Received 28 April 2021; revised 31 May 2021; accepted 21 June 2021.
Date of publication 25 June 2021; date of current version 28 September 2021.

Digital Object Identifier 10.1109/JXCDC.2021.3092436

Thermal-Aware Design Space Exploration of 3-D 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)

Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA
CORRESPONDING AUTHOR: R. MATHUR (rahul.mathur@utexas.edu)

ABSTRACT Machine learning (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. The 3-D integration packs more compute or memory in the same 2-D footprint, which can be utilized to build more powerful or energy-efficient accelerators. However, 3-D 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 3-D systolic accelerators. Using this framework, different 3-D-partitioned accelerator configurations are proposed and evaluated. The 3-D-stacked accelerator designs are modeled using the hybrid wafer bonding technique with a 1.44-μm pitch of 3-D connection. Results show that different partitioning of the systolic array and SRAM buffers in a four-tier 3-D configuration can lead to either 1.1–3.9× latency reduction or 1–3× energy reduction compared to the baseline design of the same 2-D area footprint. It is also shown that by carefully organizing the systolic array and SRAM tiers using logic over memory, the temperature rise with 3-D across benchmarks can be limited to 6 °C.

INDEX TERMS 3-D 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]–[8]. The 2-D enhanced architectures [9] place dies side-by-side and interconnect them through media, such as a silicon interposer [10] or embedded bridge [11], [12], to achieve higher interconnect densities compared to mainstream packages. The 3-D architectures, such as 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. The 3-D architectures may offer complementary gains to traditional dimensional scaling for achieving high performance, low power, high bandwidth, and fast time-to-market, all in a small footprint. Larger 2-D dies can be replaced by a few smaller ones with potentially higher manufacturing yields [15], [16]. Besides, 3-D allows heterogeneous integration of parts from different technologies instead of having to redesign every component for a specific process [17]. As 3-D technologies evolve, increasingly finer pitches of 3-D 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]; 3-D makes it possible to pack more compute or memory in the same 2-D footprint while reducing interconnect delay and power by bringing the blocks closer. Hence, 3-D provides an opportunity to build powerful and energy-efficient accelerators.

Traditional 2-D systolic array design involves careful partitioning of the silicon real estate between the processing element (PE) units and SRAM buffers to balance the throughput and external memory transfer bandwidth. The 3-D accelerator design additionally involves the optimal distribution of the increased silicon real estate available in the same 2-D footprint between the PE units and memory. Furthermore,

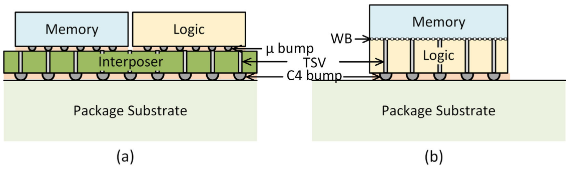


FIGURE 1. (a) 2-D enhanced: side-by-side die stacked over interposer (2.5-D). (b) 3-D: memory die stacked directly on the logic die using hybrid WB technology.

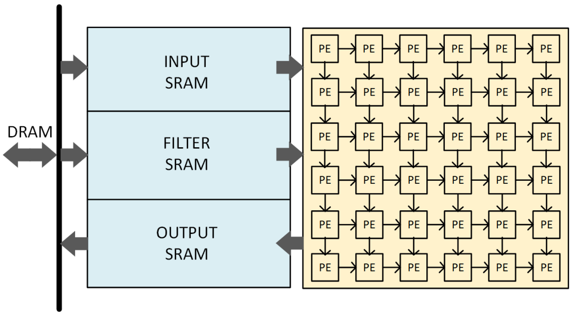


FIGURE 2. Typical systolic array-based accelerator system.

the power density of systolic accelerators is high due to their desired high computing capability and closely packed PEs. This is exacerbated in 3-D due to higher logic integration density, which may lead to worse thermal characteristics [23]. Hence, the designer must consider the thermal implications when partitioning the accelerator components among 3-D tiers. A systematic methodology for navigating the 3-D systolic accelerator design space accounting for the thermal issues is necessary. This article makes the following contributions to address this issue.

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

The remainder of this article is organized as follows. Section II provides the background and prior work on 3-D integration technologies and systolic architectures. Section III describes the 2-D baseline design and different 3-D partitioned configurations. Section IV delineates the simulation framework used to perform performance, power, and thermal analysis. Section V describes the experimental setup. Section VI presents the results from a comparative analysis of different 3-D accelerator configurations. Section VII provides concluding remarks.

II. BACKGROUND AND PRIOR WORK

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

A. OVERVIEW OF 3-D 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 2-D enhanced (also referred to as 2.5-D integration) utilizes an interposer made of silicon, glass, or ceramic for high-density communication between separate dies mounted side-by-side [see Fig. 1(a)]. The interposer may contain through-silicon vias (TSVs) [24] that are essentially holes etched out in the silicon wafer and then filled with a conductive metal such as copper.

The 3-D 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 microbumps [25] or hybrid wafer bonds (WBs) [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 [see Fig. 1(b)]. Similarly, other configurations, such as back-to-back (B2B) and face-to-face (F2F), are possible, especially when multiple dies are stacked in this manner. Compared to 2.5-D (lateral) integration, 3-D stacking worsens thermals due to increased power density with die overlap, and heat dissipation from tiers away from the heat sink is a challenge [26].

Monolithic 3-D ICs consist of multiple device layers fabricated sequentially on the same die and connecting using monolithic inter-tier vias (MIVs) that 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 3-D enables the finest pitch of 3-D 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 3-D 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 3-D stacked ICs using the hybrid wafer bonding technology to model the design of 3-D ML accelerators. Nonetheless, some of the ideas discussed in this article around efficient partitioning of an ML accelerator design into 3-D tiers can be helpful to design for other 3-D 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 are 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 (MAC), and pass the intermediate results or the original inputs to adjacent PEs. The key idea is to exploit data reuse so that fewer data transfers from memory are needed. Furthermore, purely local data movement (neighbor to neighbor) means simpler interconnect and control. PEs operating in parallel achieve high computational concurrence. Moreover, systolic architectures are modular making them easy to floorplan and scale. Fig. 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 MAC operation, while data are streamed through the array in a predefined synchronized dataflow. An example of dataflow is weight stationary where weights of the filter matrices corresponding to each DNN layer are preloaded from the filter memory into the systolic array before any matrix multiplication operation is performed. Input data are then streamed in from the input memory, and the array elements perform matrix multiplication with the weights already stored in them. The output data are 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].

TABLE 1. **List of 2-D and 3-D 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

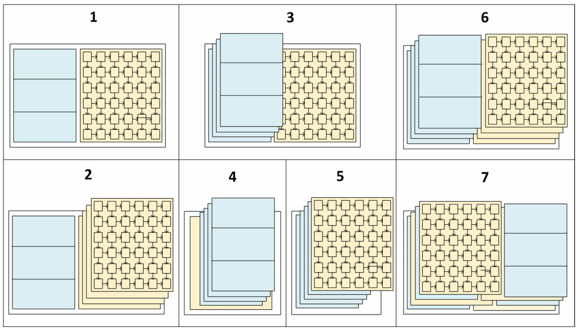


FIGURE 3. **High-level floorplan showing different approaches of partitioning SRAM buffers (blue) and PE array (yellow) in the 2-D and 3-D accelerator configurations.**

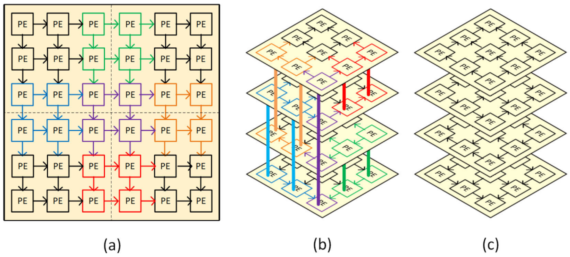


FIGURE 4. **(a) Example PE array in 2-D. (b) PE array folded in 3-D with vertical connections between PEs across tiers (configurations 2 and 6). (c) Separate smaller PE arrays operating independently (configuration 7).**

III. 2-D AND 3-D SYSTOLIC ARRAY ACCELERATORS

Traditional 2-D 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). However, again, overprovisioned SRAMs can lead to area and cost inefficiencies. In summary, designers must consider the aforementioned tradeoffs 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 2-D accelerator was selected to have a 32 × 32 PE array and 128 kB of filter, IFMAP, and OFMAP SRAM each, which is representative of common DNN inference use case [4].

The 3-D systolic accelerator design further involves distributing the additional silicon real estate available within the same 2-D 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 3-D accelerators with different partitioning styles, design points described in Table 1 are selected. The 3-D configurations considered were limited to four 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 4× larger 2-D design with increased compute or memory resources is possible, a four-tier 3-D system packs equivalent resources in the same footprint as the baseline 2-D accelerator. A 3-D system will incur lower 2-D interconnect delay and power due to shorter distances and fewer buffers compared to a 4× larger 2-D system but may incur an additional penalty in 3-D interconnects.

The 3-D configurations selected for further analysis include multiple PE array tiers (configuration 2) or multiple SRAM tiers (configurations 3–5), a scaled-up version (configuration 6), and a scaled-out version (configuration 7) of the 2-D baseline accelerator. The floorplans for all design points are shown in Fig. 3. It should be noted that configurations 3–5 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, whereas 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 Fig. 4(c), the scale-out version does

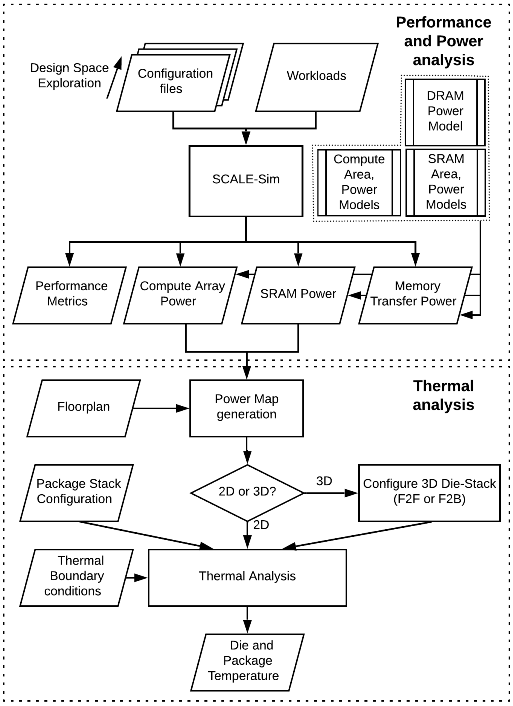


FIGURE 5. Simulation framework for 3-D systolic accelerators.

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 shown in Fig. 5. It comprises two flows that 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 files. Simulation benchmarks are translated to topology files having a layerwise 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 layerwise 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] (see Table 2). DRAM accesses can contribute a major part of the total energy [35]. For 3-D 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 12× [36]. Hence, to incorporate an average case impact of vertical interconnect energy on the overall DRAM access energy of a four-tier system, 1.35 pJ/byte (one F2F and one F2B) is added to all DRAM transfers of 3-D accelerator configurations.

TABLE 2. 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)

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

$$P_{\mathrm{PE}} = \frac{\sum_{i=1}^{n}(\mathrm{util}(i) * \mathrm{arr\_}h * \mathrm{arr\_}w * e\_\mathrm{mac} * \mathrm{cyc}(i))}{\mathrm{cycles} * \frac{1}{\mathrm{freq}} * 100} \tag{1}$$

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 and 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, and (e_mac) is the energy consumed per 8-bit MAC operation. e_mac of 0.3 pJ (Table 2) 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 16-nm process node [34].

SRAM power is calculated using the following equation:

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

where n is the number of layers in the network, srd_bw(i)*and* swt_bw(i) are the average SRAM read and write bandwidth in bytes per cycle for the execution of layer i, respectively, cyc(i) are the number of cycles taken for computing layer i, and (e_srd) and write (e_swt) are the SRAM energy consumed in access of byte-wide data. e_srd of 1.1 pJ and e_swt of 1.5 pJ (Table 2) is based on 32-kB SRAM macros generated from an industry-standard memory compiler at 16 nm and takes both dynamic and static energy into account [34].

DRAM power is calculated using the following equation:

$$P_{\mathrm{DRAM}} = \frac{\sum_{i=1}^{n}((d\_\mathrm{if}(i) + d\_\mathrm{filt}(i) + d\_\mathrm{of}(i)) * e\_\mathrm{mem} * \mathrm{cyc}(i))}{\mathrm{cycles} * \frac{1}{\mathrm{freq}}} \tag{3}$$

where n is the number of layers in the network, d_if(i), d_filt(i), and 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. e_mem of 120 pJ (Table 2) is based on off-chip DRAM accesses energy per byte assuming an LPDDR3 interface [35].

Performance in terms of latency of different 2-D and 3-D configurations is computed from the layerwise 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

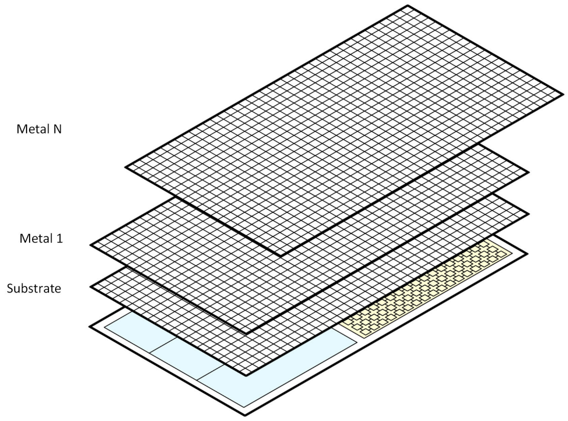


FIGURE 6. **Tile-based power map used for thermal analysis.**

in terms of throughput can be calculated in teraoperations per second (TOPS) using the following equation:

$$\text{TOPS} = \frac{\text{util} * \text{arr\_}h * \text{arr\_}w * 2}{\frac{1}{\text{freq}} * 100} \tag{4}$$

where util is the average utilization of the PE array for computing the NN (between 0 and 100), arr_h and arr_w are the PE array height and width, respectively, and freq is the frequency of operation. The factor of 2 in the numerator represents the separate MAC operations of MAC. This definition is theoretical systolic accelerator throughput with no memory-bandwidth limitations, i.e., with stall-free operation. The delay overhead of 3-D F2F vertical interconnect can be ~5 ps at nominal voltage [14]. An F2B connection (through TSVs) has a delay overhead of 3.2× over an F2F connection [36]. Hence, to incorporate a worst case impact of the vertical interconnect delay on the frequency of a four-tier system, 42 ps (two F2Fs and two F2Bs) is added to the cycle time (1/freq) of 3-D accelerator configurations.

B. THERMAL ANALYSIS FLOW

To the first order, the temperature rise in 3-D IC is primarily proportional to the effective power density in the 2-D footprint [23]. Floorplan dimensions of different 2-D and 3-D configurations are calculated based on the PE and SRAM area at 14-/16-nm technology node from Table 2. 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 floorplan dimensions. Fig. 6 shows 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 3-D configurations with relative ease.

Cadence Celsius Thermal Solver: The work [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 ten build-up layers with overall dimensions of 10 × 10 mm^2 with an 11 × 11 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, as a significant change in package thermal design power (TDP) (for instance, configurations 2–7 versus configuration 1), the heat spreader dimensions may need to be redesigned, and boundary conditions may have to be recalibrated. 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.

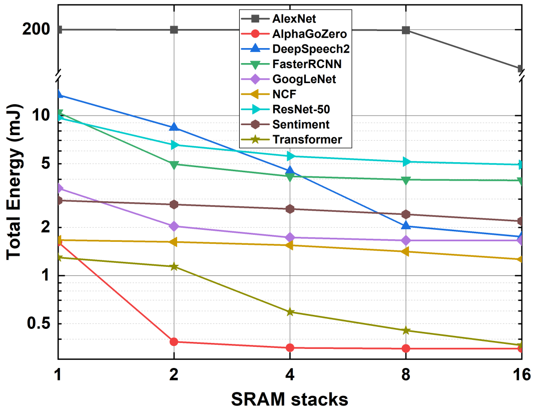


FIGURE 7. **Reduction in total energy by sweeping SRAM stacks of configuration 3 for different NN benchmarks (log scale).**

V. EXPERIMENTAL SETUP

SCALE-Sim is configured with microarchitecture features, such as PE array dimension, aspect ratio, and memory buffer sizes for different 2-D and 3-D accelerator configurations listed in Table 1. The simulator, by default, only supports a 2-D systolic configuration. The 3-D design points of configurations 2–6 can be mapped to SCALE-Sim using their respective PE and SRAM sizes, as specified in Table 1. 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 four ways 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 3-D accelerator configurations. The topology files having a layerwise description of the networks, such as input and filter dimensions, input channels, number of filters, and strides, are set up for SCALE-Sim for some common NN benchmarks such as 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

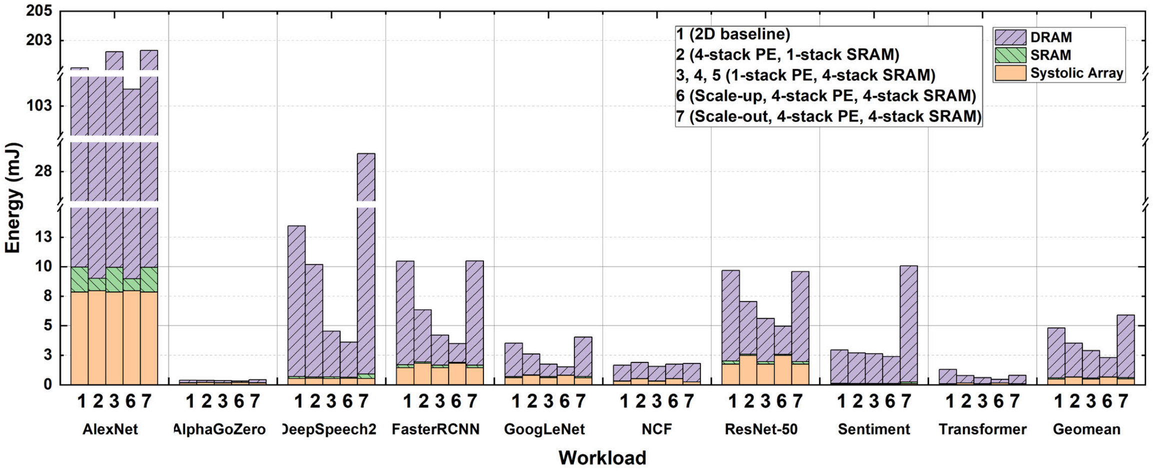


FIGURE 8. Energy comparison among configurations for different NN workloads. The DRAM energy split includes the vertical interconnect energy overhead for the 3-D configurations.

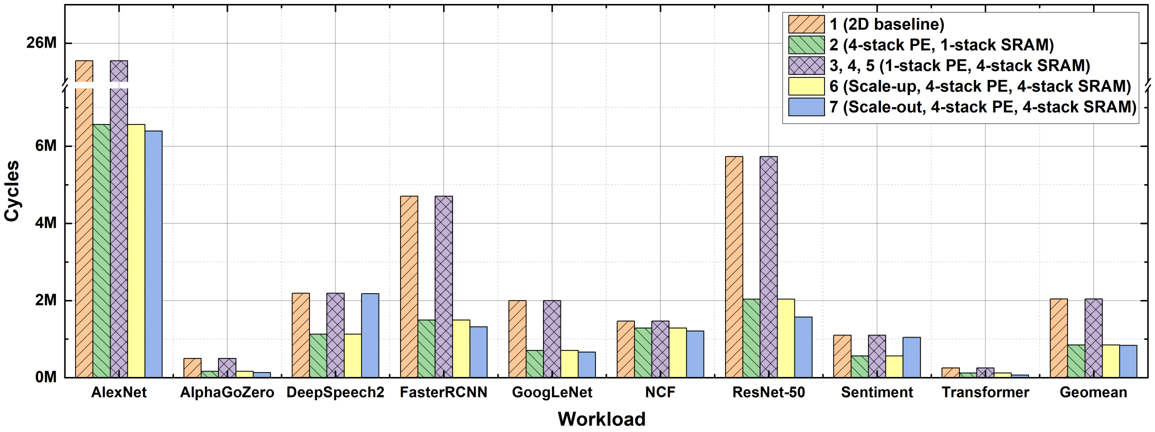


FIGURE 9. Latency comparison among configurations for different NN workloads.

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 2-D baseline.

VI. RESULTS

This section presents the simulation results comparing different 3-D 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 (see Fig. 7), especially for memory-limited networks. Fig. 8 compares the total energy of configurations 1–7 listed in Table 1 across different benchmarks and also shows the breakdown of energy between computations, SRAM, and DRAM transfers. As expected, configurations 3–6 that contain four-stack SRAM reduce the total energy to process the network compared to configuration 1 (2-D baseline). However, the energy reduction factor varies widely between benchmarks from 1.0× for NCF to 3.8× for Deep Speech 2. NCF being relatively small already fits within a single SRAM stack and additional SRAM stacks in 3-D bring no benefit. Configuration 6 (scale-up) achieves the lowest energy since it also contains four stacks of PEs along with four 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.

TABLE 3. 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

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, Fig. 9 shows that configurations 2, 6, and 7 that contain four-stack PE arrays take fewer cycles to process the network compared to configuration 1 (2-D baseline). However, the speedup varies widely between benchmarks from 1.1× for NCF to 3.9× for AlexNet. NCF has much smaller layer features such

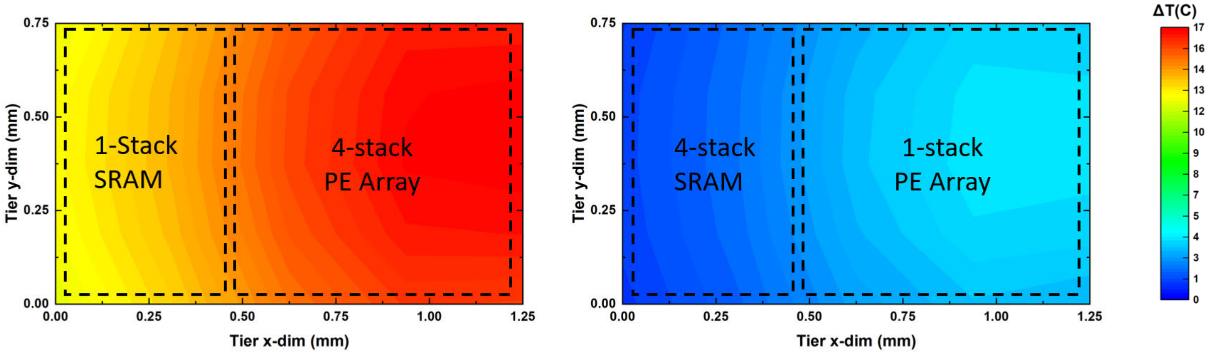


FIGURE 10. Heat maps under the ResNet-50 benchmark for (a) configuration 2 (four-stack array) and (b) configuration 3 (four-stack SRAM). The tier dimensions are in mm. All temperatures are relative to the coolest point on configuration 1 (2-D baseline) for ResNet-50.

TABLE 4. Maximum system temperature for different configurations across all benchmarks relative to the coldest point in 2-D 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

as IFMAP dimensions compared to AlexNet and is unable to utilize the additional PE tiers to achieve any more compute parallelism. Configuration 7 (4× scale-out of 2-D) shows slightly better performance than configuration 6 (4× scale-up of 2-D) 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 contains 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 3-D configurations) is presented in Table 3.

C. THERMAL

Fig. 10 shows the steady-state heat maps of configuration 2 (four-stack PE array) and configuration 3 (four-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 (2-D 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 3-D stack of PE array is higher compared to the SRAM stack. Table 4 compares the maximum temperature rise of different configurations across all benchmarks. The benchmarks have a varied size of underlying NN model, leading to different average array utilizations and SRAM accesses causing different rise of temperatures. Configurations 2, 6, and 7 that employ 3-D 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.

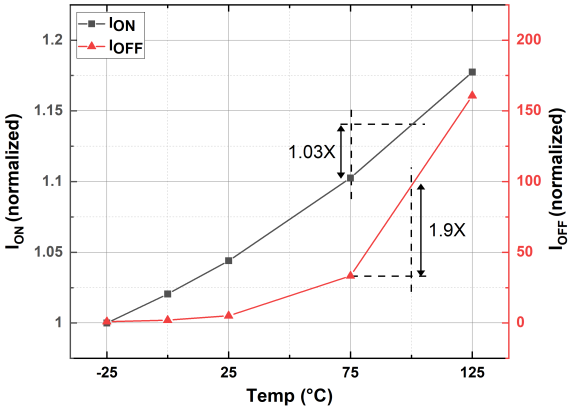


FIGURE 11. Effect of temperature rise on the ON-state current (black line) @SS/V_{NOM}-10% and ON-state current (red line) @FF/V_{NOM}+10% of a transistor with standard VTH option at 14/16 nm (0.8-V V_{NOM}).

The increase in temperature can have an impact on the overall energy of the accelerator. For example, assuming the coolest point on the 2-D 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 1.9× (see Fig. 11). Configuration 7 partially avoids overlapping hotspots by staggering the PE array and SRAM between tiers but fares only slightly better. Configurations 3 and 4 that stack multiple 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, 3-D 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. Furthermore, these configurations suffer from the worst thermal characteristics due to logic-over-logic stacking. On the other hand, stacking multiple SRAM tiers (configurations 3–6) lower 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 3-D.

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. The 3-D integration packs more compute or memory in the same 2-D 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 3-D tiers. Since different choices may have different performance, power, and thermal implications, it becomes imperative for designers to understand the tradeoffs under different application workload conditions. In this work, a systematic methodology for navigating the 3-D systolic accelerator design space is presented. Using this framework, 3-D 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 3-D systolic accelerator designs.

REFERENCES

[1] S. Kung, "VLSI array processors," *ASSP Mag.*, vol. 2, no. 3, pp. 4–22, Jul. 1985.

[2] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2017, pp. 1–12.

[3] Xilinx. (2018). *Accelerating DNNS With Xilinx Alveo Accelerator Cards*. [Online]. Available: https://www.xilinx.com/support/documentation/white_papers/wp504-accel-dnns.pdf

[4] Y.-H. Chen, T.-J. Yang, J. S. Emer, and V. Sze, "Eyeriss V2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 292–308, Jun. 2019.

[5] J. Song *et al.*, "7.1 An 11.5 tops/w 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8 nm flagship mobile SoC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 130–132.

[6] T. N. Theis and H.-S.-P. Wong, "The end of Moore's law: A new beginning for information technology," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 41–50, Mar. 2017.

[7] A. Mocuta, P. Weckx, S. Demuynck, D. Radisic, Y. Oniki, and J. Ryckaert, "Enabling CMOS scaling towards 3 nm and beyond," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 147–148.

[8] G. Yeric, "IC design after Moore's law," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2019, pp. 1–150.

[9] *Interconnects for 2D and 3D Architectures*. Accessed: May 24, 2021. [Online]. Available: https://eps.ieee.org/images/files/HIR_2020/ch22_2D-3D.pdf

[10] M.-S. Lin *et al.*, "A 7-nm 4-GHz Arm-core-based CoWoS chiplet design for high-performance computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 4, pp. 956–966, Apr. 2020.

[11] R. Mahajan *et al.*, "Embedded multidie interconnect bridge—A localized, high-density multichip packaging interconnect," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 9, no. 10, pp. 1952–1962, Oct. 2019.

[12] A. Podpod *et al.*, "A novel fan-out concept for ultra-high chip-to-chip interconnect density with 20-μm pitch," in *Proc. IEEE 68th Electron. Compon. Technol. Conf. (ECTC)*, May 2018, pp. 370–378.

[13] D. W. Fisher *et al.*, "Face to face hybrid wafer bonding for fine pitch applications," in *Proc. IEEE 70th Electron. Compon. Technol. Conf. (ECTC)*, Jun. 2020, pp. 595–600.

[14] S. Sinha *et al.*, "A high-density logic-on-logic 3DIC design using face-to-face hybrid wafer-bonding on 12 nm FinFET process," in *IEDM Tech. Dig.*, Dec. 2020, pp. 1–15.

[15] G. Yeric, "Moore's law at 50: Are we planning for retirement?" in *IEDM Tech. Dig.*, Dec. 2015, pp. 1.1.1–1.1.8.

[16] (2020). *International Roadmap for Devices and Systems*. [Online]. Available: 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 *IEDM Tech. Dig.*, Dec. 2015, pp. 25.4.1–25.4.4.

[18] A. Jouve *et al.*, "1μm pitch direct hybrid bonding with <300 nm wafer-to-wafer overlay accuracy," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Oct. 2017, pp. 1–2.

[19] M.-F. Chen, F.-C. Chen, W.-C. Chiou, and D. C. H. Yu, "System on integrated chips SoIC(TM) for 3D heterogeneous integration," in *Proc. IEEE 69th Electron. Compon. Technol. Conf. (ECTC)*, May 2019, pp. 594–599.

[20] X. Xu *et al.*, "Enhanced 3D implementation of an arm cortex-a microprocessor," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2019, pp. 1–6.

[21] B. Gopireddy and J. Torrellas, "Designing vertical processors in monolithic 3D," in *Proc. ACM/IEEE 46th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2019, pp. 643–656.

[22] A. Sayal, S. Fathima, S. S. T. Nibhanupudi, and J. P. Kulkarni, "14.4 all-digital time-domain CNN engine using bidirectional memory delay lines for energy-efficient edge computing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 228–230.

[23] J. Lau, *3D IC Integration and Packaging*. New York, NY, USA: McGraw-Hill, 2015.

[24] M. Sekiguchi *et al.*, "Novel low cost integration of through chip interconnection and application to CMOS image sensor," in *Proc. 56th Electron. Compon. Technol. Conf.*, 2006, p. 8.

[25] A. Shigetou, T. Itoh, M. Matsuo, N. Hayasaka, K. Okumura, and T. Suga, "Bumpless interconnect through ultrafine Cu electrodes by means of surface-activated bonding (SAB) method," *IEEE Trans. Adv. Packag.*, vol. 29, no. 2, pp. 218–226, May 2006.

[26] J. H. Lau, "Evolution, challenge, and outlook of tsv, 3D IC integration and 3D silicon integration," in *Proc. Int. Symp. Adv. Packag. Mater. (APM)*, pp. 462–488, 2011.

[27] S. Wong *et al.*, "Monolithic 3D integrated circuits," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, 2007, pp. 1–4.

[28] D. K. Nayak, S. Banna, S. K. Samal, and S. K. Lim, "Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Oct. 2015, pp. 1–2.

[29] L. Brunet *et al.*, "Breakthroughs in 3D sequential technology," in *IEDM Tech. Dig.*, Oct. 2018, pp. 7.2.1–7.2.4.

[30] S. K. Samal, S. Panth, K. Samadi, M. Saedi, Y. Du, and S. K. Lim, "Fast and accurate thermal modeling and optimization for monolithic 3D Ics," in *Proc. The 51st Annu. Design Autom. Conf. Design Autom. Conf.*, 2014, pp. 1–6.

[31] H. T. Kung and C. E. Leiserson, "Systolic arrays (for VLSI)," in *Sparse Matrix Proceedings 1978*, I. S. Duff and G. W. Stewart, Eds. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1979, pp. 256–282.

[32] H. T. Kung, "Why systolic architectures?" *IEEE Comput.*, vol. 15, no. 1, pp. 37–46, Jan. 1982.

[33] A. Samajdar *et al.*, "SCALE-sim: Systolic CNN accelerator simulator," 2018, *arXiv:1811.02883*. [Online]. Available: https://arxiv.org/abs/1811.02883

[34] H. Li, M. Bhargava, P. N. Whatmough, and H.-S.-P. Wong, "On-chip memory technology design space explorations for mobile deep neural network accelerators," in *Proc. 56th Annu. Design Autom. Conf.*, Jun. 2019, pp. 1–6.

[35] M. Gao *et al.*, "Tetris: Scalable and efficient neural network acceleration with 3d memory," in *Proc. 22nd Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, New York, NY, USA, 2017, pp. 751–764.

[36] C. C. Hu, M. F. Chen, W. C. Chiou, and D. C. H. Yu, "3D multi-chip integration with system on integrated chips (SoIC)," in *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. T20–T21.

[37] *Celsius Thermal Solver*. Accessed: Mar. 16, 2020. [Online]. Available: 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," in *Proc. ECTC*, Jun. 2020, pp. 541–547.

[39] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, F. Pereira, Ed. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[40] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, pp. 354–359, Oct. 2017.

[41] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and mandarin," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 173–182.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

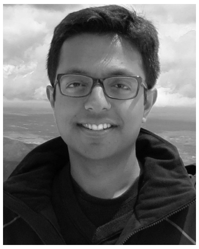
[43] C. Szegedy *et al.*, "Going deeper with convolutions," *CoRR*, abs/1409.4842, pp. 1–4, Oct. 2014.

[44] X. He *et al.*, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," 2014, *arXiv:1412.1058*. [Online]. Available: https://arxiv.org/abs/1412.1058

[47] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: https://arxiv.org/abs/1706.03762



RAHUL MATHUR (Senior Member, IEEE) 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 the Ph.D. degree with the University of Texas at Austin, Austin, TX, USA.

He has been working at Arm Ltd., Austin, since 2012. At ARM, he has led multiple memory compilers at sub-10-nm foundry platforms. He has filed 16 U.S. patents and also serves in the Patent Review Committee of Arm. His research interest is system-circuit-device design methodologies for 3-D integrated circuits (ICs).

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

AJAY KRISHNA ANANDA KUMAR (Member, IEEE) 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, Austin, TX, USA, in 2020.

From 2015 to 2018, he was an SoC Design Engineer at Qualcomm, Bengaluru, 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 microarchitectural design space exploration and focused power optimizations. His research interests include energy-efficient circuit design and CPU/GPU microarchitectural optimizations for power and performance.

LIZY JOHN (Fellow, IEEE) received the B.S. degree in electronics and communication engineering from the University of Kerala, Thiruvananthapuram, India, in 1984, the M.S. degree in computer engineering from The University of Texas at El Paso, El Paso, TX, USA, in 1989, and the Ph.D. degree in computer engineering from Pennsylvania State University, State College, PA, USA, in 1993.

In 1996, she joined The University of Texas Austin, where she currently holds the Cullen Trust for Higher Education Endowed Professorship at the Department of Electrical and Computer Engineering, The University of Texas at Austin. She holds 13 U.S. patents. She has published 16 book chapters and approximately 300 journal articles, conference papers, and workshop papers. She has coauthored books: *Digital Systems Design Using VHDL* (Cengage Publishers, 2007 and 2017) and *Digital Systems Design Using Verilog* (Cengage Publishers, 2014). She has edited a book *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 was a fellow of the National Academy of Inventors in 2020 and the Association for Computing Machinery (ACM) in 2020. She is a member of the IEEE Computer Society, ACM, and ACM Special Interest Group on Computer Architecture (SIGARCH). She was a recipient of the NSF CAREER Award in 1996, the UT Austin Engineering Foundation Faculty Award in 2001, the Halliburton, Brown and Root Engineering Foundation Young Faculty Award in 1999, the University of Texas Alumni Association Teaching Award in 2004, and the Pennsylvania State University Outstanding Engineering Alumnus in 2011. She is the Editor-in-Chief (EIC) of IEEE MICRO. She has served on the Editorial Board 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 VERY LARGE SCALE INTEGRATION SYSTEMS.

JAYDEEP P. KULKARNI (Senior Member, IEEE) received the B.E. degree from the University of Pune, Pune, India, in 2002, the M.Tech. degree from the Indian Institute of Science (IISc), Bengaluru, India, in 2004, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2009.

He worked as a Research Scientist at Intel Circuit Research Lab, Hillsboro, OR, USA, from 2009 to 2017. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA, and a fellow of Silicon Labs Chair in electrical engineering and a fellow of AMD chair in computer engineering. He has filed 36 patents. He has published two book chapters and 85 articles in refereed journals and conferences. His research is focused on machine learning hardware accelerators, in-memory computing, design technology co-optimization (DTCO) for emerging nanodevices, heterogeneous and 3-D integrated circuits, hardware security, and cryogenic computing.

Dr. Kulkarni is a Senior Member of the National Academy of Inventors. He received the 2004 Best M.Tech. Student Award from IISc, the 2008 Intel Foundation Ph.D. Fellowship Award, the 2010 Purdue School of ECE Outstanding Doctoral Dissertation Award, the 2015 IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS Best Paper Award, the 2015 Semiconductor Research Corporation (SRC) Outstanding Industrial Liaison Award, the 2018 and 2019 Micron Foundation Faculty Awards, and the 2020 Intel Rising Star Faculty Award. He has participated in the Technical Program Committee of IEEE Custom Integrated Circuits Conference (CICC), IEEE Asian Solid-State Circuits Conference (A-SSCC), Design Automation Conference (DAC), International Conference On Computer Aided Design (ICCAD), ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), and IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS) conferences. During his tenure at Intel Labs, he served as an Industrial Distinguished Lecturer for the IEEE Circuits and Systems Society and as an Industrial Liaison for SRC and NSF programs. He has served as the TPC Co-Chair and General Co-Chair for 2017 and 2018 ISLPED, respectively. He serves as an Associate Editor for IEEE SOLID-STATE CIRCUIT LETTERS and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS and a Guest Editor for IEEE MICRO. He is also serving as a Distinguished Lecturer for the IEEE Solid-State Circuit Society and the Chair for the IEEE Solid-State Circuits Society and the IEEE Circuits and Systems Society's Central Texas Joint Chapter.