On Continuing DNN Accelerator Architecture Scaling Using Tightly Coupled Compute-on-Memory 3-D ICs

Gauthaman Murali, *Graduate Student Member, IEEE*, Aditya Iyer, Lingjun Zhu, *Graduate Student Member, IEEE*, Jianming Tong, *Graduate Student Member, IEEE*, Francisco Muñoz Martínez, Srivatsa Rangachar Srinivasa, *Member, IEEE*, Tanay Karnik, *Fellow, IEEE*, Tushar Krishna, *Senior Member, IEEE*, and Sung Kyu Lim, *Fellow, IEEE*

***Abstract*— This work identifies the architectural and design scaling limits of 2-D flexible interconnect deep neural network (DNN) accelerators and addresses them with 3-D ICs. We demonstrate how scaling up a baseline 2-D accelerator in the *X/Y* dimension fails and how vertical stacking effectively overcomes the failure. We designed multitier accelerators that are 1.67× faster than the 2-D design. Using our 3-D architecture and circuit codesign methodology, we improve throughput, energy efficiency, and area efficiency by up to 5×, 1.2×, and 3.9×, respectively, over 2-D counterparts. The IR-drop in our 3-D designs is within 10.7% of VDD, and the temperature variation is within 12 °C.**

***Index Terms*— 3-D accelerator physical design, 3-D bond pitch study for accelerators, high-performance 3-D accelerator, multitier 3-D ML accelerator.**

I. Introduction

DEEP neural networks (DNNs) have become integral to several applications, such as vision, speech recognition, object detection, and autonomous driving, over the past decade. These applications need to process a large amount of data in real-time, making the DNNs increasingly compute-intensive with substantially high memory requirements. The computation and the memory requirement of DNN workloads have increased by approximately 100× and 4×, respectively [1], over the past decade. Fig. 1 shows the filter weights' memory requirements and total multiply-and-accumulate (MAC) operations of various recent CNN benchmarks.

While there has been significant improvement in the DNN accelerator architecture for modern applications, the underlying process and integration technology limit the accelerator's performance, energy efficiency, and scalability. Technology (logic/memory) downscaling deceleration, memory bandwidth (BW) scaling challenges, and increasing computational complexity have become the three main limiters of efficient scaling of DNN accelerator hardware [1]. Several scaling solutions provide multiple degrees of parallelism to compensate for process, technology, and BW limitations. However, the need for greater parallelism requires an increase in memory BW and numerous MAC units, which creates a self-perpetuating cycle. While ON-chip SRAMs provide high BW, they compromise the chip density, leading to numerous second-order design inefficiency. Therefore, several solutions incorporate small capacity and high BW ON-chip SRAM buffers.

The compute logic needs to scale according to the application trends and performance requirements. For example, every new TPU [2] version has 2× or more MAC units than the previous version to obtain higher throughput but without similar improvement in energy efficiency. Alternatively, compute-in-memory (CiM) techniques have been attractive ways to minimize the processing element (PE) complexity and energy consumption. But these techniques either alter memory bit-cell or provide analog ways of computation and compromise memory robustness or compute precision.

Three-dimensional integration techniques such as face-to-face (F2F), face-to-back (F2B), and back-to-back (B2B) involving bonding styles such as microbumping, hybrid-bonding through-silicon via (TSV), or monolithic intertier via (MIV) [3] can be used to solve several design challenges mentioned above. The integration technique, number of tiers, and bond pitch play significant roles in scaling and optimizing accelerator designs. Without careful consideration of these interdependent design aspects, the benefits of scaling the design to meet ever-increasing performance and power requirements for DNN accelerators become negligible. Hence, there is a pressing need to co-optimize DNN accelerator architectures through 3-D integration and study the effects of 3-D bond pitch and multitier integration on their performance.

Our contributions to this work are as follows.

1) We identify and study the architectural and design scaling limits of 2-D flexible interconnect DNN accelerators.
2) We present realistic 3-D architecture and technology co-optimization framework for efficient accelerator hard-

Manuscript received 30 March 2023; revised 3 July 2023; accepted 20 July 2023. Date of publication 16 August 2023; date of current version 27 September 2023. *(Corresponding author: Gauthaman Murali.)*

Gauthaman Murali, Aditya Iyer, Lingjun Zhu, Jianming Tong, Tushar Krishna, and Sung Kyu Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30308 USA (e-mail: gauthaman@gatech.edu).

Francisco Muñoz Martínez is with the Departamento de Ingeniería y Tecnología de Computadores, Facultad de Informática, Universidad de Murcia, Campus Espinardo, 30100 Murcia, Spain.

Srivatsa Rangachar Srinivasa and Tanay Karnik are with Intel Corporation, Hillsboro, OR 97124 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TVLSI.2023.3299564.

Digital Object Identifier 10.1109/TVLSI.2023.3299564

1063-8210 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

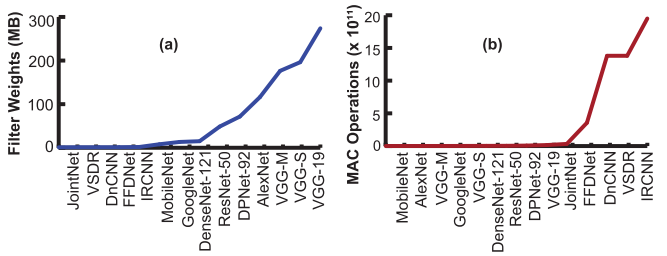

Fig. 1. (a) Memory requirements of filter weights. (b) Total computations of various recent CNN benchmarks [1].

ware scaling, with practical analysis of 3-D power delivery network (PDN) and thermal issues.

3) Our two- and three-tier designs improve throughput, energy efficiency, and area efficiency by up to 5×, 1.2×, and 3.9×, respectively, over 2-D counterparts.
4) Our 3-D designs have IR-drop within 10.7% of VDD with on-die temperature variation limited to 12 °C.

II. Background and Motivation

A. Scaling Issues in Flexible DNN Accelerators

DNN has been successfully adopted in image classification [4], object detection/recognition [5], text summarization [6], and sentiment analysis [7]. With growing workloads in these domains, we need better DNN accelerator hardware capabilities for fast and energy-efficient acceleration.

Unfortunately, most DNN accelerators involve rigid architectures with careful codesign of PEs and network-on-chip (NoC), leading to a fixed dataflow pattern. This often causes underutilization of the PEs, especially on mapping arbitrary dataflows onto them. For example, when the dimensions of a convolution layer are not multiples of the PE array dimensions, the PE array could be severely underutilized [8]. FlexSA, a flexible systolic array [8], tries to solve this by introducing flexible interconnects in systolic arrays. However, it could only improve the utilization by up to 37%. DNN accelerators need to be programmable to enable mass deployment with interconnect configurability to support the various dataflow patterns. This, on the other hand, could offer a near 100% PE utilization, which is 8%–459% higher utilization than most systolic arrays for various workloads [9].

Flexible interconnect DNN accelerators, such as MAERI [9] and SIGMA [10], permit internal BW and dataflow reconfiguration, offer better PE utilization than systolic arrays, and also work great for sparse workloads. However, unlike systolic arrays, flexible interconnect architectures have high memory BW and capacity requirements. The ON-chip memory BW heavily influences PE utilization in flexible interconnect accelerators. Often, a high ON-chip BW closer to or greater than 0.5 × #PEs (words/cycle) leads to maximum PE utilization in these accelerators due to the activation and weight sharing among the PEs, making it difficult to scale them in the X/Y dimensions.

We can improve the throughput of such memory-bound DNN accelerators by 1) increasing DRAM channels to supply sufficient data to ON-chip memory every cycle as the design scales; 2) tightly coupling PEs and ON-chip memory to reduce latency; and 3) increasing the ON-chip memory BW to keep all the MAC units busy every cycle. Among the three solutions, incorporating more DRAM channels or replacing DRAM with high bandwidth memory (HBM) leads to increased cost or die area. For a given DRAM BW, Fig. 2 shows the trend in execution time of SSD-MobileNetv1 on different configurations of MAERI [9]. Based on the maximum data flow requirement of different layers in these workloads, we observe a multifold improvement in execution times as the data BW and the number of PEs increase in the design. Of the 44 convolution layers, 12 layers show a throughput improvement of about 1.67× and the rest of them show up to 4× improvement. With 128 PEs in the design, 12 layers of SSD_MobileNetv1 show a throughput improvement of up to 1.67× and the remaining majority shows an improvement of up to 4× as the BW increases from 32 to 128 B/cycle. As the design scales from 128 to 256 PEs and the data BW increases to 256 B/cycle, the throughput improves by up to 8×, over the baseline 128 PE, 32 B/cycle BW design. Therefore, a collective scaling of compute logic, memory BW, and memory capacity is necessary to design improved DNN accelerator architectures.

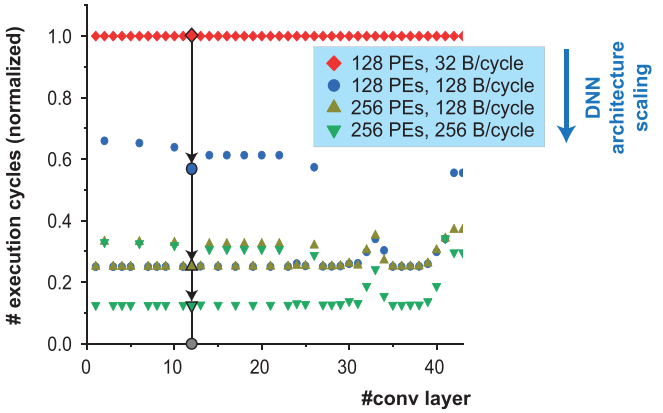

Fig. 2. Execution cycle improvements from more PEs and more memory BW. We use SSD-MobileNetv1 DNN benchmark using MAERI [9] DNN architecture.

Two-dimensional design scaling of the state-of-the-art systolic array accelerators, such as Google TPU [2] and Intel SpringHill [11], deteriorates the accelerator performance by loosely coupling the PEs and the memories. On the other hand, distributing the ON-chip memories across the design to tightly couple them with the PEs increases the data latency to the memory blocks located far away from the DRAM interface. In flexible interconnect accelerators with much higher BW, memory, and die area requirements, throughput improvement through 2-D design scaling becomes practically impossible, and efficient 3-D integration techniques are necessary to scale them.

B. Existing Works

3D-ReG [12] and AccuReD [13] perform heterogeneous 3-D integration of Re-RAM on top of GPUs to offer multifold throughput and energy efficiency improvement over traditional GPU-based acceleration with DRAMs. However, we have dedicated and improved hardware (accelerators) for training

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

DNNs and inferencing, offering better performance over GPU-based systems, such as Google TPU [2], SIGMA [10], Intel SpringHill [11], and MAERI [9].

Lu et al. [14] proposed a 3-D cross-ring DNN accelerator architecture involving 3D-SRAM cubes for energy-efficient inference by reducing DRAM access and eliminating temporal redundancy. Wang et al. [15] proposed a 3-D convolution model to improve data sharing between PEs for better throughput. These works propose new 3-D architectures or dataflow to achieve higher throughput and energy efficiency over the existing 2-D architectures. However, they do not consider the impact of multitier 3-D integration, bonding technology, frequency, and power improvement resulting from 3-D designs of the existing architectures.

Mathur et al. [16] performed thermal-aware design space exploration of multitier memory-on-compute 3-D systolic array accelerators. However, this work only analyses the power and performance of multitier designs without considering the actual physical design factors, including practical effects of power bump planning, PDN design and routing, and additional intertier vias/bonds. Ignoring them might cause prohibitive power and area overhead in actual 3-D IC implementation. Also, the multitier memory integration design in [16] offers lower throughput than their 2-D baseline. Furthermore, systolic arrays with rigid interconnections require lesser ON-chip memory BW and have low efficiency in sparse and irregular workloads, unlike flexible interconnect accelerators. But flexible accelerators offer higher efficiency through higher BW demand, which makes them better candidates for performance improvement through 3-D integration. Practical 3-D physical design implementation of flexible accelerators for better throughput, energy efficiency, and area efficiency remains an open research topic.

Unlike the works mentioned above, we do not focus on developing new 3-D accelerator designs or performing exploratory studies. Instead, our goal focuses on exploring multitier 3-D architecture and physical design co-optimization of the existing 2-D flexible accelerators to improve the overall throughput, energy efficiency, and area efficiency.

In this work, we perform actual 3-D physical designs of the accelerators using 3-D technology files in a commercial 2-D IC physical design tool to understand the real benefits of 3-D integration. We use 3-D LEF files to represent 3-D front-end-of-line (FEOL)/back-end-of-line (BEOL), 3-D LIB files for timing information of cells in different tiers, and 3-D qrctech files for *RC* parasitic information of the 3-D routing stack. We evaluate the designs using the metrics obtained from the actual physical design. However, the architectural evaluation of the accelerators is performed using modified 2-D simulators. The details of the simulator and how we modify it for our work are explained in Section III-B.

III. Design, Simulation, and EDA Methodology

A. DNN Architecture

Among several reconfigurable accelerators, we choose a state-of-the-art flexible interconnect architecture, MAERI [9], to perform and evaluate the benefits of 3-D flexible interconnect architecture designs. While our approach is generic to any accelerator architecture, our goal is to use an architecture that maximizes the usage of 3-D bonds. Flexible DNN architectures require more memory BW than systolic arrays or other rigid architectures. We believe the insights from MAERI could apply to other flexible accelerators such as SIGMA [10] as well. Fig. 3 shows the generic MAERI architecture. It has four major blocks: 1) prefetch buffers; 2) distribution tree/network (DN); 3) multiplier switches/PEs; and 4) reduction tree/network (RN). Prefetch buffers store the inputs and intermediate partial sums, while the distribution network and collection network handle the data movement between PEs and buffers. In this article, (2), (3), and (4) are collectively referred to as compute logic. We generate the required RTL using the open-source compiler [9] based on the architecture implemented.

B. Architecture Simulation Methodology

We evaluate the performance of the designs presented in this work using the STONNE [17] simulator, an open-source cycle-accurate simulator for the MAERI [9] accelerator. STONNE is a cycle-level microarchitectural simulator best suited for flexible DNN inference accelerators. It ensures that the workload dimension is always a multiple of the tile dimensions being mapped at a time to optimize reduction and distribution network usage. In addition to execution cycles of a given workload, STONNE provides static and dynamic energy numbers. The tool calculates these numbers using a lookup table (LUT) consisting of the cycle-level energy numbers of each module derived from synthesizing the MAERI RTL on Synopsys Design Compiler (DC). We updated these numbers with post-layer results from the physical design to generate accurate results.

Similar to the actual design, STONNE considers the reduction tree size to be equal to the number of the PEs. The mapping strategy used in the simulator is obtained from mRNA [18], which maximizes layer performance based on layer specifications. The intertier memory-to-logic communication is considered to be of one cycle. The cycle time is determined from the 3-D physical design. The simulator considers a 64-bit wide dual-channel DDR-5 (32 bits per channel) as the OFF-chip memory, which is connected to the ON-chip SRAM buffers. The PEs are connected using a linear topology, which share inputs and weights with each other. The adders are connected using a tree-based topology. These topologies are consistent with those used in MAERI.As the STONNE simulator does not provide any information on DRAM accesses, we use Timeloop [19] to simulate the DRAM access cycles and energy numbers.

To reduce the manual effort involved in performing simulations, we use a Python framework that incorporates every input to the simulator in the form of comma-separated files. The framework identifies the most optimal mapping for each convolution layer that results in the highest throughput. We use industry-standard inferencing and evaluation benchmarks, such as ResNet-50, SSD-MobileNetv1, GoogLeNet [20], and AlexNet [21] to evaluate our designs. The details of these workloads are shown in Table I.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DETAILS OF THE WORKLOADS USED IN THIS WORK

	ResNet-50	MobileNetv1	GoogLeNet	AlexNet
#Parameters	25.56 M	5.5M	4M	61.1M
#Conv. Layers	50	28	22	8
#Operations	4.12 B	1.14B	1.52 B	0.72 B
Top-5 accuracy	92.87%	70.89%	89.53%	79.09%

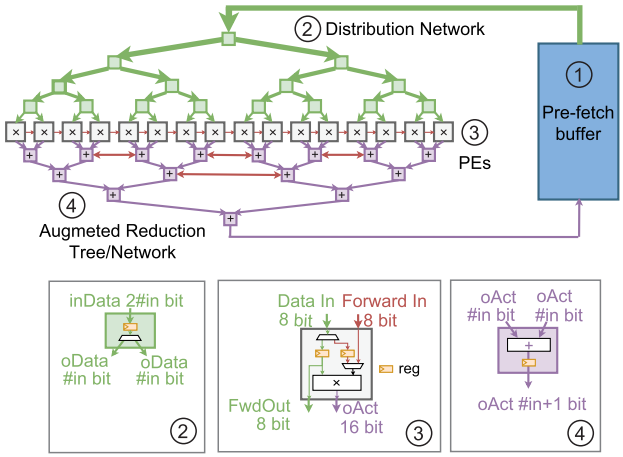

Fig. 3. MAERI [9] architecture.

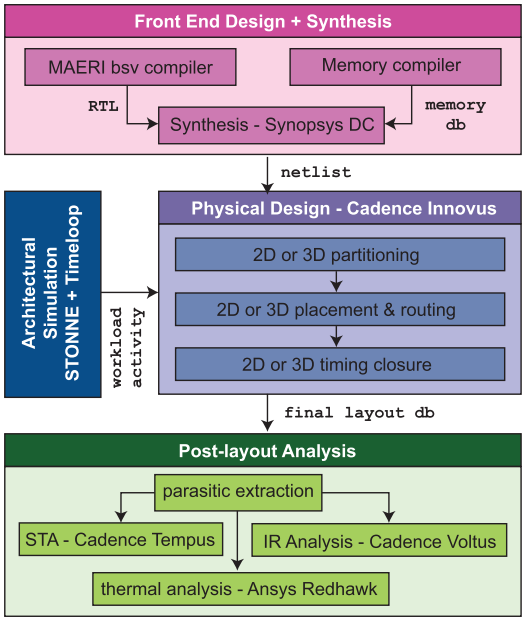

Fig. 4. EDA methodology used in this work.

C. Design Technology and EDA Methodology

We perform accelerator designs using a commercial 28-nm technology node. Our 3-D designs involve both the F2F and F2B 3-D integration. The F2F integration involves hybrid bonding, with 1-μm pitch and 0.5-μm diameter, between the corresponding tiers, and the F2B integration involves TSVs with 1-μm pitch and 0.5-μm diameter.

Fig. 4 shows the overall EDA methodology used in this work. We use MAERI blue spec Verilog (BSV) compiler [9] to generate the compute logic RTL of different MAERI accelerator configurations and a commercial 28-nm SRAM memory compiler to generate the ON-chip prefetch buffer design files. We synthesize MAERI netlist from the logic RTL and memory database (DB) files using Synopsys DC tool.

Post-synthesis, we perform place and route (P&R) using Cadence Innovus. In the case of multitier 3-D design, during the memory placement stage, we reduce the memory macros to site size (minimum possible size in a given technology node) and fix their locations on the memory tiers, as demonstrated in the Macro3D flow [22]. Since a 2-D tool cannot support multiple FEOL, we project the shrunk FEOL of all the memory tiers to the compute tier and perform standard cell placement in the compute tier.

After P&R, we evaluate our 3-D designs by performing static timing analysis (STA) using Cadence Tempus, power rail and IR-drop analysis using Cadence Voltus, and 3-D thermal analysis using Ansys Redhawk. These analyses help us understand the true performance benefits of 3-D accelerators over 2-D designs.

IV. THREE-DIMENSIONAL IC BOND PITCH IMPACT STUDY

The bond pitch used for 3-D IC integration plays a significant role in determining the feasibility of the target 3-D compute-on-memory ML accelerator designs, given the intense memory BW requirement in them. It is hard to meet the high BW requirement in these designs with μ-bumps, as they have pitches greater than 25 μm. In addition, μ-bumps require large drivers to maintain signal integrity across tiers. On the other hand, Intel has demonstrated that hybrid bonds with 10-μm pitches offer almost 5$\times$ lower bond parasitics and smaller drivers than μ-bumps with 40-μm pitches [23]. Hybrid bonds are a better candidate for compute-on-memory 3-D designs with high memory BW requirements and help avoid IO driver area and power overheads.

To understand the impact of 3-D bond pitches in accelerator physical design performance, we implemented a small 128 PE two-tier 3-D MAERI using three different bond pitches: 5, 2, and 1 μm. The diameter of these bonds is half that of the pitch size. The R/C values of the hybrid bonds used in this work with 5-μm pitch are 17 mΩ/0.1 fF, with 2-μm pitch are 44.2 mΩ/0.08 fF, and with 1-μm pitch are 97 mΩ/0.07 fF. The resistance and capacitance per unit length of the metal layers these bonds connect to are 9.05 Ω and 0.17 fF, respectively. Of these pitches, 5 μm is currently used for manufacturing, and the other two are expected to be available soon for production. We allow metal layer sharing (MLS) [24] between the two tiers so that the tool uses less capacitive routes to improve critical path timing. Therefore, the F2F pad count is higher than the logical connections between the memory and logic tier (determined by network BW). The number of nets routed using MLS is closely dependent on the F2F bonds that can be accommodated without causing a bonding violation and it reduces with increasing F2F bond pitch. Hence, choosing the right F2F pitch bond is critical to achieving the desired accelerator performance.

Based on the memory tier area, we first integrate 32 SRAM buffers, each with a 2-byte wide data bus and 8.125 kB capacity, providing a total memory capacity of 260 kB and a total memory BW of 64 B/cycle. We reconfigure the network BW on the logic tier accordingly. The variation in power and

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II
F2F PITCH IMPACT ON PPA OF MAERI [9]

3D bond pitch	5um	2um	1um	5um	2um	1um
#PEs	128	128	128	128	128	128
Buffer BW (B/cycle)	64	64	64	128	128	128
Buffer cap. (KB)	260	260	260	256	256	256
#F2F pads	8,142	39,789	48,473	6,111	60,719	89,729
Max. Freq. (GHz)	1.727	1.818	1.828	1.695	1.709	1.801
Power (W)	1.579	1.573	1.578	1.932	1.944	1.951
#F2F pad violations	752	36	0	2246	571	3

TABLE III
FOUR ARCHITECTURE CONFIGURATIONS USED IN OUR STUDY. STIFT DENOTES THE SPATIO-TEMPORAL INTEGRATED FOLDING TREE ARCHITECTURE [25]

Arch	#PEs	# Tiers	Network BW	Prefetch Buffer	STIFT [25]
Baseline 2D	1024	1	1 KB/cycle	256 KB	No
Scaled 2D	2048	1	2 KB/cycle	512 KB	Yes
2-tier 3D	2048	2	2 KB/cycle	2 MB	Yes
3-tier 3D	2048	3	2 KB/cycle	4 MB	Yes

performance with varying F2F pad pitches are summarized in Table II. Even for a BW of 64 B/cycle, 5-μm pitch introduces significantly higher F2F pad violations such as cut spacing and short violations. The 5-μm pitches are suitable for designs with BW lower than 64 B/cycle and low-frequency requirements. Pitches finer than 5 μm are necessary for higher BW and frequency. Both 2- and 1-μm bond pitches offer similar design metrics and handle more intertier connections.

We perform another set of designs with the three different bond pitches by doubling the memory BW to 128 B/cycle. We achieve this by integrating 64 SRAM buffers, each with a 2-B data bus and 4-kB memory, providing a total memory capacity of 256 kB. The reduction in memory capacity is based on the memory tier area. Again, the network BW in the logic tier is adjusted to support the new BW. The effect of varying 3-D bond pitches on the design with the 2× BW is tabulated in Table II. Even at this BW, the 1-μm pitch hardly has any 3-D bond issues. But the design with 2-μm bond pitch starts to show significant violations.

At the 28-nm node, with a support for 5-μm pitch technology, we can only design accelerators with dataflow restricted to values much lower than 64 B/cycle. To achieve a BW of 64 B/cycle, we at least require a 3-D technology that supports 2-μm bond pitch. To improve the BW beyond 128 B/cycle, we need 1 μm or finer pitches. In addition to increased BW, using finer pitches also improves design frequency. Therefore, we use 1-μm 3-D bond pitch in this work to implement our 3-D designs.

V. BASELINE 2-D ARCHITECTURE AND DESIGN

A. Architecture and Physical Design

The baseline 2-D MAERI architecture in this work has 1024 PEs, and a maximum distribution and collection network BW of 1024 B/cycle. The distribution BW is split between input activations and weights. The baseline design includes 256 B of prefetch buffer for each PE, offering a total ON-chip memory capacity of 256 kB (refer Table III).

We perform the physical design of the baseline configuration using a 28-nm technology node with six metal layers on the BEOL. Fig. 5(a) and (b) shows the generic floorplan

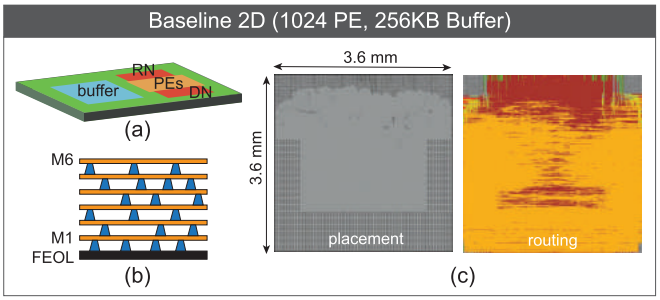

Fig. 5. Our baseline 2-D MAERI architecture with 1024 PEs and 256-kB prefetch buffer. (a) Floorplan, (b) layer stack, and (c) placement and routing GDS using a commercial 28-nm technology.

of the 2-D baseline design and its routing stack. Fig. 5(c) shows the physical design layouts of the baseline design. Our optimized memory placement brings the distribution and reduction networks closer to the prefetch buffers, and the PEs surrounds these network modules. The footprint of this design is 3.6 × 3.6 mm^2, and the maximum frequency is 600 MHz (refer Table IV).

B. Performance Analysis

We map various benchmarks listed in Section III-B onto baseline 2-D MAERI. The throughput observed ranges from 0.4 to 0.56 TOPS (tera-operations per second). The baseline 2-D design has energy efficiency in the range of 2.92–5.74 (TOPS/W). The energy efficiency reported in this work is based on the ON-chip power and does not include any OFF-chip components. The area efficiency varies from 31 to 43 (GOPS/mm^2). Table V lists workload-specific performance metrics.

VI. SCALED AND IMPROVED 2-D ARCHITECTURE AND DESIGN

A. Motivation for Architecture Scaling

Architecturally, the primary bottlenecks to performance improvement in baseline 2-D MAERI are the prefetch buffer BW, capacity, and long interconnects between the compute engines and the buffers. In addition, the MAERI architecture uses spatial buffers in the reduction tree to accumulate the partial sums generated by the PEs. The spatial-buffer-based reduction tree works fine for any DNN layer dimensions in which the controller maps input activation and weight rows required to generate an output pixel to the PEs in one cycle. However, it requires additional accumulators for bigger workloads to avoid computation stalling during partial sums' accumulation. To overcome this issue, the authors of MAERI suggest a spatio-temporal integrated folding tree (STIFT) [25]. In the case of large workloads, the spatio-temporal accumulation buffers in STIFT keep accumulating the intermediate partial sums till it generates final output pixels without stalling the computation. In other words, the STIFT acts as a near-memory adder and quantizer (NMAQ) unit and optimizes partial sum accumulation. We scale our 2-D baseline incorporating the above-mentioned architectural aspects and study the scaling impacts.

B. Architecture and Physical Design

The scaled 2-D MAERI accelerator has 2× PEs, 2× prefetch buffer BW and capacity than the 2-D baseline, and

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

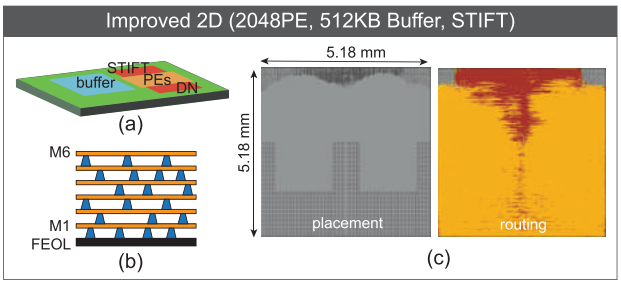

Fig. 6. Our improved 2-D MAERI architecture with 2048 PEs, 512-kB prefetch buffer, and STIFT enhancement [25]. (a) Floorplan, (b) layer stack, and (c) placement and routing GDS.

includes the STIFT reduction network. Our scaled 2-D design has 2048 PEs and offers a 512-kB prefetch buffer with a network BW of 2 kB/cycle, as shown in Table III. Fig. 6(a) and (b) shows the floorplan and routing stack, and Fig. 6(c) shows the layouts of the improved 2-D MAERI accelerator design. Similar to the baseline, the scaled 2-D design also uses six metal layers for signal routing.

The scaled 2-D design is a two-tile version of the baseline 2-D design with an STIFT reduction network. In Google TPU architectures [2], it has been shown that tiling 128 × 128 MAC unit subarray is more efficient than building bigger subarrays. We adopt a similar methodology for scaling in our work as well to ensure maximum design performance. The maximum frequency of the improved 2-D MAERI design is 580 MHz.

C. Performance Analysis

Table V shows workload-specific metrics. The throughput ranges from 0.87 to 1.30 TOPS, the energy efficiency from 2.81 to 5.46 (TOPS/W), and the area efficiency from 32 to 49 (GOPS/mm^2).

The STIFT network adds only 3% area overhead to the MAERI design while reducing computation stalling by around 30%. However, this tiny area overhead can make registers and memories even more loosely coupled than in the baseline 2-D design. We observe this in comparing different critical path groups between the baseline 2-D and scaled 2-D designs in Table VI. The scaled 2-D design has significantly higher path delays than the baseline 2-D. Integrating STIFT buffers causes more routing congestion around memories. To achieve similar operational speeds as the baseline design, scaled 2-D design needs high-strength timing buffers to optimize the critical paths, making the overall design more power-hungry. Thereby, it is not easy to have energy-efficient near-memory computing in flexible architectures as we scale them further through 2-D integration.

D. Baseline Versus Scaled 2-D Designs

The STIFT network reduces computation stalling in partial sum accumulations, as observed by more than 2× improvement in scaled 2-D design throughputs over than the baseline 2-D (see Table V). However, the architectural improvement in the scaled 2-D design leads to frequency degradation and higher design power. The energy efficiency improvement starts dropping from 7% to −5%, proving our claim that 2-D accelerator design scaling becomes less energy-efficient as the workload grows. Similarly, the area efficiency improvement drops from 41% to −15%. Therefore, 2-D design and architectural scaling of memory-intensive flexible interconnect accelerator architectures, such as MAERI, makes them less energy-efficient on larger workloads.

VII. Two-Tier 3-D Architecture and Design

A. Performance Bottlenecks in 2-D Designs

As seen in Table VI, the major performance bottleneck in 2-D accelerators is the critical paths between the SRAM buffers and the internal register logic. To optimize the mem2reg or reg2mem paths, several register-to-register (reg2reg) paths in the design are underoptimized. These collectively lead to lower design frequency as we scale the accelerators through 2-D integration. The NMAQ units also introduce several signal nets to the controller and ON-chip memory buffers. These further increase the routing overhead in the 2-D design, overshadowing the architectural benefits brought about by the NMAQ units.

B. Motivation for 3-D Integration

Commercial accelerators compensate for the shortcomings in 2-D accelerator design scaling using multiple memory hierarchies. Often, the external memory links involve BW-limited communication protocols, such as AXI, due to limited routing resources in 2-D scaling. With bigger workloads, these protocols cause computation stalling. In this section, we address the architecture and design scaling issues in 2-D accelerators through compute-on-memory 3-D integration. Three-dimensional integration [22] shows up to 28% performance and power improvement through memory-on-logic 3-D integration of processor designs with a considerable amount of ON-chip memory. Unlike processors, memory blocks are a significant component of accelerator designs, making them better candidates for compute-on-memory 3-D integration.

In 2-D MAERI, the placement of ON-chip memory around the 2-D computation causes long wires in the tree-based reduction network, which transfer data between the PEs and the ON-chip memory. However, by adopting 3-D integration, the wirelength can be reduced through the vertical stacking of ON-chip memory beneath or on top of the compute logic. This integration approach significantly enhances the scalability of the 2-D MAERI design.

C. Architecture and Physical Design

Our two-tier 3-D MAERI design has one compute tier and one memory tier integrated through F2F hybrid bonding (1-μm pitch). Fig. 7(a) shows the floorplan of the two-tier design. The external connections in the accelerator are predominantly to and from the memories. Therefore, we place the memories on the bottom tier to avoid many multitier external connections to other SoCs or package, as the external pins and C4 bumps are on the bottom tier.

The two-tier design has the same number of PEs and ON-chip BW as the scaled 2-D design, but we scale the on-chip memory capacity to 2 MB based on the additional area

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

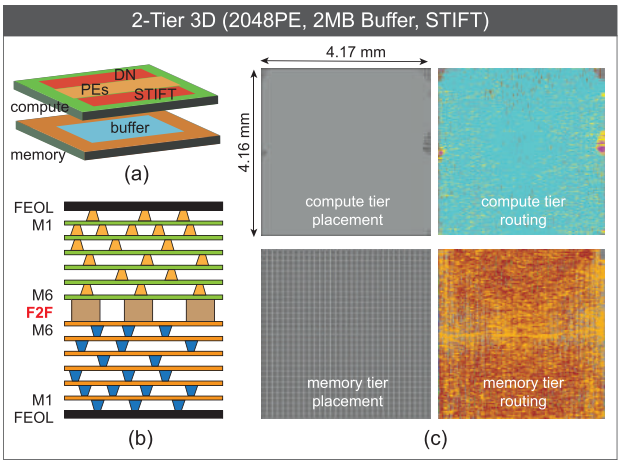

Fig. 7. Our two-tier 3-D MAERI architecture with 2048 PEs, 2-MB prefetch buffer, and STIFT enhancement [25]. (a) Floorplan, (b) layer stack, and (c) placement and routing GDS.

available in the memory tier. This additional memory is used to reduce DRAM accesses between two consecutive layers of DNN inference. Fig. 7(b) and (c) shows the routing stack and the final layouts of the two-tier design. The 3-D design has the NMAQ units and the reduction/distribution networks directly above the corresponding ON-chip memory subarray, leading to an improved frequency of 1 GHz, which is 1.67× faster than the 2-D designs.

D. Performance Analysis

The throughput varies from 1.48 to 2.24 TOPS, with the highest throughput seen on the ResNet-50 workload. The energy efficiency is in the range of 3.09–6.16 (TOPS/W). The area efficiency varies from 86 to 129 (GOPS/mm^2). With compute-on-memory 3-D integration, we improve the critical paths between the memories and the register, achieving higher design frequency. Our two-tier 3-D design offers throughput, energy efficiency, and area efficiency improvement of up to 5×, 1.2×, and 3.88×, respectively, compared with the baseline 2-D design, and up to 1.7×, 1.1×, and 2.7×, respectively, compared with the scaled 2-D design. Table V summarizes these results.

Using 3-D integration, we could not only rearchitect the design with increased compute logic and memory BW/ capacity but also bring the accumulators close to the ON-chip memory and include STIFT buffers to avoid compute stalling. Performing this in the 2-D design had a negative impact on the overall design frequency, as explained in Section VI-D. In addition to offering technology-induced benefits, 3-D integration also enables architecture improvement through faster near-memory addition and quantization (NMAQ).

VIII. Three-Tier 3-D Architecture and Design

A. Energy-Efficient Accelerator Design

With the ability to integrate high-capacity ON-chip memories in 3-D design, 3-D integration facilitates energy-efficient accelerator designs. As the workloads grow, we can scale the ON-chip memory in the accelerators further through vertical integration of additional memory tiers to reduce high-energy external memory accesses.

B. Architecture and Physical Design

Our three-tier MAERI design offers one compute tier with 1024 PEs and two memory tiers with 2 MB of ON-chip buffers per tier, totaling 4 MB. Fig. 8(a) shows the floorplan of our three-tier design. The three-tier design has 2× ON-chip memory of that of the three-tier design. The additional memory is used to reduce the total number of DRAM transfers of outputs (of the current layer)/input activations (of the next layer) while switching from one inference layer to the other.

There are two different techniques for performing three-tier 3-D designs: Monolithic 3-D involving F2B integration only or a mix of F2B and F2F integration. In F2B integration, TSVs bond multiple tiers. As TSVs penetrate through the silicon, the memory placement or the TSV usage gets restricted, especially in memory-intensive designs, such as accelerators. On the other hand, F2F integration does not pose any restriction on intertier bond usage, making it a better candidate for multitier accelerator integration. However, it is impossible to integrate all three tiers using F2F integration, and therefore, we use a combination of F2F and F2B integration to perform the three-tier MAERI design. Fig. 8(b) shows the routing stack used for this integration.

We perform a three-tier MAERI design using an 18-metal layer routing stack [Fig. 8(b)], with two memory tiers on the bottom and one compute tier on the top. (We also performed a design with F2F bonding between the compute and the middle memory tier, but its effective frequency was 11.7% lesser.) Similar to a two-tier 3-D design, we place the memories closer to the package for better connectivity to external memories. Fig. 8(c) shows the physical design layouts of our three-tier design. Even though the memory capacity doubles, the memory subarray sizes do not exactly double, leading to a smaller footprint of the three-tier design than the two-tier design.

C. Performance Analysis

Table V shows the workload-specific performance results of the three-tier 3-D design. The three-tier design offers the same throughput as that of the two-tier design, as the number of PEs, ON-chip memory BW, and design frequency are the same between the two designs. The energy efficiency is in the range of 3.07–6.03 (TOPS/W). The SRAM power increases for bigger workloads, such as ResNet50 and GoogLeNet, leading to a drop in energy efficiency. The primary energy-saving in the 3-D design stems from the reduction in DRAM accesses. For a slight increase in SRAM energy, the DRAM access energy reduces to up to 27% in the two-tier design and 46% in the three-tier design. Fig. 9 shows workload-based PE, SRAM, and DRAM energy comparisons. The PE energy in the 3-D designs is less than the scaled 2-D design due to lesser high-strength timing buffers, resulting in 72% higher frequency with only a 52%–56% increase in ON-chip power. The area efficiency of the three-tier design varies from 89 to 133 (GOPS/mm^2).

While the throughput improves in the 3-D designs, the chip power also considerably increases due to their higher frequency of operation. Therefore, the energy efficiency is consistent across all the designs. However, the footprint of

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

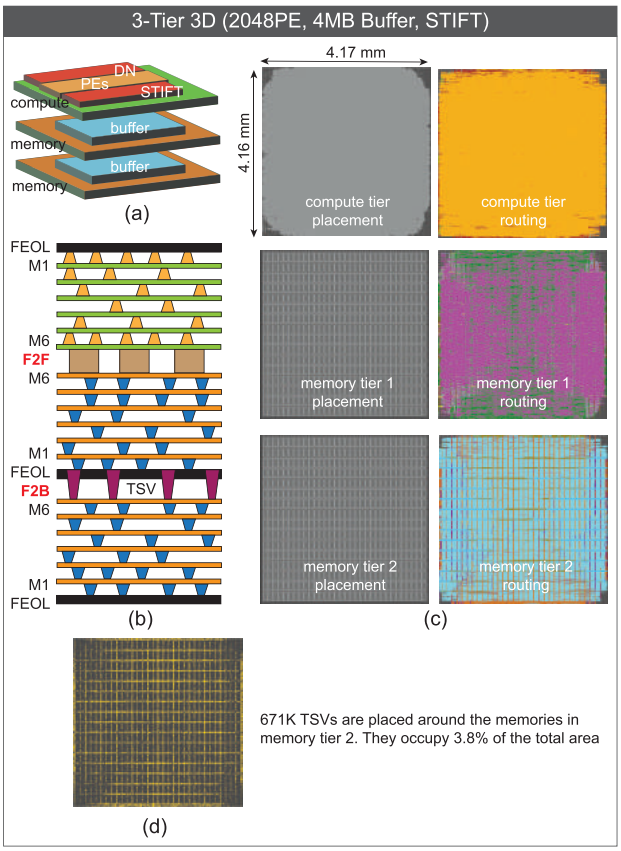

Fig. 8. Our three-tier 3-D MAERI architecture with 2048 PEs, 4-MB prefetch buffer, and STIFT enhancement [25]. (a) Floorplan, (b) layer stack, (c) placement and routing GDS, and (d) TSV locations.

TABLE IV

PHYSICAL DESIGN METRIC COMPARISON. WORKLOAD-BASED POWER NUMBERS ARE SHOWN IN TABLE V

	Baseline 2D	Scaled 2D	2-tier 3D	3-tier 3D
Footprint (mm^2)	13.01	26.78	17.36	16.80
Utilization	70.4%	73.7%	70.7%	70.2%
#Metal layers	6	6	6+6	6+6+6
Wirelength (m)	122.5	408.1	367.7	364.7
#Gates	3.6M	8.3M	8.4M	8.1M
#F2F pads	-	-	556.0K	419.5K
#TSVs	-	-	-	671.8K
Max. Freq. (MHz)	600	580	1000	1000

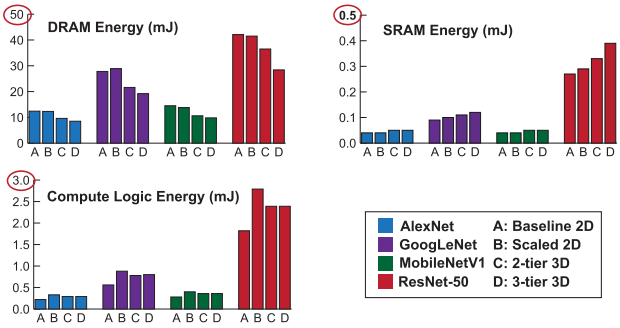

Fig. 9. Workload-based DRAM, SRAM, and PE energy comparison. Note: The energy numbers are on different scales.

the 3-D designs is much lesser than that of the scaled-2-D design (see Table IV). Therefore, the increased throughput in 3-D designs leads to a much higher area efficiency in them than the 2-D designs.

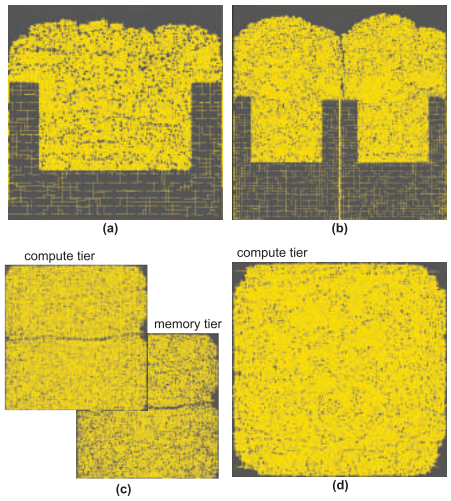

Fig. 10. Clock routing in (a) baseline 2-D, (b) scaled 2-D, (c) two-tier 3-D, and (d) three-tier 3-D MAERI designs. The routing is enhanced for visibility. The three-tier design has insignificant clock routing on memory tiers.

IX. THREE-DIMENSIONAL PHYSICAL DESIGN ANALYSIS

In this section, we analyze our 3-D physical designs and compare them against the 2-D designs.

A. Timing Critical Path Analysis

Fig. 10 shows the clock routing of the four different MAERI designs. In general, clock routing is clustered around the compute logic. In the case of the two-tier 3-D design, the F2F bonding between the tiers permits clock signals to use the memory tier's M5 and M6 metal layers for better routing. But in the case of the three-tier 3-D design, the limited space available for TSV bonds between the compute and memory tier is mainly reserved for memory signal routing optimization. The major part of the clock tree is in the compute tier. The memory tiers only have direct connections to the memory clock pins.

Table VI presents the clock metrics summary comparison. The presence of the NMAQ unit/STIFT network increases the clock latency across all three MAERI designs with them. But the latency is 9.3% and 18.6% less in the two-tier and three-tier 3-D designs, respectively, compared with the scaled 2-D design. The clock metrics of the three-tier 3-D are noticeably better than the two-tier 3-D. Splitting the memory clock pins across two tiers in the three-tier design offers better clock routability, making three-tier clock metrics better than two-tier 3-D.

Fig. 11 shows the critical paths, and Table VI shows comparison of their metrics. The critical path delays are significantly higher in the scaled 2-D design than in the baseline due to the NMAQ. In the two-tier 3-D design, we observe almost up to 34% improvement in these delays due to less routing congestion. In the case of the three-tier 3-D design, the maximum improvement is 31%, which is slightly lower than the two-tier design due to multitier vias.

B. PDN Design

In 2-D designs, we flip the whole chip and connect the package power/ground bumps to the topmost metal layer in the

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

TABLE V

WORKLOAD SIMULATION COMPARISON. ALL METRICS ARE CALCULATED BASED ON ON-CHIP COMPONENTS

	AlexNet				GoogLeNet				MobileNetv1				ResNet-50			
	base 2D	scaled 2D	2-tier 3D	3-tier 3D	base 2D	scaled 2D	2-tier 3D	3-tier 3D	base 2D	scaled 2D	2-tier 3D	3-tier 3D	base 2D	scaled 2D	2-tier 3D	3-tier 3D
Frequency	600 MHz				580 MHz				1 GHz				1 GHz			
Tera-ops per second (TOPS)	0.4	1.17	2.01	2.01	0.50	0.87	1.48	1.48	0.43	0.90	1.55	1.55	0.56	1.30	2.24	2.24
Normalized TOPS	1	2.92	5.02	5.02	1	1.74	2.98	2.98	1	2.09	3.60	3.60	1	2.32	4.00	4.00
Power (mW)	86	234	361	361	104	176	279	287	148	321	503	506	97	238	364	372
Energy efficiency (TOPS/W)	4.65	4.97	5.56	5.56	4.75	4.90	5.34	5.21	2.92	2.81	3.09	3.07	5.74	5.46	6.16	6.03
Normalized	1	1.07	1.20	1.20	1	1.03	1.12	1.10	1	0.96	1.06	1.06	1	0.95	1.07	1.05
Area efficiency (GOPS/mm^2)	31	44	116	120	38	32	86	89	33	34	90	92	43	49	129	133
Normalized	1	1.41	3.76	3.88	1	0.85	2.26	2.34	1	1.01	2.69	2.78	1	1.13	3.02	3.12

TABLE VI

CRITICAL PATH ANALYSIS

	Base 2D	Scaled 2D	2-tier 3D	3-tier 3D
clock metrics				
Max latency (ns)	0.96	1.61	1.46	1.31
Max skew (ns)	0.1	0.29	0.35	0.29
Wirelength (m)	2.51	5.39	10.01	8.64
#Buffers	57,041	125,050	127,149	119,547
mem2reg critical path				
Req. time (ns)	2.497	3.062	2.089	2.250
Arr. time (ns)	2.494	3.057	2.020	2.104
Slack (ps)	3	5	69	146
reg2mem critical path				
Req. time (ns)	2.482	2.943	2.125	2.311
Arr. time (ns)	2.477	2.905	1.959	2.002
Slack (ps)	5	38	166	309
reg2reg critical path				
Req. time (ns)	2.490	3.077	2.122	2.246
Arr. time (ns)	2.486	3.045	2.118	2.243
Slack (ps)	4	32	4	3

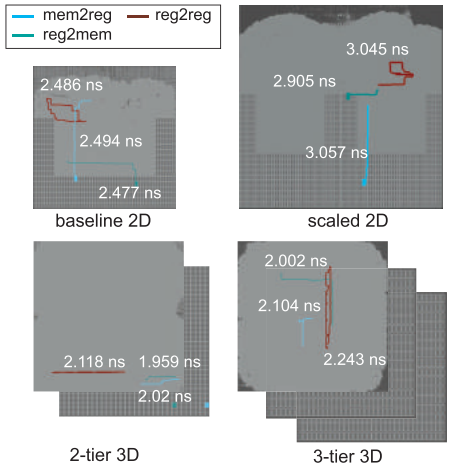

Fig. 11. Timing critical paths.

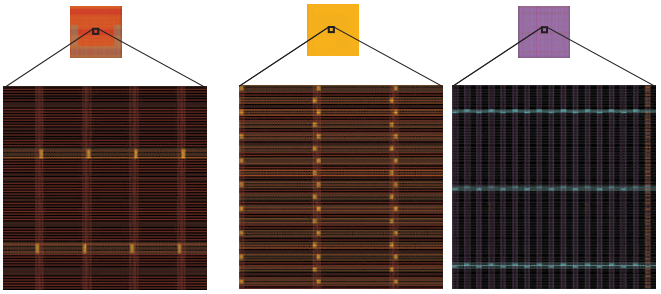

Fig. 12. PDN routing in (a) baseline 2-D and (b) two-tier 3-D MAERI designs.

TABLE VII

POWER GRID USAGE AND IR-DROP ANALYSIS. VDD = 0.9 V

	Base 2D	Scaled 2D	2-tier 3D	3-tier 3D
C4 bumps	324	624	256	256
C4 bump pitch (μm)	200	200	260	250
Hybrid bond pads	-	-	347.4K	311.6K
TSVs	-	-	-	467.8K
PDN area (mm^2)	2.01	3.98	8.45	9.17
% PDN area	2.58%	2.69%	4.10%	3.03%
PDN metal layers	2,5,6	2,5,6	Compute: 2,5,6 Memory: 4,5,6	Compute: 2,5 Memory: 4,5,6
Max. IR-drop (mV)	87.5	96.4	96.9	95.8

design. Therefore, the C4 bump pitch values can be flexible in the 2-D design. We use a C4 power/ground bump pitch of 200 μm in our 2-D designs to limit the IR-drop to be within 10% of VDD. The baseline 2-D has 162 VDD and 162 VSS bumps, and the scaled 2-D design has 312 VDD and 312 VSS bumps. Fig. 12(a) shows PDN routing in baseline 2-D. We use M2, M5, and M6 layers for PDN routing in 2-D designs.

Our 3-D designs place memories on the bottom tiers to make the external memory connections easier to meet timing. We place around 160 k (two-tier)–200 k (three-tier) external signal and clock pins on the bottom tier and restrict their location to the six metal layers on the bottom tier. These external connections are to the accelerator controller within the same package. Unlike 2-D designs, the C4 bumps for power/ground connections in the case of 3-D designs have to connect to the bottom-most metal layer of the memory tier. However, the clustered nature of memories in the bottom tier, as shown in Figs. 7 and 8(c), restricts the number of microbumps or C4 bumps' locations in the design. The key to achieving finer C4 bump pitches in memory-dominated designs is to have smaller memory subarrays and add sufficient spacing between them. With a whole tier dedicated to memory placement in 3-D designs, the slight area increase in splitting a memory block into smaller subarrays does not cause any placement congestion or routing overhead. We use power/ground bump pitches of 260 and 250 μm in our two-tier and three-tier 3-D designs, respectively, to place the C4 bumps efficiently in the gaps between the memories. Our two-tier and three-tier 3-D designs have 128 VDD and 128 VSS bumps each. Fig. 12(b) shows the PDN routing in the two-tier 3-D design. We use metal layers M2, M5, and M6 of the compute tier and M4–M6 of the memory tier for PDN routing in the two-tier design. We use layers M2 and M5 of the compute tier, M4–M6 layers

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
COMPARISON OF OUR WORK AGAINST EXISTING 3-D ACCELERATOR WORKS

Name	Technology	#PEs	Mem. Capacity	#Compute tiers	#Mem. tiers	3D:2D throughput	3D:2D energy-efficiency
Thermal-aware 3D systolic array [16]	16 nm	1,024	512 KB	1	4	0.96	1.53
Thermal-aware 3D systolic array [16]	16 nm	4,096	128 KB	4	1	2.99	1.64
3D Stacked NN accelerator [27]	7 nm	1,024	8 MB	1	4	1.9	1.4
3D Stacked NN accelerator [27]	7 nm	2,048	16 MB	1	4	3.9	1.6
3D-aCortex [28]	55nm	16,384	1 MB	1	64	0.71	0.30
Our work	28 nm	2048	4 MB	1	2	5.02	1.2

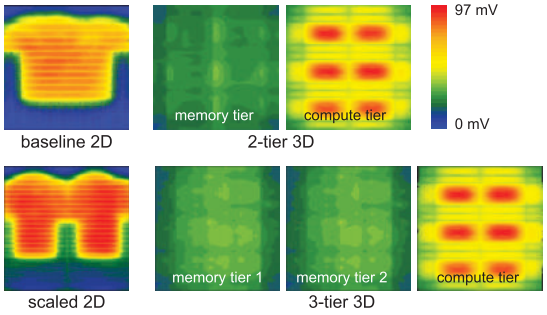

Fig. 13. IR-drop map comparison.

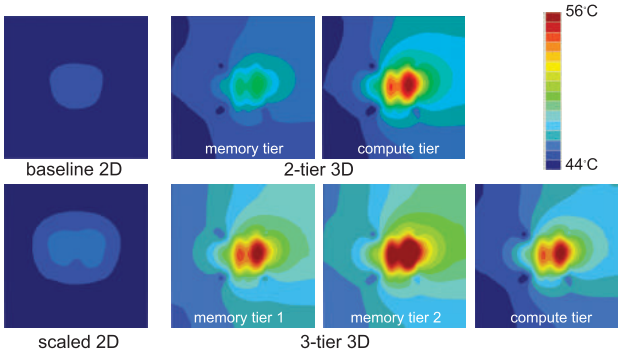

Fig. 14. Temperature map comparison.

of the middle memory tier, and M1, M4–M6 of the bottom memory tier for PDN routing in the three-tier design.

Table VII summarizes the PDN-related metrics. Our 3-D designs have a higher PDN area to limit the IR-drop. The increased 3-D PDN area does not affect signal routing, as our 3-D designs offer significant wirelength (see Table IV) and frequency (see Table V) improvements over the 2-D design.

C. IR-Drop Analysis

We use Cadence Voltus to perform power rail analysis and use them to generate the IR-drop map of our 2-D and 3-D designs using the techniques presented in [26]. We perform the IR-drop analysis of our designs considering an input switching activity of 10% to account for the worst possible scenario. The power consumption corresponding to the workloads used in this work is much less than the power corresponding to a steady-state switching activity of 10%. Fig. 13 shows the IR-drop maps of the four designs presented in this work. The maximum IR-drop occurs around the compute logic, and the minimum IR-drop occurs around the memories in both the 2-D and 3-D designs. Table VII summarizes the maximum IR-drop of all four designs. The % PDN area denotes the portion of PDN routing tracks in the total routing tracks used for routing. Despite the restricted placement of power bumps on the memory tiers of 3-D design, the IR-drop in our 3-D designs is within 10.7% of the VDD.

D. Thermal Analysis

We perform thermal analysis of our designs using Ansys RedHawk-CTA tool, based on the technique shown in [29]. We generate the compact thermal models (CTMs) of each tier at their maximum design frequency. The CTMs are then stacked on to a package model according to the integration style used. We use a dummy air-cooled package substrate with signal and power routing to simulate the thermal conductivity of an actual package. We place a heat sink on the top tier of 3-D designs. Similar to the IR-drop analysis, we consider a design power corresponding to an average steady-state switching activity of 10% to account for the worst possible scenario. The assumed switching activity corresponds to that of the average DNN workloads. The ambient temperature for thermal analysis is considered to be 25 °C. The tool divides the design into multiple tiles and calculates power density and thermal conductivity within each tile. The tile dimensions are 5 × 5 μm. As the 3-D bond sizes are much smaller than the tile dimensions, multiple 3-D bonds are within a single tile.

Fig. 14 shows the temperature maps of all four designs. The maximum on-die temperature variation in our 3-D designs is within 12 °C.

X. COMPARISON AGAINST PREVIOUS WORKS

Table VIII shows the comparison of our work against recent 3-D ML accelerator works. In this table, we have normalized the throughput and energy efficiency benefits of 3-D ML accelerators with respect to the corresponding 2-D designs presented in the works shown. We compare our work against the recent works of thermal-aware 3-D ML systolic arrays [16] (2021), 3-D stacked neural network accelerator for AR/VR application [27] (2022), and 3-D NAND flash memory-based 3D-aCortex accelerator [28] (2021), which are rigid interconnect architectures. Our analysis shows that our 3-D work offers a 1.68× throughput improvement over the best recent 3-D ML accelerator work, for just a 0.27× reduction in energy efficiency. It is also to be noted that other works are at a more recent node than our work and still offer higher throughput. If we extend our 3-D design methodology to more recent technology nodes, we can obtain further frequency, throughput, and energy efficiency improvement. This also shows that flexible interconnect accelerator designs show more

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.

improvements with 3-D integration than systolic arrays or other rigid interconnect architectures.

XI. Conclusion

Flexible interconnect DNN accelerators offer energy-efficient throughput improvement compared with systolic arrays by maximizing PE utilization through flexible data distribution and reduction. However, they also require significantly higher ON-chip memory capacity and BW than systolic arrays. Growing DNN workloads demand the corresponding scaling of DNN accelerator architectures and designs. But we see that scaling the near-memory computing architecture through NMAQ units and the compute logic in these accelerators through 2-D integration is energy-inefficient and does not use the full potential of the enhanced architecture.

On the other hand, compute-on-memory 3-D integration helps build high-speed and energy-efficient flexible interconnect DNN accelerators by enabling their full potential of near-memory computing. Unlike [16], [27], using 3-D integration, we improve the throughput, energy efficiency, and area efficiency by 5×, 1.2×, and 3.9×, respectively, in MAERI. Furthermore, increasing the total ON-chip memory capacity by integrating an additional memory tier, we also reduced DRAM access energy by up to 46%. Through efficient memory design and power planning, the IR-drop is within 10.7% of VDD. We also optimize the external memory connections by placing the memories on the bottom tier. This helps place the heat sink closer to highly active compute blocks on the top tier, thereby limiting the temperature variance across the chip to 12 °C.

References

[1] K. Siu, D. M. Stuart, M. Mahmoud, and A. Moshovos, "Memory requirements for convolutional neural network hardware accelerators," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Sep. 2018, pp. 111–121.

[2] N. P. Jouppi et al., "A domain-specific supercomputer for training deep neural networks," *Commun. ACM*, vol. 63, no. 7, pp. 67–78, Jun. 2020.

[3] J. Kim, L. Zhu, H. M. Torun, M. Swaminathan, and S. K. Lim, "Micro-bumping, hybrid bonding, or monolithic? A PPA study for heterogeneous 3D IC options," in *Proc. 58th ACM/IEEE Design Autom. Conf. (DAC)*, Dec. 2021, pp. 1189–1194.

[4] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung, "Accelerating deep convolutional neural networks using specialized hardware," *Microsoft Res. Whitepaper*, vol. 2, pp. 1–4, Feb. 2015.

[5] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. El Sallab, "YOLO3D: End-to-end real-time 3D oriented object bounding box detection from LiDAR point cloud," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 716–728.

[6] S. Gao et al., "Abstractive text summarization by incorporating reader comments," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6399–6406.

[7] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," in *Proc. 58th Annu. Meeting Assoc. for Comput. Linguistics*, 2020, pp. 2177–2190.

[8] S. Lym and M. Erez, "FlexSA: Flexible systolic array architecture for efficient pruned DNN model training," 2020, *arXiv:2004.13027*.

[9] H. Kwon, A. Samajdar, and T. Krishna, "A communication-centric approach for designing flexible DNN accelerators," *IEEE Micro*, vol. 38, no. 6, pp. 25–35, Nov. 2018.

[10] E. Qin et al., "SIGMA: A sparse and irregular GEMM accelerator with flexible interconnects for DNN training," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 58–70.

[11] O. Wechsler, M. Behar, and B. Daga, "Spring Hill (NNP-I 1000) Intel's data center inference chip," in *Proc. IEEE Hot Chips 31 Symp. (HCS)*, Aug. 2019, pp. 1–12.

[12] B. Li, J. R. Doppa, P. P. Pande, K. Chakrabarty, J. X. Qiu, and H. Li, "3D-ReG: A 3D ReRAM-based heterogeneous architecture for training deep neural networks," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 16, no. 2, pp. 1–24, Jan. 2020, doi: 10.1145/3375699.

[13] B. K. Joardar, J. R. Doppa, P. P. Pande, H. Li, and K. Chakrabarty, "AccuReD: High accuracy training of CNNs on ReRAM/GPU heterogeneous 3-D architecture," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 5, pp. 971–984, May 2021.

[14] W. Lu, P.-T. Huang, H.-M. Chen, and W. Hwang, "An energy-efficient 3D cross-ring accelerator with 3D-SRAM cubes for hybrid deep neural networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, pp. 776–788, Dec. 2021.

[15] Y. Wang, Y. Wang, C. Shi, L. Cheng, H. Li, and X. Li, "An edge 3D CNN accelerator for low-power activity recognition," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 5, pp. 918–930, May 2021.

[16] R. Mathur, A. K. A. Kumar, L. John, and J. P. Kulkarni, "Thermal-aware design space exploration of 3-D systolic ML accelerators," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 7, no. 1, pp. 70–78, Jun. 2021.

[17] F. Muñoz-Martínez, J. L. Abellán, M. E. Acacio, and T. Krishna, "STONNE: Enabling cycle-level microarchitectural simulation for DNN inference accelerators," *IEEE Comput. Archit. Lett.*, vol. 20, no. 2, pp. 122–125, Jul. 2021.

[18] Z. Zhao, H. Kwon, S. Kuhar, W. Sheng, Z. Mao, and T. Krishna, "MRNA: Enabling efficient mapping space exploration for a reconfiguration neural accelerator," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Mar. 2019, pp. 282–292.

[19] Y. N. Wu, V. Sze, and J. S. Emer, "An architecture-level energy and area estimator for processing-in-memory accelerator designs," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Aug. 2020, pp. 116–118.

[20] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1–9.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. New York, NY, USA: Curran Associates, 2012, pp. 1–9.

[22] L. Bamberg, A. Garcia-Ortiz, L. Zhu, S. Pentapati, and S. K. Lim, "Macro-3D: A physical design methodology for face-to-face-stacked heterogeneous 3D ICs," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2020, pp. 37–42.

[23] A. Elsherbini, S. Liff, J. Swan, K. Jun, S. Tiagaraj, and G. Pasdast, "Hybrid bonding interconnect for advanced heterogeneously integrated processors," in *Proc. IEEE 71st Electron. Compon. Technol. Conf. (ECTC)*, Jun. 2021, pp. 1014–1019.

[24] S. Pentapati and S. K. Lim, "Metal layer sharing: A routing optimization technique for monolithic 3D ICs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 30, no. 9, pp. 1355–1367, Sep. 2022.

[25] F. Muñoz-Martínez, J. L. Abellán, M. E. Acacio, and T. Krishna, "STIFT: A spatio-temporal integrated folding tree for efficient reductions in flexible DNN accelerators," *ACM J. Emerg. Technol. Comput. Syst.*, May 2022.

[26] L. Zhu, C. Jo, and S. K. Lim, "Power delivery solutions and PPA impacts in micro-bump and hybrid-bonding 3D ICs," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 12, no. 12, pp. 1969–1982, Dec. 2022.

[27] L. Yang et al., "Three-dimensional stacked neural network accelerator architectures for AR/VR applications," *IEEE Micro*, vol. 42, no. 6, pp. 116–124, Nov. 2022.

[28] M. Bavandpour, S. Sahay, M. R. Mahmoodi, and D. B. Strukov, "3D-aCortex: An ultra-compact energy-efficient neurocomputing platform based on commercial 3D-NAND flash memories," *Neuromorphic Comput. Eng.*, vol. 1, no. 1, Jul. 2021, Art. no. 014001, doi: 10.1088/2634-4386/ac0775.

[29] L. Zhu et al., "High-performance logic-on-memory monolithic 3-D IC designs for arm Cortex-A processors," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 6, pp. 1152–1163, Jun. 2021.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 12:29:20 UTC from IEEE Xplore. Restrictions apply.