On-Chip Memory Technology Design Space Explorations for Mobile Deep Neural Network Accelerators

Haitong Li[1], Mudit Bhargava[2], Paul N. Whatmough[2], H.-S. Philip Wong[1]
[1]Stanford University [2]Arm Research
haitongl@stanford.edu

ABSTRACT

Deep neural network (DNN) inference tasks have become ubiquitous workloads on mobile SoCs and demand energy-efficient hardware accelerators. Mobile DNN accelerators are heavily area-constrained, with only minimal on-chip SRAM, which results in heavy use of inefficient off-chip DRAM. With diminishing returns from conventional silicon technology scaling, emerging memory technologies that offer better area density than SRAM can boost accelerator efficiency by minimizing costly off-chip DRAM accesses. This paper presents a detailed design space exploration (DSE) of technology-system co-design for systolic-array accelerators. We focus on practical/mature on-chip memory technologies, including SRAM, eDRAM, MRAM, and 3D vertical RRAM (VRRAM). The DSE employs state-of-the-art optimizations (e.g., model compression and optimized buffer scheduling), and evaluates results on important models including ResNet-50, MobileNet, and Faster-RCNN. Compared to an SRAM/DRAM baseline, MRAM-based accelerators show up to 4.68× energy benefits (57% area overhead), while a 3D VRRAM-based design achieves 2.22× energy benefits (33% area reduction).

1 INTRODUCTION

Deep neural network (DNN) inference has become ubiquitous in mobile devices due to effectiveness of DNNs in applications that incorporate computer vision tasks (e.g., image classification, object detection, tracking, and segmentation). However, DNN inference imposes a large compute and storage burden, which is challenging to meet for low-power mobile SoCs. In response to this challenge, silicon vendors and IP companies have developed *domain-specific hardware accelerators* for DNN workloads [1–3], which offer improved energy efficiency and throughput compared to general-purpose mobile CPUs or GPUs.

These accelerators typically demonstrate improved throughput and energy efficiency on matrix-vector and matrix-matrix products which are heavily used in the convolutional and fully-connected layers of DNNs. One of the key goals is to achieve a high compute density, by focusing on a large datapath optimized only for the computation patterns typically required for DNN inference. Systolic MAC arrays [1] and more flexible variants thereof [4] achieve this high compute density, with narrow 8-bit integer operands. However, in addition to achieving high compute density, it is crucial to minimize data movement costs as well. Weight data can be vast, especially for a DNN model with a large number of layers, or with several fully-connected (FC) layers. Similarly, activation data traffic generated between layers can be heavy. In the case of convolutional neural networks (CNNs), those data need to be repeatedly read for each filter in a layer. Therefore, it is essential to optimize the usage of on-chip memories in order to minimize access to off-chip DRAMs, which is typically one or two orders of magnitude more expensive in energy cost. In a systolic-array DNN accelerator design with 1-MB scratchpad SRAM, running ResNet-50 inference, we find that off-chip DRAM energy can take 75% of the total system energy.

For accelerators in commercial mobile SoCs, silicon area is typically constrained to a few square millimeters or less. This means that on-chip storage is limited and expensive off-chip DRAM access is a dominant power consumer. Next-generation on-chip memory technologies promise to improve density and energy efficiency, allowing the traffic of off-chip DRAM access to be ameliorated. Among on-chip memory solutions, emerging non-volatile memory (NVM) technologies [5] have been widely explored in the context of in-memory computing for neural networks. However, most prior works to-date focused on mixed-signal (e.g., analog MAC implementations) or neuromorphic (e.g., spiking neural networks) designs. These techniques require further extensive characterization and validation of their benefits prior to broad adoption. On the other hand, large-scale deployment of DNNs on hardware for domain-specific applications call for robust and easy-to-scale digital architectures. Riding the industry's momentum of advancing NVM technologies for a variety of applications such as small-capacity embedded memory for MCU and larger-capacity last-level cache, it is highly desirable to take a deep dive into the evaluation of memory technologies for digital DNN accelerator products.

In this work, we explore a variety of on-chip memory technology solutions, including both volatile and non-volatile technology fabrics, for area-constrained systolic-array DNN accelerators targeting mobile vision applications. The contributions of this paper are summarized below:

- We develop a design space exploration (DSE) flow to benchmark systolic-array DNN accelerators with incumbent and emerging memories emphasizing practical technology characteristics.
- The extensive DSE provides a detailed trade-off analysis for practical mobile DNN accelerator designs, which leads to an energy-area-efficiency landscape for various on-chip memory technologies.
- We find that even with current performance gap between NVM and state-of-the-art SRAM, efficient allocation of chip area resources balancing dense NVM and low-power SRAM provides overall energy-efficiency benefits.

DAC '19, June 2–6, 2019, Las Vegas, NV, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6725-7/19/06…$15.00
https://doi.org/10.1145/3316781.3317874

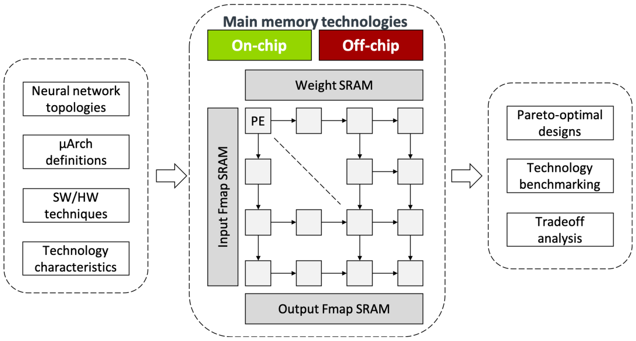


Figure 1: Schematic of the evaluation methodology for systolic-array DNN accelerators. Design space explorations for memory technology benchmarking are performed with different DNN models and software/hardware optimization techniques (model compression, static/dynamic SRAM allocation) evaluated.

2 BACKGROUND

2.1 DNN Accelerators

A wide variety of DNN accelerator designs for mobile vision applications have been published in recent years. Of these, most practical digital accelerators can be categorized into a few basic schemes: SIMD dot-product machines [6, 7], systolic arrays [1], and network-on-chip (NoC) based spatial arrays [4]. In general terms, these architectures all provide a large amount of low-precision MAC datapath, in combination with on-chip buffering for weights and activation data. The big difference between these is in the ordering of the computations, and the way the data movement is orchestrated. The latter is a real challenge with DNN acceleration because the volume of data that must be processed can be very large.

Different DNN layers have fundamentally different compute and storage requirements. Fully-connected (FC) layers are essentially matrix-vector products; the vector is reused, while the matrix is large and has no reuse. Multi-layer Perceptrons (MLPs) and the various forms of recurrent neural networks (RNNs) are built from FC layers. Convolutional neural network (CNN) layers, on the other hand, are processed as matrix-matrix products, with significant reuse of both the weight and activation (feature map) matrices. Hence, FC and CNN layers present very different challenges in terms of data movement. Nonetheless, in both cases, and variants thereof, the on-chip buffering plays a significant role in both performance and energy efficiency. Increasing on-chip buffering allows more weights and activations to be kept closer to computes, without having to access off-chip DRAM main memory. For weight data, on-chip memory allows for latency buffering in FC layers with no reuse and enables reuse across feature maps in CNNs. For activation data, on-chip memory allows feature maps to be buffered on-chip without spilling to DRAM. To get the most out of constrained storage resources, efficient buffer scheduling and model compression [8] are typically employed.

2.2 On-Chip Memory Technologies

In this section, we provide a brief discussion on various on-chip memory technologies that are evaluated as the main-memory replacement in the remainder of this paper. These include SRAM, spin-transfer-torque magnetic RAM (STT-MRAM), resistive RAM (RRAM), and embedded DRAM (eDRAM). We discuss essential technology characteristics and practical considerations, including manufacturing availability and technology maturity.

SRAM has been the mainstream on-chip memory technology due to its high-performance and logic compatibility. In fact, foundries often announced high-yielding/low-leakage SRAM in leading conferences as an indicator of technology readiness [9]. SRAMs are assumed to have unlimited endurance (the measure of number of acceptable writes cycles). However, SRAMs have low density and high static power, limiting its on-die capacity .

MRAM is a non-volatile memory technology that uses electron spin direction (up/down) to store a binary bit. MRAM offers inherently high speed, low energy, and highest endurance among NVM technologies [5], being compatible with CMOS logic (both logic voltage and fabrication processes) [10]. While several MRAM technologies exist, **STT-MRAM** is the most mature MRAM technology currently. To the best of our knowledge, all major foundries have plans to fabricate STT-MRAM and have MRAM roadmaps that include promising MRAM technologies like SOT-MRAM (or SHE-MRAM), VCMA-MRAM, and ME-MTJ MRAM. STT-MRAMs can be designed to trade off retention (the measure of non-volatility) with endurance or write energy. Working STT-MRAM in 28nm with plans down to 22nm by mid-2019 with sub-5ns cycle have been announced [11]. MRAM seems to be one of most promising technologies for large on-chip memory requirements.

RRAM is another CMOS-compatible, low-power, non-volatile memory technology [5]. Scaling-down towards sub-10-nm nodes and scaling-up in memory capacities (ranging from MByte to GByte) have been demonstrated [12]. In addition to planar 1T1R structures commonly used by various NVMs, RRAM can be fabricated in a 3D vertical architecture (3D VRRAM), similar to that of 3D vertical-NAND (VNAND), providing ultra-high density and low bit cost [5]. A four-layer 3D VRRAM array integrated with silicon CMOS logic has been demonstrated [13]. Endurance remains as a key factor to be considered when architecting RRAM into digital systems.

eDRAM technology can broadly be categorized into trench-cap (specialized process) based 1T1C eDRAM [14, 15] and the gate-cap based (but logic compatible) gain-cell eDRAM (GC-eDRAM) [16]. eDRAMs have higher density compared to SRAM (3-4x for 1T1C and 1.5-2x for GC-eDRAM). Both flavors require refresh operations to preserve data integrity. The storage node's capacitance for GC eDRAMs is much smaller and they are highly susceptible to increased leakage in scaled technology nodes. The most advanced 1T1C eDRAMs demonstration has been on 14nm CMOS technology on SOI substrate. Since the trench capacitance fabrication is more difficult to reliably manufacture on bulk CMOS technologies (the technology of choice for high performance for most leading foundries) compared to SOI technology, it is unlikely that 1T1C eDRAM will be available as a design option for future scaled high-performance designs.

2.3 Related Work

To date, there is a large body of work around DNN accelerator architectures [2–4, 6, 8]. In this work, we build our DSE and analysis on the well-known systolic array architecture [1], in order to focus on the technology-system interaction. There is also a significant research effort focused on mixed-signal arrays using analog NVMs as MAC units for DNN acceleration [17, 18]. We focus on the systolic-array digital architecture, which is a robust, easy-to-scale architecture that can incorporate the digital NVM technologies in

Table 1: Summary of technology inputs and the explored design space.

	PE	SRAM	MRAM	3D VRRAM	eDRAM	DRAM
Tech. node	14/16 nm	14/16 nm	28 nm	28 nm	28 nm	28 nm
Energy	0.3 pJ	[1.1, 1.5] pJ	Read: 4 pJ Write: 14 pJ	Read: 16 pJ Write: 48 pJ	19 pJ	120 pJ
Area	525 μm^2	32502 μm^2/32 KB	0.017 μm^2/bit	0.004 μm^2/bit	0.035 μm^2/bit	N/A
Design space	{16 × 16, 24 × 24, 32 × 32}	Weight/IFMap/OFMap: {32, 64, 128, 256, 1024} KB	MRAM-only (no off-chip DRAM)	VRRAM-only VRRAM + DRAM	eDRAM-only	LPDDR3

next-generation products. Finally, DSE for DNN accelerators has been considered before [19, 20]. Based on FPGA platform, those works are primarily throughput-oriented and do not involve detailed evaluations of on-chip memory technologies, whereas our DSE results probe into the energy-area efficiency tradeoffs for mobile DNN accelerators. Additionally, we also add important features for state-of-the-art model compression and buffer scheduling.

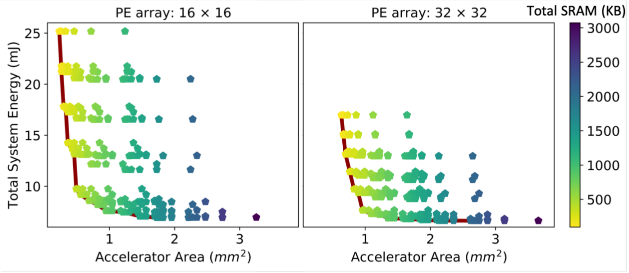
3 EVALUATION METHODOLOGY

It is of paramount importance to have a comprehensive understanding of the vast design space, from software to hardware perspectives, when evaluating and benchmarking memory technologies for DNN accelerators. Thus, the underlying simulation infrastructure needs to be built to have multi-dimensional variables exposed. These application-related and design-related parameters range from DNN models, to hardware micro-architectures, down to technology choices and characteristics. In this section, we introduce the evaluation methodology for driving technology evaluations and providing insights into accelerator design trade-offs.

3.1 System Simulation of DNN Accelerators

To capture the essential behaviors of systolic-array accelerators, we build our simulation infrastructure on top of SCALE-Sim [21], an open-source [1], cycle-accurate simulator for DNN accelerators. The SCALE-Sim simulator specifically models the systolic-array architecture, which consists of processing element (PE) arrays, on-chip scratchpad SRAM, and off-chip DRAM memory. Different mapping and data reuse strategies are supported for scheduling compute/memory operations on the array, leading to a variety of dataflows [4]. SCALE-Sim consumes DNN model definitions that describe layer-wise topology hyperparameters. It then simulates the feed-forward pass in the inference phase, producing cycle-accurate traces that lead to compute utilization and memory read/write statistics.

Figure 1 illustrates the overall structure of evaluations performed in this work. Our DNN accelerator designs consist of a systolic array (a 2-D mesh of PEs with local data movement), scratchpad SRAM blocks for buffering filter weights, input feature maps (IFMap), output feature maps (OFMap), and main memory. The main memory design choice is explored as being either off-chip (e.g., DRAM), embedded on-chip (e.g., NVM), or a hybrid configuration (DRAM/NVM). We simulate state-of-the-art DNN models that represent domain-specific workload characteristics (e.g., CNNs for computer vision tasks). For image classification workloads, the models we evaluate on DNN accelerators include ResNet-50, GoogLeNet, and MobileNet. For object detection workloads, FasterRCNN and YOLO-tiny are considered in the analysis. As these networks exhibit quite diverse characteristics (depth, topologies, and layer-wise hyperparameters), the analysis should generalize well, at least within the mobile vision domain. We pick ResNet-50 as a workload of focus to drive most design space explorations and inform technology-aware design decisions, as it is the largest among the five models in terms of memory resources required. Thus, if ResNet-50 can fit on chip, the rest of the CNN models can be accelerated without incurring additional overhead. For the mapping and scheduling of all the CNN models, output stationary (OS) dataflow is used, since it is generally more energy-efficient than other dataflows on systolic-array architectures [21].



(a) (b)

Figure 2: Baseline accelerator designs with off-chip DRAM running ResNet-50. Two PE array sizes, (a) 16 × 16 and (b) 32 × 32, are shown as examples. Pareto frontiers are extracted optimizing for energy and area efficiency.

We also extend SCALE-Sim with some new features necessary to accurately model the data movement in a state-of-the-art accelerator. First, we added data compression, which happens before mapping a DNN model onto the accelerator hardware, and aims to reduce weight load costs upfront, regardless of the underlying hardware fabric. Different weight compression ratios are analyzed (1x to 4x), similar to recent practical designs [22]. Second, as a hardware design choice, scratchpad SRAM can be designed as dedicated blocks for weights and feature maps, or as a unified SRAM with run-time buffer allocation based on layer-wise characteristics.

3.2 Design Space Exploration Parameters

Technology characteristics along with micro-architecture parameters are described in this section. For the DNN accelerator designs, the underlying hardware components include PE, SRAM, on-chip memory, and/or DRAM. Table 1 summarizes the energy consumption and area costs for the compute and memory fabrics, as well as the associated design space being explored. Optimizing for a widely-adopted 8-bit precision for CNN inference, each PE consumes 0.3 pJ

[1] www.github.com/ARM-software/SCALE-Sim

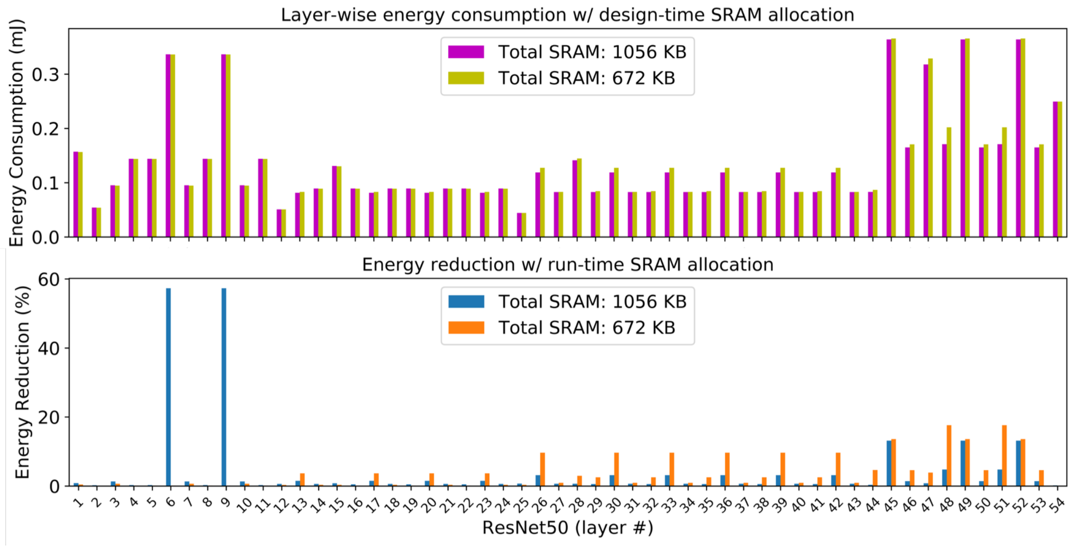


Figure 3: Energy efficiency comparison of design-time (upper) and run-time (lower) SRAM allocation schemes on ResNet-50.

per cycle at 1 GHz and occupies 525 μm^2, which are obtained from an in-house place-and-routed design. Different systolic array size is considered, ranging from 16×16 to 32×32, corresponding to different throughput targets and use cases. Industry-standard 16-nm SRAM instances are generated from an SRAM compiler, which are then used to construct small (32 KB) to large (1 MB) SRAM blocks for buffering filter weights and feature maps. Taking both dynamic and static energy into account, the size-sensitive energy consumption per byte is obtained, and the 32502 μm^2 area corresponds to one 32-KB instance from the SRAM compiler.

Two promising NVM technologies are analyzed here (Section 2. An MRAM design [23] features 4 pJ/byte for read access and 14 pJ/byte for write access, including peripherals. An effective bit area density of 0.017 μm^2/bit can be achieved. Without significant endurance limitation, no off-chip DRAM is included in the MRAM-based designs. Alternatively, 3D VRRAM provides another level of vertical scaling to achieve ultra-high density, featuring a 0.004 μm^2/bit area cost as a 4-layer main memory option [13]. Yet, read and write energy per byte access today are higher than that of MRAM designs. Without endurance optimization and resilience, 3D VRRAM may be used only for weight storage, and feature map accesses are directed to the off-chip DRAM in a hybrid configuration.

For the off-chip DRAM baseline designs, a key conservative assumption is that the DRAM DIMM modules are "free" in terms of chip area costs. In other words, the off-chip DRAM capacity is shared with mobile applications, and thereby is not included in the calculation of area costs. With an LPDDR3 interface, 120 pJ/byte is consumed by off-chip DRAM accesses [24]. Due to the the predictable dataflows in the efficient systolic-array architecture, we assume that the absolute latency of different memory technologies are hidden during the overlapped data fetching to PEs.

4 EXPERIMENTAL RESULTS

4.1 Baseline Designs with Off-Chip DRAM

In this section, we present the experimental results based on the technology-aware design space explorations of DNN accelerators with off-chip and on-chip memories. We first establish baseline designs with off-chip DRAM, by exploring design choices with variable sizes of PE arrays and different weight/IFMap/OFMap SRAM capacities, and then collect the total energy consumption and chip area of the accelerator designs running ResNet-50. Figure 2 shows the SRAM-dependency and PE-dependency trends, where we split the designs into two groups depending on the PE array size (corresponding to different throughput targets). In general, a larger PE array can better exploit data reuse within the systolic array to utilize cheap computes over expensive memory accesses. We do not explicitly explore very large PE arrays, but instead focus on a throughput-determined PE array design (e.g., 24×24 array). The scatter color-map corresponds to the total scratchpad SRAM capacity: darker scatters are collected from large SRAM designs. It can be observed that even though large SRAM designs tend to result in lower overall energy consumption due to less traffic to off-chip DRAM, there do exist "gaps" between multiple energy-efficiency clusters. This phenomenon manifests the efficiency gap between different SRAM block allocations among filter weights, IFMap, and OFMap in various designs that share a similar amount of total SRAM capacity, which is highly correlated with the nature of DNN models being accelerated (in this case, ResNet-50). Additionally, the search from 32 KB to 1 MB is relatively coarse-grained following the SRAM design practices. By optimizing energy-area efficiency, the extracted Pareto frontier provides solid baseline designs.

Extracting Pareto-optimal designs that already "distill" characteristics of all layers in a CNN model can be abstracted as a way of searching for globally efficient SRAM allocation during design time. Another way of achieving such a goal, by exploiting the deterministic nature of feedforward passes in CNNs, is to let NN compiler determine layer-wise allocation strategies and have run-time SRAM allocation. From a hardware implementation perspective, three SRAM blocks are replaced by a unified SRAM, and address mapping is performed specifically for weights, IFMap, and OFMap during layer-wise acceleration of CNN models. To compare the design-time and run-time SRAM allocation schemes, based on two Pareto-optimal designs generated previously, we run simulations

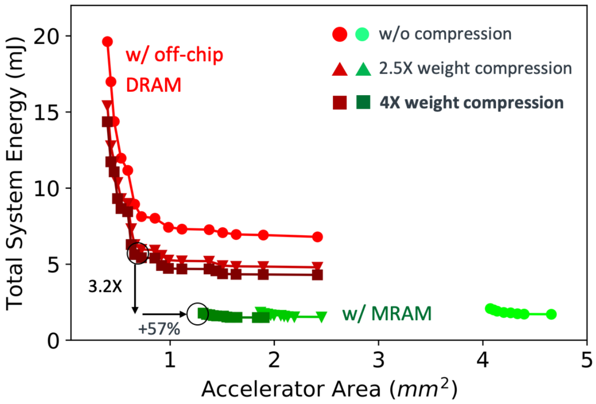


Figure 4: Energy-area tradeoffs for accelerator designs with on-chip MRAM and different weight compression rates.

to search layer-wise allocation strategies with the same total SRAM capacity between two schemes. The granularity of run-time allocation is 32 KB in the unified SRAM. Obtaining layer-wise breakdown for ResNet-50, Figure 3 shows the original energy consumption using design-time allocation and the layer-wise energy savings using run-time allocation. Early convolutional layers (e.g., layer #6 and layer #9) benefit most when total SRAM capacity exceeds 1 MB. Compiler would tend to allocate more SRAM resources for IFMap (> 800 KB) and only 32 KB SRAM for filter wights. In another case with 672 KB total SRAM capacity, towards the final layers of ResNet-50, >500 KB SRAM would be allocated to filter weights in order to benefit from the "layer evolution" of ResNet-50 (ratio between feature maps and weights). It is worth noting that the final fully-connected layer does not see a noticeable difference between design-time and run-time allocation schemes. Overall, a 8% overall energy saving is obtained using run-time allocation. Since this work is not focused on DNN accelerator compilers, we use design-time allocation scheme (i.e., having weight SRAM, IFMap/OFMap SRAM) for the following design space explorations and analysis.

4.2 Energy-Area Tradeoffs with MRAM

Embedding MRAM on-chip to eliminate off-chip DRAM accesses is first analyzed, based on the evaluation methodology and the initial baseline results introduced previously. The designs use 24 × 24 PE size for iso-throughput DSE. Both weights and activations are handled by on-chip MRAM, including read and write accesses.

The direct cost of off-loading the model storage capability from off-chip DRAM to on-chip NVM is the area overhead. With 8-bit precision, ResNet-50 requires 25 MB memory resources assuming sequential layer processing. Without hurting inference accuracy as the quality of service metric, weight compression can be utilized during the compilation phase. Leveraging the freedom of designing embedded memories tailored for model compression techniques, instead of sticking with large DIMM modules off-chip, area overhead can be reduced. Meanwhile, the reduced weight traffic benefits both off-chip DRAM designs and on-chip NVM designs. Figure 4 illustrates these two trends by analyzing the accelerator energy-area tradeoffs, where all data points are extracted from Pareto frontiers after full design space explorations for the respective memory technologies. Comparing two Pareto-optimal designs, 3.2× energy benefits (defined as the ratio between total energies of on-chip NVM designs versus off-chip DRAM designs) and 57% area overhead are obtained for MRAM designs incorporating 4× weight compression, given that off-chip DRAM baselines also benefit from reduced weight-traffic energy consumption. Owing to improved weight compression rates during NN compilation, on-chip NVM designs and off-chip DRAM designs see horizontal shifts and vertical shifts, respectively, in the area-energy metric space.

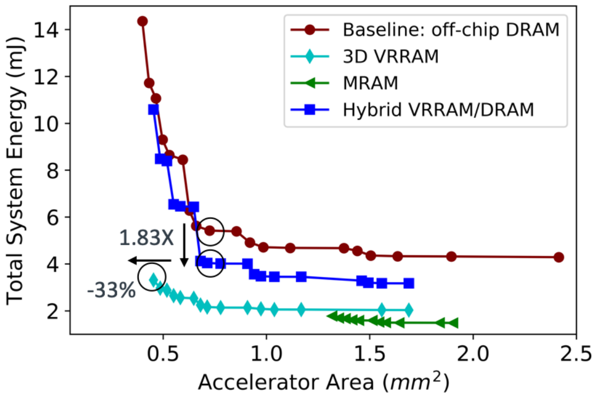


Figure 5: Energy-area tradeoffs for endurance-aware architectures: embedding 3D VRRAM with endurance resilience or using a hybrid VRRAM/DRAM configuration.

4.3 Endurance-Aware 3D RRAM Architectures

3D VRRAM provides the vertical scaling opportunity with ultra-high density for on-chip memories. Compared to MRAM, relative increase in read and write energy accesses and the endurance constraints would lead to interesting energy-area tradeoffs in design decisions, sometimes requiring us to re-architect the accelerator memory hierarchy.

We first consider the 3D VRRAM design with endurance resilience techniques incorporated (e.g., wear leveling) [25], such that activation data also go to on-chip NVM, eliminating off-chip DRAM accesses. The designs use 24×24 PE size for iso-throughput DSE. As indicated by Figure 5, vertical scaling (from planar NVM to multi-layer, 3D NVM) plays an important role in effectively bringing down the area cost. In fact, 1.83× energy benefits and 33% area reduction are achieved simultaneously compared to a Pareto-optimal baseline design, employing 4× weight compression for ResNet-50. This result implies that, even with a technology performance gap between NVM and state-of-the-art SRAM, good allocation of chip area resources can lead to a more efficient accelerator design balancing dense NVM and low-power SRAM .

Aiming for endurance-aware designs that employ a wider range of device stacks and designs without strong endurance resilience, we consider a hybrid configuration with both DRAM and 3D VRRAM (hybrid-VRRAM). Filter weights utilize ultra-dense 3D VRRAM for read-only accesses, whereas IFM and OFM activation data are directed to DRAM for read and write accesses. Minimum one-time weight writes are scheduled for 3D VRRAM. In this scenario, 1.36× energy benefits can be obtained at nearly iso-area (only 3% area overhead).

Table 2: Energy-area-efficiency landscape of on-chip memory solutions accelerating mobile vision applications.

Baseline: DRAM	All-SRAM			MRAM			VRRAM			Hybrid-VRRAM			eDRAM		
DNN models [*]	RN	MN	FRN	RN	MN	FRN	RN	MN	FRN	RN	MN	FRN	RN	MN	FRN
Energy benefits	3.48×	5.33×	3.14×	3.19×	4.68×	2.89×	1.83×	2.22×	1.88×	1.36×	1.30×	1.38×	1.81×	2.19×	1.64×
Area overhead		+733%			+ 57%			$-33\%(savings)$			+ 0.03% (nearly iso-area)			+ 172%	

[*] RN: ResNet-50. MN: MobileNet. FRN: Faster-RCNN.

4.4 Energy-Area-Efficiency Landscape

In this section, we present energy-area-efficiency benefits for on-chip memory solutions over off-chip DRAM baseline designs, leveraging the design space explorations and extracted designs of interest. In addition to the NVM technology explorations (MRAM and 3D VRRAM), we analyze another two potential solutions: all-SRAM designs and eDRAM-based designs. All-SRAM designs at 16-nm node utilize low-energy read and write accesses, yet with large sizes of SRAM the leakage energy becomes dominant. eDRAM-based designs feature 20 pJ per access with 0.035 μm^2/bit density [15]. Comparing Pareto-optimal designs generated from on-chip memory solutions and off-chip DRAM baselines, we summarize the key energy and area numbers in 2, for accelerating ResNet-50, MobileNet, and Faster-RCNN. eDRAM-based designs may provide energy benefits replacing off-chip DRAM, but still suffer from high area cost (172%), while its future scaling remains challenging. While the all-SRAM design are area-costly even with 16-nm technology, it offers the highest energy benefits across various solutions for all three state-of-the-art DNN models (3.14× to 5.33× energy benefits over the off-chip DRAM baseline). The MRAM-based design shows a comparable energy efficiency versus the all-SRAM design, even though an MRAM cell consumes higher write/read energy. This is mainly due to the trend that leakage energy becomes significant in the scaled-up all-SRAM design, diminishing the energy benefit of the advanced SRAM technology. The 57% area overhead in the MRAM design is a 12× reduction over the all-SRAM design. Without aggressive circuit optimization, the 3D VRRAM design provides a simultaneous reduction in both accelerator energy (1.83× to 2.22×) and area (33% less). Furthermore, without endurance resilience at the worst-case scenario, a hybrid VRRAM/DRAM configuration where 3D VRRAM serves as dense weight storage still provides energy benefits at near-iso-area. The key finding here is that ultra-dense NVM technologies (e.g., 3D VRRAM) provide ample opportunities optimizing accelerator area resource allocation.

5 CONCLUSION

In this paper, we perform technology-system design space explorations that examine the efficacy of on-chip memory technologies for practical DNN accelerator designs in the mobile vision domain. State-of-the-art optimization techniques such as model compression and run-time buffer allocation are included by extending the SCALE-Sim systolic-array simulator. We emphasize memory technologies that have practical high-volume manufacturing and system integration capabilities (SRAM, eDRAM, MRAM, RRAM). Energy benefits (up to 4.68× with MRAM) and area reduction (up to 33% savings with 3D VRRAM) can be attained over Pareto-optimal baseline designs. Even with today's performance gap between maturing NVM and state-of-the-art SRAM, efficient allocation of chip area resources balancing between dense NVM vs. low-power SRAM provides overall energy-efficiency benefits to mobile DNN accelerators.

6 ACKNOWLEDGEMENTS

This work was supported in part by ASCENT, one of six centres in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, and in part by Stanford NMTRI and Stanford SystemX Alliance. We acknowledge the fruitful discussions with Greg Yeric, Brian Cline, Matthew Mattina, YK Chong from Arm, and Priyanka Raina from Stanford University.

REFERENCES

[1] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *ISCA*, 2017.

[2] "Arm Machine Learning Processor." [Online]. Available: https://developer.arm.com/products/processors/machine-learning/arm-ml-processor

[3] "NVIDIA Deep Learning Accelerator (NVDLA)." [Online]. Available: http://nvdla.org/primer.html

[4] Y.-H. Chen *et al.*, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," in *ISCA*, 2016.

[5] H.-S. P. Wong *et al.*, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, p. 191, 2015.

[6] T. Chen *et al.*, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," in *ASPLOS*, 2014.

[7] P. N. Whatmough *et al.*, "DNN engine: A 28-nm timing-error tolerant sparse deep neural network processor for IoT applications," *JSSC*, 2018.

[8] S. Han *et al.*, "EIE: efficient inference engine on compressed deep neural network," in *ISCA*, 2016.

[9] M. Clinton *et al.*, "A 5GHz 7nm L1 cache memory compiler for high-speed computing and mobile applications," in *ISSCC*, 2018.

[10] Q. Dong *et al.*, "A 1Mb 28nm STT-MRAM with 2.8 ns read access time at 1.2 V VDD using single-cap offset-cancelled sense amplifier and in-situ self-write-termination," in *ISSCC*, 2018.

[11] M. Mendicino, "eMRAM: Winning the IoT and AI Applications." [Online]. Available: https://mramdeveloperday.com/English/Conference/Keynotes.html

[12] H.-S. P. Wong *et al.*, "Stanford memory trends," *tech. report*, 2016.

[13] F.-K. Hsueh *et al.*, "First fully functionalized monolithic 3D+ IoT chip with 0.5 V light-electricity power management, 6.8 GHz wireless-communication VCO, and 4-layer vertical ReRAM," in *IEDM*, 2016.

[14] C. Lin *et al.*, "High performance 14nm SOI FinFET CMOS technology with 0.0174 μm^2 embedded DRAM and 15 levels of Cu metallization," in *IEDM*, 2014.

[15] K. Huang *et al.*, "A high-performance, high-density 28nm eDRAM technology with high-K/metal-gate," in *IEDM*, 2011.

[16] R. Giterman *et al.*, "An 800-MHz mixed-V_T 4T IFGC embedded DRAM in 28-nm CMOS bulk process for approximate storage applications," *JSSC*, 2018.

[17] P. Chi *et al.*, "Prime: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *ISCA*, 2016.

[18] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *ISCA*, 2016.

[19] C. Zhang *et al.*, "Optimizing FPGA-based accelerator design for deep convolutional neural networks," in *FPGA*, 2015.

[20] N. Suda *et al.*, "Throughput-optimized OpenCL-based FPGA accelerator for large-scale convolutional neural networks," in *FPGA*, 2016.

[21] A. Samajdar *et al.*, "SCALE-Sim: Systolic CNN accelerator," *arXiv preprint arXiv:1811.02883*, 2018.

[22] I. Bratt, "Arm's First-Generation Machine Learning Processor," *Hot Chips*, 2018. [Online]. Available: https://www.hotchips.org/hc30/2conf/2.07_ARM_ML_Processor_HC30_ARM_2018_08_17.pdf

[23] C. Park *et al.*, "Systematic optimization of 1 Gbit perpendicular magnetic tunnel junction arrays for 28 nm embedded STT-MRAM and beyond," in *IEDM*, 2015.

[24] M. Gao *et al.*, "Tetris: Scalable and efficient neural network acceleration with 3d memory," in *ASPLOS*, 2017.

[25] T. F. Wu *et al.*, "43pJ/cycle non-volatile microcontroller with 4.7 μs shutdown/wake-up integrating 2.3-bit/cell resistive RAM and resilience techniques," in *ISSCC*, 2019.