SAND2022-17053C

Analog Neural Network Inference Accuracy in One-Selector One-Resistor Memory Arrays

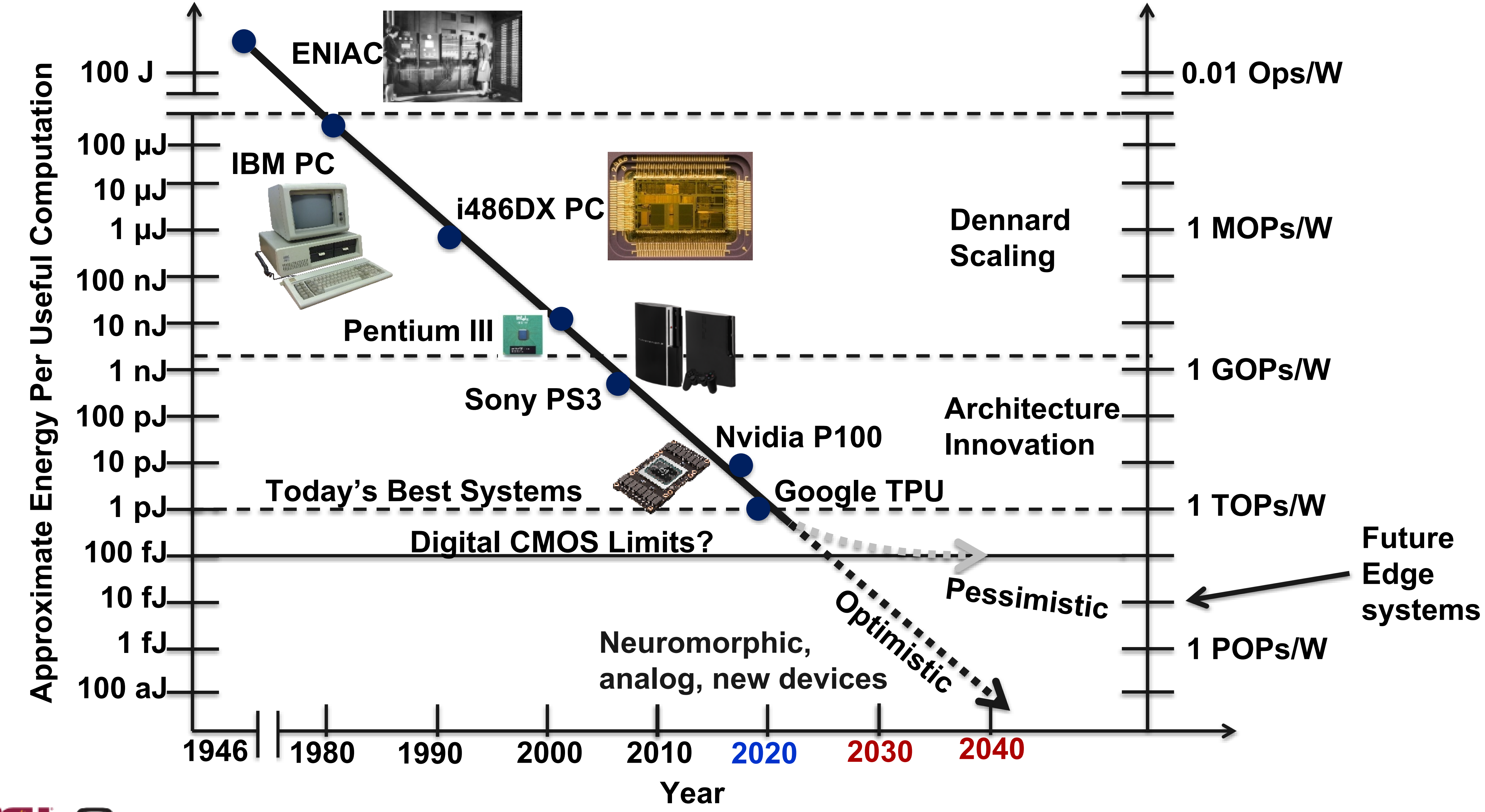
Joshua E. Kim[1*], T. Patrick Xiao[2], Christopher H. Bennett[2], Donald Wilson[1], Matthew Spear[1], Maximilian Siath[1], Ben Feinberg[2], Sapan Agarwal[2], Matthew J. Marinella[1]

1 – School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe AZ
2 – Sandia National Laboratories, Albuquerque, NM
jekim7@asu.edu

December 8, 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Sec

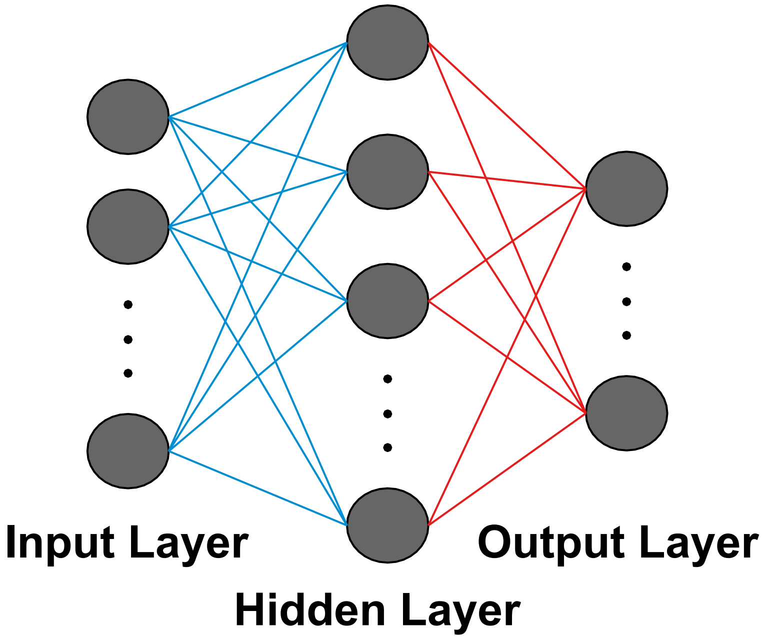
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

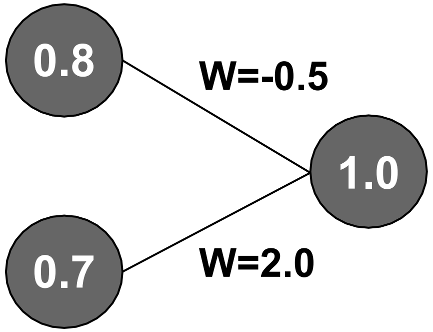


Adapted from: **Marinella and Agarwal, Nature Electronics 2, 437, 2019**

Neural Networks and Analog Accelerators

Neural Networks Basics:

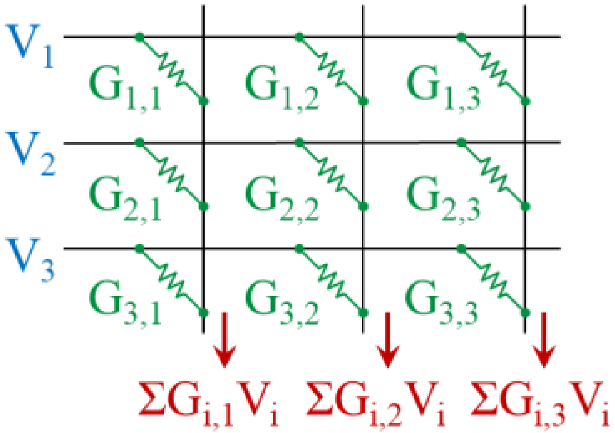




Matrix-vector multiplication:

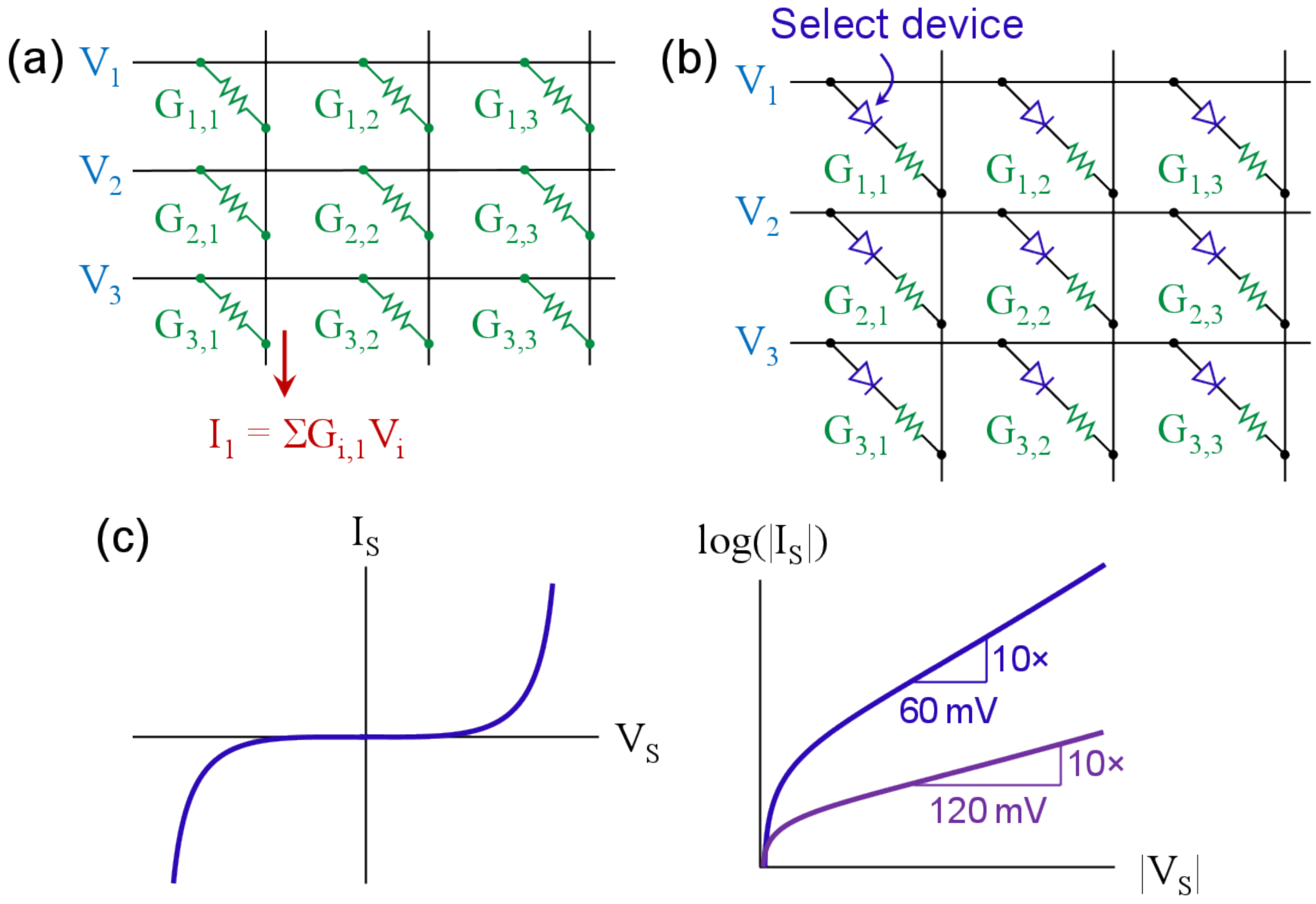
Ax

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^T \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix} = \begin{bmatrix} \Sigma A_{i,1}x_i & \Sigma A_{i,2}x_i & \Sigma A_{i,3}x_i \end{bmatrix}$$



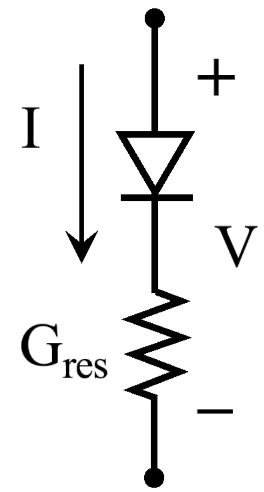
Select Device Necessary for Write Operation

- Many devices “half-selected” during write operation
- Draws current ruining power efficiency
- Select devices are a solution
- 1T1R memory arrays solve this better than 1S1R arrays do
- However, 1T1R is less compact and may be incompatible with back-end-of-line integration of dense memory arrays

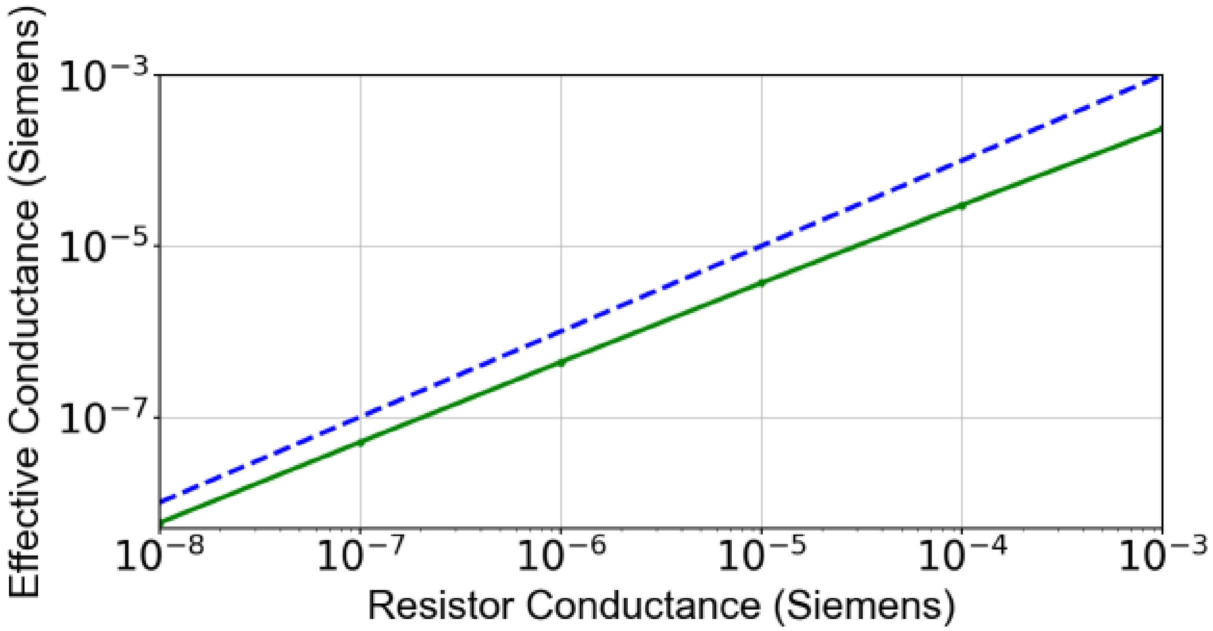


Effective Conductance of 1S1R

- Some voltage is dropped across the select device
- To achieve correct output current, a higher voltage must be applied
- This additional voltage changes as a function of the conductance

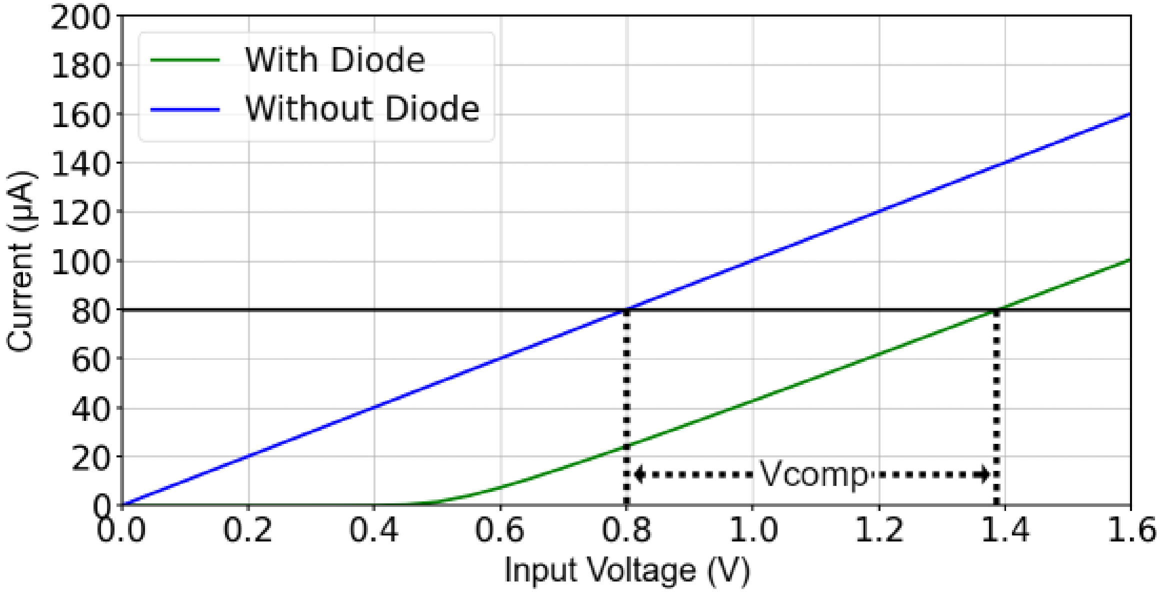


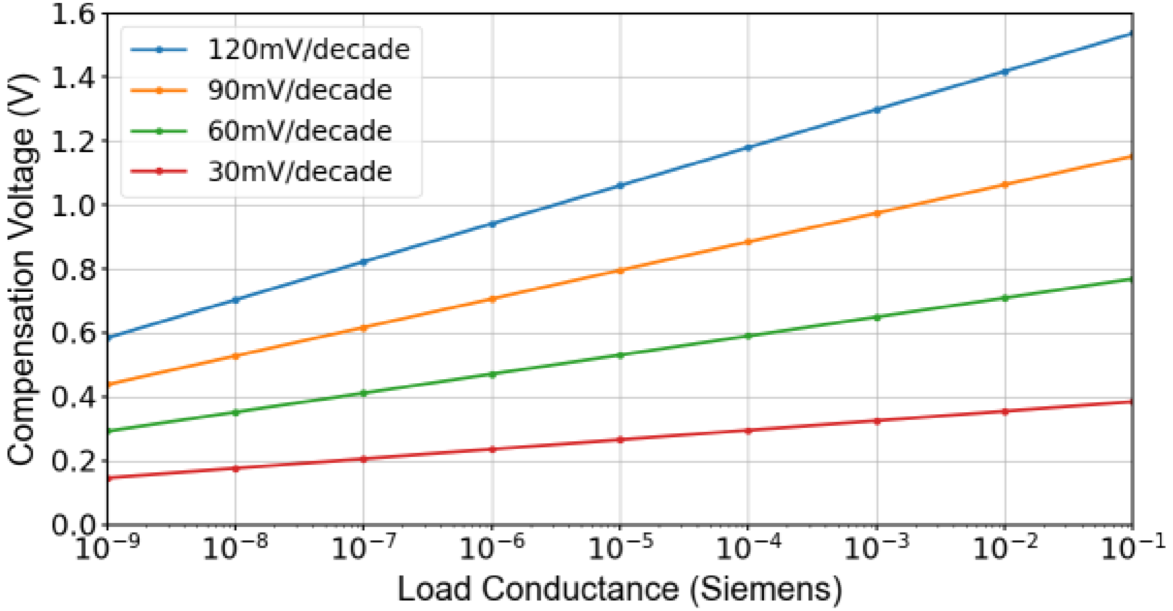
$G_{eff} = I / V$



The blue dashed line represents y=x or the conductance of the resistor by itself.

Compensation Voltage





- Compensation voltage can be found perfectly for a single cell
- Function of select device steepness, resistor conductance, and “nominal” voltage
- Nominal voltage is the voltage across the resistor necessary to output the correct current
- Not practical to implement individual compensation for each cell
- Pick one compensation voltage for entire array with goal of minimizing error

Weighted vs Unweighted Compensation

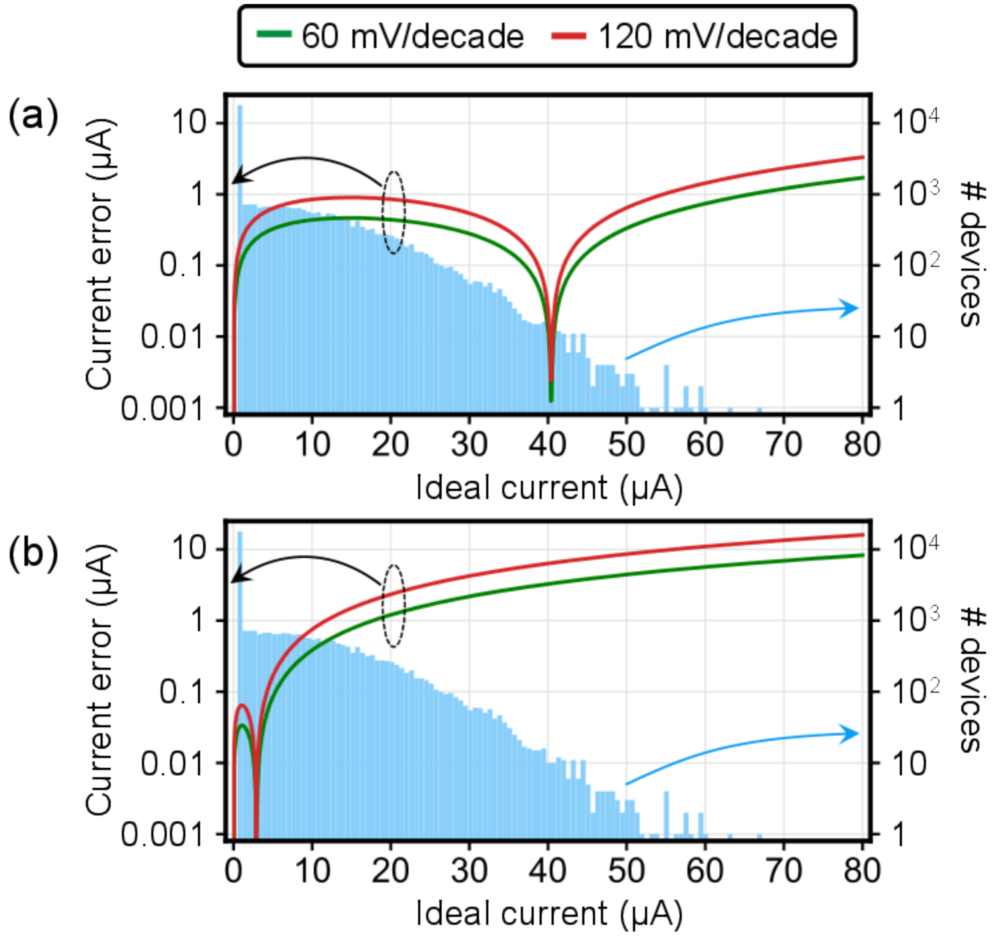
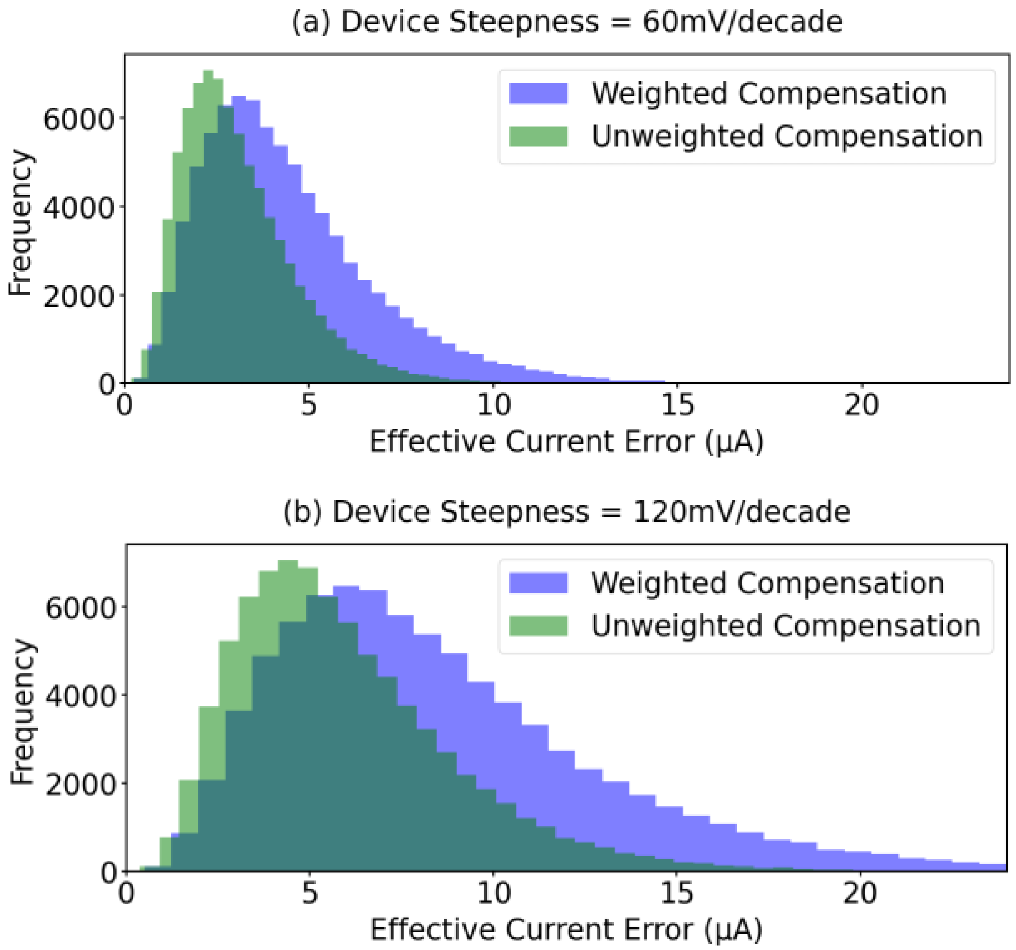
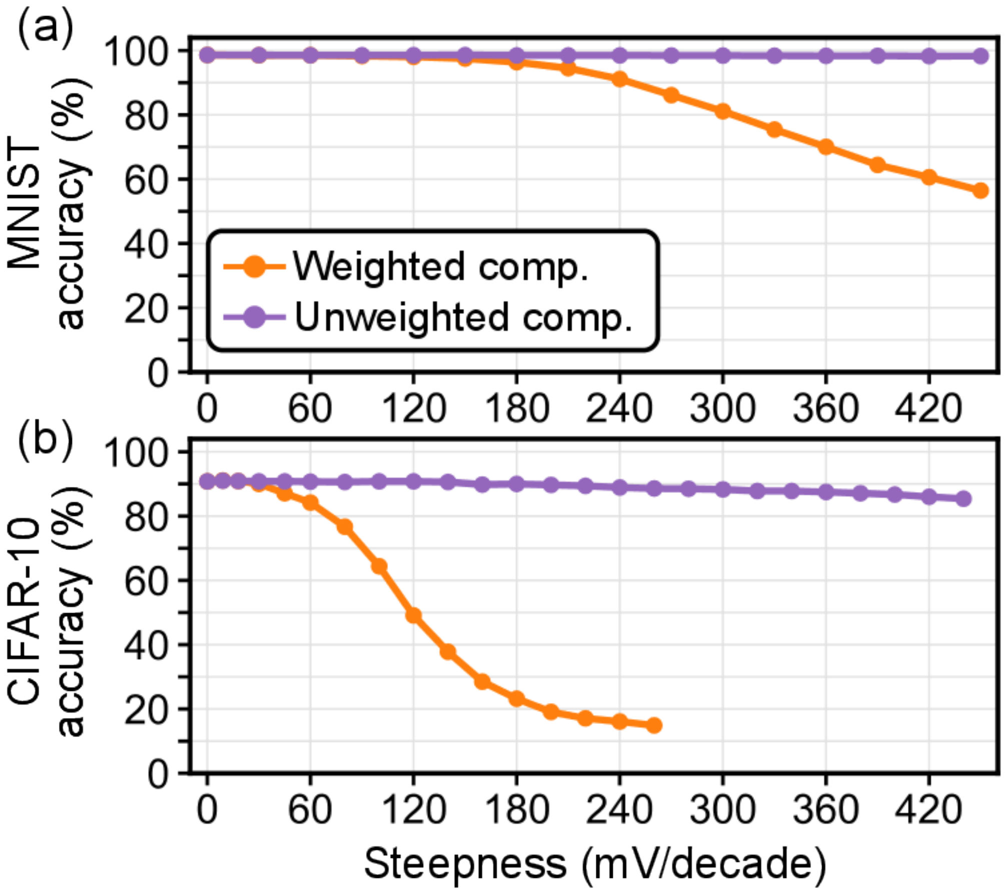


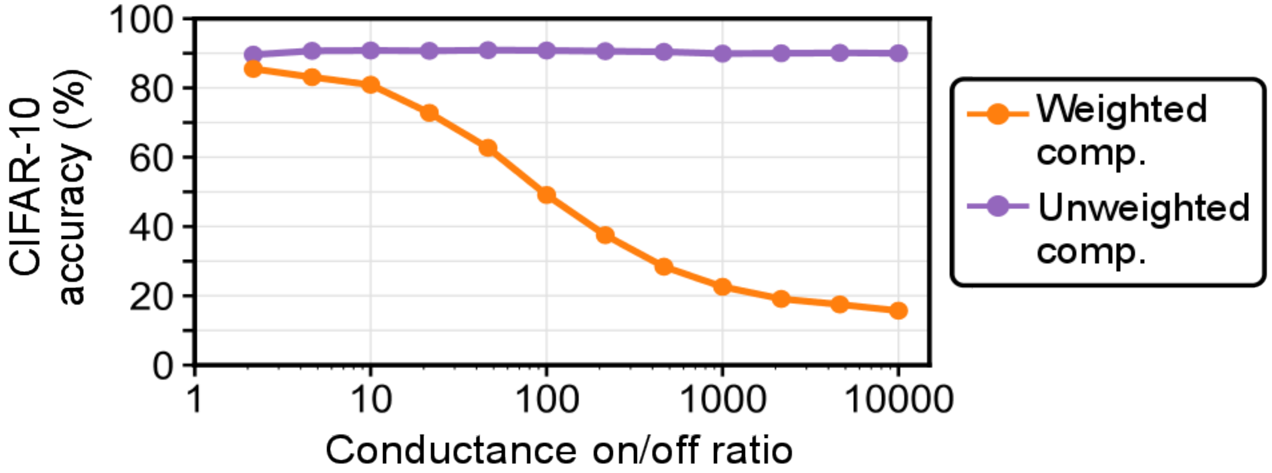
Fig. 5. (a) unweighted compensation (b) weighted compensation



ROSS SIM Results



- Unweighted compensation accuracy on both MNIST and CIFAR-10 is minimally affected by device steepness and conductance on/off ratio
- Assuming a realistic 60mV/dec, unweighted compensation achieves an accuracy of 90.29%
- This is only 0.44% below ideal floating-point results



Conclusions

- Non-volatile memory arrays with a 1S1R topology can be made compatible with accurate neural network inference if the errors induced by the select device are appropriately compensated
- Showed that a single compensation voltage, applied uniformly across the entire system, can effectively reduce these errors to enable accurate inference
- With this compensation, a CIFAR-10 accuracy that is within 0.44% of the floating-point digital result can be achieved using a realistic selector with 60 mV/decade steepness
- The accuracy is insensitive to the memory device On/Off ratio
- These results are promising for the use of dense 1S1R arrays for analog neural network inference
- Future work should investigate how selector-induced errors interact with other sources of analog error, and how these different errors can be mitigated together
  - These include parasitic resistance, process variations, other non-idealities, etc.

Acknowledgements

- This work was funded by:

- Sandia National Laboratories Laboratory Directed Research and Development (LDRD)