# NeuroSim+: An Integrated Device-to-Algorithm Framework for Benchmarking Synaptic Devices and Array Architectures

Pai-Yu Chen, Xiaochen Peng, and Shimeng Yu*
School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, USA, *Email: shimengy@asu.edu

***Abstract*—**NeuroSim+ is an integrated simulation framework for benchmarking synaptic devices and array architectures in terms of the system-level learning accuracy and hardware performance metrics. It has a hierarchical organization from the device level (transistor technology and memory cell models) to the circuit level (synaptic array architectures and neuron periphery) and then to the algorithm level (neural network topologies). In this work, we study the impact of the "analog" eNVM non-ideal device properties and benchmark the trade-offs of SRAM, digital and analog eNVM based array architectures for online learning and offline classification. The source code of NeuroSim+ version 1.0 is publicly available at https://github.com/neurosim/MLP_NeuroSim.

## I. Introduction

The off-chip memory access is the bottleneck for designing the machine learning hardware accelerators (as shown in the design of Eyeriss [1]). Analog emerging non-volatile memory (eNVM) has been proposed as compact synaptic devices to replace on-chip SRAM synapses due to its excellent scalability and multilevel conductance states. An ideal weight update behavior of analog eNVM assumes a linear update of the conductance (or weight) with programming voltage pulses. However, the realistic devices reported in literature do not follow such ideal trajectory (Fig. 1), exhibiting "non-ideal" properties such as nonlinear and noisy weight increase/decrease, limited precision and finite conductance ON/OFF ratio, etc. (Fig. 2). The impact of these non-ideal device properties of eNVM has been studied in online learning scenarios where the "accuracy" is the primary performance metric [2, 3]. A comprehensive analysis on the circuit-level performance (e.g. area, latency, energy) remains unexplored. Therefore, it is necessary to develop a hierarchical simulation framework that covers from the device to algorithm to investigate the design trade-offs with different neuro-inspired architectures. In this work, we use a 2-layer multilayer perceptron (MLP) neural network (NN) with MNIST handwritten digits [4] as the training and testing dataset (Fig. 3(a)) to implement the online learning and offline classification. In online learning, the simulator emulates hardware to train NN with images randomly picked from the training dataset (60k images) and classify the testing dataset (10k images). In offline classification, NN is pre-trained by software, and the simulator only emulates hardware to classify the testing dataset.

## II. NeuroSim+ Framework for Building MLP NN

We developed NeuroSim, a C++ based circuit-level macro model that can be used to estimate the area, latency, dynamic energy and leakage power of SRAM and eNVM based neuro-inspired architectures to facilitate the design space exploration, following the principle of CACTI [5] for SRAM cache and NVSim [6] for NVM. NeuroSim can support various mainstream and emerging memory technologies, and NeuroSim can also support various machine learning NNs to form a complete simulation framework, namely "NeuroSim+". In this 2-layer MLP NN (Fig. 3(a)), the MNIST input images are cropped and converted to black and white (1-bit) data to reduce the complexity of input encoding. For design simplicity, each neuron will truncate the weighted sum to 1-bit output value for the input of next neuron node. In this way, offline classification, which is purely feed forward (FF) can be realized with low-precision. However, the computation on the back propagation (BP) of weight update still needs high precision [2]. Such a simple network can achieve 96~97% in online learning in the software baseline, which is not as high as the reported ones (>98%) [7] due to the simplicity made for hardware implementation, where the MLP NN weights are mapped to the synaptic cores (Fig. 3(b)) – the computation units designed for performing weighted sum and weight update. Synaptic cores can be categorized into two types: the digital synapses and the analog synapses (Fig. 3(c)). The digital synapses are more mature schemes that exploit only the ON and OFF binary states of memory cells and a few of these cells are grouped together to represent the precision of a synapse in a binary format. SRAM and digital 1T1R eNVM synaptic cores are in this category. For analog synapses, the representative one is the pseudo-crossbar array architecture [8], which resolves the issue of weight update disturbance with the cell access transistor and has less energy than the crossbar one. In NeuroSim+ framework (Fig. 4), the MLP NN simulator can call the proposed circuit architectures (Fig. 3(b)) with the support of NeuroSim, while the simulator itself is developed in an algorithm+device fashion that could study the synaptic array parasitics (Fig. 5) and realistic device properties (Fig. 2) in detail as in prior works [2, 3]. Evaluation in NeuroSim+ is instruction-accurate. Whenever a weighted sum or weight update instruction is given during FF and BP, the instruction will be passed to the synaptic array and device behavioral model for calculation of computation error, as well as passed to NeuroSim for evaluation of the circuit-level performances. To validate NeuroSim, we have performed the area calibration of a synaptic core using layout at FreePDK 45nm [9] (Fig. 6) and the calibration of other performance metrics in main circuit modules using SPICE (Fig. 7).

## III. Impact of Non-Ideal Device Properties

For analog eNVM based architecture, after the weight update amount is obtained in software, it will be translated by control logic circuitry into number of programming pulses and then applied to the synaptic devices. To quantify the impact of

these realistic device properties (Fig. 2), we performed sensitivity analyses in the both learning modes (Fig. 8), which suggests a) the weight precision should be at least 6 bits for online learning to achieve high accuracy ~95%, while 1 or 2-bits are sufficient for offline classification; b) limited conductance ON/OFF ratio<50 will degrade the accuracy of offline classification, while the network may adapt itself to this limited ON/OFF ratio during online learning; c) the accuracy loss will become noticeable when σ of the read noise is beyond 15~25%; d) high learning accuracy can only be guaranteed with small nonlinearity (0.5~1); e) even though the network is generally resilient to the device-to-device variation [2, 3], its impact becomes prominent at high nonlinearity (>1); f) small cycle-to-cycle variation (<2%) can alleviate the degradation of learning accuracy by high nonlinearity. However, too large variation (>2%) overwhelms the deterministic update amount defined by BP thus is harmful to the accuracy. For b) and c), the accuracy drop in online learning is much sharper once the non-ideal effect is beyond the tolerance range, probably because the network will deviate more from its correct form with both erroneous FF and BP results. Overall, we listed the reported devices with the extracted realistic device parameters (Table 1). Today's analog eNVM devices are problematic to be used in online learning as well as offline classification. Therefore, we propose to set up the targeted eNVM specs that is able to achieve a good accuracy of 90%. As a reference, eNVM device with ideal specs is also listed, which is supposed to be as perfect as SRAM to achieve the same high accuracy of 94~95% with 6-bit weights for online learning (Fig. 8(a)).

## IV. Circuit-Level Performance Benchmark

In NeuroSim+, the area and leakage power can be directly obtained given the configuration of architecture, while the latency and energy consumption will be calculated at the run-time of learning (Fig. 4). The benchmark is performed at 14 nm node. We estimated the total area of SRAM, digital and analog eNVM based architectures for the online learning (Fig. 9(a)). The analog eNVM one can achieve the smallest area due to smaller cell size and multiple bits per cell. The leakage power of these architectures is also estimated (Fig. 9(b)). Unlike SRAM, the eNVM based synaptic cores do not need power supply to maintain the data in memory cells thus their leakage power is much smaller. The digital eNVM has less leakage power than the analog one because a large portion of leakage power in the analog one comes from the SL/BL switch matrix. However, SRAM has more advantages over digital and analog eNVM on the latency and energy consumption for online learning (Fig. 10). Although analog eNVM's parallel weighted sum operation in FF is much faster than SRAM's and digital eNVM's row-by-row read, most of the latency is dominated by the analog eNVM's slow weight update, which requires 2 phases (2 voltage polarities for weight increase and decrease) and multiple pulses per phase. Even the write pulse width of ideal analog eNVM is assumed to be 10 ns, the weight update latency of ideal eNVM is still ~28X slower than that of digital eNVM, and SRAM can be even faster than the digital eNVM by ~36X because the write latency of a single SRAM cell (<ns) is much less than that of a single digital eNVM (10 ns) and we assume that only part of the row is selected at a time for weight update in digital eNVM. For the energy consumption, the difference is less significant. As there is only a small portion of weights being updated at each cycle, analog eNVMs with different pulse widths do not show any difference in weight update energy. In fact, most of the energy is consumed at biasing the unselected columns. For offline classification, accuracy>93% can be achieved using either 2-bit SRAM and digital eNVM (equivalently Fig. 8(a)) or 2-bit target/ideal analog eNVM (Table 1). Without any training process, the analog eNVM based architecture can be superior to the other two designs in terms of latency and energy (Table 2).

## Conclusion

The NeuroSim+ framework is helpful for device engineers to benchmark different synaptic device candidates and array architectures for the machine learning. In a small MLP NN, SRAM is preferred over eNVM for online learning due to its excellent performance in weight update but with an overhead of much larger area and leakage power. Capable of parallel weighted sum operation, the analog eNVM based architecture is suitable for the read-intensive applications such as offline classification, while the digital eNVM one may be a better choice for low standby power design. A summary of design optimization for these architectures is provided in Table 3.

## Acknowledgment

This work is supported by NSF-CCF-1552687.

## References

[1] Y.-H. Chen *et al.*, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 262-263, 2016.

[2] P.-Y. Chen *et al.*, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," *ACM/IEEE International Conference on Computer-Aided Design (ICCAD)*, pp. 194-199, 2015.

[3] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 Synapses) using phase-change memory as the synaptic weight element," *IEEE Transactions on Electron Devices,* vol. 62, no. 11, pp. 3498-3507, 2015.

[4] Y. LeCun *et al.*, "The MNIST database of handwritten digits," 1998.

[5] CACTI, http://www.hpl.hp.com/research/cacti/

[6] X. Dong *et al.*, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 31, no. 7, pp. 994-1007, 2012.

[7] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[8] S. Yu *et al.*, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," *IEEE International Electron Devices Meeting (IEDM)*, pp. 451-454, 2015.

[9] FreePDK45, https://www.eda.ncsu.edu/wiki/FreePDK45:Contents

[10] L. Gao *et al.*, "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology,* vol. 26, no. 45, pp. 455204, 2015.

[11] S. Park *et al.*, "Neuromorphic speech systems using advanced ReRAM-based synapse," *IEEE International Electron Device Meeting (IEDM)*, pp. 625-628, 2013.

[12] S. H. Jo *et al.*, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Letters,* vol. 10, no. 4, pp. 1297-1301, 2010.

[13] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using $AlO_x/HfO_2$ bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters,* vol. 37, no. 8, pp. 994-997, 2016.

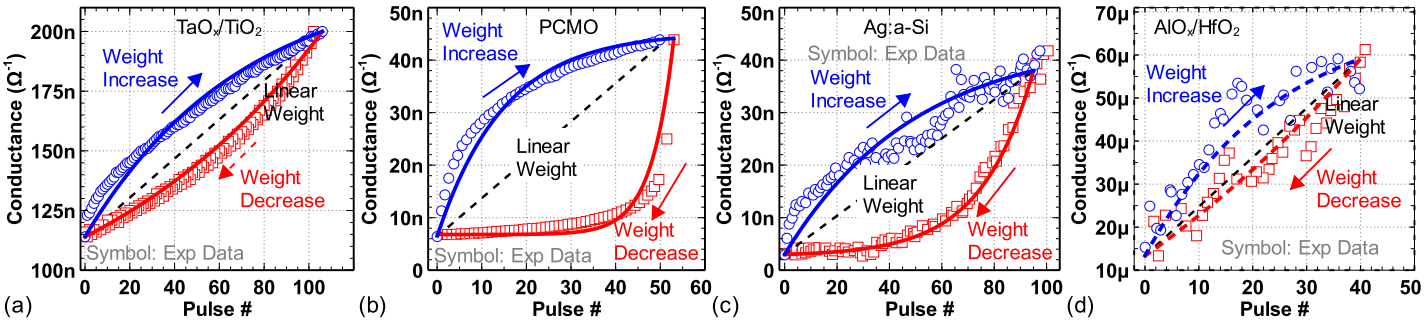


Fig. 1 Weight update curve of the reported (a) $TaO_x/TiO_2$ [10], (b) PCMO [11], (c) Ag:a-Si [12] and (d) $AlO_x/HfO_2$ [13] type analog eNVMs.

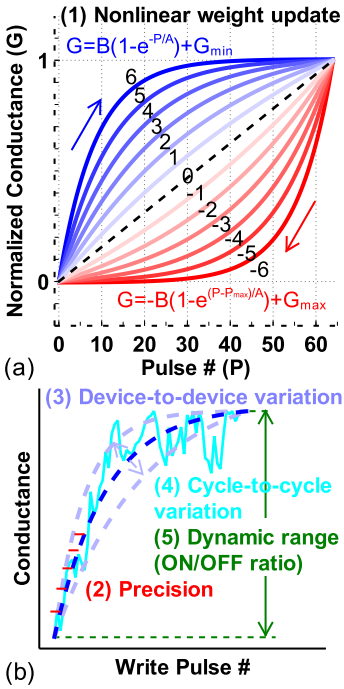


Fig. 2 Summary of non-ideal analog eNVM device properties modeled in this work. Exponential functions are used to model the nonlinear weight update behaviors, where the nonlinearity is labeled from -6 to 6. Device-to-device and cycle-to-cycle variations are the variation in the nonlinearity baseline and conductance change, respectively.

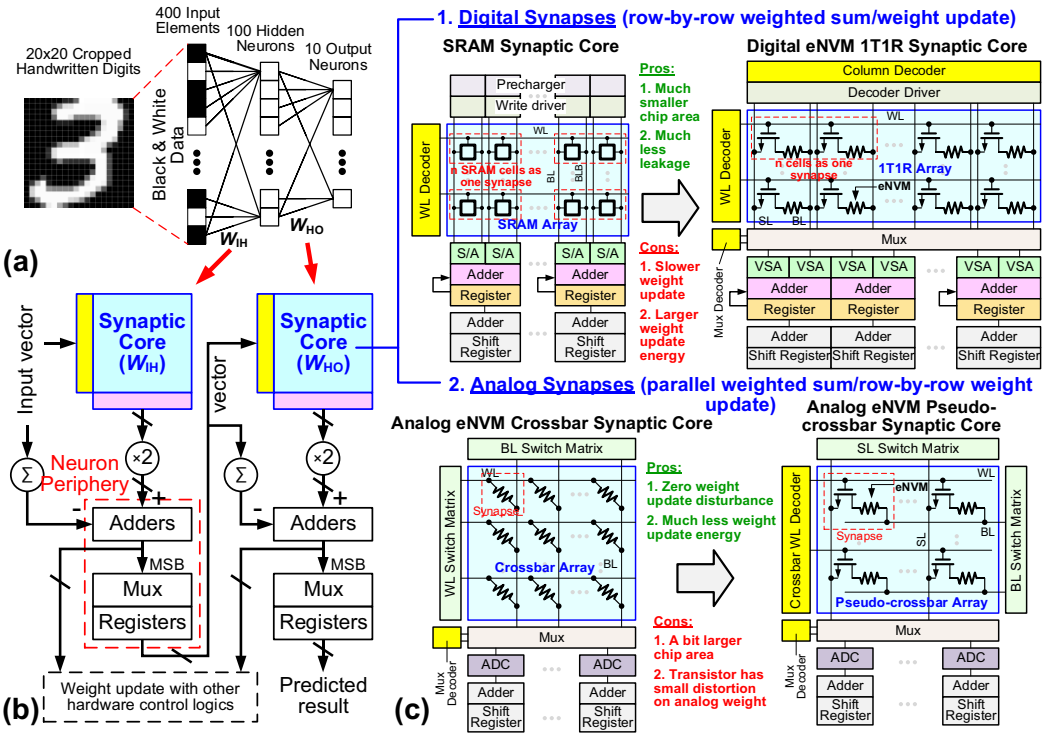


Fig. 3 (a) The 2-layer multilayer perceptron (MLP) neural network (NN). The input MNIST images are cropped and encoded into black and white data for simplification on hardware implementation. (b) Circuit block diagram for hardware implementation of the 2-layer MLP NN. The weights are mapped to synaptic cores, which are computation units designed for performing weighted sum and weight update. (c) 2 categories of the synaptic core architectures: digital synapses and analog synapses. For digital synapses, the digital eNVM 1T1R synaptic core can achieve much smaller area and leakage than the SRAM one. For analog synapses, the analog eNVM pseudo-crossbar synaptic core is highly preferred over the crossbar one due to zero weight update disturbance and less weight update energy.

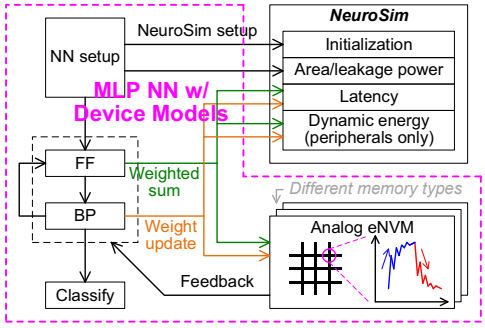


Fig. 4 The NeuroSim+ framework. At the run-time of NN, the weighted sum and weight update instructions will be given to both the synaptic array/device model and NeuroSim for evaluation of computation error and circuit-level performances, respectively.

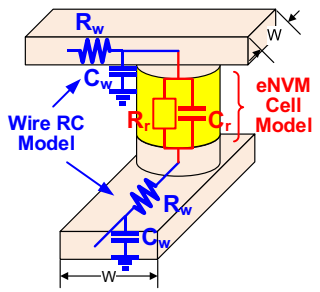


Fig. 5 The SPICE sub-circuit module for a single eNVM cell in the crossbar array. The wire resistance and parasitic capacitance are included, which can also be applied to the pseudo-crossbar array structure.

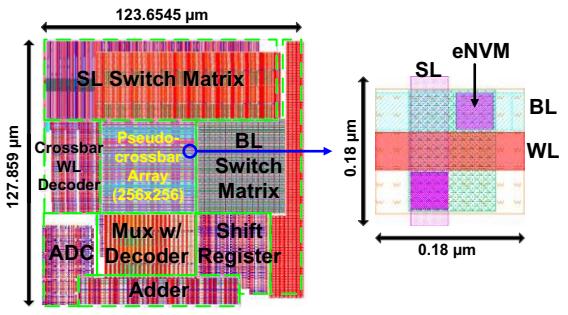


Fig. 6 Example layout of the eNVM pseudo-crossbar synaptic core (256×256 array size) at FreePDK 45 nm. The entire layout area is measured to be 15,810 $\mu m^2$, with a cell size of 0.0324 $\mu m^2$ (4F× 4F), while the area estimation by NeuroSim is 15,454 $\mu m^2$, achieving an error rate of -2.5%.

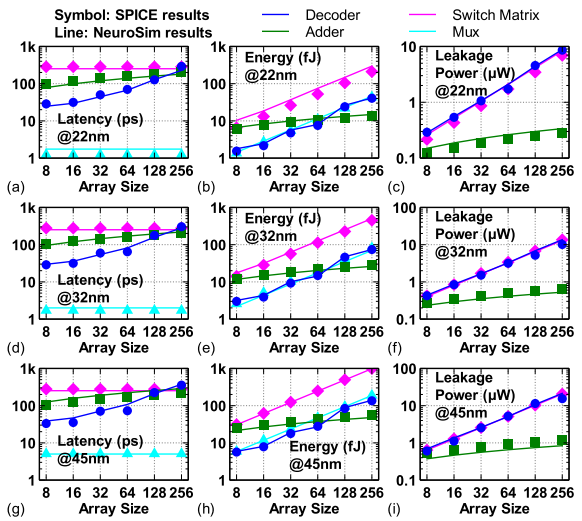


Fig. 7 SPICE validation of latency, dynamic energy, and leakage power on main circuit modules (decoder, switch matrix, adder, mux) of NeuroSim with different synaptic array sizes at 22 nm, 32 nm and 45 nm technology nodes.

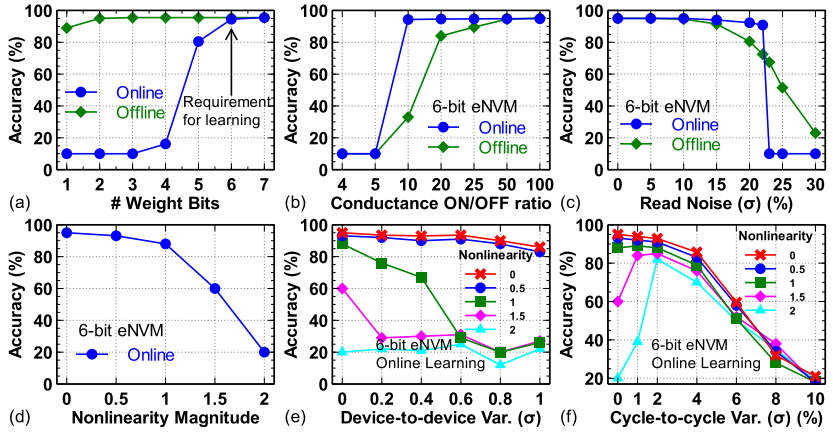


Fig. 8 The impact of (a) weight precision, (b) conductance ON/OFF ratio, (c) read noise, (d) weight update nonlinearity, (e) device-to-device variation and (f) cycle-to-cycle variation in online learning and/or offline classification. With both erroneous FF and BP results, the accuracy drop in online learning is usually much sharper than that of offline classification. For the weight update, high accuracy can only be guaranteed with small nonlinearity (<0.5~1), while small cycle-to-cycle variation (<2%) can help alleviate the degradation of learning accuracy by high nonlinearity.

Table 1. Specs and Learning Accuracy of Reported and Desired Analog eNVMs

| | Reported eNVMs for learning | | | | Desired eNVMs for learning | |
|---|---|---|---|---|---|---|
| Analog eNVM type | $TaO_x/TiO_2$ [10] | PCMO [11] | Ag:a-Si [12] | $AlO_x/HfO_2$ [13] | Targeted eNVM | Ideal eNVM |
| # of conductance states | 102 | 50 | 97 | 40 | 64 (6 bits) | 64 (6 bits) |
| Nonlinearity (weight increase/decrease) | 0.66/-0.69 | 3.68/-6.76 | 2.4/-4.88 | 1.94/-0.61 | 1.0/-1.0 | 0/0 |
| $R_{ON}$ | 5 MΩ | 23 MΩ | 26 MΩ | 16.9 kΩ | 200 kΩ | 200 kΩ |
| ON/OFF ratio | 2 | 6.84 | 12.5 | 4.43 | 50 | 50 |
| Weight increase pulse | 3V/40ms | -2V/1ms | 3.2V/300μs | 0.9V/100μs | 2V/100ns | 2V/10ns |
| Weight decrease pulse | -3V/10ms | 2V/1ms | -2.8V/300μs | -1V/100μs | 2V/100ns | 2V/10ns |
| Weight update cycle-to-cycle variation (σ) | <1% | <1% | 3.5% | 5% | 2% | 0% |
| Accuracy for online learning | ~10% | ~10% | ~73% | ~41% | 90% | 94.8% |
| Accuracy for offline classification | ~10% | ~20% | ~63% | ~10% | 94.5% | 94.5% |
| Area | 1071.3 $\mu m^2$ | 1071.3 $\mu m^2$ | 1072.0 $\mu m^2$ | 3657.2 $\mu m^2$ | 1247.3 $\mu m^2$ | 1247.3 $\mu m^2$ |
| Latency for online learning (1M images) | 3.57E10 s | 7.00E8 s | 4.20E8 s | 5.60E7 s | 8.82E4 s | 8.82E3 s |
| Energy for online learning (1M images) | 65.86 mJ | 29.4 mJ | 87.94 mJ | 150 mJ | 29.80 mJ | 29.80 mJ |

Green: good attributes, Red: major cause of learning failure

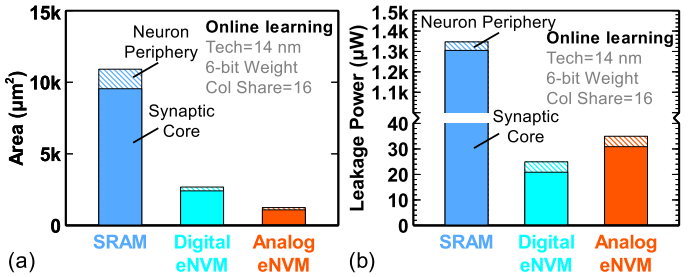


Fig. 9 (a) Area and (b) leakage power estimation of MLP NN architecture with SRAM, digital and analog eNVM synaptic cores in online learning. Analog eNVM can achieve the smallest area, while digital eNVM has the least leakage power consumption.

Table 2. Benchmark for Offline Classification

| | 2-bit SRAM | 2-bit digital eNVM | 2-bit analog eNVM |
|---|---|---|---|
| Area | 4450.8 $\mu m^2$ | 1071.2 $\mu m^2$ | 1247.3 $\mu m^2$ |
| Latency | 32.997 ms | 10.39 ms | 0.25 ms |
| Dynamic Energy | 16.939 μJ | 7.30 μJ | 3.38 μJ |
| Leakage Power | 475.67 μW | 22.89 μW | 35.29 μW |

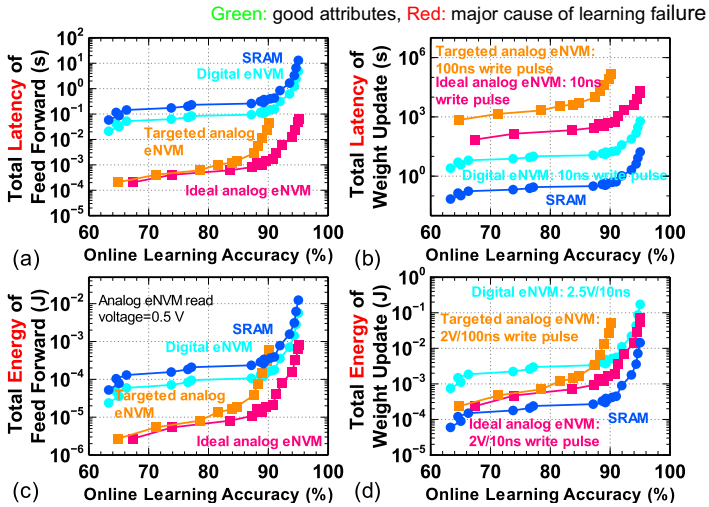


Fig. 10 Dynamic trace of latency and energy consumption in feed forward (FF) and weight update during online learning. Despite analog eNVM's parallel weighted sum in FF, SRAM is still more powerful than both eNVMs in terms of the overall speed and energy efficiency for small 2-layer NN.

Table 3. Summary of Design Optimization for Different Learning Modes

| | Accuracy Opt. | Area Opt.* | Latency Opt.* | Energy Opt.* | Leakage Opt.* | **Overall Recom.** |
|---|---|---|---|---|---|---|
| Online Learning | SRAM, eNVMs** | Analog eNVM | SRAM | SRAM | Digital eNVM | **SRAM** |
| Offline Classification | SRAM, eNVMs** | Digital eNVM | Analog eNVM | Analog eNVM | Digital eNVM | **Analog eNVM** |

*Bottom line: accuracy>90%, **eNVMs referred to as both digital and analog eNVMs