SAPIENS: A 64-Kbit RRAM-Based Non-Volatile Associative Memory for One-Shot Learning and Inference at the Edge

Haitong Li, *Student Member, IEEE,* Wei-Chen Chen, Akash Levy, Ching-Hua Wang, Hongjie Wang, Po-Han Chen, Weier Wan, Win-San Khwa, Harry Chuang, Y.-D. Chih, Meng-Fan Chang, *Fellow, IEEE,* H.-S. Philip Wong, *Fellow, IEEE,* and Priyanka Raina

***Abstract*—Learning from a few examples (one/few-shot learning) on the fly is a key challenge for on-device machine intelligence. We present the first chip-level demonstration of one-shot learning with SAPIENS, a resistive RAM (RRAM) based non-volatile associative memory (AM) chip that serves as the backend for memory-augmented neural networks. The 64-kbit fully-integrated RRAM-CMOS AM chip performs long-term feature embedding and retrieval, demonstrated on a 32-way one-shot learning task on the Omniglot dataset. Using only one example per class for 32 unseen classes during on-chip learning, SAPIENS achieves 79% measured inference accuracy on Omniglot, comparable to edge software model accuracy using 5-level quantization (82%). It achieves an energy-efficiency of 118 GOPS/W at 200 MHz for in-memory L1 distance computation and prediction. Multi-bank measurements on the same chip show that increasing the capacity from 3 banks (24 kb) to 8 banks (64 kb) improves the chip accuracy from 73.5% to 79%, while minimizing the accuracy excursion due to bank-to-bank variability.**

***Index Terms*—One-shot learning, memory-augmented neural networks, associative memory, resistive random access memory (RRAM, ReRAM).**

I. Introduction

On-device machine intelligence requires continuous real-time learning of never-before-seen data/events [1]. Memory-augmented neural networks (MANNs) aim to address this demand by utilizing an explicit associative memory to augment the feature learning capabilities of neural networks (NNs) with scarce data [2]–[4]. A MANN consists of a frontend neural feature extractor (such as a convolutional neural network) and a backend associative memory (AM). In the backend AM, real-time learning occurs by embedding (storing) new features into the memory, and inference occurs through similarity-based retrieval of features. In this work, we demonstrate a chip for accelerating the MANN backend, using an RRAM-based non-volatile associative memory that naturally enables long-term feature embedding and efficient feature retrieval. In our MANN system, the frontend NN is initially trained offline (meta-training), after which its weights are fixed and do not need to be updated. During k-shot learning, novel features (from unseen classes that are not included in NN meta-training) are mapped into the associative memory, using only k examples per class (where $k \geq 1$ is a small number). During inference, the query samples initiate similarity-based retrieval from the associative memory, which makes predictions based on similarity. At a device-level, prior work has explored the use of novel devices such as ferroelectric memories for emerging MANN workloads with a combination of device-level experiments and simulations [5]. At an architecture-level, a crossbar-based accelerator for a variant of MANNs [6] was previously proposed and analyzed through system simulations [7]. However, chip-level implementation and characterization are missing, and thus, the impact of device non-idealities and circuit design limitations on the performance of MANNs remains to be elucidated. Demonstrating the integration of a new device technology at the chip-level, with real-time characterization of the target workloads, as illustrated in this work, is of paramount importance in accelerating technology development through accelerator design and optimization.

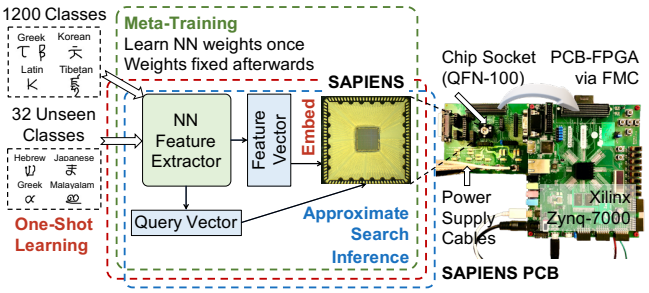

Fig. 1. Overview of memory-augmented neural network (MANN) workflow: meta-training phase, (one-shot) learning phase, and inference phase. One-shot learning and inference are demonstrated on SAPIENS in this work.

Here, we present SAPIENS (Stanford Associative memory for Programmable, Integrated Edge iNtelligence via life-long learning and Search), a 64-kbit fully-integrated RRAM-CMOS associative memory (AM) chip as the backend of MANNs [8]. Leveraging the monolithic integration of RRAM on top of CMOS [9], the single-chip AM core occupies an area of 0.2 mm^2 in TSMC 40 nm RRAM technology [10], [11]. SAPIENS supports the key one-shot learning and inference operations

Manuscript received July 2021. This work is supported in part by SRC JUMP ASCENT Center, and in part by Stanford SystemX Alliance. A. Levy is supported by NSF GRFP.

Haitong Li, Wei-Chen Chen, Akash Levy, Ching-Hua Wang, Hongjie Wang, Po-Han Chen, Weier Wan, H.-S. Philip Wong, Priyanka Raina are with Department of Electrical Engineering, and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305, USA.

Win-San Khwa, Harry Chuang, Y.-D. Chih, and Meng-Fan Chang are with Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu, Taiwan.

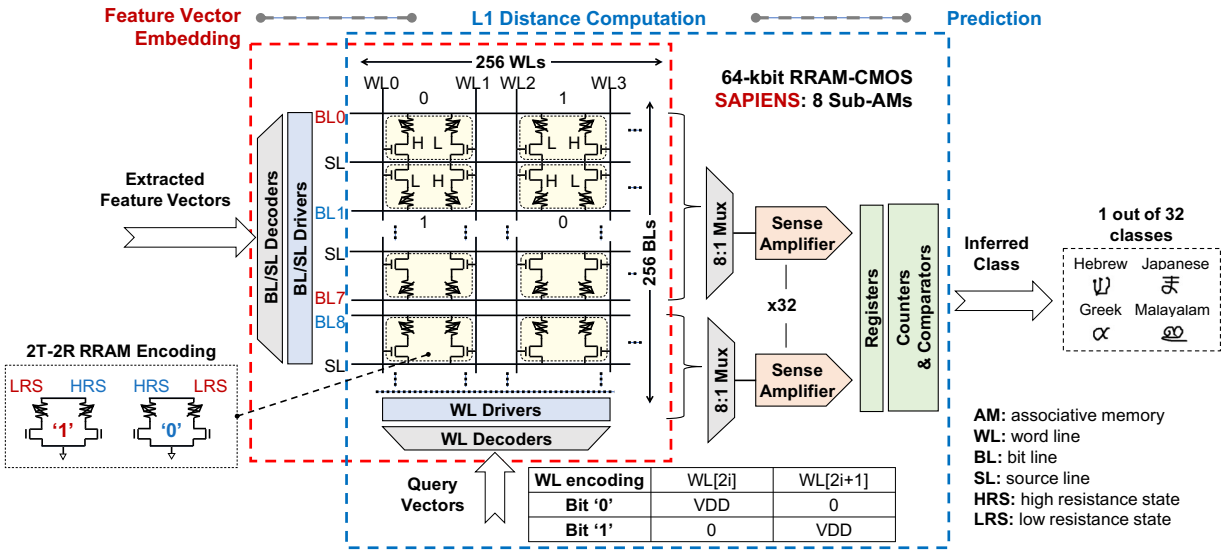

Fig. 2. SAPIENS architecture, data encoding, and dataflow. Feature vector embedding and L1 distance computation are the main operations during one-shot learning and inference phases.

needed for a MANN model via two modes: (1) feature vector embedding in AM (learning); and (2) L1 distance computation between query set (test images) and support set (embedded novel features) for similarity-based prediction (inference).

The rest of the paper is organized as follows. Section II presents a brief introduction to MANN models and the role of SAPIENS. Section III discusses the chip architecture and operation schemes. Section IV presents experimental measurements of SAPIENS on one-shot learning workloads. Finally, the conclusion in Section V provides forward-looking research directions based on insights from this work.

II. MANN for One-Shot Learning

Memory-augmented neural networks such as [3] enable learning continuously from one or few examples on the fly, also known as one-shot or few-shot learning. Figure 1 shows the major components and dataflow of a MANN model, where a frontend neural network with feature learning capabilities feeds into a backend associative memory for similarity measurement. A MANN model, as a whole, is heterogeneous because of the inherently different structures of the frontend neural network and the backend associative memory. A typical NN accelerator architecture optimized for multiply-accumulate (MAC) kernels does not address the data movement and the operations associated with an external associative memory efficiently. Memory-centric hardware implementations are needed for MANN models to enable efficient learning and inference at the edge [12]. As a case study, we use the Omniglot dataset [13] for our one-shot and few-shot learning experiments. The dataset contains over 1,600 classes of characters from worldwide alphabets, with only 20 examples per class drawn by different people. The dataset represents a challenging task even for humans as it requires fast learning and recognition using a few examples [13].

A MANN workflow consists of three phases (Figure 1). A subset of 1,200 classes is used for the initial meta-training phase [3]. This phase is performed offline once to train the NN feature extractor. The weight parameters of the feature extractor are fixed after offline training and are not updated in the following phases. Next, for the one-shot learning phase, a number of unseen classes, that were not presented to the MANN before, are used as the support set and fed into the previously-trained feature extractor. Only one image example is used per class for one-shot learning. (In general, k image examples are used per class for k-shot learning.) The extracted feature vectors coming out of the feature extractor are embedded into the backend associative memory. In this work, the embedding is performed by programming the feature vectors into the RRAM array of the SAPIENS chip. The embedded features stay within the associative memory without consuming standby power, due to the non-volatility of RRAM. For the inference phase, the frontend feature extractor takes new input samples (e.g., Omniglot images) and produces query vectors that represent the input samples. These query vectors have the same format as the feature vectors in the learning phase. They are sent to the associative memory, where similarity measurement between the query vectors and the embedded feature vectors is performed. This approximate search process identifies the closest class the test image belongs to, yielding the inference result. This operation of the AM is distinct from most conventional uses of content-addressable memories (e.g. in IP routers) where exact matches are required [14], [15].

III. SAPIENS Architecture and Operations

As shown in Figure 2, each unit memory cell has a two-transistor, two-resistor (2T-2R) structure, leveraging the monolithic integration of RRAM directly on top of CMOS. Following the complementary encoding scheme discussed in [16], each 2T-2R cell encodes '0' with high resistance state (HRS)-low resistance state (LRS), or '1' with LRS-HRS. A pair of wordlines (WLs) that control the select transistors in the 2T-2R cell encodes '0' with VDD-GND biases, or '1' with

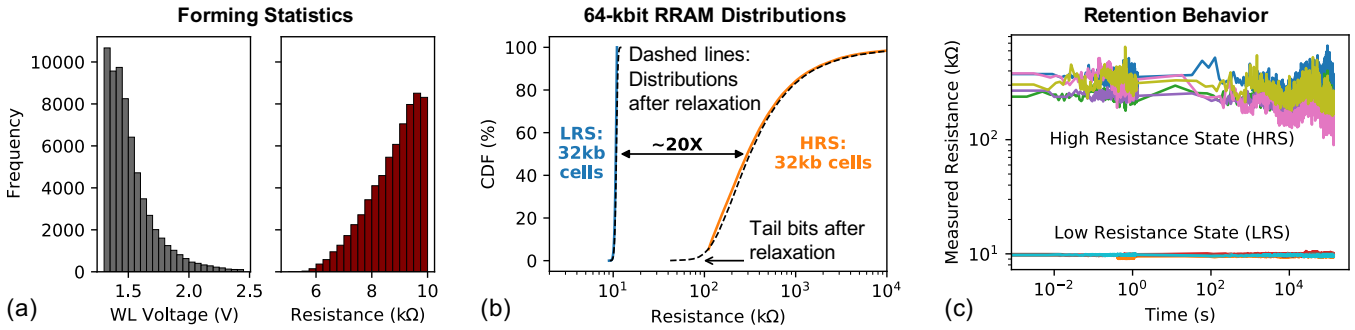

Fig. 3. (a) Chip-level forming statistics characterized by the WL voltage distribution and the RRAM resistance distribution after the forming operation. (b) Measured high resistance state (HRS) and low resistance state (LRS) distributions from the 64-kbit RRAMs, after embedding a set of features from Omniglot dataset for a one-shot learning workload. HRS tail bits resulting from relaxation are captured as well for the inference phase. (c) Retention measurements for HRS and LRS. Each colored line shows a different cell's behavior.

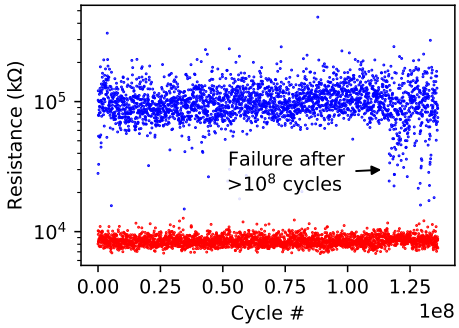

Fig. 4. RRAM two-level endurance characterization. Each red/blue point represents an LRS/HRS state measured at the corresponding cycle number on the x-axis (one read measurement was recorded every 40,000 cycles). A SET/RESET pulse width of 1 µs was used. Early signs of cell failure are observed after $> 10^8$ cycles.

GND-VDD biases. Addressing, programming, and reading are enabled with the decoder and driver peripheral circuits, whereas additional analog/digital circuitry performs sensing, accumulation, and comparison functions that are essential to the execution of different MANN phases. I/O transistors are used within the driver circuitry while the rest of the decoding, sensing, and post-processing peripherals use core transistors. The AM core is partitioned into 8 sub-AMs, where every 8 BLs are grouped into a sub-AM bank (i.e., 32 rows per bank). The chip supports feature vector embedding and L1 distance computation as two major operations.

A. Programming and Characterization of RRAM

A standard RRAM forming operation is needed before feature embedding: the WL voltage is ramped from 1.3 V with an increment of 0.05 V, and the bitline (BL) voltage is kept at 3.3 V with all source lines (SLs) grounded. A pulse width of 1 ms is used for all forming pulses. The forming operation only needs to be performed once before the RRAM cells are used. Chip-level statistical distributions for the forming operation are shown in Figure 3(a).

In the learning phase, the 256 × 256 RRAM array and the decoder and driver peripherals are activated. The extracted feature vectors are sent into the AM. Each 128-bit feature vector gets programmed into one entire row of the RRAM array (along each BL). The core is partitioned into 8 sub-AMs that can store the same or different support set features. For each sub-AM, a write-verify scheme [17] is used for feature embedding. For SET operations, BL voltage is 3.3 V while SLs are grounded, and for RESET operations, SLs are biased at 3.5 V while BLs are grounded. SET pulses have 1 µs pulse width, and RESET pulses have 100 µs pulse width. After full-chip programming, another two iterations of verification are applied, where cell re-programming may or may not occur based on how much relaxation it experiences. During verification, read voltage on the BLs is 0.2 V while the WL voltage is 2.5 V.

Figure 3(b) shows the measured resistance distributions from the 64-kbit RRAM, after embedding a set of features from the Omniglot dataset for a one-shot learning workload. Resistance relaxation effect is sometimes observed after programming, and may be correlated with the underlying oxygen vacancies re-generation and re-combination processes [18]. The relaxation behavior after array programming is characterized on SAPIENS. Less than 5% of the cells that are programmed to the HRS drift below 100 kΩ, as shown by the tail bits in Figure 3(b). As shown in Figure 3(c), retention measurements at room temperature indicate stable non-volatile behavior necessary for inference operations. Figure 4 further shows the endurance characterization. The relatively large cell resistance state variations from cycle to cycle are due to fast programming pulses (1 µs) without write verification to speed up the endurance measurement. The retention and endurance characterizations show robust device behaviors necessary for the learning and inference experiments that will be discussed in Section IV.

B. Sensing Operation for Inference

As the associative memory core is non-volatile, the learned features remain in memory for the subsequent inference operations. For inference, test samples are presented to the frontend feature extractor from which the query vectors are generated. The 128-bit query vectors are sent into the AM core through the WL circuitry. In the sensing circuitry, there are 32 sense amplifiers (SAs) in total, and each SA is shared by 8 BLs. This means that there are 32 rows being activated while the L1 distance computation is performed in parallel between the query vector and 32 feature vectors. Within each clock cycle, 4 WLs are activated in parallel, where each pair of WL pulses

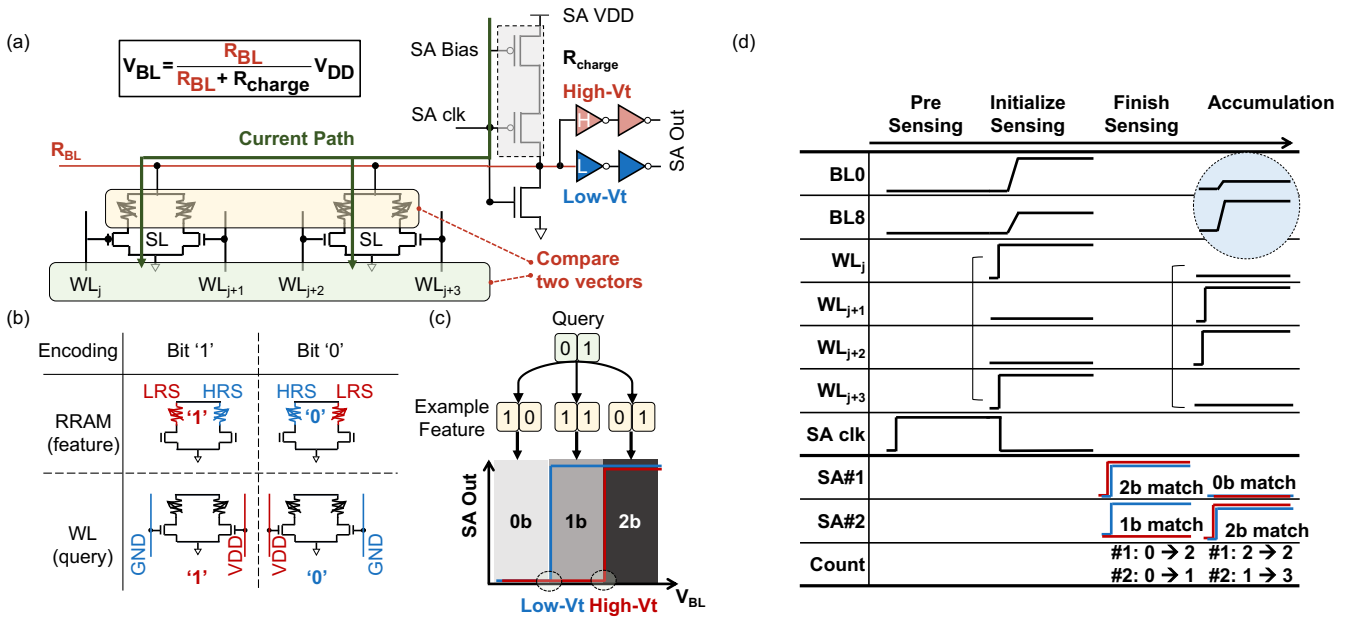

Fig. 5. (a) Illustration of sensing on one BL. (b) Data encoding for the features stored in RRAMs (resistance representation) and the query data sent in through WLs (voltage representation). (c) Three examples for query-feature vector comparison and sensing results based on the corresponding BL voltages. (d) An example of input biases and SA outputs throughout the sensing operation for L1 distance computation.

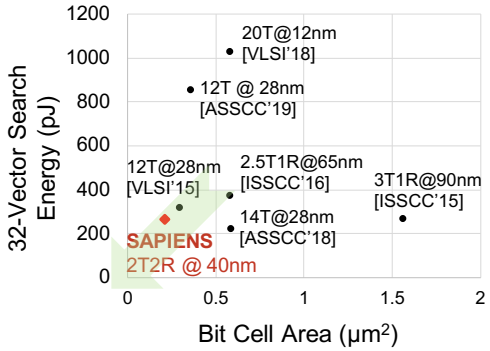

Fig. 6. Energy and area efficiency benchmarking for an approximate search workload on memory chips with search functionalities.

along a 2T-2R unit cell encodes 1 bit in the query vector. Activating WLs that control the gates of select transistors creates a current path from the SA charger, through the BLs and the 2T-2R cells, to the SL ground. As a result, the BL potentials are set by voltage division between the activated 2T-2R cells and the SA charger, as shown in Figure 5(a). The degree of match between the voltage-encoded WL query vector and the resistance-encoded feature vector (Figure 5(b)) determines the relative pull-down and pull-up strengths on the BL. As a result, different WL inputs (query data), when compared with previously stored features in RRAM cells, lead to different SA input voltage levels. Figure 5(c) shows three cases of query and feature vector comparison, each of which falls under a different BL voltage window based on the voltage division behavior illustrated in Figure 5(a). Finally, as the upper and lower SA buffers set two thresholds, the BL voltage can be sensed to yield the SA output, which is configurable for 1-bit per cycle (0.4 GOPS) or 2-bit per cycle (0.8 GOPS) for sensing. An example of sensing operation from initialization to accumulation is provided in Figure 5(d). To initialize sensing, SA clk is driven low, while the partial query vectors are sent through WL0 to WL3 as voltage pulses. Towards the end of each sensing cycle, the SA outputs for the corresponding BLs (i.e., classes) are captured by the registers and then accumulated with counters. A larger number of matching bits indicates a smaller L1 distance between the query vector and the feature vector. Such operation is repeated for the remaining sets of WLs encoding the rest of query vector. Finally, a comparator tree takes the accumulator outputs from 32 classes and identifies the maximum degree of match, corresponding to the smallest L1 distance between input query sample and stored class.

The width of the partial WL vector (i.e., activating 4 WLs per cycle at 200 MHz) is a design choice based on (1) the RRAM ON/OFF ratio distribution at the array-level, and (2) the design complexity and energy/area penalties of the sensing circuitry at a high frequency when activating many WLs in parallel. In our BL-parallel, WL-sequential operation scheme, the L1 distance computation between a full query vector and 32 feature vectors takes 640 ns. This latency can be hidden in the end-to-end pipeline, where the frontend image capture and CNN feature processing take 8.33 ms even on a 120-fps CMOS image sensor (CIS) integrated with a digital signal processing (DSP) core at 262.5 MHz [19] .

To benchmark the sensing circuitry with respect to previous memory chips that support in-memory search, we take the measured data reported in previous work, and extrapolate to the same workload of approximate search among 32 classes of 128-bit vectors. As shown in Figure 6, SAPIENS enables energy- and area-efficient approximate search, while other chips with exact search capability would not handle this type of workload efficiently [20]–[24]. As opposed to designs with complex cell structures, our compact cell array at 40 nm node

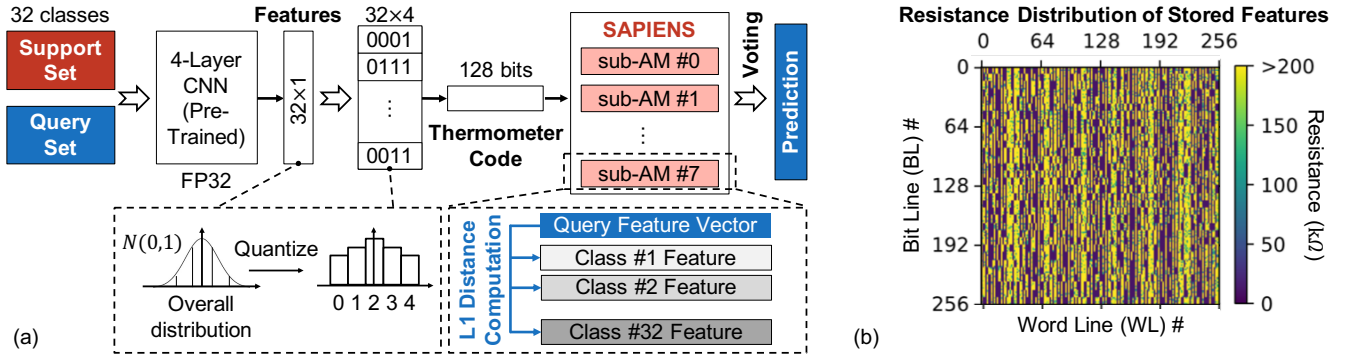

Fig. 7. (a) Flow of on-chip one-shot learning and inference experiments. (b) Measured 64-kbit data pattern (post-relaxation resistance distributions) after 32-way, 1-shot learning on chip. 32 novel features are broadcast to 8 sub-AMs and programmed as vectors along the bit lines (BLs).

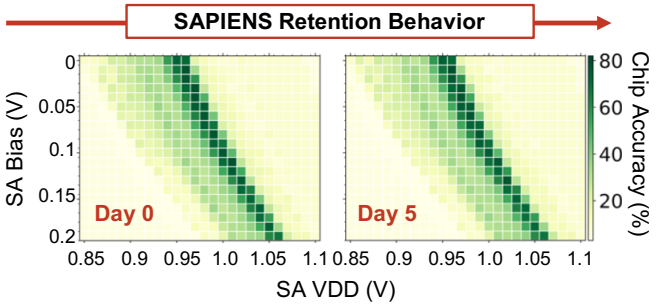

Fig. 8. Optimization of 32-class Omniglot inference accuracy on SAPIENS by modulating sense amplifier (SA) biasing voltages. 79% hardware accuracy is achieved and is maintained during the chip retention test.

reduces total wire length, and as a result, the dynamic energy is reduced while activating multiple rows and columns.

IV. On-Chip One-Shot Learning with SAPIENS

In this section, we discuss the data flow and the measured results. In our demonstration, we conduct 32-way, 1-shot learning experiments on the Omniglot dataset. For the frontend feature extractor in the MANN model, we use a 4-layer CNN that works with both support set and query set having 32 classes [3]. As shown in Figure 7(a), the extracted features are 32×1 vectors with 32-bit floating point values. On all the values in a feature vector, a linear quantization is applied to obtain 5 discrete levels. The quantized values are encoded as 4-bit thermometer codes. A larger quantized value indicates more '1's in the code, which resembles a thermometer reading. The 4-bit codes of the 32 elements in the feature vector are concatenated to form a single 128-bit vector which is sent into the AM. The AM computes the L1 distance as the number of matching bits between this 128-bit query vector and the vectors stored in the AM. For one sub-AM, L1 distance computation is performed in parallel between the query vector and all 32 classes. SAPIENS makes use of an ensemble of multiple sub-AM banks (up to 8) to embed the the same set of features, and the final prediction is made via voting among the sub-AM banks. The 64-kbit data pattern shown in Figure 7(b) is measured from SAPIENS after the 32-way, 1-shot learning on the chip. The 32 features are broadcast to 8 sub-AMs and remain there after power-off due to non-volatility of the RRAM. The resistance distributions are measured post resistance relaxation. The data shown in Figure 7(b) also reflects the cycle-to-cycle and device-to-device programming variations given the identical set of feature vectors among the 8 sub-AM banks. The tail bits of the high resistance state can lead to narrower sensing windows during the inference operations due to lowered ON/OFF ratio in the worst case.

Using the MANN flow described earlier, we then conduct inference operations on SAPIENS and collect the hardware inference results from measurements using the Omniglot test set (32 classes with 320 images per class). The chip consumes 0.21 mW power at 10 MHz frequency and 3.39 mW at 200 MHz frequency. For L1 distance computation and class prediction, an energy efficiency of 118 GOPS/W is measured. Each operation is defined as the L1 distance computation between the 128-bit query vector and all 32 128-bit feature vectors, and the final prediction is factored into the overall energy efficiency as well. The precision for vector elements is 4 bits using thermometer encoding as described earlier. Static power of RRAM cells in the activated sub-AM bank contributes about 16% of the total power during inference, which is determined by RRAM resistances. RRAM characteristics would also impact design choices such as encoding scheme, sensing circuitry, and dataflow, which largely affect the dynamic and other leakage components of the total power.

We further analyze and calibrate the sense amplifier (SA) voltage biases for chip inference operation. The VDD supply (SA VDD) and the PMOS charger bias (SA Bias) as discussed in Figure 5(a) modulates the relative strength or the effective resistance of SA compared to the 2T-2R cells on a BL. As a result, degradation in sensing accuracy that results from device variations, resistance relaxation, and circuit non-idealities such as IR drop can be compensated by adjusting the voltage biases. Figure 8 shows the correlation between SA voltage modulation and the overall inference accuracy, with 79% inference accuracy obtained. The chip retention behavior at room temperature is characterized, by powering off SAPIENS for five days and performing the same set of experiments afterwards. The non-volatile nature of SAPIENS plays an important role in ensuring life-long learning and inference on chip. For the 32-class inference tests on the Omniglot dataset, the accuracy reduces from 86% to 82% after quantizing the floating-point model to a 4-bit representation for edge systems.

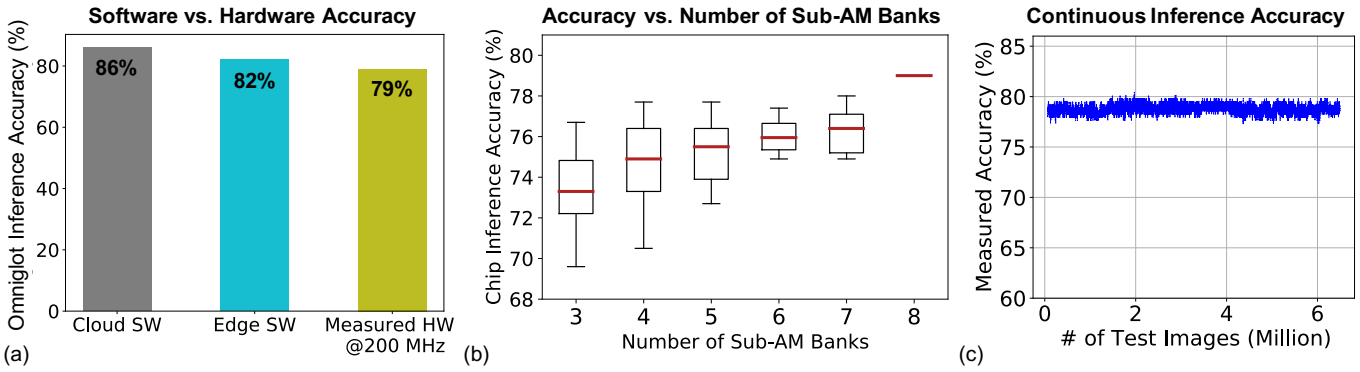

Fig. 9. (a) Omniglot inference accuracy using a cloud model without quantization, edge model with quantization, and SAPIENS at 200 MHz. (b) Measured chip inference accuracy with different number of sub-AM banks activated for voting. (c) Continuous inference with SAPIENS using over 6 million images.

By mapping the edge software model onto SAPIENS with SA biasing optimizations, we obtain a 79% accuracy at 200 MHz frequency from hardware measurements (Figure 9(a)).

With the multi-bank architecture, we characterize and analyze the trade-off between hardware resources and the quality of results (QoR) of the MANN. Specifically, the measured chip inference accuracy with various number of sub-AM banks is shown in Figure 9(b). The statistical behavior indicated by the error bars reflects different possible combinations of sub-AM banks used, given a certain number of sub-AM banks. This shows how device- and circuit-level variations translate into final model predictions. A larger number of banks leads to less variability in the prediction results as well as higher chip inference accuracy, trading hardware resources for improved QoR. When using the full chip capacity, i.e., all the 8 sub-AM banks, the highest chip accuracy is achieved, using the globally optimized SA bias condition from Figure 8. By enabling fine-grained SA bias modulation in future work, accuracy with fewer number of banks may be further improved.

From a chip reliability perspective, we also probe the robustness of SAPIENS via continuous inference tests. More than 6 million test images are sent into the chip while monitoring the inference accuracy from measurements. Figure 9(c) shows that the RRAM-based SAPIENS chip is robust against read disturb and conductance drift that may be encountered over long-term operation.

For future system-technology co-optimization (STCO) using RRAM and other non-volatile memories (NVMs) in one-shot learning applications, key device characteristics such as HRS-to-LRS ratio and resistance uniformity play a critical role, similar to the conventional digital memory and data storage use cases. However, spatial variations and the interaction with sensing peripherals are aspects unique to SAPIENS-like associative memory hardware. On one hand, as SA mismatches and RRAM variations co-exist, tightening the HRS distribution above 100 kΩ would leave a larger voltage margin for SA bias modulation to reduce the accuracy loss from hardware non-idealities. On the other hand, reducing RRAM variations globally as in the typical cases of on-chip memories and data storage may not suffice, if such optimizations are done without awareness of spatial information. From the experiments of capacity-accuracy trade-off, we find that the spatial variations across sub-AM banks play an important role. After the one-shot learning phase, certain 1 or 2 sub-AM banks may exhibit relatively higher device-to-device (D2D) variations tied to lower tail bits of HRS distributions after relaxation. As each row in a sub-AM bank represents a unique class feature and each sub-AM bank equally contributes to the inference result, such spatial variability becomes a source of hardware accuracy loss as well as accuracy excursion that are not captured in typical memory designs. This presents new challenges and requirements for spatial-aware hardware design and optimization with RRAM or other NVMs.

V. Conclusion

SAPIENS highlights the importance of new technology integration and characterization at a chip-level, for emerging AI workloads that may benefit from new hardware architectures and new device technologies. Using the experimental characterizations of one-shot learning on chip, we explore various hardware-software implications, from device and circuit non-idealities to capacity-accuracy trade-offs. The experimental demonstrations and analysis attained from this work point to possible realization of the more ambitious goal of lifelong learning applications in the future, where new hardware technologies and architectures can offer energy-efficient, privacy-centered solutions. From a hardware perspective, this requires exploring STCO techniques to address non-idealities and fully utilize the benefits of non-volatile memories with tight logic integration in 3D and analog programmability. From an application perspective, this in return leads to new opportunities around incorporating the unique hardware characteristics (e.g., SAPIENS measurements and non-idealities) into the design of lifelong learning models, as well as domain adaptation in a dynamic environment.

Acknowledgment

The authors would like to thank TSMC for the chip fabrication and support of the project. We also thank Guenole Lallement, Robert Radway, Kartik Prabhu, Christopher Torng, Prof. Subhasish Mitra, and Prof. Boris Murmann of Stanford University for technical assistance and discussions.

REFERENCES

[1] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[2] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[3] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[4] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," *arXiv preprint arXiv:1605.06065*, 2016.

[5] K. Ni, X. Yin, A. F. Laguna, S. Joshi, S. Duenkel, M. Trentzsch, J. Müller, S. Beyer, M. Niemier, X. S. Hu *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019.

[6] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[7] A. Ranjan, S. Jain, J. R. Stevens, D. Das, B. Kaul, and A. Raghunathan, "X-MANN: A crossbar based architecture for memory augmented neural networks," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[8] H. Li, W.-C. Chen, A. Levy, C.-H. Wang, H. Wang, P.-H. Chen, W. Wan, H.-S. P. Wong, and P. Raina, "One-shot learning with memory-augmented neural networks using a 64-kbit, 118 GOPS/W RRAM-based non-volatile associative memory," in *2021 IEEE Symposium on VLSI Technology*. IEEE, 2021, pp. 1–2.

[9] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, pp. 191–194, 2015.

[10] C.-F. Lee, H.-J. Lin, C.-W. Lien, Y.-D. Chih, and J. Chang, "A 1.4 Mb 40-nm embedded ReRAM macro with 0.07um2 bit cell, 2.7 mA/100MHz low-power read and hybrid write verify for high endurance application," in *2017 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. IEEE, 2017, pp. 9–12.

[11] C.-C. Chou, Z.-J. Lin, P.-L. Tseng, C.-F. Li, C.-Y. Chang, W.-C. Chen, Y.-D. Chih, and T.-Y. J. Chang, "An N40 256K$\times$ 44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2018, pp. 478–480.

[12] J. R. Stevens, A. Ranjan, D. Das, B. Kaul, and A. Raghunathan, "Manna: An accelerator for memory-augmented neural networks," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 794–806.

[13] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[14] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: A tutorial and survey," *IEEE journal of solid-state circuits*, vol. 41, no. 3, pp. 712–727, 2006.

[15] H. J. Chao, "Next generation routers," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1518–1558, 2002.

[16] J. Li, R. Montoye, M. Ishii, K. Stawiasz, T. Nishida, K. Maloney, G. Ditlow, S. Lewis, T. Maffitt, R. Jordan *et al.*, "1Mb 0.41 μm 2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing," in *2013 Symposium on VLSI Technology*. IEEE, 2013, pp. C104–C105.

[17] B. Q. Le, A. Levy, T. F. Wu, R. M. Radway, E. R. Hsieh, X. Zheng, M. Nelson, P. Raina, H.-S. P. Wong, S. Wong, and S. Mitra, "RADAR: A Fast and Energy-Efficient Programming Technique for Multiple Bits-Per-Cell RRAM Arrays," *IEEE Transactions on Electron Devices*, pp. 1–7, 2021.

[18] C. Wang, H. Wu, B. Gao, L. Dai, N. Deng, D. Sekar, Z. Lu, M. Kellam, G. Bronner, and H. Qian, "Relaxation effect in RRAM arrays: demonstration and characteristics," *IEEE Electron Device Letters*, vol. 37, no. 2, pp. 182–185, 2015.

[19] R. Eki, S. Yamada, H. Ozawa, H. Kai, K. Okuike, H. Gowtham, H. Nakanishi, E. Almog, Y. Livne, G. Yuval *et al.*, "A 1/2.3 inch 12.3 Mpixel with On-Chip 4.97 TOPS/W CNN Processor Back-Illuminated Stacked CMOS Image Sensor," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64. IEEE, 2021, pp. 154–156.

[20] M. Yabuuchi, M. Morimoto, K. Nii, and S. Tanaka, "12-nm Fin-FET 3.0 G-search/s 80-bit$\times$ 128-entry Dual-port Ternary CAM," in *2018 IEEE Symposium on VLSI Circuits*. IEEE, 2018, pp. 19–20.

[21] S. Jeloka, N. Akesh, D. Sylvester, and D. Blaauw, "A configurable TCAM/BCAM/SRAM using 28nm push-rule 6T bit cell," in *2015 Symposium on VLSI Circuits (VLSI Circuits)*. IEEE, 2015, p. C272.

[22] C.-C. Lin, J.-Y. Hung, W.-Z. Lin, C.-P. Lo, Y.-N. Chiang, H.-J. Tsai, G.-H. Yang, Y.-C. King, C. J. Lin, T.-F. Chen *et al.*, "A 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14$\times$ improvement in wordlength-energyefficiency-density product using 2.5T1R cell," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2016, pp. 136–137.

[23] M.-F. Chang, C.-C. Lin, A. Lee, C.-C. Kuo, G.-H. Yang, H.-J. Tsai, T.-F. Chen, S.-S. Sheu, P.-L. Tseng, H.-Y. Lee *et al.*, "17.5 A 3T1R nonvolatile TCAM using MLC ReRAM with sub-1ns search time," in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*. IEEE, 2015, pp. 1–3.

[24] C.-X. Xue, W.-C. Zhao, T.-H. Yang, Y.-J. Chen, H. Yamauchi, and M.-F. Chang, "A 28mn 320Kb TCAM Macro with Sub-0.8 ns Search Time and 3.5+ x Improvement in Delay-Area-Energy Product using Split-Controlled Single-Load 14T Cell," in *2018 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. IEEE, 2018, pp. 127–128.