# Capacitor-based Cross-point Array for Analog Neural Network with Record Symmetry and Linearity

Y. Li, S. Kim, X. Sun, P. Solomon, T. Gokmen, H. Tsai, S. Koswatta, Z. Ren,
R. Mo, C. C. Yeh, W. Haensch and E. Leobandung
IBM T. J. Watson Research Center, Yorktown Heights, NY10598, USA E-mail: yulongl@us.ibm.com

## Abstract

We report a capacitor-based cross-point array that can be used to train analog-based Deep Neural Networks (DNNs), fabricated with trench capacitors in 14nm technology. The fundamental DNN functionalities of multiply-accumulate and weight-update are demonstrated. We also demonstrate the best symmetry and linearity ever reported for an analog cross-point array system. For DNNs, the capacitor leakage does not impact learning accuracy even without any refresh cycle, as the weights are continuously updated during training. This makes capacitor an ideal candidate for neural network training. We also discuss the scalability of this array using optimized low-leakage DRAM technology.

## Introduction

Analog-based neural network (NN) accelerators have the potential to achieve orders of magnitude improvement in training time and energy consumption compared to conventional CPU/GPU systems [1,2]. Non-volatile memory (NVM) based cross-point arrays have achieved promising results for inference tasks [3,4]. 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 1000 update states) and symmetric (so that positive and negative updates (or pulses) balance on average) [2, 5]. To address these shortcomings, the concept of capacitor-based cross-point array has been proposed [6] but not demonstrated. In a capacitor, charge can be added or subtracted continuously if the number of electrons is high, so analog and symmetric weight update can be achieved. In this paper, we report such a capacitor-based array using trench capacitor in 14nm technology, demonstrating new records in update symmetry and linearity. We also investigate the feasibility of scaling this to large arrays for accelerating the training of large-scale DNNs.

## Array design

As shown in Fig. 1, a capacitor can serve as an analog memory, connected to the gate of a “readout” pFET. This capacitor is charged/discharged by two “current source” FETs, as controlled by two analog inverters and one digital inverter. During readout, the synaptic weight can be accessed by measuring the conductance of the readout FET. During weight update, the YW signal controls the analog inverters to drive the current sources for positive or negative updates. When these current sources are in saturation, the update currents become independent of capacitor voltage. The updated charge is determined by the length of YW pulse multiplies the current from the current source FET controlled by the XW_P/XW_N voltage. Fig. 2 shows the schematic of a 4×5 array. The layout of the array is shown in Fig. 3, and Fig. 4 shows the cross-sectional TEM image of the trench capacitor [7].

## Results and Discussions

Weight update is demonstrated in Fig. 5. with multiple positive and negative update pulses applied to a single cell. Device current was measured after each pulse from the readout FET showing clear modulation by the pulses. Fig. 6(a) and (b) show 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. 400 intermediate conductance states were achieved. This number is limited by the measurement parameters and not inherent to the circuit. Circuit simulations show the capacitor voltage (Vcap) range where expected asymmetry between positive and negative update is better than 10% (Fig. 7). Experimental results in Fig. 6 are consistent with simulation (Fig. 8). Fig. 9 compares the experimental non-linearity-update factors [8] for our capacitor-based analog synapse against other NVM technologies. To the best of our knowledge, the capacitor-based unit cell provides the best symmetry and linearity demonstrated to date. Cell to cell variation causes extra asymmetry (Fig. 10), and better than 15% asymmetry was achieved for most cells within the 4×5 array. Multiply-accumulate operation is demonstrated on a 2×2 array (Fig.11). Randomized weights were measured individually, and then different input voltages were applied to the two columns. The current was measured on each row showing errors <0.5%. Fig. 12 demonstrates parallel weight update on a 2×2 array. Different XW_P voltages were applied on each row, with different YW pulse widths on each column. No disturb was observed during parallel update and read.

The capacitor retention time was measured by first charging the Vcap to 0.8 V, and then observing the change in readout current after turning off the current sources, which are the dominating leakage path. The corresponding Vcap change is shown in Fig. 13, with retention time in the order of seconds. To illustrate the effect of retention time on training, Fig. 14 shows test error of a simulated 784×256×128×10 fully-connected network trained on the MNIST dataset by stochastic gradient descent and backpropagation, assuming weights constantly decaying with different RC time constant τ [6]. Assuming the training cycle length per layer (forward+backward+update) is 200ns [2], the penalty in training accuracy due to capacitor charge-loss becomes negligible when τ >0.2s ($10^6$ × the training cycle length). Since weights are implicitly updated continuously during training, refresh cycles are not necessary. Retention requirements for a convolutional neural network (CNN) [10] are larger, due to the weight sharing (reuse) in convolutional layers (Fig. 15). The scalability of this capacitor-based array as a function of leakage is shown in Fig. 16. Since NN training can tolerate ~5-10% bad cells, the leakage spec can be significantly relaxed (2 sigma vs. worse case in DRAM). With high leakage technology such as current 14nm logic technology, large capacitor is needed. On the other hand, DRAM technology with leakages of 1 fA/cell [11] requires less than 1 fF capacitance/cell. This can be achieved with a DRAM stack capacitor, and the cell area will be dominated by FETs in control circuitry. For CNN, larger capacitances would be needed (~100 fF), requiring multiple stack capacitors and larger cell area. The scalability to larger input and more layers needs further study.

Device-to-device variation (Fig. 17) and non-ideal device characteristics may affect training accuracy (Fig. 18 (a) to (c)). While training is less sensitive to readout variation (σ=7% in this work), the variation of current sources causes mismatches between programmed charge updates (σ=6% in this work) and finite output conductance of current sources causes Vcap-depended asymmetry (Fig. 7, <10% in this work). With all these non-ideality, estimated MNIST test accuracy of 97% is achievable based on this work.

## Conclusion

A capacitor-based cross-point array was demonstrated on 14nm technology using deep trench capacitors. We demonstrated record linear and symmetric weight update with 400 conductance states. This capacitor-based array, with continuous weight updates and no refresh cycle, has potential to be an ideal candidate for accelerating the training of deep neural networks.

978-1-5386-4218-4/18/$31.00 ©2018 IEEE

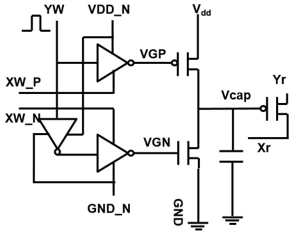

Fig. 1. Unit cell schematic of a capacitor-based cross-point array.

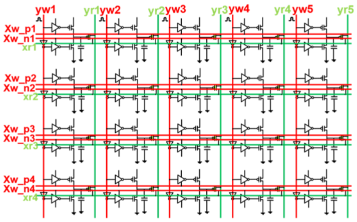

Fig. 2. Schematic of a 4 by 5 array.

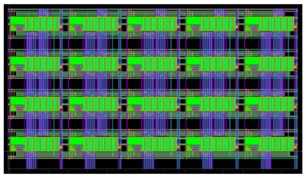

Fig. 3. Layout of the 4 by 5 array.

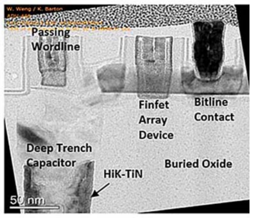

Fig. 4. TEM image of the deep trench capacitor [7].

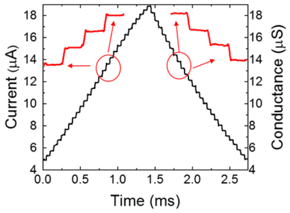

Fig. 5. Update of a single unit cell with multiple positive and negative update pulses. Pulse width: 500 ns, period: 50 µs.

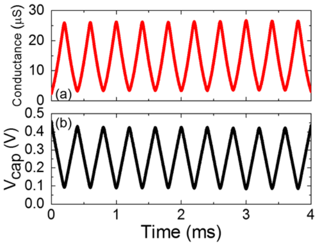

Fig. 6. (a) Experimental results for updating single-cell with 8000 pulses. (b) Corresponding capacitor voltage change. Pulse width 50 ns, period: 500 ns.

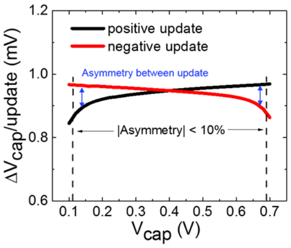

Fig. 7. Simulated asymmetry between positive and negative update as a function of Vcap. Asymmetry = (1-$\Delta Vcap_{positive}$/ $\Delta Vcap_{negative}$)

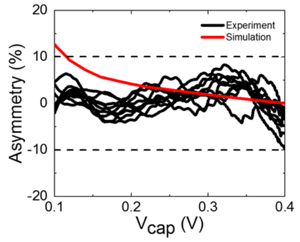

Fig. 8. Experimental update asymmetry and compared with simulation result.

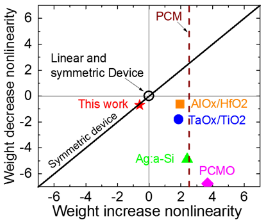

Fig. 9. Conductance non-linearity of this work compared with other NVM technologies [8]. When positive update nonlinearity equals to negative update nonlinearity, the update is symmetric. PCM (extracted from [3]) is shown as a dash line since its conductance can only be modulated gradually in one direction.

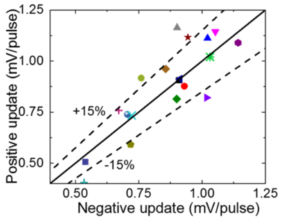

Fig. 10. Statistical results of updating single cells. Dashed line: +/-15% asymmetry. Different dots represent different unit cell in the 4×5 array.

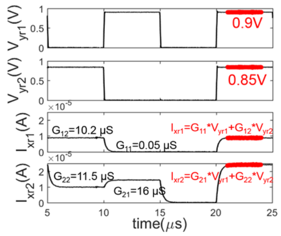

Fig. 11. Multiplication and add operation done by 2×2 array. Different input voltages were applied to different columns (yr1 and yr2). The current was measured on each row (xr1 and xr2). The measured error on each row was less than 0.5%.

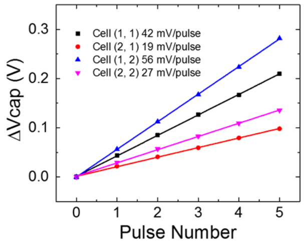

Fig. 12. Experimental parallel weight update for a 2×2 array with 5 pulses. The slope difference between cell 11 and cell 12 (also cell 21 and cell 22) is determined by different YW pulse widths in each column. The slope difference between cell 11 and cell 21 (also cell 12 and cell 22) is determined by different XW_P voltages on each row. Results match with simulation.

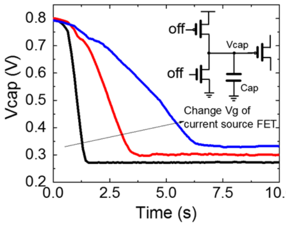

Fig. 13. Retention measurement. The retention time is in the order of seconds and can be maximized by turning the current sources deeply off, which are the dominating leakage path.

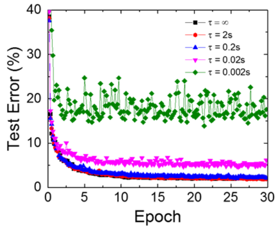

Fig. 14. Simulated test error of MNIST data set, assuming weights decay continuously with different RC time constant τ, 200ns training cycle length.

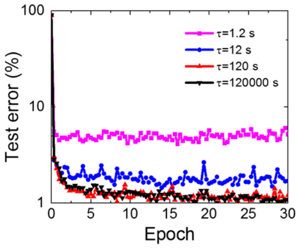

Fig. 15. Simulated retention time requirement for capacitor-based array to train convolutional neural network. 200ns training cycle length.

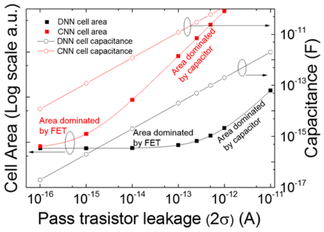

Fig. 16. Trend of required unit cell capacitance and area as a function of leakage current. When leakage is small, the area is limited by FETs in the control circuitry. Assume 200ns training cycle length, τ=0.2 s for DNN and τ=120 s for CNN. Vcap range 1V.

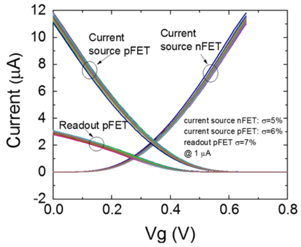

Fig. 17. Measured device variation of current source FETs and readout FETs in the 4×5 array.

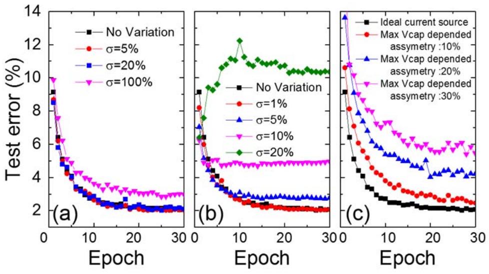

Fig. 18. Simulated test error of MNIST data set with (a) different amount of readout FET variation, σ=7% in this work, (b) different amount of current source variation σ=6% in this work, (c) non-ideal current sources, Vcap depended asymmetry <10 % in this work. τ =0.2s, 200ns training cycle length in simulation.

**Reference:**

[1] C. Merkel, *Computer*, vol. 49, pp. 56-64, 2016. [2] T. Gokmen, *Frontiers in neuroscience*, vol. 10, 2016. [3] G. W. Burr, *IEDM*, 2014. [4] S. Yu, *IEEE TED*, vol. 58, pp. 2729-2737, 2011. [5] S. Agarwal, *IJCNN*, pp. 929-938 2016. [6] S. Kim, MWSCAS, 2017. [7] G. Freeman, IEEE JSSC 2016. [8] P-Y. Chen, *ICCAD*, pp. 194-199 2015. [9] J. Liu, *ISCA*, pp. 60-71, 2013. [10] T. Gokmen, arXiv:1705.08014. [11] D. Chidambarrao, *VLSI-TSA*, 2003.