Heterogeneous Mixed-Signal Monolithic 3-D In-Memory Computing Using Resistive RAM

Gauthaman Murali, Xiaoyu Sun, Shimeng Yu, *Senior Member, IEEE*,
and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*— Resistive random access memory (RRAM)-based compute-in-memory architecture helps overcome the bottleneck caused by large memory transactions in the convolutional neural network (CNN) accelerators. However, their deployment using 2-D IC technology faces challenges, as today's RRAM cells remain at legacy nodes above 20 nm due to high programming voltages. Besides, power-hungry analog-to-digital converter (ADC) units limit the throughput of RRAM accelerators. In this article, we present the first-ever heterogeneous (multiple technology nodes) mixed-signal monolithic 3-D IC designs of the RRAM CNN accelerator. Our RRAM remains at legacy 40-nm nodes in one tier, but CMOS periphery scales toward advanced 28/16 nm in another tier. Our 3-D designs overcome the bottleneck caused by ADCs and offer up to 4.9× improvement in energy efficiency in TOPS/W and up to 50% reduction in footprint area over 40-nm 2-D IC designs. Compared with existing 2-D works, our 3-D architecture offers up to 28.6× improvement in energy efficiency.**

***Index Terms*— Hybrid integrated circuits, neural network, resistive random access memory (RRAM), three-dimensional integrated circuits.**

I. Introduction

DEEP neural networks (DNNs) have shown remarkable success in various intelligent applications, including computer vision, speech recognition, and natural language processing. The state-of-the-art DNNs tend to aggressively increase the depth and size of the model to achieve incremental accuracy improvement, which poses grand challenges to the hardware implementations on the edge platforms with stringent power/area budget. To address the challenges, CMOS ASIC-based accelerators have been developed, such as Google's TPU [1] and MIT's Eyeriss [2], which demonstrates higher area and energy efficiency, compared with the mainstream graphics processing units (GPUs). However, the massive data communications between memory and processing elements (PEs) remain as a roadblock to further improving the throughput and energy efficiency. Therefore, it is of great interest to embed the computation into the memory array itself, namely computing-in-memory (CIM), to minimize the data communications. To implement CIM, SRAM-based designs have been proposed [3]. However, one SRAM cell consumes a significant area (>150 F2, F is the technology node) and suffers from the increasing leakage power with CMOS scaling, thus making the emerging nonvolatile memories (eNVMs) more preferable for inference applications. Among the eNVM candidates, including resistive random access memory (RRAM) [4], spin-transfer-torque magnetic random access memory (STT-MRAM) [5], and phase-change memory (PCM) [6], we adopt RRAM for its simple and low-cost integration with CMOS process.

The multiply-and-accumulate (MAC) operation in CIM is the dominant computation in DNNs. It is initiated by asserting multiple or all the rows of the memory array simultaneously. Then, the elementwise multiplication is realized as the multiplication between the input voltage and cell conductance, and the sum is added up as the column current following Kirchhoff's law. Analog-to-digital converters (ADCs) are typically deployed at the edge of the memory array to convert the analog current/voltage-to-digital code for further processing (e.g., shift-and-add, activation, and pooling). To avoid a large footprint, several columns of a memory array share the ADCs through column multiplexing. This sharing of ADCs notably undermines the superiority of CIM in terms of throughput and energy efficiency [4]. To mitigate the ADC bottleneck, monolithic 3-D (M3D) IC stacking is an attractive solution that could not only spare more space to accommodate the ADCs but also reduce the interconnect power consumption by effectively reducing the wire length.

In M3D, minuscule monolithic intertier vias (MIVs) (diameter $<$ 400 nm) are used for the connections crossing the tiers, leading to highly denser vertical integration compared with through-silicon vias (TSVs) [7]. Unlike TSVs, MIVs are highly reliable, as they are similar to metal-to-metal vias. MIVs are as robust to defects as any metal-to-metal vias. However, in the case of defective MIV, it is impossible to single out them and fix them individually like any other vias, and such defects are usually included in the yield calculation. The work [8] pioneered an M3D-based CIM design with a two-layer reconfigurable SRAM macro, but without implementing complete functionalities for DNN inference. Studies are carried out to use 3-D vertical RRAM for memory applications [9]. However, M3D-based CIM design with CMOS and RRAM needs further exploration.

Manuscript received August 9, 2020; revised November 2, 2020; accepted November 22, 2020. Date of publication December 10, 2020; date of current version January 28, 2021. This work was supported in part by Semiconductor Research Corporation under the Computer-Aided Design and Test thrust (Task ID: 2929) and in part by the National Science Foundation under Award NSF-CCF-1903951. *(Corresponding author: Gauthaman Murali.)*

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: gauthaman@gatech.edu; xiaoyusun@gatech.edu; shimeg.yu@ece.gatech.edu; limsk@ece.gatech.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TVLSI.2020.3042411.

Digital Object Identifier 10.1109/TVLSI.2020.3042411

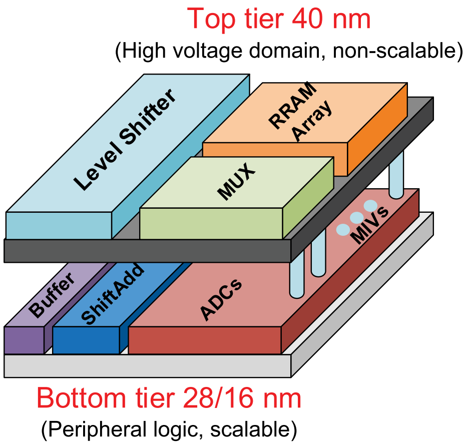

Fig. 1. Our heterogeneous mixed-signal M3D-based RRAM CIM architecture. The top tier contains RRAM at 40-nm process, while the bottom tier is scaled down to 28 nm and then to 16 nm.

The additional area obtained due to tier partitioning in M3D design allows using one ADC per column of RRAM array, thereby unleashing the maximum potential of the parallel CIM architecture. Furthermore, the scaling of RRAM lags behind that of traditional CMOS (the state-of-the-art 40-nm RRAM process from TSMC [10] and 22-nm RRAM process from Intel [11] compared with 7-nm CMOS [12]). Based on this fact, for the first time, we propose an M3D-based CIM design with CMOS and RRAM at hybrid nodes to further improve the area and energy efficiency. In this M3D design, we retain the RRAM tier at the relatively dated process and scale the tier that holds CMOS peripheral circuits to a more advanced node, as illustrated in Fig. 1.

The main contributions of this work are as follows.

1) We implement 2-D and M3D designs of full-chip CIM tile consisting of nine PEs using multiple technology nodes (TSMC 40, 28, and 16 nm).
2) We present a new mixed-signal physical design flow for heterogeneous M3D integration involving sequential integration of two different technology nodes at back-end-of-line (BEOL).
3) We present a comprehensive comparison between 2-D and heterogeneous M3D designs in terms of power, performance, and area (PPA).
4) We perform a detailed thermal analysis of M3D CIM tiles and generate thermal maps to understand their thermal behavior.

Our 3-D designs offer a staggering 3.4–4.9× improvement in energy efficiency in TOPS/W and 43%–50% reduction in footprint area over 40-nm 2-D IC counterpart, measured with commercial-grade GDS designs and sign-off quality simulations.

II. Related Works

Unlike most of the previous works, our design is not a simple memory on logic M3D design. The prior designs of CMOS and RRAM integration assume only one layer of active transistors at the substrate, and the RRAM cells are integrated at BEOL metal vias, as shown in Fig. 2(a). There are rare M3D-based CIM designs reported with CMOS and RRAM that employ two layers of active transistors. The fabrication of two layers of active transistors is feasible by the recent wafer-scale experimental demonstrations at advanced technology nodes (e.g., using the laser-annealing technique to crystallize the top-tier transistors [13], [14]), thus offering new design opportunities for M3D integration. The tier consisting of RRAM transistors also consists of CMOS transistors to implement nonscalable analog peripheries to the memory, which is discussed further in Section V-A.

Though the concept of RRAM + CMOS integration has been familiar for a long time, this is the first detailed work on such a heterogeneous implementation of M3D RRAM-based CIM architecture integrating RRAM and CMOS transistors together and providing complete PPA and thermal analysis results. Our design stands out from traditional CMOS and RRAM integration by placing CMOS cells in the top tier along with RRAM cells, as shown in Fig. 2(b). This combined placement of RRAM and CMOS cells in the same layer helps in grouping cells operating at a particular voltage domain and also in reducing the critical path delay by placing the analog components (RRAM cell and controlling mux) closer to each other in the same tier.

III. Architecture and Design Flow

A. Our CIM Architecture

Fig. 3(a) shows the diagram of our RRAM-based synaptic subarray consisting of 128 × 128 one-transistor-one-resistor (1T1R) RRAM cells and CMOS peripheries. Though multilevel RRAM cells have been demonstrated in academia [15], [16], the reliability and uniformity remain critical issues that could introduce significant accuracy degradation for DNN inference. Therefore, it is more practical to use a technologically more mature binary RRAM that has been demonstrated at giga-bit chip-level in the industry. In our design, every eight cells in a row are grouped as one weight to implement 8-bit weights, and eight sequential cycles of input vectors of different significances are used to implement 8-bit neuron activation. The input vectors are encoded as the digital voltages ("0" or "1") and applied to the wordlines (WLs). WL switch matrix is used to assert all the WLs simultaneously to initiate parallel MAC computation, and level shifters shift up the voltage domain for RRAM. Partial sums are added up as the column currents along the bitlines (BLs) accumulate following Kirchhoff's law. Next, they go through the ADCs and thermometer-to-binary encoders, resulting in digital binary codes. It is important to note that multiple columns of the RRAM subarray share one ADC due to the large footprint of ADCs. In our 2-D baseline design, eight columns share one ADC. The shift-and-add modules shift and accumulate the partial sums from eight weight columns during eight cycles to obtain the final outputs.

Fig. 3(c) shows the diagram of one RRAM-based CIM tile that is capable of performing one typical layer of convolutional neural network (CNN) inference. The tile integrates nine PEs to implement 3×3 convolution kernels using the mapping strategy in [17], and each PE consists of 16 subarrays, as shown in Fig. 3(b). At the tile level, accumulators

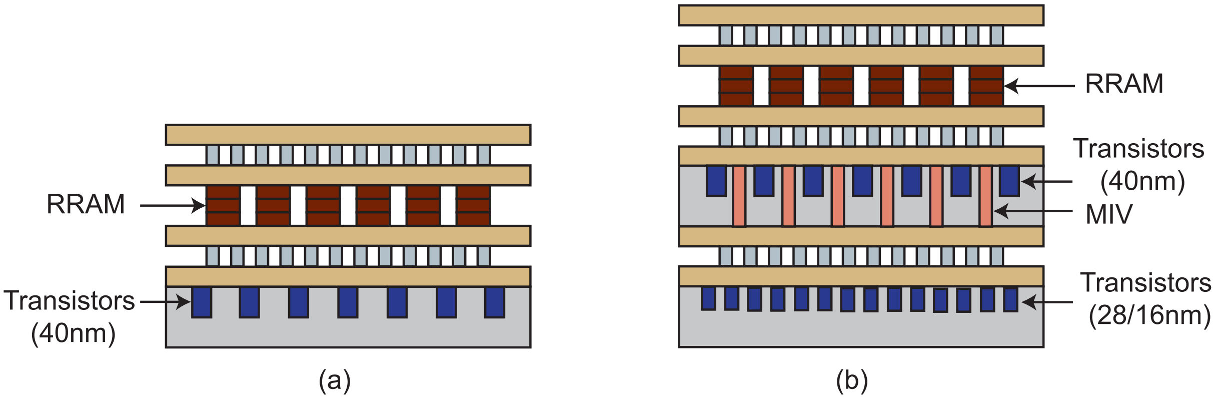

Fig. 2. (a) Existing commercial integration of RRAM and CMOS. In this "pseudo-3D" integration, there exist no CMOS transistors in the "RRAM tier." (b) Our "true" monolithic integration of RRAM and CMOS M3D integration. CMOS transistors are present in the "RRAM tier," and our RRAM module is partitioned across the two tiers.

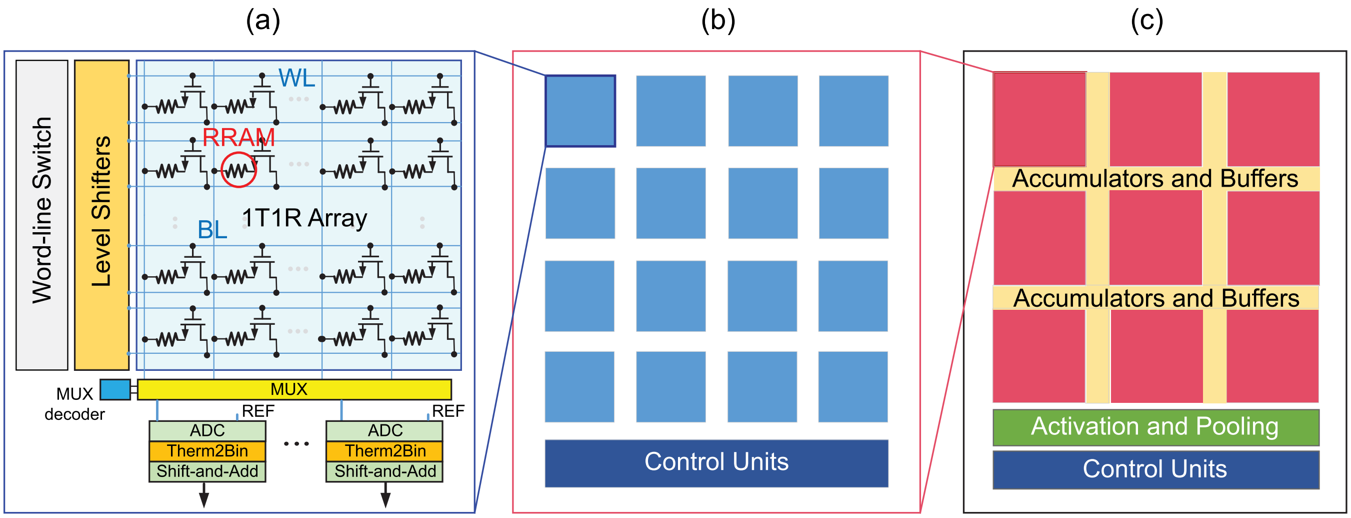

Fig. 3. Diagram of our (a) RRAM-based synaptic subarray, (b) RRAM-based CIM PE consisting of 16 subarrays and control units, (c) RRAM-based CIM tile consisting of nine PEs, control units, buffers, and a complete set of logic modules, including accumulators, activation modules, and pooling modules.

sum up the partial sums from each PE. Using these sums, activation (ReLU) and max-pooling modules compute the final output feature maps.

1) Programming RRAM Resistance for Efficient Computation: The RRAM resistance should be accurately programed to achieve high inference accuracy; more specifically, the LRS state is more critical than the HRS state as the BL current is dominated by LRS cells. We typically enforce a closed-loop write-verify protocol during programming to tighten the distribution of LRS state, meaning that multiple iterative write operations will be performed to push the resistance of the selected cell into the desired range. Since the basic RRAM operations are not the focus of this work, please refer to our prior work on a fabricated RRAM CIM chip for more details on the programming scheme and related write-verify measurement results [4].

2) Parasitic Effect on RRAM Computation: The R_{ON} resistance of RRAM is 10 KΩ, and the wire resistance of WLs and BLs for a 40-nm 128 × 128 RRAM array is approximately 141 Ω. Thus, the wire resistance being much smaller than the RRAM ON resistance, IR drop is not a major concern. Also, there have been prior works on addressing the IR drop issue. For example, Liu *et al.* [18] proposed an adaptive compensation method during training from algorithm end, and Woo *et al.* [19] proposed to use a selector device to suppress the impact of IR drop. Such optimizations are out of the scope of this work, and the contribution of this work is to overcome the ADC bottleneck by enabling more ADCs through an M3D integration.

B. EDA for Mixed-Signal M3D ICs

Our heterogeneous mixed-signal M3D design flow is shown in Fig. 4. The analog blocks (RRAM cells, level shifters, and ADC) in the CIM tile are custom-made using Cadence Virtuoso. The RRAM cells and the level shifters remain at the older 40-nm node, whereas the ADCs scale down to advanced nodes of 28/16 nm. Integrating digital and analog blocks requires a mixed-signal design flow. We generate the abstract models of analog blocks and integrate them as hard macros with digital control logic in Cadence Innovus. A major

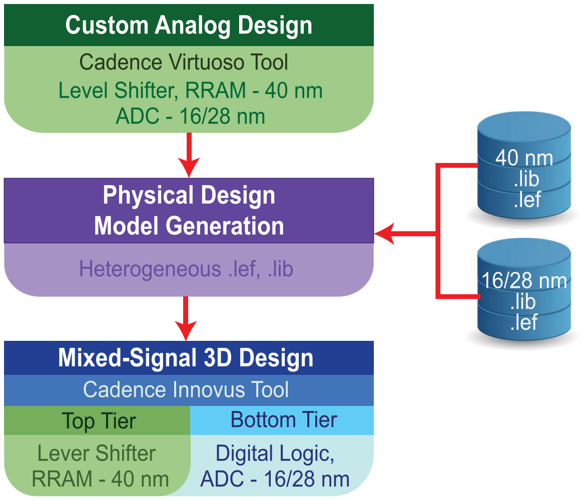

Fig. 4. Our EDA flow for heterogeneous mixed-signal M3D ICs.

challenge is in generating accurate physical design models for heterogeneous integration. Our 3-D flow uses a unique technology Library Exchange Format (LEF) file to enable the integration of two different technology nodes. We create our technology LEF file by merging the 2-D technology LEF files of both nodes, resulting in a hybrid 14 metal layer stack 3-D technology. For example, our combined heterogeneous BEOL stack consists of eight metal layers from the TSMC 28-/16-nm node and six metal layers from the 40-nm node.

We also modify the standard cell LEF files according to this new heterogeneous 3-D technology BEOL LEF file. The goal is to use the appropriate metal layers for cell designs and cell delay/power characterization. Our macro designs and characterizations are done in the same way. Besides, the timing and power optimization tools need a technology file that contains the parasitics (resistance and capacitance) information of each metal layer and the dielectrics between them. For this purpose, we generate a heterogeneous technology file by merging the corresponding parasitic details from both the top- and bottom-tier technology nodes.

Once the heterogeneous 3-D library is ready, we perform top-level physical design. The top tier only consists of custom analog macros. We restrict the placement of any standard cells on this tier. This way, the top tier contains 40-nm RRAM cells and other custom analog blocks, while the bottom tier contains digital logic, ADCs, and timing buffers. As commercial 3-D tools are unavailable, we perform the 3-D design using a modified version of Shrunk-2-D flow [20], which uses 2-D IC commercial tools to implement 3-D designs.

It is important to note that 2-D IC tools do not use multiple tiers to place cells and macros. Instead, they put them into a single tier. Each technology node has minimum allowed cell dimensions (the smallest possible rectangle) called the site size. We shrink the transistor layer of top-tier analog macros into the site size at the bottom-tier technology node. However, we retain the original pin locations in the 3-D BEOL structure. This way, the placement of shrunk top-tier cells becomes a familiar-looking 2-D placement problem, while the routing connects pins from both tiers in an optimized fashion.

C. M3D Integration

Several 3-D integration techniques have been proposed [13], [14], [21], [22]. Among these, the 3-D sequential integration technique [13] provides a way to stack multiple layers with a nanoscale resolution, leading to smaller 3-D interconnects between multiple tiers. The 3-D sequential integration ensures the stability of the bottom tier by limiting the thermal budget of fabrication of the top tier to less than 500 °C. With CEA-Leti moving 3-D sequential integration closer to commercialization (CoolCube), this technique can be employed to manufacture our proposed design.

D. Testing M3D RRAM Arrays

It is a general perception that the external I/O pins are to be restricted for M3D IC. Nevertheless, in our design, the I/O pins for RRAM cell-by-cell operations (SET/RESET/Forming) have to be kept as those pins are needed to program the weight pattern to the array before the deployment of the chip for inference. Though testing circuits are not a part of our current implementation, in future designs, the built-in-self-test (BIST) on-chip module could be integrated to perform the write-verify in principle.

IV. Baseline 40-nm 2-D CIM Tile

We design our 2-D CIM tile using a bottom-up approach. First, we implement a PE consisting of 16 128 × 128 RRAM subarrays. We then integrate nine such PEs to create a CIM tile, capable of performing a typical layer of CNN inference. Each PE consists of a 4 × 4 array of RRAM subarrays. A single subarray consists of a digital periphery, level shifters, and custom analog blocks, including RRAM cells, high-voltage muxes, and ADCs. Fig. 5(a) shows the layouts of these blocks.

RRAM cells and their muxes require a voltage of 3.3 V for their operation, as described in TSMC 40-nm RRAM process [10]. Besides, the smallest technology node that offers such high-voltage I/O transistors is 40 nm. Hence, we use TSMC 40 nm to implement our 2-D CIM tile. The digital periphery in an RRAM subarray generates the control signals to the RRAM cells and muxes while operating at 3.3 V in the analog partition. However, the standard cells in the digital block operate at 0.81 V. Hence, we place a level shifter block between the digital and analog blocks of an RRAM subarray, as highlighted in Fig. 5(a). One of the major components determining the efficiency of a PE is its ADC. However, the ADCs are power-hungry and occupy a considerable amount of space in the design. Hence, using one ADC for each column in an RRAM subarray leads to a huge increase in the system power and area, thereby reducing the efficiency. We achieve a tradeoff between the system efficiency and power dissipation by using one ADC for every eight columns of the RRAM subarray, similar to the prior reported tape-out [4]. Therefore, there are 16 ADCs for each RRAM subarray and 256 ADCs in a PE. Fig. 5(b) shows the final layout of 2-D RRAM PE.

We integrate nine PEs into a single CIM tile to implement a 3 × 3 convolution kernel in a typical CNN topology [17].

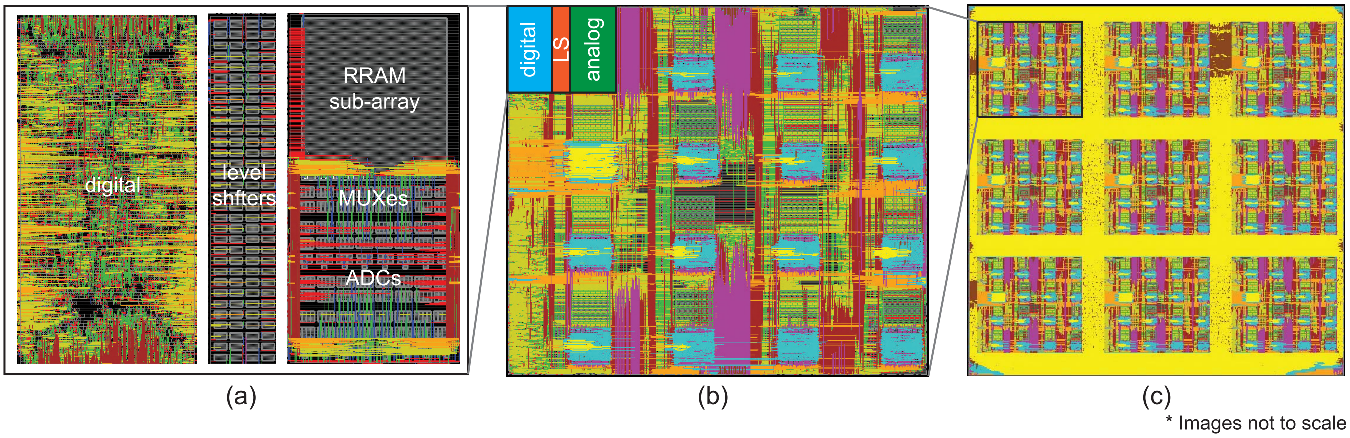

Fig. 5. Our mixed-signal 2-D CIM tile design using a commercial 40-nm technology. (a) Single subarray (166 μm×140 μm). (b) Single PE (625 μm×650 μm). (c) Overall tile design (2.4 mm × 2.4 mm).

We evenly distribute: 1) accumulators that calculate the partial sum from each PE; 2) activation modules; and 3) max-pooling modules among the nine PEs in the top-level tile design. Our 2-D RRAM tile operates at a maximum frequency of 100 MHz. It has a footprint of 2.4 mm × 2.4 mm and uses ten metal layers to route the entire design. The total number of ADCs in this 2-D tile is 2304. Fig. 5(c) shows the final layout of our 2-D RRAM tile.

V. Proposed 28-+ 40-nm 3-D CIM Tile

Our 2-D tiles restrict the number of ADCs used due to their high power and area overhead. This restriction, in turn, reduces the throughput of CNNs to a great extent. Scaling RRAM cells to a smaller geometry is extremely difficult as of today due to the high programming voltage requirement. Our solution is to adopt heterogeneous M3D IC technology to effortlessly integrate multiple technology nodes together and scale scalable modules to a more advanced node.

A. Advantages of Heterogeneous 3-D CIM Tile

In our CIM tile, the digital control logic and the ADCs are capable of scaling down to a more advanced technology node. This ability enables us to build a heterogeneous M3D CIM tile with: 1) digital and ADC blocks on the bottom tier at 16-/28-nm technology node and 2) level shifters, RRAM cells, and muxes on the top tier using 40-nm technology. There is a massive number of connections (55 296) between the ADC and the digital control logic. Thus, scaling ADCs down to a smaller geometry helps place ADCs as close as possible to the digital block. In our 2-D CIM design described in Section IV, the area occupied by the digital control logic and ADCs is equal to the combined area of level shifters, RRAM mux, and RRAM cells.

In our heterogeneous M3D design, we scale the digital modules along with ADCs down to a 28-nm node. This causes the bottom 28-nm tier to occupy a smaller area than the top 40-nm tier. We leverage this area imbalance to add additional ADCs in the design and improve the throughput of the CIM tile. The area imbalance between the two tiers allows adding 8× more ADCs into the design. This way, each column in an RRAM subarray has a dedicated ADC for itself, offering more parallelism. Note that attempting this ADC addition in a 2-D design or a homogeneous M3D design would cost us more area and power consumption. In contrast, our heterogeneous 3-D stacking helps keep the area overhead to a minimum due to the smaller geometry in 28 nm.

B. Heterogeneous 3-D CIM Tile Design

Similar to the 2-D design, we follow a bottom-up approach for our 3-D design. We split a single PE consisting of 16 RRAM subarrays into two tiers and design the two tiers separately. Fig. 6(a) and (b) shows the bottom-tier (28 nm) and top-tier (40 nm) layouts of 3-D RRAM PE. This 3-D PE has 2048 ADCs in it, which is 8× more than its 2-D counterpart. Fig. 7 illustrates this difference.

Using the heterogeneous mixed-signal M3D flow described in Section III-B, we integrate nine 3-D PEs into a single 3-D tile. In addition to the PEs, we integrate the tile-level digital logic, comprising of the accumulator, activation, and max-pooling blocks, into the bottom tier. Fig. 6(c) and (d) shows the bottom-tier (28 nm) and top-tier (40 nm) layouts of our 3-D CIM tile. Our 3-D tile has a footprint of 1.7 mm × 1.7 mm and uses 14 metal layers to route the entire design. The bottom tier uses eight metal layers, and the top tier uses six metal layers. The total number of ADCs in the entire 3-D tile is 18 432. Despite the 8× increase in the number of ADCs, the total internal power of the design is 40% less than that of the corresponding 2-D design, as we scaled the ADCs down to 28 nm. However, the switching power is 2.5× higher than the corresponding 2-D design due to eightfold increase in ADC count and an increase in the capacitance of wires connecting to the ADCs. More details are provided in Table I.

VI. Proposed 16 + 40-nm 3-D CIM Tile

A. Motivation

To further reduce the power consumption and improve the energy efficiency of 3-D CIM tiles, we scale the digital logic and ADCs down to 16 nm. Once again, this scaling

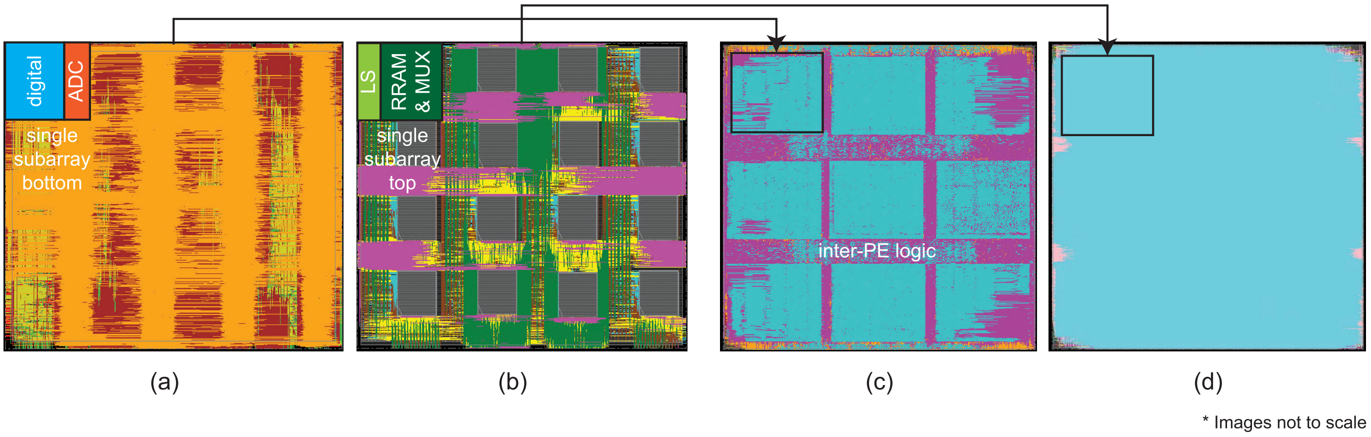

Fig. 6. Our heterogeneous mixed-signal 3-D CIM tile design using a commercial 28- and 40-nm technologies. (a) Single PE bottom tier (28 nm). (b) Single PE top tier (40 nm). (c) CIM tile bottom tier (28 nm). (d) CIM tile top tier (40 nm).

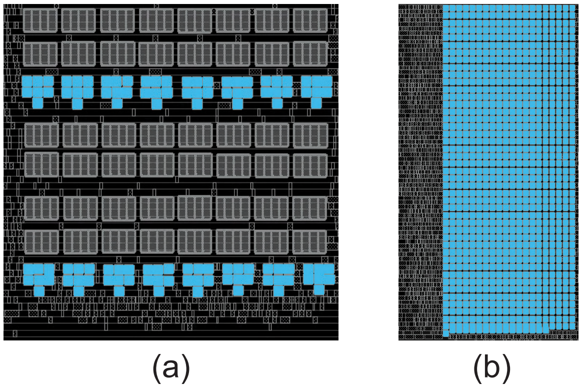

Fig. 7. GDS layouts of ADCs (highlighted in blue) in (a) 2-D CIM tile (40 nm) and (b) 3-D CIM tile (28 nm). We fit 8× more ADCs due to the size reduction in 28 nm.

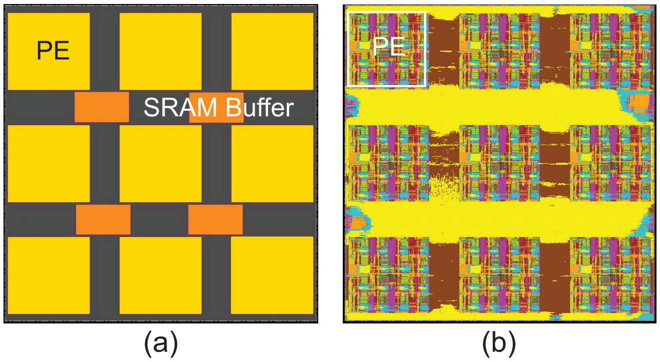

Fig. 8. (a) Floorplan (2.59 mm × 2.59 mm). (b) Final layout of our 2-D CIM tile. We add input SRAM buffers unlike our previous design shown in Fig. 5(b). We use a commercial 40-nm technology.

introduces an area imbalance between the top and bottom tiers in the 3-D design. Currently, each convolutional kernel in a CNN topology has to access an external DRAM to receive the activation inputs and store the result of the computation. However, accessing external DRAM is energy-consuming, especially for a large volume of input feature maps on CNN.

We leverage the additional area available in the 16-nm bottom tier to integrate four SRAM buffers within a tile to store the input–output feature maps. Each buffer has a capacity of 32 kB. The size of the input feature map determines the capacity of SRAM buffers. Based on the data obtained from Eyeriss accelerator [2], the normalized energy per bit of DRAM is 33× higher than that of SRAM. Thus, integrating SRAM buffers within the tile helps us achieve 33× energy savings. For a fair comparison, we add SRAM buffers to both 2-D and M3D versions of our CIM tile.

B. 2-D CIM Tile Extension

Fig. 8 shows the floorplan and layout of the 2-D CIM tile with the input buffers. We place SRAM input buffers generated by a TSMC 40-nm memory compiler, as shown in Fig 8(a). The 2-D tile has a footprint of 2.59 mm × 2.59 mm and uses ten metal layers to route the entire design. Similar to the 2-D CIM tile without SRAM buffers, eight columns of each RRAM subarray share one ADC to achieve a tradeoff between power dissipation and energy efficiency. Hence, the total number of ADC in this 2-D CIM tile is 2304.

C. 3-D CIM Tile With SRAM Input Buffers

Our 3-D CIM tile with SRAM buffer integrates digital control logic, ADCs, and SRAM buffers at the 16-nm tier, level shifters, RRAM cells, and their muxes at the 40-nm tier. We generate 16-nm SRAM buffers using a TSMC 16-nm memory compiler. The top tier of this CIM tile and its PEs, and the number of ADCs (18 432) remains unaltered from those of the 3-D CIM tile without the buffers presented in Section V-B. Similar to the 3-D tile design without SRAM, we begin designing the 3-D tile by designing the PEs. Fig. 9(a) and (b) shows the bottom-tier (16 nm) and top-tier (40 nm) layouts of a single PE. We then integrate nine 3-D PEs to implement a 3×3 convolution kernel. Besides, we place the SRAM buffers and tile-level digital control logic among the nine PEs in the bottom tier.

Fig. 9(c) and (d) shows the bottom-tier (16 nm) and top-tier (40 nm) layouts of the 3-D CIM tile with SRAM input buffers. The footprint of this 3-D tile is 1.7 mm × 1.7 mm, which is the same as that of the 3-D CIM tile without buffers. The

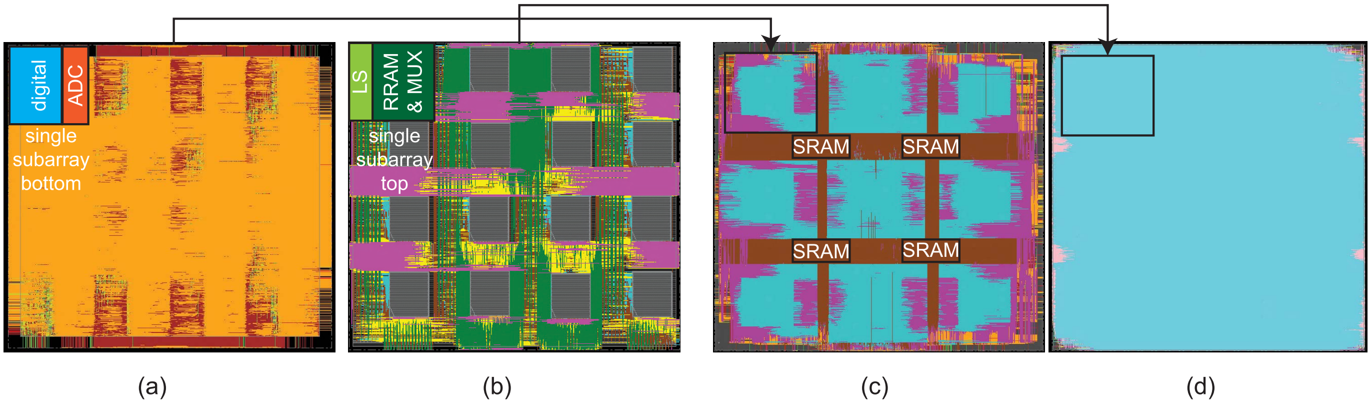

Fig. 9. Our heterogeneous mixed-signal 3-D CIM tile design using a commercial 16- and 40-nm technologies. (a) Single PE bottom tier (16 nm). (b) Single PE top tier (40 nm). (c) CIM tile bottom tier (16 nm). (d) CIM tile top tier (40 nm).

footprint of the 3-D designs is mainly determined by the top-tier dimensions, as the top tier is at an older technology node. The 3-D CIM tile with input buffers uses 14 metal layers to route the entire design, eight metal layers for the bottom tier, and six metal layers for the top tier. From Table I, it can be observed that scaling down the bottom tier to 16 nm not only allows packing additional gates and SRAM buffers but also reduces the wirelength by 14.7% compared with the 3-D design without buffers (28-/40-nm CIM tile).

VII. Experimental Results

A. Energy Efficiency Calculation

Each MAC operation in the CNN involves the multiplication of an 8-bit weight with 8-bit input. We store the weights in every 128 × 128 RRAM subarray, and therefore, a single RRAM subarray can hold 128 × 16 weights. Each MAC computation involves two operations. Thus, the total number of operations occurring in a subarray is 128 × 16x2. The total number of operations in a PE with 16 subarrays is 128 × 16x2 × 16, and the total number of operations in a tile with nine PEs is 128 × 16x2 × 16x9. To perform a multiplication operation between 8-bit weights and 8-bit inputs, we feed in a 1-bit input vector to each ADC and every clock cycle. In the case of a 2-D tile, which involves 8-to-1 ADC sharing, processing 1-bit input takes eight cycles to read out the columnwise partial sums for shift-and-add. Therefore, 64 cycles are needed in total to perform the MAC operation on 8-bit input. However, the dedicated ADCs for each column in 3-D tile reduce the read-out latency to eight clock cycles. The following equation gives the energy efficiency of a tile

$$\mathrm{EE} = \frac{\#\text{ of ops}}{\text{Total \ Energy}} = \frac{\#\text{ of ops}}{\text{cycles} \times \text{period} \times \text{Total Power}}. \quad (1)$$

Table I shows the energy efficiency of both variants of 2-D and 3-D tiles calculated using (1). In the case of RRAM tiles without input buffers, the 3-D design is 3.4× energy-efficient than the corresponding 2-D design. With input buffers, the 3-D design is 4.9× energy-efficient than its 2-D counterpart. The power breakdown of the CIM tiles is shown in Fig. 10.

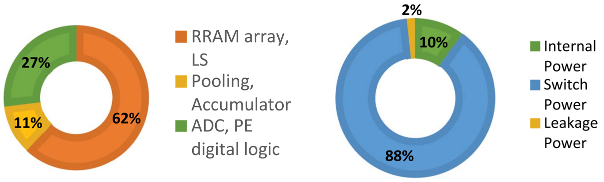

Fig. 10. Power breakdown of CIM tiles.

B. Overall 2-D Versus 3-D CIM Tile Comparison

Table I summarizes the design details of two variants of our 2-D and 3-D CIM tiles (without versus with SRAM input buffers). The footprint of a 3-D tile without input buffers is 43% smaller than that of its 2-D counterpart. The footprint of a 3-D tile with input buffers is 50% of that of its 2-D counterpart. Despite the huge reduction in footprint, 3-D tiles are still able to integrate 8× more ADCs than the 2-D tiles. However, this increases the total routing length and power dissipation of the 3-D tiles. However, at the same time, the number of operations performed by 3-D tiles is several folds higher than the 2-D ones. Therefore, we compare the number of operations performed per unit power dissipated (energy efficiency) instead of comparing the absolute power dissipated. In summary, our 3-D design (no SRAM case) overcomes the bottleneck caused by ADCs. It offers a staggering 3.4× improvement in energy efficiency in TOPS/W and 43% reduction in footprint area over 40-nm 2-D IC counterpart, measured with commercial-grade designs and sign-off quality simulations.

The critical path delay of the proposed design is 8.926 ns, which limits the frequency to 100 MHz. The critical path mainly consists of the latency for activating RRAM WLs, voltage developing on BLs, sensing of ADCs, and thermometer-to-binary encoding. Such a memory-access delay is typical in the reported RRAM-based prototype chips. The frequency could potentially be boosted by inserting additional pipeline stages (e.g., between ADCs and encoders) or using a smaller array to reduce the *RC* parasitic of WLs and BLs. However, those design optimizations are out of the scope of this work.

TABLE I
COMPARISON OF 2-D AND 3-D CIM TILES

Metric	2D w/o buffers	3D w/o buffers	2D w/ buffers	3D w/ buffers
Technology	$40nm$	$40nm$ top tier $28nm$ bottom tier	$40nm$	$40nm$ top tier $16nm$ bottom tier
Input precision	8b	8b	8b	8b
Weight precision	8b	8b	8b	8b
Footprint	$2.4\ mm \times 2.4\ mm$	$1.7\ mm \times 1.7\ mm$	$2.59\ mm \times 2.59\ mm$	$1.7\ mm \times 1.7\ mm^2$
Gate Count	3,875,000	4,651,278	4,461,638	6,190,350
ADC Count	2,304	18,432	2,304	18,432
Total ADC Area	$0.047\ mm^2$	$0.25\ mm^2$	$0.047\ mm^2$	$0.21\ mm^2$
Metal Layers	10	14	10	14
Total Wirelength	$95.6\ m$	$159.8\ m$	$99.2\ m$	$136.3\ m$
MIV Count	N/A	373,656	N/A	266,164
Frequency	$100\ MHz$	$100\ MHz$	$100\ MHz$	$100\ MHz$
Total Power	$412.7\ mW$	$975.5\ mW$	$428.4\ mW$	$695.27\ mW$
Throughput (TOPS)	0.92	7.37	0.92	7.37
Energy-efficiency	2.23 TOPS/W	7.56 TOPS/W	2.15 TOPS/W	10.6 TOPS/W

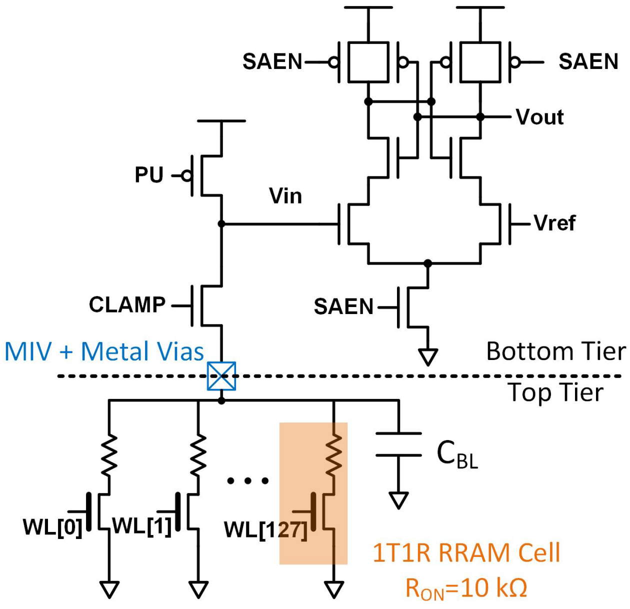

Fig. 11. Schematic of the VSA-based sensing scheme. We consider the parasitics of additional metal vias incurred by M3D integration to analyze their impact on the sensing latency. The HSPICE simulation results show a negligible increase in sensing latency (∼8 ps).

C. Impact of 3-D Integration

To analyze the impact of the parasitics of additional metal vias incurred by M3D integration on the sensing latency of ADCs, we run HSPICE simulations with TSMC 40-/16-nm PDKs. Fig. 11 presents the schematic of the voltage-mode sense amplifier (VSA)-based sensing scheme, which is the basic unit of ADCs. The pull-up transistor (PU), clamping transistor (CLAMP), and BL essentially form a voltage-divider structure, where the voltage level of node Vin is determined by the number of activated low-resistance state cells (RON) along the corresponding BL. Then, Vin is compared with the reference voltage Vref at the assertion of enable signal (SAEN) to generate the final output Vout. We add additional parasitics (including the metal vias between each metal layer and the MIV, e.g., $R_{\mathrm{MIV}} = 12.5\ \Omega$ and $C_{\mathrm{MIV}} = 0.037$ fF) between two partitions of the circuit, as depicted in blue in the figure. We run the simulations for the worst case where only one R_{ON} is activated in the column, and the HSPICE simulation results show a negligible increase in sensing latency (∼8 ps), meaning that the latency overhead induced by M3D integration on ADC sensing is not a problem. This is as expected as the value of parasitics of vias is much less than the BL capacitance (∼15 fF) and RRAM cell resistance (RON assumed to be 10 KΩ).

VIII. THERMAL ANALYSIS OF M3D CIM TILES

One of the primary concerns of M3D designs is the thermal performance of 3-D ICs. Integrating RRAM and CMOS on different technology nodes into an M3D IC is an emerging methodology to implement CIM-based CNN accelerators. Thus, we performed thermal analysis on our two different versions of 3-D CIM tiles to inspect if our designs have any heating issues. In our thermal analysis flow, we divide the entire 3-D design into several small thermal cubes and associate them with a consolidated power density and thermal conductivity based on different cells, and gate and interconnect layers within the thermal cubes. The 3-D stack used to model our CIM tiles for thermal analysis is shown in Fig. 12.

We use Ansys Fluent to evaluate the temperature gradient across the entire 3-D design by inputting the thermal cube structure associated with the power density and thermal conductivity values. Based on the temperature gradient across different areas of the die, Ansys Fluent generates a thermal map of the entire design. The thermal maps of the bottom and top tiers of 28-/40-nm CIM tile are shown in Fig. 13, and those of 16-/40-nm CIM tile are shown in Fig. 14.

The 28-/40-nm CIM tile reaches a maximum temperature of 309.58 K, and the 16-/40-nm CIM tile reaches a maximum temperature of 305.50 K, when operated at a frequency of 100 MHz. The major heat spots in both the designs are at the center of the die, as expected in any M3D design. However, the overshoot in the temperature is still under control, thereby making our M3D designs less susceptible to thermal issues.

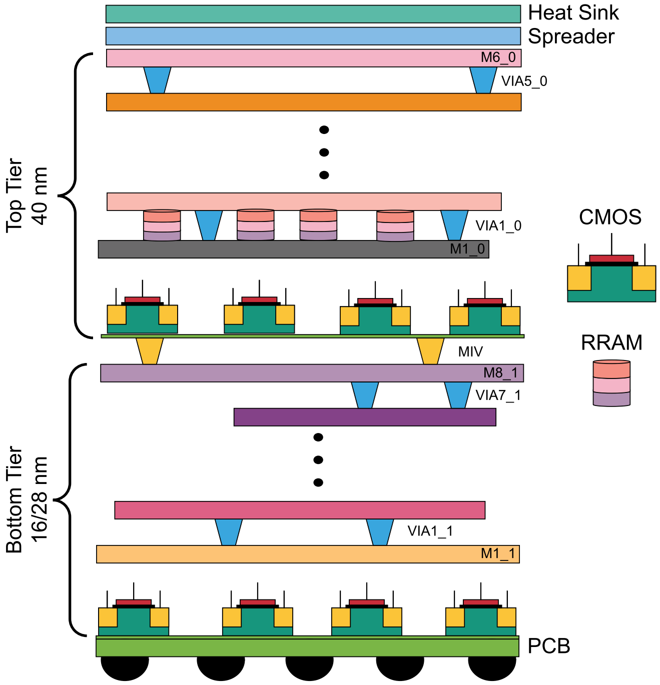

Fig. 12. 3-D die stack model used for thermal analysis.

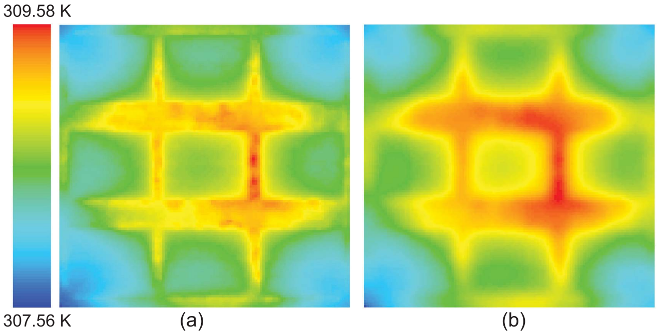

Fig. 13. Thermal maps of 3-D CIM tile without SRAM buffers (28/40 nm). (a) Bottom tier (28 nm). (b) Top tier (40 nm).

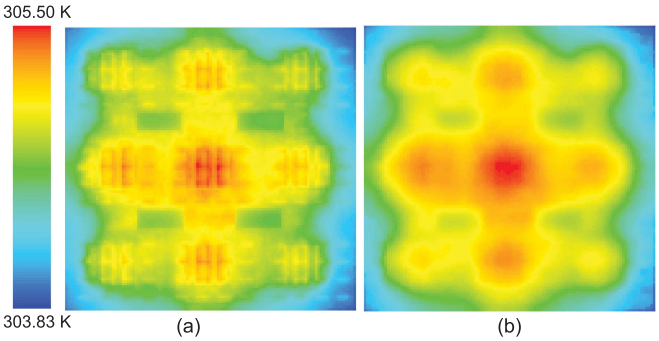

Fig. 14. Thermal maps of 3-D CIM tile with SRAM buffers (16/40 nm). (a) Bottom tier (28 nm). (b) Top tier (40 nm).

This small increase in the temperature across the M3D dies is due to low power dissipation. The overall power consumed by both the designs does not exceed 1 W in total, which is very small compared with the large scale of the design. It is important to note that small variations in temperature across the die do not significantly affect the read/write cycles of RRAM. Also, RRAM retention is usually affected by drift in resistance at a temperature of 100 °C or greater [23]. As the temperature across our designs does not exceed 37 °C, RRAM retention capability remains unaffected.

IX. Comparison With Existing Works

Several RRAM-based CIM prototypes have been proposed in recent years. Table II compares two such previous works with our 3-D CIM tile. Xue *et al.* [24] designed a macro of size 256×512 on a 55-nm process that showed an energy efficiency of 53.17 TOPS/W with 1-bit activation and 2-bit weights. Since only nine rows of the memory array are activated simultaneously during the MAC operation, the parallelism was significantly undermined. Yin *et al.* [4] presented a macro at the 90-nm node that activated all the 128 rows simultaneously to improve the throughput. Nevertheless, since every eight columns shared one ADC in this design, the throughput was still limited. The energy efficiency was reported as 24.1 TOPS/W with 1-bit activation inputs and weights. The MAC operation in this array is between the 1-bit weights and activation compared with 8-bit activation and weights in our design. Besides, the other two designs are much smaller 2-D designs, whereas our design is a complete 3-D system, including PEs, pooling/activation modules, and interconnects. The reported energy efficiencies from [4] and [24] only account for one macro array. If 8-bit activation and weights are considered for [4] and [24], then their energy efficiencies would drop approximately by 64 and 32 times to 0.37 and 1.66 TOPS/W, respectively.

It should be pointed out that here we are evaluating an 8 bit × 8 bit MAC operations in this study; 1 OPS is 2 MAC (8 bit × 8 bit). Some articles in the literature may report >100 TOPS/W for low-precision MAC (even binary neural network). If we convert our energy efficiency for binary neural network, it will be 640 TOPS/W. To make a fair comparison, the precision should be kept the same. Table II contains several such comparisons. It can be seen that, when the energy efficiencies are scaled for 8-bit activation and 8-bit weights, our work provides the highest energy efficiency.

In addition to this, the energy consumption of our design is dominated by ADCs. On the one hand, increasing the ON-resistance of RRAM cells can effectively reduce the BL current during ADC sensing, thus reducing the energy consumption. On the other hand, we used the traditional Flash ADC in this work without any low-power techniques. The deployment of such low-power ADCs [31] can also help on improving energy efficiency. In addition, given the typical high sparsity in both weights and activations in DNNs, leveraging this high sparsity by skipping "0"-involved computations can also improve the energy efficiency [32]. These potential optimizations are all applicable to our proposed architecture.

X. Implementing Practical DNN

Fig. 15 shows the top-level diagram of our RRAM CIM architecture that exploits multiple PE tiles to implement practical DNNs. Activation units and pooling units are needed

TABLE II
COMPARISON OF OUR 3-D CIM TILES WITH PREVIOUS WORKS

Metric	Xue et al. [24]	Yin et al. [4]	14.5 Envision [25]	Sticker [26]	Sayal et al [27]	SNAP [28]	UNPU [29]	QUEST [30]	Our 16/40nm 3D CIM tile
IC Type	2D	2D	2D	2D	2D	2D	2D	3D	3D
Technology	CMOS + RRAM	CMOS + RRAM	FDSOI	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS + RRAM
Technology Node	55nm	90nm	28nm	65nm	40nm	16nm	65nm	40nm	40nm top tier 16nm bottom tier
Mode of Measurement	Test chip	Test chip	Test chip	Test chip	Test chip	Test chip	Test chip	Test chip	Simulation
Input Bits (A)	1	1	3	8 (5%)	4	16 (10%)	1	2	8
Weight Bits (W)	2	1	4	8 (5%)	1	16 (10%)	1	2	8
Energy-eff. (TOPS/W)	53.17	24.1	10	62.1	12.08	21.55	50.6	7.49	10.6
Scaled Eff. (TOPS/W) A:8, W:8	1.66	0.37	1.875	0.155	0.755	0.86	0.79	0.47	10.6

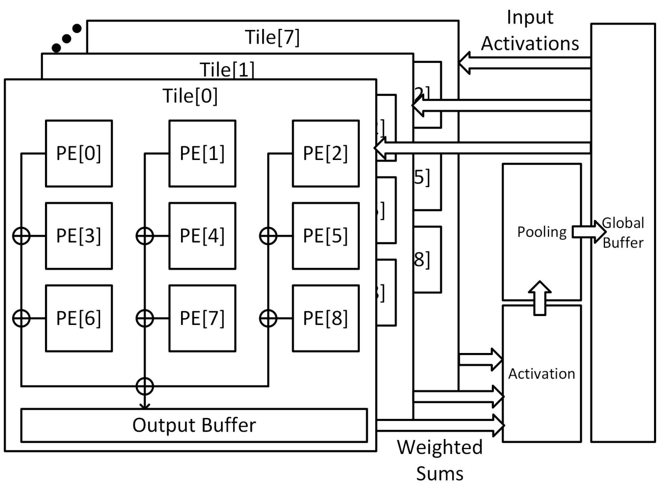

Fig. 15. Top-level diagram of RRMA CIM architecture consisting of PE tiles, activation units, pooling units, and global buffers. As an example, eight tiles are utilized to implement the largest layer of ResNet-18.

TABLE III
BENCHMARK RESULTS ON RESNET-18

	2D (40nm)		3D (16/40nm)	
	Energy (nJ)	Area (mm^2)	Energy (nJ)	Area (mm^2)
PE Tiles	12840.48	354.69	2604.44	152.81
Global Buffer	8763.12	3.99	1657.89	0.39
Pooling Units	0.041	0.12	0.0082	0.012
Activation Units	37.63	0.014	7.53	0.0014
Total	21641.27	358.82	4269.86	153.21
TOPS/W	1.28		6.47	

to postprocess the weighted sum results after the accumulation of the partial sums from each tile. Global buffers are used to store intermediate results between each layer. As a case study, we benchmark the hardware performance of our 2-D design (40 nm) and M3D design (40/16 nm) on implementing ResNet-18 [33], a representative CNN designed for ImageNet classification. Table III summarizes the energy and area consumption of two design options to map the whole network. The results for PE tiles are based on the synthesis results, as shown in Table I, and the results for the other modules are from NeuroSim [34], a benchmarking framework that supports flexible CIM array design options with different memory technologies with various peripheral circuit modules. It can be seen that PE tiles and global buffers are dominating energy and area consumption. The overall energy efficiency drops to 1.28 and 6.47 TOPS/W, respectively, due to the additional energy consumption of activation units, pooling units, and global buffers at the top level. We note that the energy-efficiency improvement over 2-D baseline design further increases from 4.93× to 5.05× as the additional top-level modules (i.e., activation/pooling units and global buffers) are scaled to 16 nm that induces more savings.

Our proposed design implements nine PEs to intrinsically accommodate up to 3×3 kernels, which covers the most typical kernel sizes in practice. However, a larger kernel can be implemented by stacking multiple 3×3 kernels, as introduced in [35]. Therefore, our proposed design is capable of supporting larger kernels as well.

XI. CONCLUSION

Our 3-D CIM tiles not only offer a smaller footprint than 2-D tiles but also show a manyfold improvement in energy efficiency. The number of ADCs limits the performance of the 2-D CIM tile. However, a heterogeneous M3D technology allows scaling of the digital logic and ADCs irrespective of any changes in the RRAM technology node. With the 16-nm control logic tier, we were able to integrate the activation input buffers within the tile. Further performance improvements can be achieved by scaling down the digital logic tier and integrating weight input buffers within the design.

REFERENCES

[1] N. P. Jouppi *et al.*, “In-datacenter performance analysis of a tensor processing unit,” 2017, *arXiv:1704.04760*. [Online]. Available: http://arxiv.org/abs/1704.04760

[2] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, “Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks,” *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, May 2017.

[3] X. Si *et al.*, “24.5 A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

[4] S. Yin *et al.*, “Monolithically integrated RRAM- and CMOS-based in-memory computing optimizations for efficient deep learning,” *IEEE Micro*, vol. 39, no. 6, p. 54–63, Jan. 2019.

[5] C.-C. Chang *et al.*, “NV-BNN: An accurate deep convolutional neural network based on binary STT-MRAM for adaptive AI edge,” in *Proc. 56th Annu. Design Autom. Conf.*, Jun. 2019, pp. 1–6.

[6] S. Ambrogio *et al.*, “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature*, vol. 558, no. 7708, p. 60, 2018.

[7] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Design and CAD methodologies for low power gate-level monolithic 3D ICs,” *Proc. Int. Symp. Low power Electron. Des.*, 2014, pp. 171–176.

[8] S. Srinivasa *et al.*, “Monolithic 3D -IC based Reconfigurable Compute-in-Memory SRAM Macro,” *2019 Symp. VLSI Technol.*, page T32–T33, 2019, doi: 10.23919/vlsit.2019.8776506.

[9] F. Zokaee, M. Zhang, X. Ye, D. Fan, and L. Jiang, “Magma: A monolithic 3D vertical heterogeneous ReRAM-based main memory architecture,” in *Proc. 56th Annu. Des. Autom. Conf.*, Jun. 2019, pp. 1–6.

[10] C.-C. Chou *et al.*, “An N40 256×44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 1–8.

[11] P. Jain *et al.*, “13.2 A 3.6Mb 10.1Mb/mm^2 embedded non-volatile ReRAM macro in 22nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5ns at 0.7 V,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, p. 212.

[12] J. Chang *et al.*, “12.1 A 7nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 1–6.

[13] L. Brunet *et al.*, “Breakthroughs in 3D sequential technology,” in *IEDM Tech. Dig.*, Dec. 2018, pp. 1–4.

[14] C.-C. Yang *et al.*, “Footprint-efficient and power-saving monolithic IoT 3D+ IC constructed by BEOL-compatible sub-10nm high aspect ratio (AR>7) single-grained Si FinFETs with record high ion of 0.38 mA/μm and steep-swing of 65 mV/dec. And Ion/Ioff ratio of 8,” in *IEDM Tech. Dig.*, Dec. 2016, pp. 1–5.

[15] J. Park, M. Kwak, K. Moon, J. Woo, D. Lee, and H. Hwang, “TiOx-based RRAM synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing,” *IEEE Electron Device Lett.*, vol. 37, no. 12, pp. 1559–1562, Aug. 2016.

[16] Q. Liu *et al.*, “33.2 A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, p. 5.

[17] X. Peng, R. Liu, and S. Yu, “Optimizing weight mapping and data flow for convolutional neural networks on RRAM based processing-in-memory architecture,” in *Proc. ISCAS*, May 2019, pp. 1–5.

[18] B. Liu *et al.*, “Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems,” in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD)*, Nov. 2014, pp. 63–70.

[19] J. Woo, X. Peng, and S. Yu, “Design considerations of selector device in cross-point rram array for neuromorphic computing,” in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–4.

[20] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Shrunk-2-D: A physical design methodology to build commercial-quality monolithic 3-D ICs,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1716–1724, May 2017.

[21] L. Brunet *et al.*, “First demonstration of a CMOS over CMOS 3D VLSI CoolCube integration on 300 mm wafers,” in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2016, pp. 1–2.

[22] C.-C. Yang *et al.*, “Enabling low power BEOL compatible monolithic 3D+ nanoelectronics for IoTs using local and selective far-infrared ray laser anneal technology,” in *IEDM Tech. Dig.*, Dec. 2015, pp. 1–8.

[23] Z. Fang *et al.*, “Temperature instability of resistive switching on HfO$_x$-based RRAM devices,” *IEEE Electron Device Lett.*, vol. 31, no. 5, pp. 476–478, Jul. 2010.

[24] C. X. Xue *et al.*, “A 1Mb multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN based AI edge processors,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Aug. 2019, p. 388–390.

[25] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, “14.5 envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 246–247.

[26] Z. Yuan *et al.*, “Sticker: A 0.41-62.1 TOPS/W 8Bit neural network processor with multi-sparsity compatible convolution arrays and online tuning acceleration for fully connected layers,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2018, pp. 33–34.

[27] A. Sayal, S. S. T. Nibhanupudi, S. Fathima, and J. P. Kulkarni, “A 12.08-TOPS/W all-digital time-domain CNN engine using bi-directional memory delay lines for energy efficient edge computing,” *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 60–75, Jan. 2020.

[28] J. Zhang, C. Lee, C. Liu, Y. S. Shao, S. W. Keckler, and Z. Zhang, “Snap: A 1.67–21.55tops/w sparse neural acceleration processor for unstructured sparse deep neural network inference in 16nm cmos,” in *Proc. Symp. VLSI Circuits*, Dec. 2019, pp. 306–307.

[29] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, “UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision,” *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.

[30] K. Ueyoshi *et al.*, “QUEST: Multi-purpose log-quantized DNN inference engine stacked on 96-MB 3-D SRAM using inductive coupling technology in 40-nm CMOS,” *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 186–196, Jan. 2019.

[31] L. Chen, X. Tang, A. Sanyal, Y. Yoon, J. Cong, and N. Sun, “A 0.7-V 0.6-μW 100-kS/s low-power SAR ADC with statistical estimation-based noise reduction,” *IEEE J. Solid-State Circuits*, vol. 52, no. 5, pp. 1388–1398, May 2017.

[32] J.-H. Kim, J. Lee, J. Lee, H.-J. Yoo, and J.-Y. Kim, “Z-PIM: An energy-efficient sparsity aware processing-in-memory architecture with fully-variable weight precision,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2020, pp. 1–2.

[33] 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, p. 770.

[34] X. Peng, S. Huang, Y. Luo, X. Sun, and S. Yu, “DNN+neuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies,” in *IEDM Tech. Dig.*, Dec. 2019, p. 5.

[35] M. Mao, X. Peng, R. Liu, J. Li, S. Yu, and C. Chakrabarti, “MAX2: An ReRAM-based neural network accelerator that maximizes data reuse and area utilization,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 398–410, Jun. 2019.