Capacitor-based architecture for AI hardware accelerators

July 10 2018, by Yulong Li, Effendi Leobandung

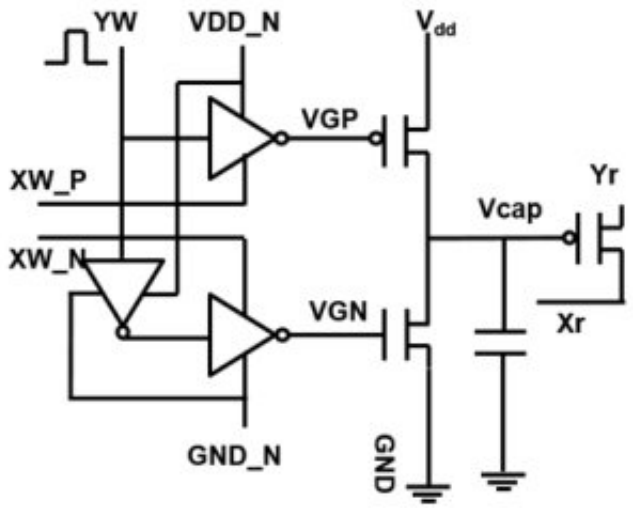

Figure 1. Unit cell schematic of a capacitor-based cross-point array. Credit: IBM

IBM is reaching beyond digital technologies with a capacitor-based cross-point array for analog neural networks, exhibiting potential orders of magnitude improvements in deep learning computations. Analog computing architectures exploit the storage capability and physical attributes of certain memory devices not just to store information, but also to perform computations. This has the potential to greatly reduce the time and energy required by computers because data doesn't need to be shuttled between the memory and processor. The drawback could be a reduction in computational accuracy, but for systems that do not require high accuracy, it is the right trade-off.

In analog neural networks (NN), non-volatile memory (NVM) based cross-point arrays have achieved promising results for inference tasks. However, training NNs to high accuracy is difficult for NVM devices, since successful training depends on keeping the incremental changes in NN weight small (requiring roughly 1,000 update states) and symmetric (so that positive and negative updates balance on average). Such issues can be addressed by using capacitors. Since charge can be added or subtracted continuously if the number of electrons is high, analog and symmetric weight update can be achieved. We presented a capacitor-based cross-point array for analog neural networks at the 2018 VLSI Technology Symposium. The new architecture achieved record symmetry and linearity for weight update.

Figure 1 shows the unit cell schematic of a capacitor-based cross-point array. The key component is the capacitor which is connected to a readout field effect transistor (FET). The charge on the capacitor represents the synaptic weight and the capacitor is charged and discharged with two current source FETs. Figure 2 shows the measured change in the conductance of the readout FET of a single cell, and corresponding capacitor voltage respectively, by applying ten cycles of 400 positive updates followed by 400 negative updates. Figure 3 compares the experimental non-linearity-update factors for our capacitor based analog synapse against other NVM technologies. The capacitor-based unit cell provides the best symmetry and linearity demonstrated to date. Figure 4 demonstrates parallel weight update on a 2×2 array.

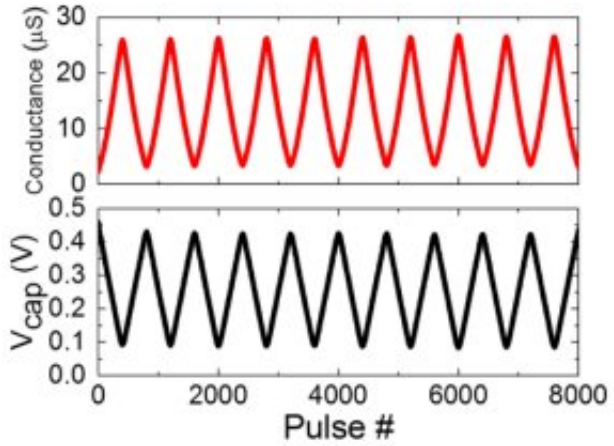

Figure 2. (a) Experimental results for updating single-cell with 8000 pulses. (b) Corresponding capacitor voltage change. Pulse width 50 ns, period: 500 ns. Credit: IBM

Even though capacitors are volatile, the leakage could be compensated during weight update. Since training repeatedly goes through forward, backward and weight update cycles, weights after decay in previous cycle are used in training for next cycle and get updated. Therefore, no intentional refresh cycles are needed. We tested the effect of retention time on training, using a fully-connected network. It has one input layer, two hidden layers, and one output layer (Figure 5) and was trained on the MNIST dataset by stochastic gradient descent and backpropagation. Assuming the training cycle length per layer (forward+backward+update) is 200 ns and synaptic weight decays with RC time constant τ, we found that penalty in training accuracy due to capacitor charge-loss becomes negligible when $\tau > 106 \times$ the training cycle length (Figure 6). We also tested the retention time requirement for a convolutional network. Our test network has two convolutional layers with two pooling layer and two fully connected layers (Figure 7). Due to the weight sharing (reuse) in convolutional layers, the retention requirements for a convolutional neural network (CNN) are about 600

larger (Figure 8).

We estimate the scalability of this capacitor-based array as a function of leakage for both fully connected and convolutional neural networks (Figure 9). Circle data points shows that the capacitor linearly scales with pass transistor leakage. Square data points show that when the leakage is large, the cell area is dominated by the capacitors; when the leakage current is small, the area will be dominated by FETs in the cell. For DRAM technology with leakage of 1 fA/cell requires capacitor network/algorithm optimization could reduce capacitor requirement.

IBM is now working on novel ideal memory with optimized analog behavior. These capacitors will allow analog AI core to be implemented on an accelerated schedule, since the technology and process are available.

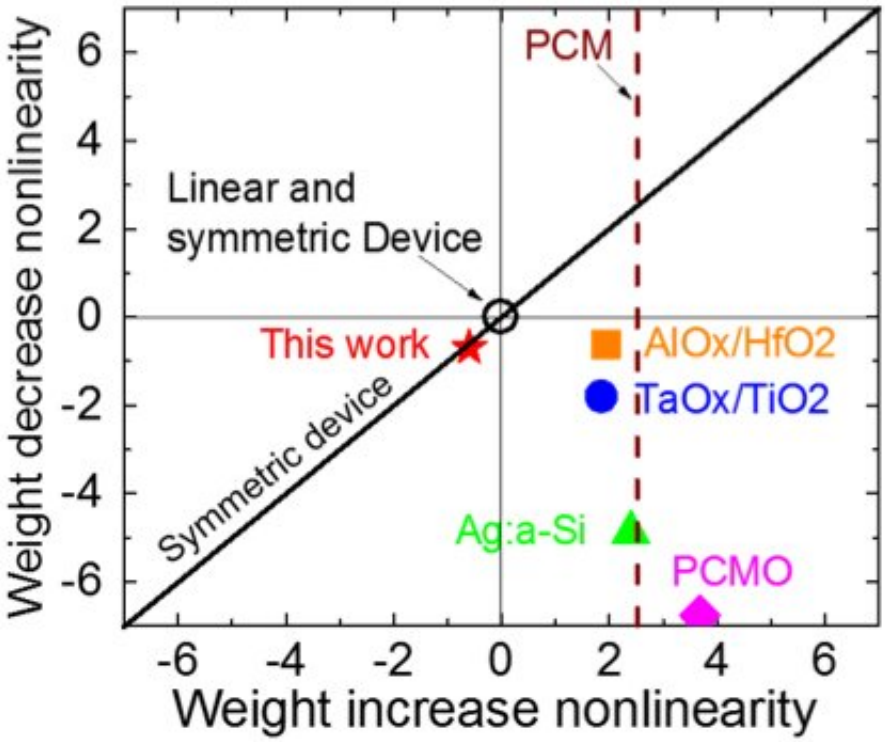

Figure 3. Conductance non-linearity of this work compared with other NVM technologies. Credit: IBM

In addition to our capacitor approach, IBM is exploring other novel elements for analog memory and computation such as phase change memory (PCM) and resistive RAM (RRAM). These elements vary in term of cell areas, retention, symmetry, and maturity. Analog accelerators are one component of IBM Research AI's pipeline of AI hardware accelerators. The pipeline starts with getting the most from existing GPU accelerators, followed by innovative digital AI cores exploiting approximate computing.

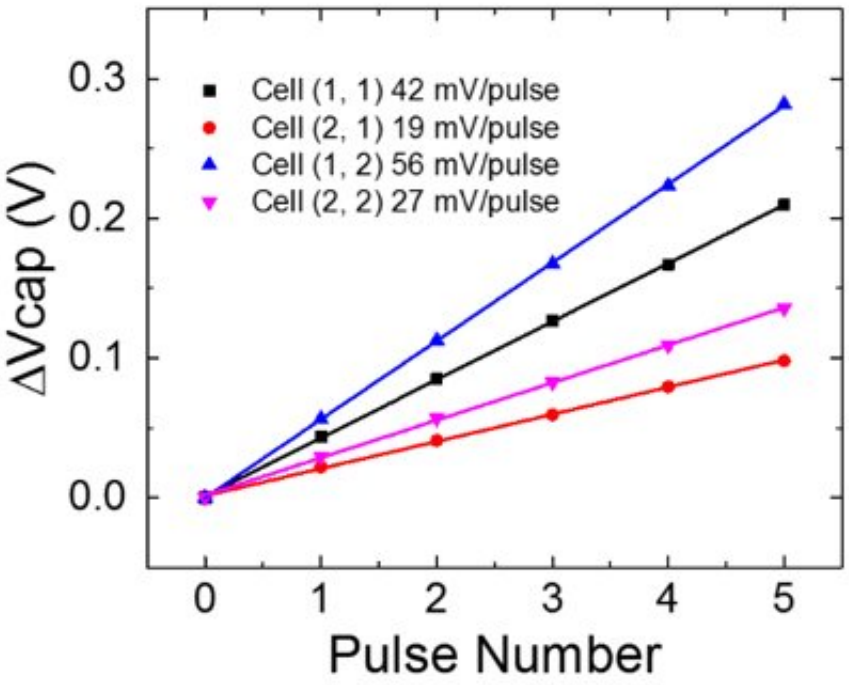

Figure 4. Parallel weight update on a 2×2 array. Credit: IBM

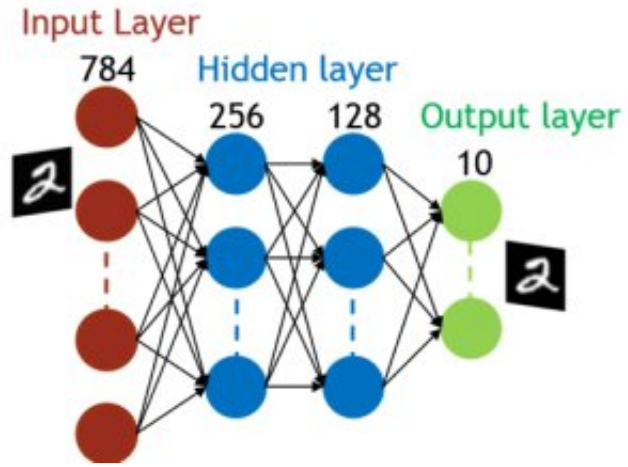

Figure 5. Simulated structure for fully connected neural network. Credit: IBM

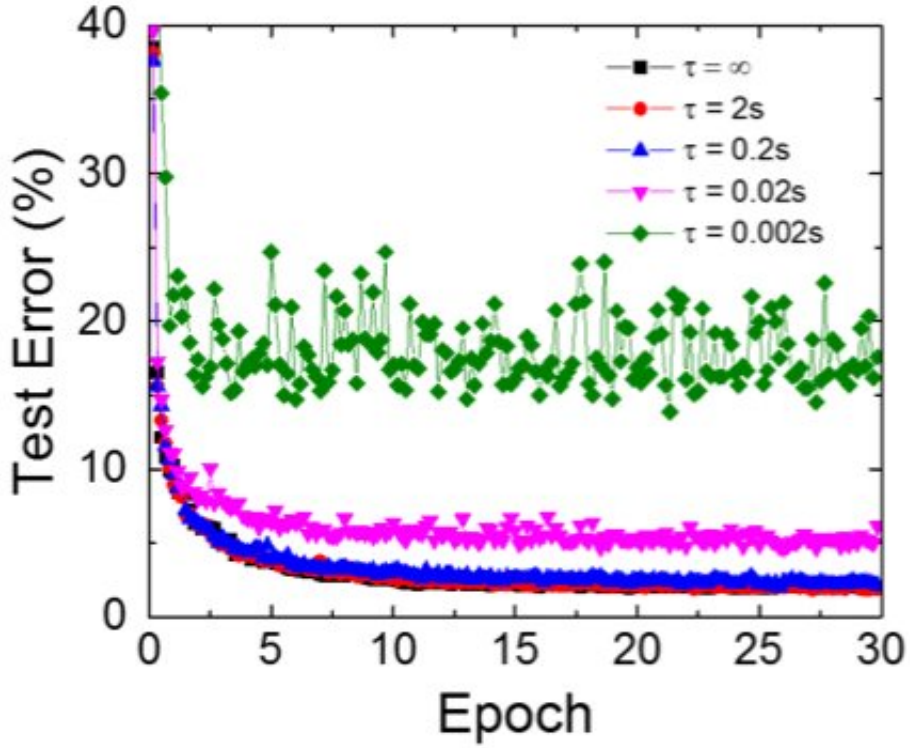

Figure 6. Simulated test error of MNIST data set, assuming weights decay continuously with different RC time constant τ, 200ns training cycle length. Credit: IBM

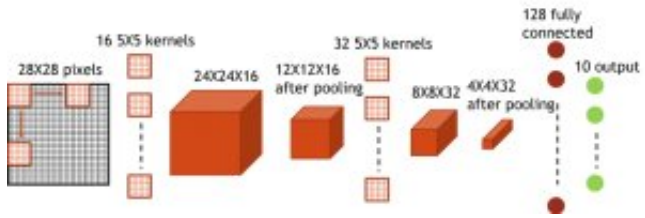

Figure 7. Simulated structure for convolutional neural network. Credit: IBM

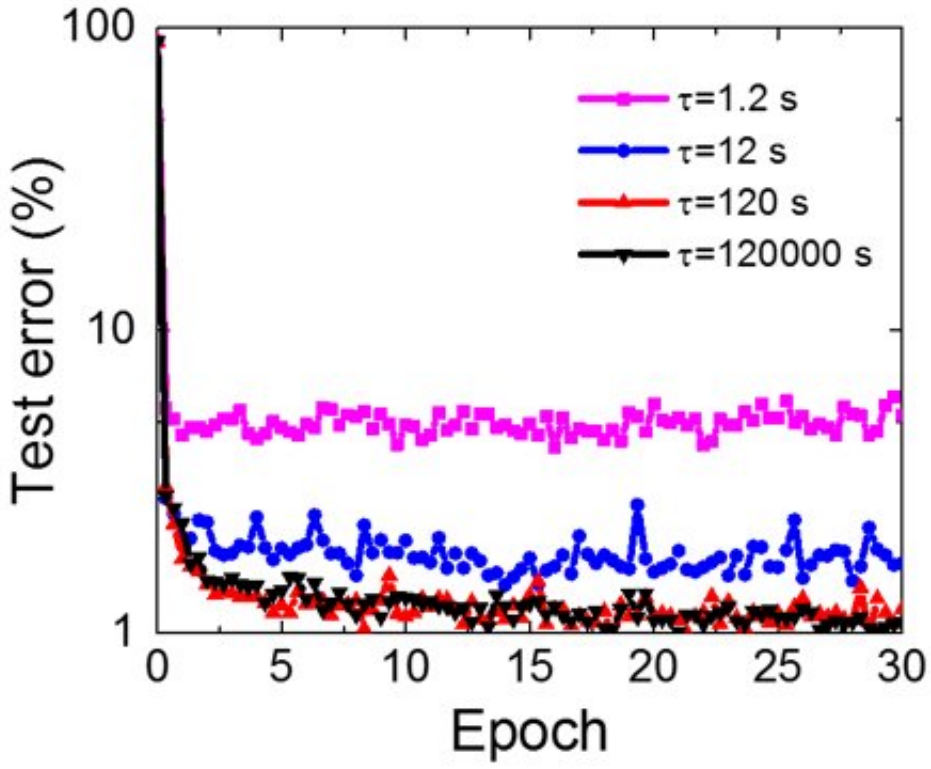

Figure 8. Simulated retention time requirement for this capacitor-based array to train convolutional neural network. Credit: IBM

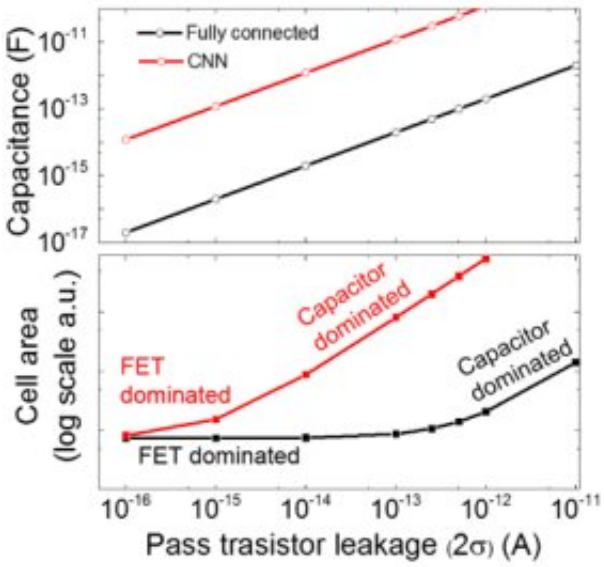

Figure 9. Scalability of this capacitor-based array as a function of leakage for both fully connected and convolutional neural networks. Credit: IBM

More information: References:

T. Gokmen, Front. Neurosci., vol. 10, 2016.
G. W. Burr, IEDM, 2014.
S. Kim, MWSCAS, 2017.
T. Gokmen, Front. Neurosci., Oct. 2017.
D. Chidambarrao, VLSI-TSA, 2003.
P-Y. Chen, IEEE TCAD, 2018.

Provided by IBM