MARVEL: A Vertical Resistive Accelerator for Low-Power Deep Learning Inference in Monolithic 3D

Fan Chen, Linghao Song, Hai "Helen" Li, Yiran Chen
Department of Electrical and Computer Engineering, Duke University, Durham NC, U.S.A.
Email: {fan.chen, linghao.song, hai.li, yiran.chen}@duke.edu

***Abstract*—Resistive memory (ReRAM) based Deep Neural Network (DNN) accelerators have achieved state-of-the-art DNN inference throughput. However, the power efficiency of such resistive accelerators is greatly limited by their peripheral circuitry including analog-to-digital converters (ADCs), digital-to-analog converters (DACs), SRAM registers, and eDRAM buffers. These power-hungry components consume 87% of the total system power, despite of the high power efficiency of ReRAM computing cores. In this paper, we propose MARVEL, a monolithic 3D stacked resistive DNN accelerator, which consists of carbon nanotube field-effect transistors (CNFETs) based low-power ADC/DACs, CNFET logic, CNFET SRAM, and high-density global buffers implemented by cross-point Spin Transfer Torque Magnetic RAM (STT-MRAM). To compensate for the loss of inference throughput that is incurred by the slow CNFET ADCs, we propose to integrate more ADC layers into MARVEL. Unlike the CMOS-based ADCs that can only be implemented in the bottom layer of the 3D structure, multiple CNFET layers can be implemented using a monolithic 3D stacking technique. Compared to prior ReRAM-based DNN accelerators, on average, MARVEL achieves the same inference throughput with 4.5× improvement on performance per Watt. We also demonstrated that increasing the number of integration layers enables MARVEL to further achieve 2× inference throughput with 7.6× improved power efficiency.**

***Index Terms*—accelerator, monolithic 3D, DNNs, inference**

I. INTRODUCTION

Deep Neural Networks (DNNs) have demonstrated supreme performance accuracy on a variety of applications including computer vision [1], [2], natural language processing [3], and autonomous systems [4]. These performance advantages come at the cost of outstanding computing and storage requirements. For instance, VGG16 [2] involves 138M parameters, 15G operations, and >60MB memory storage, to perform inference on a single RGB image with only 224×224 pixels. Moreover, the network depth and model size continue to grow, which poses a major challenge for the deployment of DNN models on hardware, especially on end devices with limited available resources and constrained power budgets.

Given the diminishing benefits from general-purpose processors [5], various accelerators [6]–[11] have been proposed to accommodate the growing computing and storage demands of DNN models. Among these candidate designs, the emerging resistive memory (ReRAM) opened up a new horizon for *in-situ* processing massive matrix-vector multiplication (MVM)—the computation kernel in DNNs—with 1,000× to 10,000× energy-delay efficiency improvement compared to custom digital MVM engines [12]. Such in-ReRAM MVM is essentially a form of analog computation. To communicate with other digital components in the system, the inputs of the ReRAM crossbars are converted into analog voltages using digital-to-analog converters (DACs), and the output analog currents of the ReRAM crossbars are digitized using multi-bit analog-to-digital converters (ADCs). The CMOS ADCs and DACs account for ∼58% of the power consumption and ∼31% of the chip area in a typical resistive DNN accelerator [9]. Therefore, state-of-the-arts ReRAM-based accelerators [9]–[11] suffer from huge power consumption and area overhead of CMOS ADCs/DACs and can only retain <10% of the energy-delay efficiency benefits of the ReRAM MVM engines at system-level.

3D integration technology that allows multiple stacked active layers connected with dense, high-speed interface has been explored to reduce energy consumption, silicon area, and wire delays in general-purpose processors [13], [14] as well as customized accelerators [15], [16]. The original 3D integration technology bonds pre-fabricated dies using through-silicon-vias (TSVs) [13] achieved simultaneously 15% performance improvement and 15% power reduction in a high performance microprocessor. The advanced monolithic 3D (M3D) technology using ultradense inter-layer vias (ILVs) allows multiple layers of devices to be sequentially fabricated on the same substrate, showing up to three orders of magnitude improvement on energy and performance compared to traditional 2D systems [14]–[16]. Such 3D circuits are modeled using custom CAD tools capable of 3D floorplanning [17], and the fabrication is compatible with existing CMOS ICs [14], [15]. The feasibility of the M3D approach has already been demonstrate in several hardware prototypes [16], [18]–[20].

Prior ReRAM-based DNN accelerators focus on (i) parallel DNN computation in morphable ReRAM arrays [10]; (ii) pipelined DNN processing with smart mapping and coding schemes [9]; (iii) general-purpose in-ReRAM processor with instruction set and compiler support [11]. None of them targeted to reduce the power consumption and area overhead of CMOS ADCs/DACs in order to retain the energy-delay benefits of ReRAM MVM engines. In this paper, we examine how to address this challenge by leveraging M3D integration and advanced nanotechnologies. The following highlights the main challenges and outlines our solutions and contributions.

1) **M3D integration reduces chip area but naïvely stacked resistive accelerator violates the thermal limits in mobile systems.** Even with advanced heat dissipation solutions, the chip peak temperature will increase significantly due to the high power consumption of the bottom ADC layer, which will increase the refresh energy of the eDRAM global buffers, thereby reducing the system power efficiency. To address the thermal challenge, we propose to replace the CMOS ADCs with carbon nanotube field-effect transistors

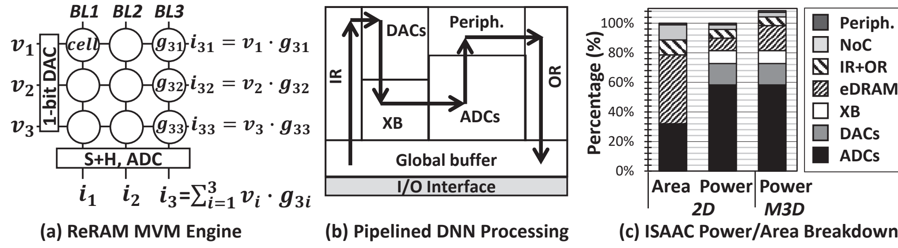


Fig. 1. ReRAM-based DNN processing.

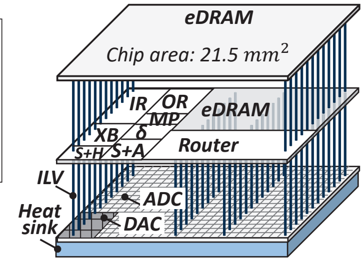


Fig. 2. M3D integrated ISAAC.

(CNFETs) based low-energy designs.

2) **Utilizing low-power CNFET ADCs is inevitable but it reduces DNN inference throughput.** Although the CNFET ADCs consums less energy, they are slow. To compensate the system performance, we explored the design space of CNFET ADCs in frequency/area/energy and propose to vertically implement multiple ADCs layers.

3) **The CMOS SRAM input/output registers and eDRAM buffers consume 56% of chip area and limit the system area efficiency.** We propose to realize SRAM registers with CNFET CMOS SRAM, because it provides acceptable access latency but occupies 3× smaller chip footprint. We also replace the eDRAM buffer with high-density cross-point Spin Transfer Torque Magnetic RAM (STT-MRAM). STT-RAM has $4F^2$ cell area, $<$10 *ns* latency and requires no refresh. To establish the effectiveness of our architectural innovation, we evaluate MARVEL against various DNN accelerators on the state-of-the-arts networks. On average, MARVEL delivers the same DNN inference throughput with 4.5× enhanced performance efficiency.

II. Background and Motivation

A. ReRAM-based DNN Accelerators

ReRAM MVM engine. An example MVM engine using a 3×3 ReRAM crossbar is shown in Figure 1 (a). Each ReRAM cell, denoted as a circle, stores an element value of the matrix as ReRAM conductance. The vector elements are converted to read voltages and applied onto the horizontal wordlines (WLs). Based on Ohm's law, the read current for each cell represents the multiplication between corresponding elements of the vector and the matrix, *e.g.,* $i_{31} = v_1 \cdot g_{31}$. The bitlines (BLs) aggregate the currents passing through all the cells on the same BL, *e.g.,* $i_3 = \sum_{i=1}^{3} v_i \cdot g_{3i}$, based on Kirchhoff's law. All the BLs simultaneously produce currents summation, and hence the multiplication between a $1 \times N$ vector and a $N \times N$ matrix can be parallel processed in a $N \times N$ ReRAM crossbar in analog domain. To facilitate communication with other digital components, a 1-bit DACs serially convert the digital input into driving voltages, while a multi-bit ADC connected to the BL converts the summation current to digital signals. Prior ReRAM-based MVM engine [12] using 512×512 crossbars and 4-bit ADCs achieved up to 10,000× speed-energy efficiency improvement compared to ASICs.

Pipelined *in-situ* DNN processing. DNN processing is essentially vectorized operation with well-defined dataflow. Most ReRAM-based DNN accelerators [9]–[11] are implemented with the weight-stationary dataflow [8], where the input streams flow through processing engines dedicated for different layers in a pipelined fashion for high system throughput. Figure 1 (b) conceptually illustrates a typical DNN processing pipeline in resistive accelerators. The major operations can be classified into (1) data access to *global buffer*, *input register (IR)*, and *output register (OR)*; (2) digital/analog conversion in *DACs* and *ADCs*; (3) analog MVM in *ReRAM crossbars (XB)*; and (4) supporting operations in *peripheral circuitry (periph.)*. The processing time of each stage varies but the cycle time of the pipeline is essentially bottlenecked by the ADCs. In this work, we use ISAAC [9] as the 2D baseline design. It employs a 8-bit 1.28 giga-samples-per-second (GSps) ADC shared among 128 BLs, resulting a 100 *ns* cycle time.

Power/area breakdown. The left two bars in Figure 1 (c) show the area and power breakdown in ISAAC [9]. The major bottleneck comes from the CMOS ADCs and DACs, consuming ∼73% of the power consumption and ∼32% of the chip area. As the ReRAM crossbar size and device resolution continues increasing, the power efficiency and area efficiency are aggravated to an extent that renders resistive DNN accelerators impractical. Moreover, the IR/OR and eDRAM buffers consume 56% of the chip area and significantly limit the system area efficiency. Overall, the power-consuming and area-hungry CMOS ADC/DACs, IR/OR, and eDRAM buffers yield the core in-ReRAM processing insignificant.

B. Monolithic 3D (M3D) Vertical Processor

M3D integration. The advanced monolithic 3D (M3D) technology is fundamentally different from the conventional 3D because of its three unique characteristics. First, M3D sequentially fabricated multiple active device layers on the same substrate. In the contrast, conventional through-silicon-vias (TSVs) based 3D (TSV3D) fabricates different wafers in parallel followed by a stacking or bonding step. Second, the tightly integrated upper layers in M3D can only be implemented with technologies that are compatible with low-temperature ($<$ 300°C) fabrication process. Third, M3D achieves fine-grained vertical integration through denser nano-scale inter-layer vias (ILVs). Such ILVs have the same pitch and dimensions as conventional metal layer vias in today's IC, providing orders of magnitude smaller area overhead than TSVs. We show the detailed comparison of ILV-based M3D and two designs of TSV3D in terms of integration level, via diameter/height/pitch and capacitance in Table I. The ILV model is adopted from [21]. We obtain the 5μm TSV model from [22]. An aggressive mini-TSV model with 60% diameter reduction [20] compared to conventional

TABLE I
THE TSV3D AND M3D COMPARISON [21]–[23].

	Level	Diameter	Height	Pitch	Cap.
ILV	device	50nm	310nm	80nm	0.05fF
TSV	wafer	5μm	30μm	20μm	30.8fF
mini-TSV		2μm	20μm	6μm	3.8fF

TSV is also considered in this work. The numbers of mini-TSV are estimated based on the high aspect ratio via model [23] using the TSV parameters in [21]. As shown in the table, ILVs are shorter and thinner than TSVs, resulting 616× (76×) smaller capacitance compared to TSV (mini-TSV). Hence, the delay and power of the gates that drive an ILV are substantially reduced. A M3D SRAM prototype demonstrated 7.81% energy savings and 2.15× latency reduction [24] at circuits level.

M3D-based Processors. Recent research efforts [14], [15], [19] monotonically stack circuits layers through fine-grained ILVs to improve system power efficiency. For CMOS based designs, upper layers have lower-performance transistors due to the manufacturing low-temperature thermal constraints. A prior work [14] examined partitioning scheme for logic and storage structures in a processor for M3D and showed that a M3D processor run 25% better with 41% less energy than a 2D core. The Nano-Engineered Computing Systems Technology (N3XT) approach [15] advocate heterogeneous integration of logic devices with emerging nonvolatile memory in a M3D architecture to accommodate the computing demands of abundant-data applications. The N3XT approach has been demonstrated in several hardware prototypes [15], [16], [18], [25] including a general 64-core CPU system [15] and a DNN accelerator [25].

C. Carbon Nanotube Field-Effect Transistor

Carbon nanotube (CNT) field-effect transistors (CNFETs) have a higher-than-CMOS current density, 10× improved energy-delay product (EDP) compared to CMOS, and scalability down to 3 *nm* and beyond [26]. Due to the low-temperature back-end-of-line (BEOL)-compatible CNT-specific processing, CNFET circuits can be directly fabricated on upper layers in a M3D structure [18]. A systematic evaluating approach with a realistic yield model is presented in [27]. Recent works have demonstrated various low-power CNFET prototypes including both digital systems [18], [28] and analog [29] circuits.

D. Motivation

Naïve M3D baseline. In this paper, we take ISAAC [9] as a 2D baseline to explore how to leverage 3D integration technology in resistive DNN accelerator designs. We explored different partitions and identifies the best candidate solution as shown in Figure 2. We place the most energy-consuming ADCs/DACs at the bottom layer, taking into account that the bottom layer is closest to the heat sink. We partition the most area-consuming eDRAM buffers into two portions, allocating a smaller piece to the ReRAM layer to obtain a uniformly distributed $\sim 21.5mm^2$ layer footprint. We consider TSV, mini-TSV, and ILV as three candidates of inter-layer connections. In ISAAC, each 128×128 ReRAM crossbar consumes $25\mu m^2$ area and requires connections to 128 1-bit DAC, 1 8-bit ADC, and other peripheral circuits. The required vertical interconnection density is $2 \times 10^6/mm^2$. However, current TSV interconnections achieves a maximum density of 10^5 vias/mm^2 [22], [23]. Hence, our baseline 3D resistive accelerator is integrated in M3D through ILVs.

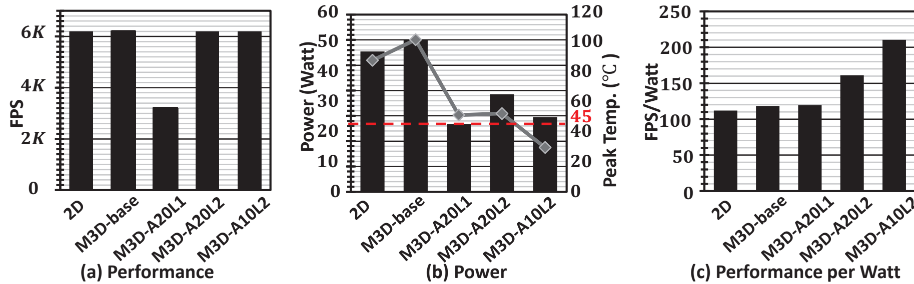
Thermal/performance analysis. We use 3D-ICE [30] for thermal simulations. Results show that the peak temperature and the average power density in baseline M3D ISAAC are 103 °C, and 200 $Watt/cm^2$, significantly violating the 45°C skin temperature limit specifications set for mobile systems [31]. In addition, the retention time of the 8 MB eDRAM decreases as temperature increases. Based on our simulation on CACTI [32], the 8 MB eDRAM in ISAAC has $100\mu s$ retention time at 103°C, resulting in 2× more refresh energy than room temperature [33]. The right-most bar in Figure 1 (c) shows the energy consumption of baseline M3D design normalized to 2D ISAAC. Unfortunately, the energy reduction brought by wire delays is offset by the increase in refresh energy of the eDRAM. *Compared to a 2D design, the naïve M3D integrated accelerator cannot reduce system energy consumption, but causes an energy consumption increase of about 10%.*

III. MARVEL

The baseline 3D resistive accelerator cannot benefit from the ultradense vertical connections from M3D integration but increases the system power, while requiring advanced thermal management to support significant heat dissipation for the bottom ADC layer. In this section, we propose MARVEL, a resistive vertical accelerator architecture for low-power DNN inference that can be integrated using monolithic 3D integration through low-temperature fabrication.

A. Low-Power CNFETs based ADCs

The first goal of MARVEL is to limit the heat dissipation of the bottom layer consisting DACs and DACs within the mobile system requirement. ISAAC is composed of 168 tiles connected with on-chip mesh. Each tile consists of 12 processing engines (PEs). Each engine has 8 ReRAM crossbars, 8 8-bit 1.28 GSps ADCs, 8×128 1-bit DACs, and some supporting circuits. The DACs are essentially invertors with negligible hardware overheads. We simply replace them with CNFET invertors. For the power-consuming ADCs, we explored to replace them with the recent low-power CNFET ADC [29]. However, the successive-approximation-register (SAR) ADC in [29] is implemented in 10 nm technology, and extremely optimized for energy with 10 nW energy consumption from 1.1 V supply with 20 kHz clock frequency. Despite the low power consumption, the low sampling rate prevents it from being used in DNN accelerators. To obtain the power and area for the SAR ADC with higher sample rate, we linearly scale the power and area of the vref buffer, memory, and clock in the original SAR ADC [29] , while scale the power and area of the capacitive DAC exponentially. Due to the page limit, we omit the detailed scale approach and refer interested readers to [34]. We obtained one 8-bit CNFET ADC with 310 μm^2 area, 20 MSps sample rate, and 140 μW energy consumption. The $21.5mm^2$ layer area can accommodate total $32 \times 12 \times 168$ such 8-bit ADCs. Therefore, each ADC has to be shared among 4 BLs within the 128×128 ReRAM crossbar. We denote such a design as *M3D-A20L1*,



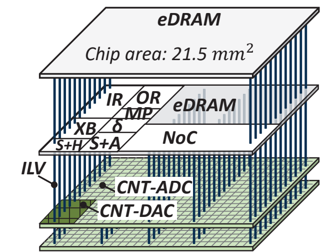


Fig. 3. The performance, power, and performance per Watt comparison between 2D and 3D designs.

Fig. 4. M3D integrated ISAAC with two-layer CNFET ADC.

TABLE II
THE 10 MHZ AND 20MHZ CNTFET ADC COMPARISON [29], [34].

Freq.	Power	Area	Num.	Total Area
20 MHz	140 μW	310 μm^2	$32 \times 12 \times 168$	20 cm^2
10 MHz	35 μW	150 μm^2	$64 \times 12 \times 168$	19.4 cm^2

representing we replace the original CMOS ADC with one layer of 20 MSps CNFET ADC. We compare the performance, power, and performance per Watt of *M3D-A20L1* against 2D and M3D baseline designs in Figure 3. All the experiments are run on AlexNet [1]. The experimental setup is detailed in Section IV. The results show that the power consumption of the bottom ADC layer has significantly reduced compared to the M3D baseline (*i.e., M3D-base*), resulting a 54.5 $Watt/cm^2$ average power density and 52°C peak temperature. However, the performance of *M3D-A20L1* reduced by 2× compared to 2D and 3D baseline design. This is because *M3D-A20L1* shares a 20 MSps ADC unit among 4 BLs, leading to 200 *ns* pipeline cycle, while the 2D and 3D baseline have a 100 *ns* cycle time.

B. Multi-layer CNFETs ADCs

In order to improve the performance of *M3D-A20L1*, we inevitably have to construct multiple layers of CNFET ADCs. Thanks to the low-temperature (<300 °C) fabrication property of CNFETs, they are ideal for M3D integration to provide homogeneous layer-wise performance. We first explore to integrating two layers of the 20 MSps ADC unit. Moreover, we perform a design space exploration of CNFET ADCs, and report the power and area in Table II. We consider two candidates: the aforementioned 20 MSps design and a 10 MSps ADC unit. For both designs, it requires two layers of implementation to obtain 100 *ns* cycle time. We denote the designs with 2-layer 20 MSps ADCs and 2-layer 10 MSps ADCs as *M3D-A20L2* and *M3D-A10L2*, respectively. The comparisons on performance, power, and performance per Watt are exhibited in the right-most two bars in Figure 3. *M3D-A20L2* and *M3D-A10L2* both have the same throughput as 2D and 3D baselines, but *M3D-A10L2* achieves 1.3× better performance per Watt than *M3D-A20L2*. This can be explained by the comparisons in Table II. When we scale the frequency of ADCs from 20 MHz to 10 MHz, the area reduced by 2× while power reduced by 4×. The unproportionate reduction makes the M3D design favor the 10 MHz at the system-level. For the following evaluations, all the results are reported by using the 10 MSps ADC unit.

C. Low-Power CNFET SRAM registers and High Density Global Buffer using 3D Cross-Point STT-MRAM

The IR, OR, and eDRAM buffers are the area bottleneck in 2D and M3D baselines. To increase the system area efficiency, we realize IR/OR with CNFET-based CMOS SRAM, since it provides <100 *ns* access time, occupies 3× smaller chip area, requires no fresh, and hence consumes lower power than eDRAM [35]. The global buffer is implemented with high-density cross-point Spin transfer torque magnetic RAM (STT-MRAM) which has a 4 F^2 cell area, <10 *ns* operation speed, >10 year data retention, and $> 10^6$ endurance [36]. We adopt the SRAM parameters from a recent CNFET based CMOS 1 Kbit 6T SRAM prototype [28]. We scale the SRAM size to accommodate the 4 MB IR and 1 MB OR requirements in ISAAC for a fair comparison. The memory system performance, energy, and area are simulated with CACTI [32] and reported in Table V. The device parameters of cross-point STT-MRAM are adopted from a commercial chip [37]. We implemented a 8 MB STT-RAM buffer within 4.648 cm^2 chip area in three layers. The circuit-level performance, energy, and area are simulated using modified NVSim [38] and reported in Table V.

D. Overall MARVEL Organization

The overview of the MARVEL architecture is illustrated in Figure 5. All the peripheral circuits are implemented with CNFETs. We use the Stanford University Virtual Source CNFET Model [39] for device simulation. For digital circuit simulation we use the Variation-Aware Nanosystem Design Kit [40], which is compatible with standard synthesis and place-and-route tools while accounting for variations in CNT density and CNT diameters. For apples-to-apples comparison against the 2D ISAAC baseline design, we adopt the same ReRAM device parameters as in ISAAC. The leakage power, dynamic energy, latency and area of MARVEL is modeled with the same 32nm process technology. The ReRAM MVM engine is operated at 10 MHz. Given the simulated area and power consumption, where the MVM engines are sandwiched by the upper and lower ADC layers to maximize the benefits of high-throughput vertical connections. The overall architecture consumes 14.53 Watt power and occupies 13.383mm^2 chip area. We summarize the detailed power consumption and area overhead of MARVEL in Table V.

IV. EVALUATION

Benchmark. We evaluate the MARVEL architecture using four state-of-the-art DNNs listed in Table III. LeNet [41] is trained with MNIST for simple hand-written digits classification, while AlexNet [1], VGG16 [2], ResNet-18 [42], and MobileNet [43] are trained with ImageNet for complex

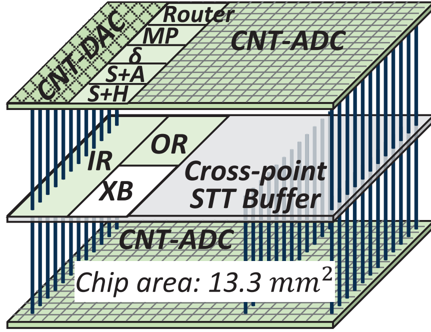


Fig. 5. The MARVEL architecture.

classification tasks. All models are trained in TensorFlow. We quantized both the activations and weights of all CNNs with 16-bit of precision as the same in ISAAC [9].

Scheme. We compared MARVEL against five counterparts: an Intel Xeon E5-2630 V3 8-core CPU, an Nvidia GTX 1080 GPU, a Xilinx Virtex7 FPGA [7], one ASIC chip Google TPU [6], and 2D ISAAC [9]. Google TPU comprises four chips, each of which can achieve larger throughput but consume more power. We describe the configuration and list the power consumption of each counterpart in Table IV. We denote the 3-layer MARVEL architecture shown in Figure 5 as MARVEL-*A*. To further evaluate the potential benefits of multi-layer M3D integration, we also constructed a 5-layer MARVEL architecture denoted as MARVEL-*B*. In MARVEL-*B* we doubled the 10 MSps ADC layers. The structure is similar as MARVEL-*A* except that the middle in-ReRAM processing layers are sandwiched by two upper and two lower ADC layers.

Simulation. The run times of CPU/GPU platforms are measured by Tensorflow and the energy costs are measured on real hardware. The FPGA numbers are scaled and calculated based on the original paper. We build a in-house simulator to model the performance of TPU. We build a simulator based on NVSim [38] to evaluated the inference throughput, power, and energy consumption of ReRAM-based accelerators including ISAAC, MARVEL-A, and MARVEL-B.

Performance. Figure 6 compares the CNN inference throughput among different designs. In general, domain-specific accelerators achieve higher frame per second (FPS) than general-purpose CPU and GPU due to the specialized hard-

TABLE III
THE CNN BENCHMARKS (ACC: ACCURACY; C: CONVOLUTIONAL LAYER; P: POOLING LAYER; F: FULLY CONNECTED LAYER).

name	database	topology	Ops
LeNet	MNIST	3C,2P,1F	< 1M
AlexNet	ImageNet	5C,3P,2F	724.4M
VGG16	ImageNet	13C,5P,3F	15.5G
ResNet-18	ImageNet	18C,2P,1F	11.3G
MobileNet	ImageNet	10C,1P,1F	569.4M

TABLE IV
THE SCHEME COMPARISON (NORMALIZED TO $32nm$).

Name	Description	$Power\ (W)$
CPU	Intel Xeon E5-2630 V3	85
GPU	Nvidia Tesla P100	250
FPGA	Xilinx Virtex7 VX485T	40
TPU	4-chip ASIC	384
ISAAC	ReRAM PIM	65.8

TABLE V
3-LAYER MARVEL-A PARAMETERS.

Component	Params	Spec	Power (mW)	Area (mm^2)
First Layer (Bottom Layer)				
CNFET ADC	resolution frequency number	8 10 MHz 178,356	2477.2	13.377
Total			**2.48 W**	**13.377**
Second Layer (Middle Layer)				
ReRAM	size bit precision number	128×128 2 12,288	4838.4	0.403
IR	size	4 MB	833.28	1.411
OR	size	1 MB	248.64	0.697
3-layer STT-RAM Buffer	size num_bank bit-width number	64 KB 2 256 168	1159.2	4.648
STT-RAM bus	width number	384 168	392	5.04
Total			**7.47 W**	**12.2**
Third Layer (Top Layer)				
CNFET ADC	resolution frequency number	8 10 MHz 79,692	1106.8	5.977
CNFET DAC	resolution number	1 16.5 M	2688	0.3472
Router	flit_size num_port	32 8	588	6.342
S+H	number	16.5 M	6.72	0.081
S+A	number	8064	137.2	0.494
maxpool	number	168	22.4	0.0403
sigmoid	number	336	29.12	0.101
Total			**4.58 W**	**13.383**
MARVEL-A Total			**14.53 W**	**13.383**
MARVEL-B Total			**18.91 W**	**13.383**
ISAAC [9] Total			**65.8 W**	**62.5**

ware and optimized data management. Intuitively, as hardware resources increase, the accelerators will achieve higher performance. For instance, the 4-chip Google TPU achieves better performance compared to FPGA accelerator and even ReRAM-based ISAAC in most benchmarks. The only exception is MobileNet on ImageNet, this is because MobileNet has a compact structure with less data reuse compared to previous CNN models. In this case, ISAAC outperforms TPU in FPS thanks to its low-cost data access. Among all the benchmarks. MARVEL-A achieves similar-to-ISAAC FPS since they have the equivalent hardware resources and cycle time. On the contrary, MARVEL-B improves the inference throughput by 2×, because it has more CNFET ADCs compared to MARVEL-A, resulting in a 2× reduction in cycle time.

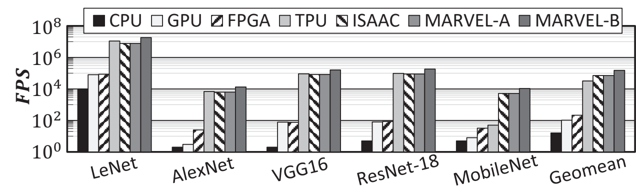


Fig. 6. The performance of different designs.

Performance per Watt. Figure 7 exhibits the CNN inference performance per Watt. The higher the bars are, the more power efficient the corresponding architectures are. In general, the FPS per Watt of different designs share the similar trend with the throughput shown in Figure 6. The significantly low power efficiency of CPUs and GPUs excludes them from low-power DNN inference applications on edge devices. The 4-chip

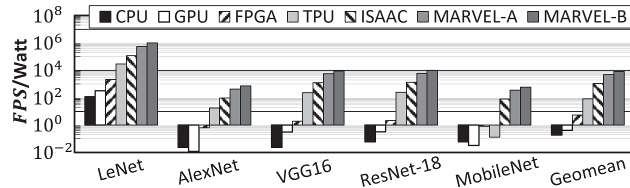


Fig. 7. The performance per Watt of different designs.

TPU achieve only ∼20% power efficiency compared to 2D ISAAC in the execution of conventional CNN models. For the recent compact MobileNet, TPU obtains <1% power efficiency compared to 2D ISAAC. The 3-layer MARVEL-A achieves 4.5× improvement on performance efficiency compared to 2D baseline. As we increase the integration layer to five, MARVEL-B further achieves 7.6× improvement on performance per Watt compared to the 2D ISAAC.

V. CONCLUSION

In this paper, we present MARVEL for low-power DNN inference. We first propose to integrate multiple CNFET ADC layers to reduce power consumption while maintaining system throughput. We then leverage emerging cross-point STT RAM and CNFET SRAM to replace the area- and energy-consuming eDRAM and CMOS SRAM. Compared to the 2D baseline, a 3-layer MARVEL can achieve the same inference throughput with 4.5× improved throughput per Watt. We further demonstrated a 5-layer MARVEL to show the potential of M3D integrated vertical resisitice DNN accelerators. On average, a 5-layer MARVEL achieves 2× throughput with 7.6× throughput per Watt compared to the 2D baseline.

ACKNOWLEDGMENT

This work was supported by NSF EECS-2023752, NSF 1955246, NSF 1910299, and ARO W911NF-19-2-0107. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of grant agencies or their contractors.

REFERENCES

[1] A. Krizhevsky *et al.*, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.
[2] K. Simonyan *et al.*, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, 2015.
[3] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv e-prints*, 2018.
[4] M. Bojarski *et al.*, "Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car," *arXiv e-prints*, 2017.
[5] H. Esmaeilzadeh *et al.*, "Dark Silicon and the End of Multicore Scaling," in *ISCA*, 2011.
[6] N. P. Jouppi *et al.*, "In-Datacenter Performance Analysis of a Tensor Processing Unit," in *ISCA*, 2017.
[7] C. Zhang *et al.*, "Optimizing FPGA-Based Accelerator Design for Deep Convolutional Neural Networks," in *FPGA*, 2015.
[8] Y.-H. Chen *et al.*, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in *ISSCC*, 2016.
[9] A. Shafiee *et al.*, "ISAAC: A Convolutional Neural Network Accelerator with in-Situ Analog Arithmetic in Crossbars," in *ISCA*, 2016.
[10] L. Song *et al.*, "PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning," in *HPCA*, 2017.
[11] D. Fujiki *et al.*, "In-memory data parallel processor," in *ASPLOS*, 2018.
[12] M. Hu *et al.*, "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *DAC*, 2016.
[13] B. Black *et al.*, "Die Stacking (3D) Microarchitecture," in *MICRO*, 2006.
[14] B. Gopireddy *et al.*, "Designing Vertical Processors in Monolithic 3D," in *ISCA*, 2019.
[15] M. M. S. Aly *et al.*, "The N3XT Approach to Energy-Efficient Abundant-Data Computing," in *Proceedings of the IEEE*, 2018.
[16] T. F. Wu *et al.*, "Brain-inspired computing exploiting carbon nanotube FETs and resistive RAM: Hyperdimensional computing case study," in *ISSCC*, 2018.
[17] S. Bobba *et al.*, "Cell Transformations and Physical Design Techniques for 3D Monolithic Integrated Circuits," *J. Emerg. Technol. Comput. Syst.*, 2013.
[18] M. M. Shulaker *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *IEDM*, 2014.
[19] L. Brunet *et al.*, "First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers," in *VLSI*, 2016.
[20] S. K. Samal *et al.*, "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology," in *S3S Conference*, 2016.
[21] J. Shi *et al.*, "A 14nm FinFET transistor-level 3D partitioning design to enable high-performance and low-cost monolithic 3D IC," in *IEDM*, 2016.
[22] Z. Xu *et al.*, "Through-Silicon-Via Fabrication Technologies, Passives Extraction, and Electrical Modeling for 3-D Integration/Packaging," *IEEE Transactions on Semiconductor Manufacturing*, 2013.
[23] S. Van Huylenbroeck *et al.*, "Small Pitch, High Aspect Ratio Via-Last TSV Module," in *ECTC*, 2016.
[24] S. Srinivasa *et al.*, "Compact 3-d-sram memory with concurrent row and column data access capability using sequential monolithic 3-d integration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2018.
[25] W. Hwang *et al.*, "Special session paper 3D nanosystems enable embedded abundant-data computing," in *CODES+ISSS*, 2017.
[26] A. D. Franklin *et al.*, "Sub-10 nm carbon nanotube transistor," *Nano letters*, 2012.
[27] S. Banerjee, A. Chaudhuri, and K. Chakrabarty, "Analysis of the impact of process variations and manufacturing defects on the performance of carbon-nanotube fets," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 6, pp. 1513–1526, 2020.
[28] P. S. Kanhaiya *et al.*, "Carbon Nanotube-Based CMOS SRAM: 1 kbit 6T SRAM Arrays and 10T SRAM Cells," *IEEE Transactions on Electron Devices*, 2019.
[29] A. G. Amer, R. Ho, G. Hills, A. P. Chandrakasan, and M. M. Shulaker, "Sharc: Self-healing analog with rram and cnfets," in *ISSCC*, 2019.
[30] A. Sridhar *et al.*, "3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs," *IEEE Transactions on Computers*, 2014.
[31] V. Chiriac *et al.*, "A figure of merit for mobile device thermal management," in *ITherm*, 2016.
[32] S. Thoziyoor *et al.*, "Cacti 5.1," tech. rep., Technical Report HPL-2008-20, HP Labs, 2008.
[33] C. Wilkerson *et al.*, "Reducing Cache Power with Low-Cost, Multi-Bit Error-Correcting Codes," *SIGARCH Comput. Archit. News*, 2010.
[34] M. Saberi *et al.*, "Analysis of power consumption and linearity in capacitive digital-to-analog converters used in successive approximation ADCs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2011.
[35] A. Bachtold *et al.*, "Logic circuits with carbon nanotube transistors," *Science*, 2001.
[36] H. Yang *et al.*, "3D Cross-Point Spin Transfer Torque Magnetic Random Access Memory," in *Spin*, 2017.
[37] Y. Huai *et al.*, "High density 3D cross-point STT-MRAM," in *IEEE International Memory Workshop*, 2018.
[38] X. Dong *et al.*, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2012.
[39] C. . Lee *et al.*, "A Compact Virtual-Source Model for Carbon Nanotube FETs in the Sub-10-nm Regime—Part I: Intrinsic Elements," *IEEE Transactions on Electron Devices*, 2015.
[40] G. Hills, "Variation-Aware Nanosystem Design Kit (NDK)," 2015.
[41] Y. Lecun *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.
[42] K. He *et al.*, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.
[43] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *CoRR*, vol. abs/1704.04861, 2017.