Neuro-inspired computing with emerging memories: where device physics meets learning algorithms

Haitong Li, Priyanka Raina, and H.-S. Philip Wong
Department of Electrical Engineering, Stanford University
Stanford, CA 94305, USA

ABSTRACT

Modern cognitive computing workloads require computing systems tailored to the applications, where the underlying hardware fabrics should naturally match the characteristics of learning algorithms and compute kernels. With emerging memory technologies (e.g., resistive RAM (RRAM), magnetic RAM (MRAM)), we design neuro-inspired computing systems that exploit technology characteristics such as rich device physics, circuit architecture, and integration capabilities with CMOS and beyond-CMOS technologies. Our methodology is built upon a combination of experimental characterization, cross-stack modeling, and system integration, illustrated by case studies for neural networks and high-dimensional (HD) computing. Finally, we discuss the prospects of heterogeneous learning machines that emphasize the integration of compute kernels and learning algorithms, as well as the integration of emerging nanotechnologies.

Keywords: Non-volatile memories (NVM), resistive RAM (RRAM), neuro-inspired computing, device-algorithm co-design, domain-specific architectures, machine learning, 3D integration

1. INTRODUCTION

The explosion of data and the growing computational complexity in vision, speech, control, health and other cognitive applications have far exceeded the storage and processing capabilities of today's computing solutions. The application characteristics and requirements are evolving towards more sophisticated learning, personalization and domain adaptation, and the computing and learning activities are spreading from cloud to edge. Isolated improvements in technologies, architectures, or computational models have diminishing returns by themselves and is insufficient for addressing the grand challenge of developing energy-efficient learning hardware for wide adoption.

Neuro-inspired computing, where the fundamental fabrics of memory and computation function closely and dynamically with deployed learning algorithms, offers unique opportunities towards domain-specific architectures with desired functionalities, versatility, and efficiency, where both ends of the spectrum are exposed to cross-stack design and optimization. Emerging non-volatile memories, such as phase-change memory (PCM), resistive RAM (RRAM), magnetic RAM (MRAM), and ferroelectric RAM (FeRAM), are becoming key technology enablers [1]. Device-level properties including analog programmability in resistive memories and nonlinear dynamics in spintronics, combined with circuit architectures (e.g., crossbar array), have been explored for hardware realizations of neural networks [2]-[11].

In this paper, we emphasize co-designing neuro-inspired computing systems with the technology characteristics, by presenting several experimental case studies using RRAM. Beyond data storage, RRAM used in those case studies serve as "nanokernels" for key operations in target applications. We first review the modeling of RRAM, which serves as the basis for linking experimental characterizations with circuit and system analysis. We then discuss the interaction between intrinsic stochasticity of RRAM and learning behaviors of neural networks. Next, we introduce high-dimensional (HD) computing as a robust neuro-inspired model mimicking activities and associativity in high-dimensional neural circuits of the human brain [12], followed by a description of a native realization of essential HD kernels within 3D vertical RRAM. Finally, we probe into the future prospects of learning systems that capitalize on the integration of emerging technologies, driven by the integration of computational models and learning algorithms to meet the diverse needs of applications, such as continuous and life-long learning of computing systems.

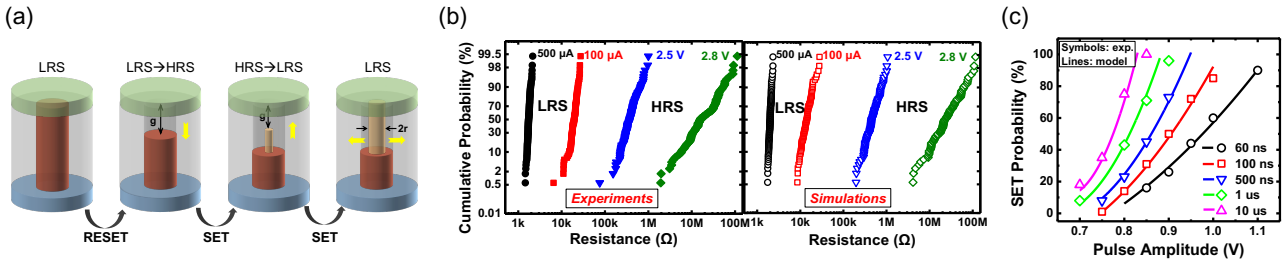

Figure 1. (a) Modeling conductive filament evolutions in metal-oxide RRAM. (b) Measured and modeled statistical distributions of low resistance states (LRS) and high resistance states (HRS) under different SET and RESET programming conditions. (c) Measured and modeled stochastic SET behaviors. Overall SET probability is a function of pulsing conditions.

2. MODELING AND EXPERIMENTAL CASE STUDIES

2.1 Variability-aware modeling of RRAM

The fundamental switching mechanisms of resistive memories, RRAM, lead to statistical behaviors during programming. This phenomenon is characterized by physical parameters such as resistances and voltages. These statistical behaviors need to be properly characterized and modeled for a better understanding of their implications and roles in neuro-inspired computing systems. For metal-oxide RRAM devices, switching of states is governed by conductive filament evolutions. A full cycle of filament growth and rupture can be described by a suite of oxygen vacancy generation and recombination processes, induced and maintained by electrical field and temperature effects.

Building upon these basic physical understandings, RRAM models [13], [14] have been developed to capture essential SET and RESET switching behaviors between low resistance states (LRS) and high resistance states (HRS), as illustrated in Fig. 1(a). The RRAM models are publicly available for download and use [15]. These models account for the cycle-to-cycle variability of oxygen-vacancy-based filament evolutions under certain programming conditions. As shown in Fig. 2(b), analog resistance distributions that are measured from HfO_x-based RRAM devices are reproduced under various voltage and current conditions. When RRAM is operated at a voltage below the SET and RESET thresholds, it exhibits a different stochastic switching behavior. Fig. 1(c) shows measured statistical results for overall SET probability of RRAM devices starting from HRS states, using different strength of pulsing conditions characterized by pulse amplitude and pulse width. Capturing the statistical distributions within different regimes of operations is important for utilizing and optimizing RRAM in a computing system that can harness the device characteristics. One example is a recent work that uses analog programmability and statistical distributions to store encoded analog information within RRAM arrays [16].

2.2 Stochastic synapses in neural networks

Probabilistic switching of RRAM as a function of pulse voltage and time is as characterized in Fig. 2(a). The probabilistic switching leads to a weight primitive in neural networks, where weight values are masked with stochasticity. The sparsity is tunable depending on the pulsing conditions seen by RRAM weight storage devices. As an illustration for the interaction between device stochasticity and learning characteristics, a simple fully-connected neural network under an unsupervised feature learning setting is simulated with stochastic weight updates and winner-take-all (WTA) mechanism [17]. Dominant receptive fields encoded by RRAM resistances are formed after learning on images with noises. Synaptic stochasticity assists the convergence of the learning process while smoothing out the impact of input noises. Using the variability-aware RRAM model described in the previous subsection, we can further study implications of such device-algorithm interactions from an energy perspective. Fig. 2(c) shows the simulated network energy consumption, with contributions from probabilistic switching of RRAM and current summation, as a function of pulsing conditions. For a given switching probability, one can trade off overall energy consumption with speed of convergence by picking different combinations of pulse width and amplitude. Moderately short pulses result in the lowest energy consumption. The trend holds true across a wide range of switching probabilities, which are one of the design choices that lead to tradeoffs between convergence speed and energy consumption.

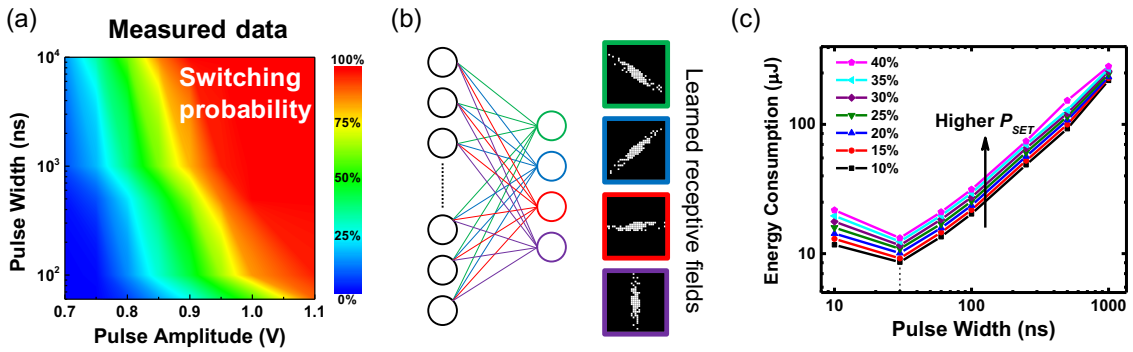

Figure 2. (a) Experimental mapping of pulsing conditions onto switching probabilities (P_{SET}). (b) Illustration of feature learning in neural networks with synaptic weights masked by stochasticity. (c) Optimizing neural network energy consumption by tuning pulse conditions.

In addition to learning robust features, the synaptic stochasticity has been investigated as a bridge between spiking networks and backprop-based deep neural networks (DNNs), with the intention of understanding the role of synaptic uncertainty observed in the cortex and the role of regularization in deep neural networks. For example, in Synaptic Sampling Machines (S2Ms) which can be configured as either non-spiking or spiking models [18], it is found that synaptic noise plays an important role of a regularizer during learning, akin to the effect of the DropConnect technique being applied to DNN training for regularization and decorrelation [19]. Robustness to pruning (80% of weight connections) can be obtained in S2M.

2.3 3D RRAM as nanokernels for HD computing

In addition to device physics, the unique physical structure and circuit architecture of NVM have been exploited for computing. Here we elaborate with an example that utilizes a vertical 3D RRAM architecture for high-dimensional (HD) computing [20].

Learning is about data representations and associated operations that form knowledge upon those representations. Drawing inspiration from how the brain computes with patterns of neural activities not readily associated with scalar numbers, a robust learning framework called HD computing was developed based on high-dimensional vector representations (when the dimensionality is in thousands) [12]. Cognitive applications demonstrated using HD computing framework range from recognition and visual question answering, to bio-signal processing for human-machine interface and healthcare (e.g., electromyography or EMG, electroencephalography or EEG) [21]. HD computing is built upon rich and subtle mathematical properties of high-dimensional space. The vectors sampled from HD space (i.e., HD vectors) are (pseudo)random with independent and identically distributed (i.i.d.) components, where information is distributed equally among all the components. As a result, each individual HD vector forms a powerful and robust representation that is resilient to errors in the components.

HD vectors are initialized randomly to represent symbolic information. In an example of language recognition task, letters in an alphabet are paired with HD vectors that are naturally near-orthogonal. Learning in HD space does not involve weight update or tuning. Instead, the HD vectors corresponding to sampled inputs are piped through a set of compute kernels, namely multiply, accumulate, permute (MAP), which preserve, bind, and compose low-level symbolic information into more complex and richer representations, in the same form as the inputs (i.e., the outputs are still HD vectors of the same dimensionality). This is similar to how low-level features are combined into high-level features in deep convolutional neural networks (CNNs). In HD computing, what's being learned is the underlying structure, relationship, or pattern associated with the sampled inputs for the specific task. This is akin to a compression process, where the encoded HD vector contains richer information than the sampled inputs. Using a simplified comparison for understanding, learning in neural networks produces new weight matrices whereas learning in HD framework produces new HD vectors. HD inference can be seen as decoding with inquiry vectors, or equivalently, similarity comparison between learned vectors and inquiry vectors.

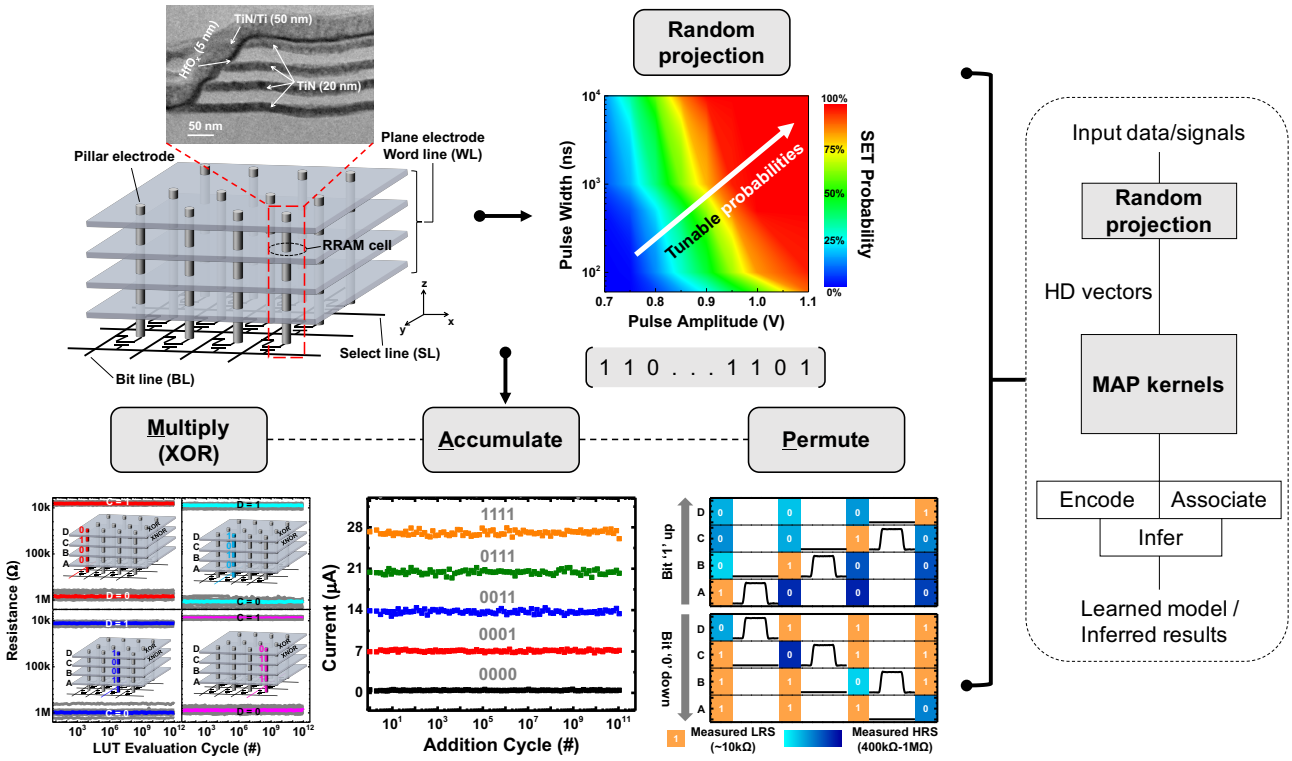

Figure 3. Experimental demonstrations of native multiply, accumulate, permute (MAP) kernels for HD computing within 4-layer 3D vertical RRAMs.

We approach the memory-intensive HD operations with a vertical 3D architecture of RRAM where memory cells are around vertical pillars across multiple layers [22], [23]. Due to the nonvolatile and stochastic nature of programming, binary vectors can be produced directly within RRAM cells with inherent randomness required for the initial random projection. Afterwards, these binary vectors stay within memories, where MAP kernels are implemented by exploiting the circuit-level properties of the vertical 3D architecture. Fig. 3 illustrates the MAP operations on a 4-layer 3D vertical RRAM integrated with FinFET select transistors. The multiply operations on binary vectors are equivalent to bit-wise XOR. We leverage the voltage dividers formed by the RRAM cells and select transistors underneath to construct non-volatile XOR/XNOR look-up tables. Details of operation schemes are discussed in [13]. This architecture design results in a few initial write operations for creating the XOR look-up tables, while most subsequent operations are read-only, without the need of re-programming RRAM. We measured multiply operations for 10^{12} cycles without errors. For accumulate operations, current summing is performed along vertical pillars with each RRAM cell contributing to the total current. We measured up to 10^{11} cycles without disturb errors. Permute operations simply shift bit '1' or '0' in a vector, which are realized through bit copy operations within 3D RRAM. In summary, algorithm-level characteristics (e.g., error resilience in HD representations) and technology-level characteristics (stochasticity and 3D connectivity of RRAM) are exploited together for an RRAM-centric HD computing system design.

3. HETEROGENEOUS LEARNING MACHINES

We have been witnessing the growing need for providing hardware support in the cloud and at the edge for increasingly complex learning and inference workloads, while accommodating the diverse compute kernels found in them [24]. Combined with the opportunities of natural and native "nanokernel" realizations with emerging device technologies, this leads us to co-designing emerging device technologies and computing architectures to create scalable, efficient, and secure heterogeneous learning machines (Fig. 4). Here, integration of technologies is crucial. In the case of energy-efficient neural network acceleration at the edge, NVM technologies such as MRAM and 3D RRAM, integrated fully on chip with CMOS-based accelerators, address the inefficiencies of off-chip DRAM [25]. Technology-system design

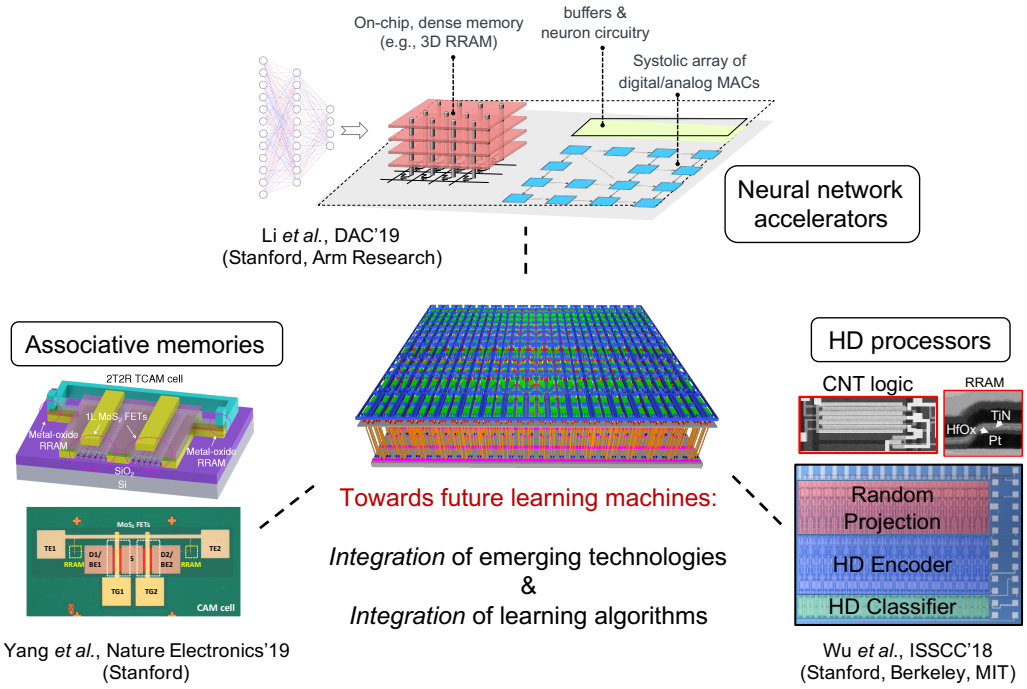

Figure 4. Illustration of a heterogeneous learning system with on-device learning, inference, and long-term memory associations. Examples of modules are based on separate studies that are relevant to these key features.

space explorations of NVM-embedded accelerators are conducted using important vision workloads such as ResNet, MobileNet and Faster-RCNN. We find that high-density on-chip NVM (e.g., 3D RRAM) enables more aggressive pareto optimizations and provides simultaneous energy and area benefits for accelerators, which is not achievable with today's embedded DRAM (eDRAM) or large SRAM. In the case of running MobileNet, compared to accelerators with off-chip DRAM, integrating 3D RRAM on chip provides 2.22× overall energy benefits with 4× less on-chip SRAM buffers, resulting in 33% accelerator area savings at the same time.

As discussed earlier, HD computing serves as a simple yet powerful learning template with sequences and signals. Through technology integration, benefits and characteristics of individual device components can be combined and utilized. For example, an HD nanosystem was built with monolithic 3D integration of RRAM and carbon nanotube transistors (CNTs) [26]. Running language recognition on the HD nanosystem, a 7.6× energy benefit is projected over a silicon CMOS implementation, as a result of energy efficiency of RRAMs/CNTs, and new designs that exploit device-level properties (e.g., CNT's variability, RRAM's analog programmability).

Finally, for a learning machine that targets continuous, lifelong learning, an energy- and area-efficient associative memory (AM) module is crucial for long-term knowledge storage and fast retrieval, which also helps to overcome potential catastrophic forgetting issues in a learning system. A hardware AM realization that leverages the integration of HfO_x RRAM and MoS_2 FETs has been reported [27]. In this work, low leakage and robust current control lead to high search capacity and energy efficiency. Owing to low temperature fabrication, the combination of RRAM and MoS_2 can be further integrated into a high-rise monolithic 3D system, approaching a closer emulation of human memories in terms of ultra-dense connectivity for learning and memory functionalities.

4. CONCLUSIONS

In this paper, we present a device-to-algorithm analysis of neuro-inspired computing with emerging non-volatile memories, specifically using RRAM as an example. Our methodology is built upon a combination of experimental characterization, cross-stack modeling, and system integration. There is plenty of room at the bottom, for exploiting inherent device characteristics as well technology integration opportunities. There is also plenty of room at the top, for

exploring diverse models and learning algorithms. Innovations in computing architectures and efficient hardware realizations in the middle will be necessary to build a heterogeneous learning machine for future computing workloads.

ACKNOWLEDGEMENTS

We acknowledge the support from ASCENT, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, NSF/NRI/GRC E2CDA, Stanford NMTRI, and Stanford SystemX Alliance. H. Li would like to thank Tony Wu, Abbas Rahimi, Weier Wan, Prof. Rui Yang, Prof. Subhasish Mitra, Prof. Jan Rabaey, Prof. Sayeef Salahuddin, Prof. Bruno Olshausen, Pentti Kanerva, Wen-Kuan Yeh, Mudit Bhargava, Paul Whatmough, Brian Cline, and Greg Yeric, for fruitful collaborations and discussions related to the works described in this paper.

REFERENCES

[1] Wong, H.S.P. and Salahuddin, S., 2015. Memory leads the way to better computing. Nature nanotechnology, 10(3), p.191.
[2] Yu, S., 2018. Neuro-inspired computing with emerging nonvolatile memorys. Proceedings of the IEEE, 106(2), pp.260-285.
[3] Ielmini, D. and Wong, H.S.P., 2018. In-memory computing with resistive switching devices. Nature Electronics, 1(6), p.333.
[4] Islam, R., Li, H., Chen, P.Y., Wan, W., Chen, H.Y., Gao, B., Wu, H., Yu, S., Saraswat, K.C. and Wong, H.S.P., 2018. Device and materials requirements for neuromorphic computing. Journal of Physics D: Applied Physics.
[5] Burr, G.W., Shelby, R.M., Sebastian, A., Kim, S., Kim, S., Sidler, S., Virwani, K., Ishii, M., Narayanan, P., Fumarola, A. and Sanches, L.L., 2017. Neuromorphic computing using non-volatile memory. Advances in Physics: X, 2(1), pp.89-124.
[6] Torrejon, J., Riou, M., Araujo, F.A., Tsunegi, S., Khalsa, G., Querlioz, D., Bortolotti, P., Cros, V., Yakushiji, K., Fukushima, A. and Kubota, H., 2017. Neuromorphic computing with nanoscale spintronic oscillators. Nature, 547(7664), p.428.
[7] Grollier, J., Querlioz, D. and Stiles, M.D., 2016. Spintronic nanodevices for bioinspired computing. Proceedings of the IEEE, 104(10), pp.2024-2039.
[8] Salahuddin, S., Ni, K. and Datta, S., 2018. The era of hyper-scaling in electronics. Nature Electronics, 1(8), p.442.
[9] Sengupta, A. and Roy, K., 2017. Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing. Applied Physics Reviews, 4(4), p.041105.
[10] Fang, Y., Gomez, J., Wang, Z., Datta, S., Khan, A.I. and Raychowdhury, A., 2019. Neuro-mimetic Dynamics of a Ferroelectric FET Based Spiking Neuron. IEEE Electron Device Letters.
[11] Jerry, M., Chen, P.Y., Zhang, J., Sharma, P., Ni, K., Yu, S. and Datta, S., 2017, December. Ferroelectric FET analog synapse for acceleration of deep neural network training. In 2017 IEEE International Electron Devices Meeting (IEDM) (pp. 6-2). IEEE.
[12] Kanerva, P., 2009. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. Cognitive computation, 1(2), pp.139-159.
[13] Li, H., Wu, T.F., Mitra, S. and Wong, H.S.P., 2017. Resistive RAM-centric computing: Design and modeling methodology. IEEE Transactions on Circuits and Systems I: Regular Papers, 64(9), pp.2263-2273.
[14] Jiang, Z., Wu, Y., Yu, S., Yang, L., Song, K., Karim, Z. and Wong, H.S.P., 2016. A compact model for metal–oxide resistive random access memory with experiment verification. IEEE Transactions on Electron Devices, 63(5), pp.1884-1892.
[15] Stanford RRAM Model, https://nano.stanford.edu/stanford-rram-model, accessed August 16, 2019.
[16] Zheng, X., Zarcone, R., Paiton, D., Sohn, J., Wan, W., Olshausen, B. and Wong, H.S.P., 2018, December. Error-Resilient Analog Image Storage and Compression with Analog-Valued RRAM Arrays: An Adaptive Joint Source-Channel Coding Approach. In 2018 IEEE International Electron Devices Meeting (IEDM) (pp. 3-5). IEEE.
[17] Li, H., Li, K.S., Lin, C.H., Hsu, J.L., Chiu, W.C., Chen, M.C., Wu, T.T., Sohn, J., Eryilmaz, S.B., Shieh, J.M., Yeh, W.K., and Wong, H.S.P., 2016, June. Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing. In 2016 IEEE Symposium on VLSI Technology (pp. 1-2). IEEE.

[18] Neftci, E.O., Pedroni, B.U., Joshi, S., Al-Shedivat, M. and Cauwenberghs, G., 2016. Stochastic synapses enable efficient brain-inspired learning machines. Frontiers in neuroscience, 10, p.241.
[19] Wan, L., Zeiler, M., Zhang, S., Le Cun, Y. and Fergus, R., 2013, February. Regularization of neural networks using dropconnect. In *International conference on machine learning*(pp. 1058-1066).
[20] Li, H., Wu, T.F., Rahimi, A., Li, K.S., Rusch, M., Lin, C.H., Hsu, J.L., Sabry, M.M., Eryilmaz, S.B., Sohn, J., Chiu, W.C., Chen, M.C., Wu, T.T., Shieh, J.M., Yeh, W.K., Rabaey, J.M., Mitra, S., and Wong, H.S.P., 2016, December. Hyperdimensional computing with 3D VRRAM in-memory kernels: Device-architecture co-design for energy-efficient, error-resilient language recognition. In 2016 IEEE International Electron Devices Meeting (IEDM) (pp. 16-1). IEEE.
[21] Rahimi, A., Kanerva, P., Benini, L. and Rabaey, J.M., 2018. Efficient biosignal processing using hyperdimensional computing: Network templates for combined learning and classification of exg signals. Proceedings of the IEEE, 107(1), pp.123-143.
[22] Chen, H.Y., Yu, S., Gao, B., Huang, P., Kang, J. and Wong, H.S.P., 2012, December. HfOx based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector. In 2012 International Electron Devices Meeting (pp. 20-7). IEEE.
[23] Hsueh, F.K., Shen, C.H., Shieh, J.M., Li, K.S., Chen, H.C., Huang, W.H., Wang, H.H., Yang, C.C., Hsieh, T.Y., Lin, C.H. and Chen, B.Y., 2016, December. 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 2016 IEEE International Electron Devices Meeting (IEDM) (pp. 2-3). IEEE.
[24] LeCun, Y., 2019, February. Deep Learning Hardware: Past, Present, and Future. In 2019 IEEE International Solid-State Circuits Conference-(ISSCC) (pp. 12-19). IEEE.
[25] Li, H., Bhargava, M., Whatmough, P.N. and Wong, H.S.P., 2019, June. On-Chip Memory Technology Design Space Explorations for Mobile Deep Neural Network Accelerators. In Proceedings of the 56th Annual Design Automation Conference 2019 (p. 131). ACM.
[26] Wu, T.F., Li, H., Huang, P.C., Rahimi, A., Hills, G., Hodson, B., Hwang, W., Rabaey, J.M., Wong, H.S.P., Shulaker, M.M. and Mitra, S., 2018. Hyperdimensional Computing Exploiting Carbon Nanotube FETs, Resistive RAM, and Their Monolithic 3D Integration. IEEE Journal of Solid-State Circuits, 53(11), pp.3183-3196.
[27] Yang, R., Li, H., Smithe, K.K., Kim, T.R., Okabe, K., Pop, E., Fan, J.A. and Wong, H.S.P., 2019. Ternary content-addressable memory with MoS 2 transistors for massively parallel data search. Nature Electronics, 2(3), p.108.