram of a typical systolic system with an array of PEs and scratchpad memory for storing input, filter, and output.

DNN computation is a highly parallel workload of dense matrix multiplication operations between the input matrix (or the output of the previous layer) and the filter matrix. Systolic array architectures can effectively leverage the abundant data reuse opportunity in DNNs by using their local data shifting movement and keeping the PEs busy to provide high throughput. Each PE performs a simple multiply-and-accumulate (MAC) operation, while data is streamed through the array in a pre-defined synchronized dataflow. An example of dataflow is weight stationary where weights of the filter matrices corresponding to each DNN layer are pre-loaded from the filter memory into the systolic array before any matrix multiplication operation is performed. Input data is then streamed in from the input memory and the array elements perform matrix multiplication with the weights already stored in them. The output data is continuously accumulated, passed through activation and/or quantization functions before eventually being stored in the output memory. The cost of fetching data from memory is amortized over several compute cycles leading to high energy-efficiency. The systolic array has been utilized as the underlying fabric to achieve orders of magnitude gains in performance and energy-efficiency over traditional CPUs, and GPUs for DNN acceleration [2] [3] [5].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

III. 2D AND 3D SYSTOLIC ARRAY ACCELERATORS

Traditional 2D systolic array design involves selecting an appropriate size and dimension of the PE array as well as the size of memory, which would store the NN input feature maps (IFMAP SRAM), filters (FILTER SRAM), and output feature maps (OFMAP SRAM). In theory, a designer can choose an arbitrary number of PEs. One would expect that a large number of PEs improves the local data reuse, especially for compute-limited (or large) networks. This may lead to an increase in the throughput of operations, thereby reducing the number of total cycles (latency) needed to process the network. However, for applications targeting small networks, a large PE array can increase the latency of NN computation as inputs have to traverse the entire length and height of the array before the output is ready. Regarding buffer sizing, a larger SRAM would minimize expensive data transfers to main memory (DRAM). But again, over-provisioned SRAMs can lead to area and cost inefficiencies. In summary, designers must consider the aforementioned trade-offs for both the PE array and SRAM buffer sizes, keeping in mind the target application workload to achieve an optimal design. For this study, the baseline 2D accelerator was selected to have a 32x32 PE array and 128KB of Filter, IFMAP, and OFMAP SRAM each, which is representative of common DNN inference use-case [4].

3D systolic accelerator design further involves distributing the additional silicon real estate available within the same 2D footprint between PE elements or SRAM buffers to balance network throughput and external memory transfers. Moreover, the partitioning method of the PE array and SRAM buffers among the vertical tiers may have thermal implications. In order to evaluate and compare 3D accelerators with different partitioning styles, design points described in Table I are selected. The 3D configurations considered were limited to 4 stacks of PE array or SRAM. Increasing SRAM stacks has diminishing returns in energy reduction, and increasing PE stacks leads to worsening thermals, as explained in section VI. It must be noted that while a 4x larger 2D design with increased compute or memory resources is possible, a 4-tier 3D system packs equivalent resources in the same footprint as the baseline 2D accelerator. The physical design of a 3D system will incur lower 2D interconnect delay and power due to shorter distances and fewer buffers compared to a 4x larger 2D system but may incur an additional 3D interconnects delay and power.

TABLE I: List of 2D and 3D accelerator configurations.

#	PE	SRAM	Description
1	32x32	128KB	2D Baseline
2	64x64	128KB	4-Stack PE next to 1-stack SRAM
3	32x32	512KB	1-stack PE next to 4-stack SRAM
4	32x32	512KB	1-stack PE under 4-stack SRAM
5	32x32	512KB	1-stack PE over 4-stack SRAM
6	64x64	512KB	scale-up, 4-stack PE, 4-stack SRAM
7	4x(32x32)	4x(128KB)	scale-out, 4-stack PE, 4-stack SRAM

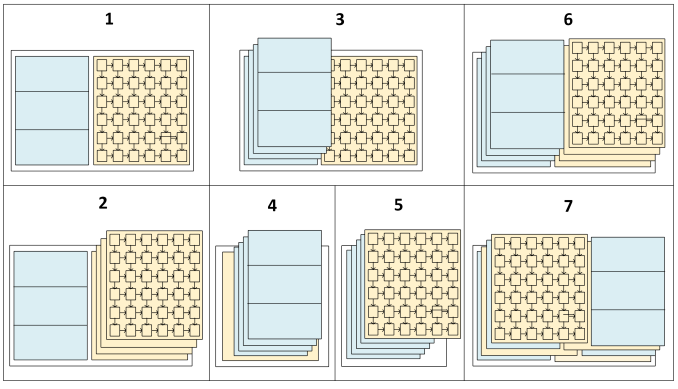


Fig. 3: High-level floorplan showing different approaches of partitioning SRAM buffers (blue) and PE array (yellow) in the 2D and 3D accelerator configurations.

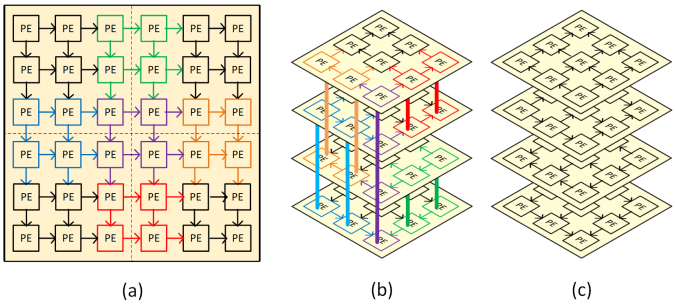


Fig. 4: (a) An example PE array in 2D (b) PE array folded in 3D with vertical connections between PEs across tiers (config 2 and 6) (c) Separate smaller PE arrays operating independently (config 7).

3D configurations selected for further analysis include multiple PE array tiers (configuration 2) or multiple SRAM tiers (configuration 3-5), a scaled-up version (configuration 6), and a scaled-out version (configuration 7) of the 2D baseline accelerator. The floorplans for all design points are shown in Figure 3. It should be noted that configurations 3, 4, and 5 have the same amount of overall compute and memory resources but differ in the method of how these resources are partitioned among vertical tiers. Scaling-up simply means a larger system folded into multiple tiers, while scaling-out means multiple smaller systems in separate tiers [33]. In contrast to configuration 6, the different tiers in configuration 7 do not share the same SRAM and only share an off-chip DRAM. As shown in figure 4(c), the scale-out version does not require any connections in the vertical direction between PE elements in different tiers as the four systolic arrays operate independently in this configuration. Vertical connections would still be needed to transfer the data from DRAM to SRAM in different tiers and for power and ground lines.

IV. SIMULATION FRAMEWORK

The simulation framework developed and used in this work is depicted in Figure 5. It comprises two flows which are explained in this section.

A. *Power and performance analysis flow*

An open-source simulator SCALE-Sim [33] is chosen for the power and performance analysis. Accelerator design parameters such as PE array dimensions, SRAM buffer sizes, and dataflow can be selected and mapped to a list of configuration

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

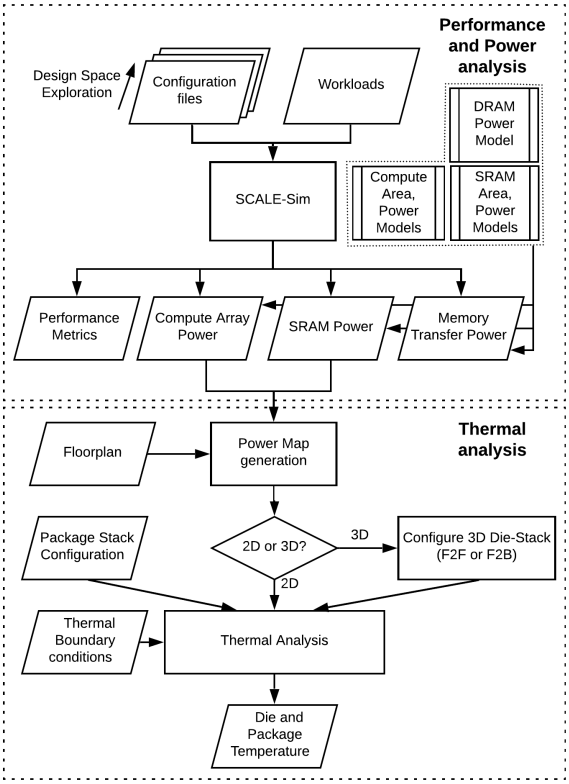


Fig. 5: Simulation framework for 3D systolic accelerators.

files. Simulation benchmarks are translated to topology files having a layer-wise description of the network. The simulator runs a stall-free DNN inference and after processing the entire network, reports the latency in cycles, array utilization, SRAM accesses, DRAM accesses, and DRAM bandwidth requirements.

The power of different configurations is computed from the layer-wise average utilization of the PEs and average bandwidth for SRAM and DRAM reads/writes provided by SCALE-Sim in conjunction with the technology data from [34] (Table II). DRAM accesses can contribute a major part of the total energy [35]. For 3D accelerators, the DRAM transfers may incur an additional energy overhead in transferring data to accelerator components in different tiers. The energy per bit overhead for F2F is reported as 0.013 pJ at nominal voltage [14]. The energy overhead of F2B over F2F is reported as 12X [36]. Hence, to incorporate an average case impact of vertical interconnect energy on the overall DRAM access energy of a 4-tier system, 1.35 pJ per byte (one F2F, one F2B) is added to all DRAM transfers of 3D accelerator configurations.

Power consumed in the PE array is calculated using the following equation:

TABLE II: Technology data from [34] used in conjunction with SCALE-Sim outputs for power calculations.

	PE	SRAM	DRAM
Tech. node	14/16 nm	14/16 nm	28 nm
Energy	0.3 pJ	[1.1, 1.5] pJ	120 pJ
Area	525 um2	32502 um2/32 KB	N/A (off-chip)

$$P_{PE} = \frac{\sum_{i=1}^{n}(util(i) * arr\_h * arr\_w * e\_mac * cyc(i))}{cycles * \frac{1}{freq} * 100} \quad (1)$$

where, n is the number of layers in the network, util(i) is the average utilization of the PE array for computing layer i (between 0-100), cyc(i) is the number of cycles taken for computing layer i, arr_h and arr_w are the PE array height and width respectively, freq is the frequency of operation, (e_mac) is the energy consumed per 8-bit multiply-accumulate (MAC) operation. The e_mac of 0.3 pJ (Table II) is per cycle energy consumed in the PE at 1 GHz, based on a place-and-routed design of an 8-bit precision MAC in 16nm process node [34].

SRAM power is calculated using the following equation:

$$P_{SRAM} = \frac{\sum_{i=1}^{n}((srd\_bw(i) * e\_srd + swt\_bw * e\_swt) * cyc(i))}{cycles * \frac{1}{freq}} \quad (2)$$

where, n is the number of layers in the network, srd_bw(i), swt_bw(i) are the average SRAM read and write bandwidth in bytes per cycle for the execution of layer i, cyc(i) are the number of cycles taken for computing layer i, (e_srd) and write (e_swt) are the SRAM energy consumed in access of byte-wide data. The e_srd of 1.1 pJ and e_swt of 1.5 pJ (Table II) is based on 32KB SRAM macros generated from an industry-standard memory compiler at 16nm and takes both dynamic and static energy into account [34].

DRAM power is calculated using the following equation:

$$P_{DRAM} = \frac{\sum_{i=1}^{n}((d\_if(i) + d\_filt(i) + d\_of(i)) * e\_mem * cyc(i))}{cycles * \frac{1}{freq}} \quad (3)$$

where, n is the number of layers in the network, d_if(i), d_filt(i), d_of(i) are the average bandwidth to access input feature map, filter, and store output feature map in DRAM for the layer i respectively, and (e_mem) is the DRAM energy consumed per byte access. The e_mem of 120 pJ (Table II) is based on off-chip DRAM accesses energy per byte assuming an LPDDR3 interface [35].

Performance in terms of latency of different 2D and 3D configurations is computed from the layer-wise cycle count provided by SCALE-Sim. For configurations 1-6, the total number of cycles to complete the entire benchmark is computed by summing the cycles taken to complete each network layer. Since the computation in the vertical tiers is parallel in configuration 7, the sum of cycles per network layer can be directly computed by simulating a single tier. Performance in terms of throughput can be calculated in Tera Operations Per Second (TOPS) using the following equation:

$$TOPS = \frac{util * arr\_h * arr\_w * 2}{\frac{1}{freq} * 100} \quad (4)$$

where, util is the average utilization of the PE array for computing the NN (between 0-100), arr_h and arr_w are the PE array height and width, respectively, and freq is the frequency of operation. The delay overhead of 3D F2F vertical interconnect can be $\sim$5 ps at nominal voltage [14]. The energy overhead of a F2B connection (through TSVs) over F2F is 3.2X [36]. Hence, to incorporate a worst-case impact of the vertical interconnect delay on the frequency of a 4-tier system, 42 ps (two F2F, two F2B) is added to the cycle time (1/freq) of 3D accelerator configurations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

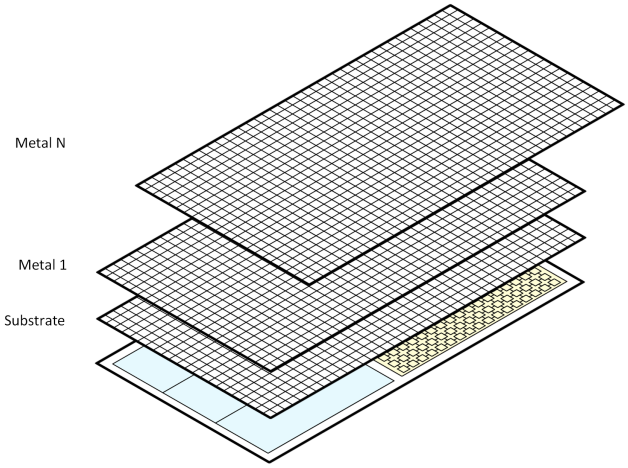


Fig. 6: Tile-based power map used for thermal analysis.

B. Thermal analysis flow

To the first order, the temperature rise in 3DIC is primarily proportional to the effective power density in the 2D footprint [23]. Floor plan dimensions of different 2D and 3D configurations are calculated based on the PE and SRAM area at 14/16 nm technology node from table II. A spatial tile-based power map is created for each tier by using the power data computed for PE and SRAM regions in conjunction with the respective floor plan dimensions. Figure 6 depicts a typical tile-based power map which is essentially a division of the entire tier into equal-sized tiles. The power of each tile is the sum of the power associated with the blocks within the tile. The power map contains the metal density and thermal conductivity properties of all the layers in the BEOL stack. Abstracting the power consumed by the PEs and SRAM in terms of per-tier power maps allows us to mix and match different tiers and build and analyze thermal characteristics for different 3D configurations with relative ease.

Cadence Celsius Thermal Solver [37] is used to run static thermal simulations. The tool uses the power map file along with a complete physical description of the package stack-up, bumps, molding compound, lid, thermal-interface material (TIM), and a detailed description of the vertical stack, i.e., devices, interconnects, and dielectrics along with their thermal conductivity properties. The package comprises 10 build-up layers with overall dimensions of 10x10 mm^2 with an 11x11 mm^2 copper lid on top. TSVs of diameter 5 µm are modeled at every 50 µm in the die stack-up. Thermal simulations are run for different benchmarks with the same package and die size assumptions maintained for all the configurations for a fair comparison. However, a significant change in package thermal design power (TDP) (for instance, configurations 2-7 vs. configuration 1), the heat spreader dimensions may need to be redesigned, and boundary conditions may have to be re-calibrated. Setting up realistic boundary conditions for the tool is critical for getting accurate results. Thermal boundary conditions calibrated with actual hardware measurement data using on-die temperature sensors are sourced from [38]. The tool generates thermal heat maps and maximum temperature data of different dies in each configuration.

V. Experimental Setup

SCALE-Sim is configured with micro-architecture features like PE array dimension, aspect ratio, memory buffer sizes for different 2D and 3D accelerator configurations listed in Table I. The simulator, by default, only supports a 2D systolic configuration. 3D design points of configurations 2-6 can be mapped to SCALE-Sim using their respective PE and SRAM sizes as specified in table I. Configuration 7 is equivalent to four separate systolic systems and can be mapped to SCALE-Sim with PE and SRAM size of configuration 1 with the benchmarks split 4-way along their output channels. The dataflow is set to weight stationary. Although this limits the design space explored, it still enables for a like to like comparison between different 3D accelerator configurations. The topology files having a layer-wise description of the network like input and filter dimensions, input channels, number of filters, and strides are setup for SCALE-Sim for some common NN benchmarks like AlexNet [39], AlphaGo Zero [40], Deep Speech 2 [41], Faster R-CNN [42], GoogLeNet [43], Neural Collaborative Filtering (NCF) [44], ResNet-50 [45], Sentiment Seq-CNN [46], and Transformer [47]. The geometric mean of results from all benchmarks is included to illustrate the overall difference between configurations across all benchmarks. The metric for performance is the number of cycles required to process the benchmark (measure of latency) and TOPS (measure of throughput). The metric for energy-efficiency is TOPS/W. The metric for thermal is the maximum temperature increase in °C relative to the coolest point of the 2D baseline.

VI. Results

This section presents the simulation results comparing different 3D accelerator configurations. Insights are drawn for optimal partitioning strategy for energy, performance and thermal for different network workloads.

A. Energy

Intuitively, it can be said that stacking multiple SRAM tiers would lower the DRAM transfers bringing down the total energy (Figure 7), especially for memory-limited networks.

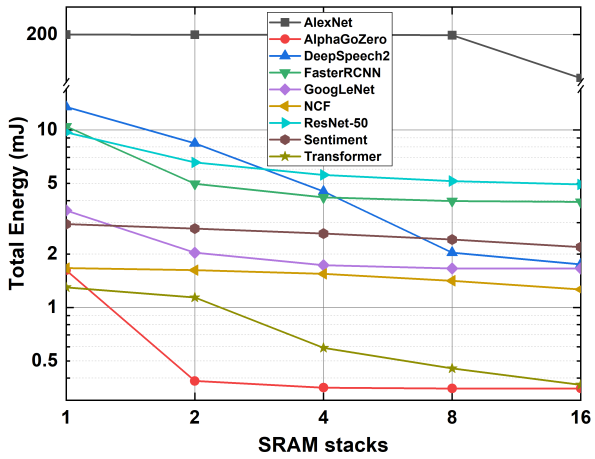


Fig. 7: Reduction in total energy by sweeping SRAM stacks of configuration 3 for different NN benchmarks (log scale).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

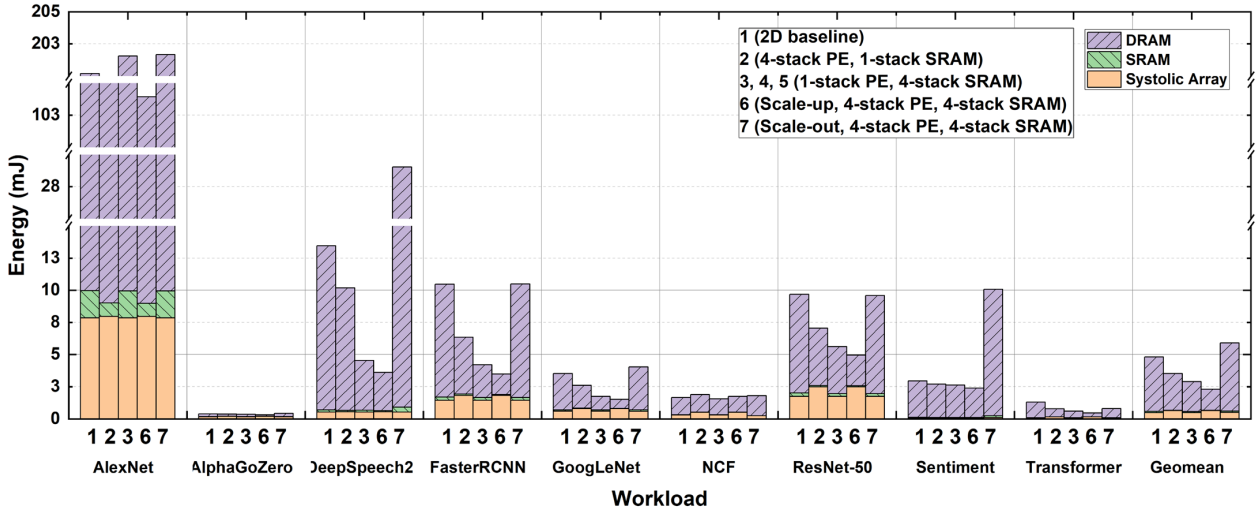


Fig. 8: Energy comparison among configurations for different neural network workloads. The DRAM energy split includes the vertical interconnect energy overhead for the 3D configurations.

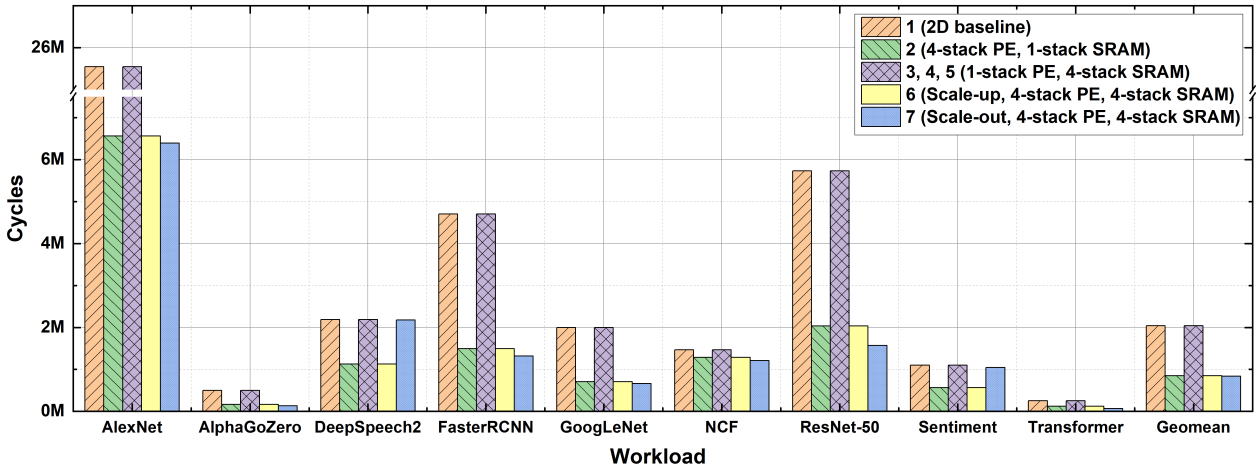


Fig. 9: Latency comparison among configurations for different neural network workloads

Figure 8 compares the total energy of configurations 1-7 listed in Table I across different benchmarks and also illustrates the breakdown of energy between computations, SRAM, and DRAM transfers. As expected, configurations 3-6 which contain 4-stack SRAM reduce the total energy to process the network compared to configuration 1 (2D baseline). However, the energy reduction factor varies widely between benchmarks from 1.0x for NCF to 3.8x for Deep Speech 2. NCF being relatively small already fits within a single SRAM stack and additional SRAM stacks in 3D bring no benefit. Configuration 6 (scale-up) achieves the lowest energy since it also contains 4 stacks of PEs along with 4 stacks of SRAMs increasing the local data reuse within the PEs hence minimizing both SRAM and DRAM transfers. Configuration 7 (scale-out) operating on partitioned output channel requires input feature maps to be duplicated in the SRAMs, causing multiple DRAM accesses to fetch the same input data leading to high total energy.

B. Performance

The number of cycles taken to complete a benchmark should decrease with the increase in the number of PEs, especially for compute-limited (large) networks. As expected, figure 9 shows that configurations 2, 6, and 7 which contain 4-stack PE arrays take fewer cycles to process the network compared to configuration 1 (2D baseline). However, the speedup varies widely between benchmarks from 1.1x for NCF to 3.9x for AlexNet. NCF has much smaller layer features like IFMAP dimensions compared to AlexNet and is unable to utilize the additional PE tiers to achieve any more compute parallelism. Configuration 7 (4x scale-out of 2D) shows slightly better

TABLE III: Power-Performance comparison of accelerator configurations for geomean of all benchmarks

Configuration	TOPS	TOPS/W
1 (2D baseline)	1.59	0.64
2 (4-PE, 1-SRAM)	4.76	1.05
3, 4, 5 (1-PE, 4-SRAM)	1.53	0.98
6 (scale-up: 4-PE 4-SRAM)	4.76	1.53
7 (scale-out: 4-PE 4-SRAM)	3.74	0.50

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

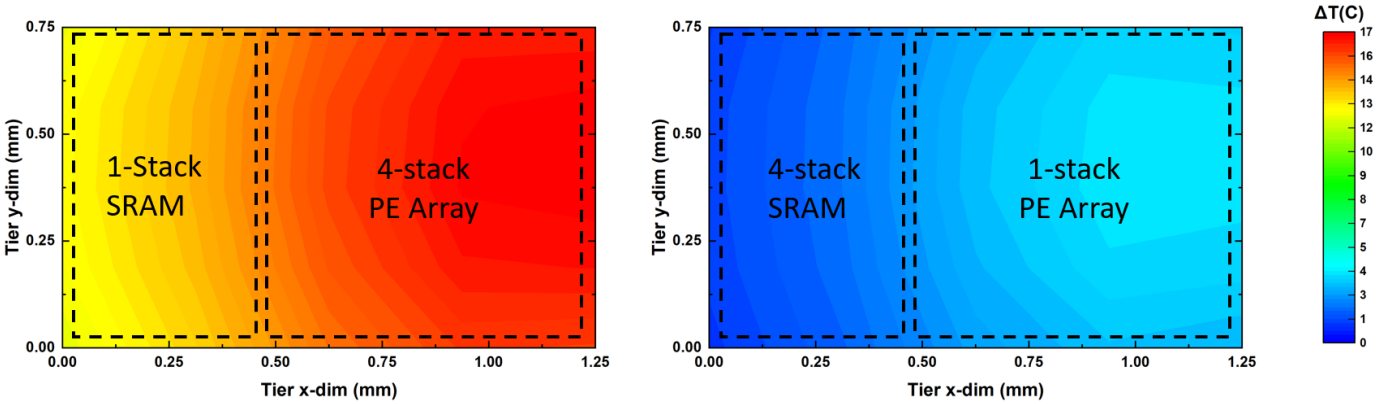


Fig. 10: Heat maps under the ResNet-50 benchmark for (a) Configuration 2 (4-stack array) (b) Configuration 3 (4-stack SRAM). The tier dimensions are in mm. All temperatures are relative to the coolest point on configuration 1 (2D baseline) for ResNet-50.

TABLE IV: Maximum system temperature for different configurations across all benchmarks relative to the coldest point in 2D Baseline for Sentimental Seq-CNN benchmark

Configuration	ΔT(°C)								
	AlexNet	AlphaGoZero	DeepSpeech2	FasterRCNN	GoogLeNet	NCF	ResNet	Sentiment	Transformer
1 (2D baseline)	4.4	4.0	3.3	4.0	3.9	2.1	3.9	0.3	4.1
2 (4-PE next to 1-SRAM)	23.5	21.8	9.1	22.4	20.4	6.5	22.3	2.3	22.3
3 (1-PE next to 4-SRAM)	7.0	6.5	5.3	6.6	6.4	3.8	6.4	0.8	6.3
4 (1-PE under 4-SRAM)	7.2	6.6	5.5	6.7	6.6	3.9	6.6	0.8	6.5
5 (1-PE over 4-SRAM)	5.6	5.1	4.2	5.2	5.0	2.9	5.1	0.5	4.9
6 (scale-up 4-PE 4-SRAM)	24.8	21.5	9.0	22.2	20.3	6.5	22.1	2.1	21.9
7 (scale-out 4-PE 4-SRAM)	23.4	21.4	5.8	20.0	16.2	2.4	19.9	2.8	20.9

performance than configuration 6 (4x scale-up of 2D) for some benchmarks such as AlexNet, AlphaGo Zero, and ResNet-50. This is due to fewer cycles for filling up the smaller independent PE arrays of configuration 7 compared to a single larger folded PE array of configuration 6 which suffers from this overhead at the start of computation of each layer. For other benchmarks such as Deep Speech 2, which contain a small number of output channels and large input feature maps, configuration 7 loses its advantage and suffers from low PE utilization. The power-performance in TOPS and TOPS/W (including the delay and energy overheads of the vertical interconnects for 3D configurations) is presented in table III.

C. Thermal

Figure 10 shows the steady-state heat maps of configuration 2 (4-stack PE array) and configuration 3 (4-stack SRAM) to highlight the difference in thermal characteristics of logic-over-logic and memory-over-memory. Both configurations are running the ResNet-50 benchmark. The temperature values are relative to the coolest point on configuration 1 (2D baseline). The heat maps clearly emphasize that the PE array part of the die runs hotter by around 5°C. It can be further observed that the maximum temperature of configuration 2 is about 13°C higher than configuration 3. This is because the average power density of the 3D stack of PE array is higher compared to the SRAM stack. Table IV compares the maximum temperature rise of different configurations across all benchmarks. The benchmarks have varied size of underlying NN model leading to different average array utilization and SRAM accesses causing different rise of temperatures. Configurations 2, 6, and 7 which employ 3D stacking of the PEs (logic-over-logic) suffer from a temperature rise of up to 24.8°C relative to the coolest point on configuration 1.

The increase in temperature can have an impact on the overall energy of the accelerator. For example, assuming the coolest point on the 2D baseline to be 75°C, an increase in temperature by 25°C has a marginal effect on transistor on-state current but increases the off-state current by 1.9X (Figure 11). Configuration 7 partially avoids overlapping hotspots by staggering the PE array and SRAM between tiers but fares only slightly better. Configuration 3 and 4 which stacks mul-

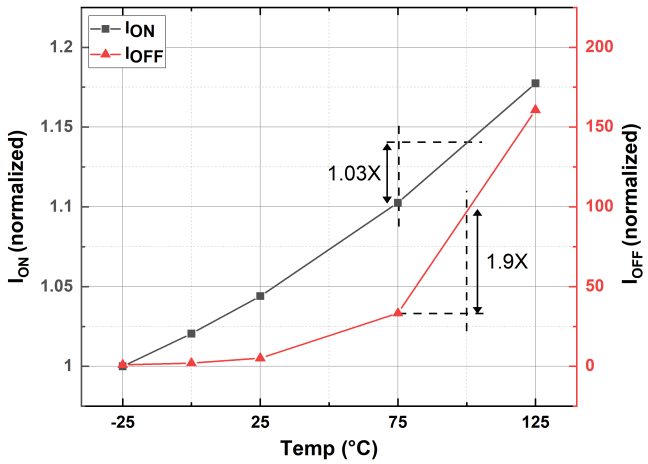


Fig. 11: Effect of temperature rise on the on-state current (black line) @SS/V_{NOM}-10% and off-state current (red line) @FF/V_{NOM}+10% of a transistor with standard V_{TH} option at 14/16nm (0.8V V_{NOM})

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3092436, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

tiple tiers of SRAM are only up to 7.2°C hotter. Furthermore, changing the ordering and stacking the PE array on top of the SRAM stack as in the case configuration 5 (logic-over-memory) limits the temperature rise to only up to 5.6°C making it the best choice from a thermal standpoint. The reason behind this is that the tier containing PE array is significantly hotter than ones containing SRAM and placing it on top reduces its relative proximity to the heat sink.

In summary, 3D stacking of PE arrays (configurations 2, 6, and 7) can reduce the latency of the network computation, but the speedup depends on the network size. Further, these configurations suffer from the worst thermal characteristics due to logic-over-logic stacking. On the other hand, stacking multiple SRAM tiers (configurations 3, 4, 5, and 6) lowers the DRAM transfers making them a good choice where energy-efficiency is important. Furthermore, stacking PE array on top of the SRAM stack (configuration 5) in a logic-over-memory fashion can not only achieve low energy but also mitigate the thermal impact of 3D.

VII. CONCLUSION

Systolic accelerators have been deployed for training and inference, on edge devices as well as on the cloud for a wide variety of workloads. These use cases may constrain accelerator requirements for latency, energy, and area differently. 3D integration packs more compute or memory in the same 2D footprint allowing more powerful and energy-efficient accelerators. However, it also presents more options to the designer for partitioning the PE array and memory among 3D tiers. Since different choices may have different performance, power, and thermal implications, it becomes imperative for designers to understand the trade-offs under different application workload conditions. In this work, a systematic methodology for navigating the 3D systolic accelerator design space is presented. Using this framework, 3D configurations with different partitioning styles are evaluated and compared providing several insights and takeaways for designers. This work can pave the pathway for future thermal aware 3D systolic accelerator designs.

REFERENCES

[1] S. Kung. Vlsi array processors. *IEEE ASSP Magazine*, 2(3):4–22, 1985.

[2] N. P. Jouppi et al. In-datacenter performance analysis of a tensor processing unit. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1–12, 2017.

[3] Xilinx. accelerating dnns with xilinx alveo accelerator cards. technical report, 2018. https://www.xilinx.com/support/documentation/white$_p aper s/wp504 - accel - dnns.pdf$.

[4] Y. Chen et al. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2):292–308, 2019.

[5] J. Song et al. 7.1 an 11.5tops/w 1024-mac butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile soc. In *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 130–132, 2019.

[6] T. N. Theis and H. . P. Wong. The end of moore's law: A new beginning for information technology. *Computing in Science Engineering*, 19(2):41–50, 2017.

[7] A. Mocuta et al. Enabling cmos scaling towards 3nm and beyond. In *2018 IEEE Symposium on VLSI Technology*, pp. 147–148, 2018.

[8] G. Yeric. Ic design after moore's law. In *2019 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–150, 2019.

[9] Interconnects for 2d and 3d architectures (hir). https://eps.ieee.org/images/files/HIR$_2$020/$ch22_2 D - 3D.pdf$.

[10] M. Lin et al. A 7-nm 4-ghz arm¹-core-based cowos¹ chiplet design for high-performance computing. *IEEE Journal of Solid-State Circuits*, 55(4):956–966, 2020.

[11] R. Mahajan et al. Embedded multidie interconnect bridge—a localized, high-density multichip packaging interconnect. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 9(10):1952–1962, 2019.

[12] A. Podpod et al. A novel fan-out concept for ultra-high chip-to-chip interconnect density with 20-µm pitch. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*, pp. 370–378, 2018.

[13] D. W. Fisher et al. Face to face hybrid wafer bonding for fine pitch applications. In *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, pp. 595–600, 2020.

[14] S. Sinha 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)*, Dec 2020.

[15] G. Yeric. Moore's law at 50: Are we planning for retirement? In *2015 IEEE International Electron Devices Meeting (IEDM)*, pp. 1.1.1–1.1.8, Dec 2015.

[16] International roadmap for devices and systems, 2020. https://irds.ieee.org.

[17] T. Wu et al. Low-cost and tsv-free monolithic 3d-ic with heterogeneous integration of logic, memory and sensor analogy circuitry for internet of things. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pp. 25.4.1–25.4.4, 2015.

[18] A. Jouve et al. 1m pitch direct hybrid bonding with <300nm wafer-to-wafer overlay accuracy. In *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1–2, Oct 2017.

[19] M. Chen et al. System on Integrated Chips (SoIC(TM) for 3D Heterogeneous Integration. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pp. 594–599, May 2019.

[20] X. Xu et al. Enhanced 3D Implementation of an Arm® Cortex®-A Microprocessor. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 1–6, July 2019.

[21] B. Gopireddy and J. Torrellas. Designing vertical processors in monolithic 3d. In *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*, pp. 643–656, 2019.

[22] A. Sayal et al. 14.4 all-digital time-domain cnn engine using bidirectional memory delay lines for energy-efficient edge computing. In *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 228–230, 2019.

[23] J. Lau. 3D IC Integration and Packaging. chapter 9. McGraw-Hill Education, 2015.

[24] M. Sekiguchi et al. Novel low cost integration of through chip interconnection and application to cmos image sensor. In *56th Electronic Components and Technology Conference 2006*, pp. 8 pp.–, 2006.

[25] A. Shigetou et al. Bumpless interconnect through ultrafine cu electrodes by means of surface-activated bonding (sab) method. *IEEE Transactions on Advanced Packaging*, 29(2):218–226, 2006.

[26] J. H. Lau. Evolution, challenge, and outlook of tsv, 3d ic integration and 3d silicon integration. In *2011 International Symposium on Advanced Packaging Materials (APM)*, pp. 462–488, 2011.

[27] S. Wong et al. Monolithic 3d integrated circuits. In *2007 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, pp. 1–4, 2007.

[28] D. K. Nayak et al. Power, performance, and cost comparisons of monolithic 3d ics and tsv-based 3d ics. In *2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1–2, 2015.

[29] L. Brunet et al. Breakthroughs in 3d sequential technology. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 7.2.1–7.2.4, 2018.

[30] S. K. Samal et al. Fast and accurate thermal modeling and optimization for monolithic 3d ics. In *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2014.

[31] H. T. Kung. Systolic arrays for vlsi. *Introduction to VLSI Systems*, 1980.

[32] H. T. Kung. Why systolic architectures? *IEEE Computer*, 15(1):37–46, 1982.

[33] A. Samajdar et al. Scale-sim: Systolic cnn accelerator simulator. 2018.

[34] H. Li et al. On-chip memory technology design space explorations for mobile deep neural network accelerators. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2019.

[35] M. Gao et al. Tetris: Scalable and efficient neural network acceleration with 3d memory. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and*

Operating Systems, ASPLOS '17, pp. 751–764, New York, NY, USA, 2017. Association for Computing Machinery.

[36] C. Hu et al. 3d multi-chip integration with system on integrated chips (soic™). *2019 Symposium on VLSI Technology*, pp. T20–T21, 2019.

[37] Celsius thermal solver. https://www.cadence.com/en_US/home/tools/system-analysis/thermal-solutions/celsius-thermal-solver.html.

[38] R. Mathur. Thermal analysis of a 3d stacked high-performance commercial microprocessor using face-to-face wafer bonding technology. ECTC '20, 2020.

[39] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In F. Pereira et al., editors, *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[40] D. Silver et al. Mastering the game of go without human knowledge.

[41] D. Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 173–182. JMLR.org, 2016.

[42] S. Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

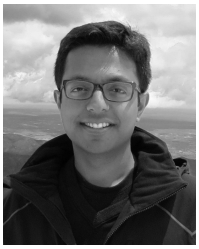
[43] C. Szegedy et al. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[44] X. He et al. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[45] K. He et al. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[46] R. Johnson and T. Zhang. Effective use of word order for text categorization with convolutional neural networks. 12 2014.

[47] A. Vaswani et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017.



Rahul Mathur received the B.E. degree in electrical and electronics engineering from Panjab University, Chandigarh, India, in 2009 and the M.E. degree in electrical engineering from Texas A&M University, College Station, TX, USA in 2012. He is currently pursuing his Ph.D. at the University of Texas at Austin, TX, USA.

He is pursuing a Ph.D. part-time while working at ARM Austin where he has worked since 2012. At Arm, he has led multiple memory compilers at sub-10nm foundry platforms. He has filed 15 US patents and also serves in the Patent Review Committee of Arm. His research interest is System-Circuit-Device Design Methodologies for 3D IC.

Mr. Mathur was a recipient of the University Gold Medal at Panjab University in 2009, the JK Pal Memorial Best Student Award from IEEE Delhi Section in 2009, and the International Education Fee Scholarship from Texas A&M University in 2011. He is a senior member of IEEE.

Ajay Krishna Ananda Kumar received the B.E. degree in electronics and communication engineering from Anna University, Chennai, India, in 2015 and the M.S. degree in electrical engineering from the University of Texas at Austin, TX, USA in 2020.

From 2015 to 2018, he was an SoC Design Engineer at Qualcomm, India. He was a graduate research assistant at The University of Texas at Austin from 2019 to 2020 where he worked on the application of Machine Learning methods for building accurate power models of CPUs to enable power-aware micro-architectural design space exploration and focused power optimizations. His research interests include energy-efficient circuit design and CPU/GPU micro-architectural optimizations for power and performance.

Dr. Lizy John received the B.S. degree in electronics and communication engineering from the University of Kerela, India, in 1984, the M.S. degree in computer engineering from The University of Texas at El Paso, TX, USA, in 1989, and the Ph.D. degree in computer engineering from The Pennsylvania State University, PA, USA, in 1993.

She joined The University of Texas Austin faculty in 1996 where she now holds the Cullen Trust for Higher Education Endowed Professorship in the Department of Electrical Computer Engineering at The University of Texas at Austin. She holds 13 US patents and has published 16 book chapters, approximately 300 journal, conference, and workshop papers. She has coauthored books on Digital Systems Design using VHDL (Cengage Publishers 2007, 2017), Digital Systems Design using Verilog (Cengage Publishers, 2014) and has edited a book on Computer Performance Evaluation and Benchmarking (CRC Press). She has also edited three books on workload characterization. Her research is in the areas of computer architecture, multicore processors, memory systems, performance evaluation and benchmarking, workload characterization, and reconfigurable computing.

Dr. John is recipient of the NSF CAREER award (1996), UT Austin Engineering Foundation Faculty Award (2001), Halliburton, Brown and Root Engineering Foundation Young Faculty Award (1999), University of Texas Alumni Association Teaching Award (2004), The Pennsylvania State University Outstanding Engineering Alumnus (2011), etc. She is the Editor-in-Chief (EIC) of IEEE MICRO and has served on the editorial boards of IEEE Transactions on Computers, ACM Transactions on Architecture and Code Optimizations (TACO), IEEE Computer Architecture Letters, IEEE Transactions on Sustainable Computing, and IEEE Transactions on VLSI. She is a member of IEEE, IEEE Computer Society, ACM, and ACM SIGARCH. She was named a Fellow of IEEE in 2009, a Fellow of the National Academy of Inventors in 2020, and a Fellow of the ACM in 2020.

Dr. Jaydeep Kulkarni received the B.E. degree from University of Pune, India in 2002, M. Tech degree from Indian Institute of Science (IISc), India in 2004 and Ph.D. degree from Purdue University, USA in 2009.

He worked as a Research Scientist at Intel Circuit Research Lab in Hillsboro, OR from 2009-2017. Currently, he is an assistant professor in the department of electrical and computer engineering at the University of Texas at Austin and a fellow of Silicon Labs Chair in electrical engineering and a fellow of AMD chair in computer engineering. He has filed 36 patents, published 2 book chapters, and 85 papers in refereed journals and conferences. His research is focused on machine learning hardware accelerators, in-memory computing, DTCO for emerging nano-devices, heterogeneous and 3D integrated circuits, hardware security, and cryogenic computing.

Dr. Kulkarni received 2004 best M. Tech student award from IISc Bangalore, 2008 Intel Foundation Ph.D. fellowship award, 2010 Purdue school of ECE outstanding doctoral dissertation award, 2015 IEEE Transactions on VLSI systems best paper award, 2015 SRC outstanding industrial liaison award, 2018, 2019 Micron Foundation Faculty Awards, and 2020 Intel Rising Star Faculty Award. He has participated in technical program committees of CICC, A-SSCC, DAC, ICCAD, ISLPED, and AICAS conferences. During his tenure at Intel Labs, he served as an industrial distinguished lecturer for IEEE Circuits and Systems Society and as an industrial liaison for SRC, NSF programs. He has served as a TPC co-chair and general co-chair for 2017 and 2018 ISLPED respectively, and currently serves as an associate editor for IEEE Solid State Circuit Letters, IEEE Transactions on VLSI Systems and a guest editor for IEEE Micro. He is currently serving as a distinguished lecturer for the IEEE Solid State Circuit Society and also serving as the chair of IEEE solid state circuits society and circuits and systems society central Texas joint chapter. He is a senior member of IEEE and National Academy of Inventors.