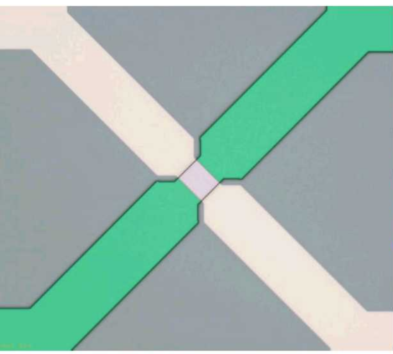
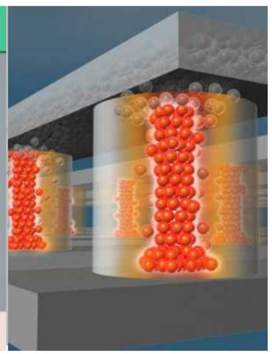
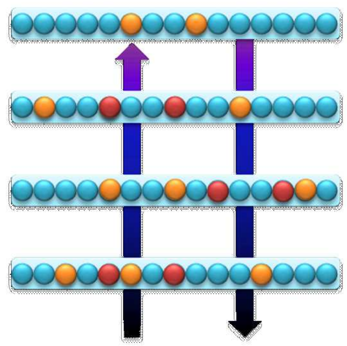
SAND2019-6931C





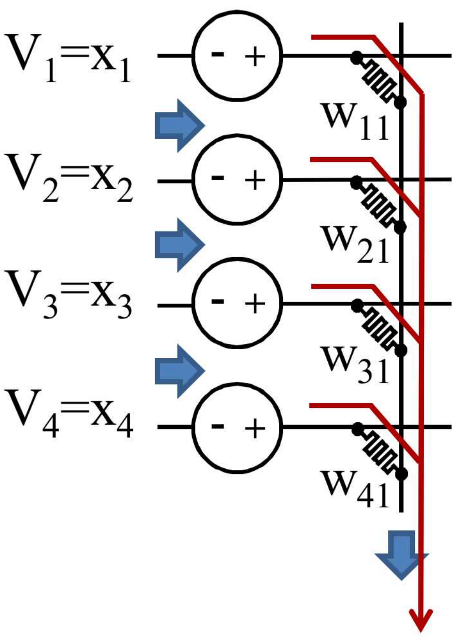


Designing and Modelling Analog Neural Network Training Accelerators

Sapan Agarwal, Robin B. Jacobs-Gedrim, Christopher Bennett, Alex Hsia, Michael S. Van Heukelom, David Hughart, Elliot Fuller, Yiyang Li, A. Alec Talin, Matthew J. Marinella
Sandia National Laboratories

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 Security Administration under contract DE-NA0003525.

Need to Use Analog to Efficiently Discard Precision



Sum 1024 8 bit weights X 8 bit inputs:

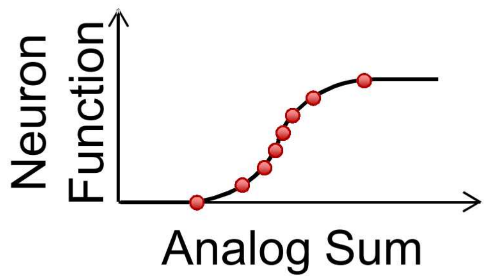
- Result has 26 bits of information!
- A 26 bit ADC would eliminate any analog advantage!

The sum can be done at full precision in analog, but a lower precision approximation is needed when digitizing

- i.e. digitize only 8 bits or fewer

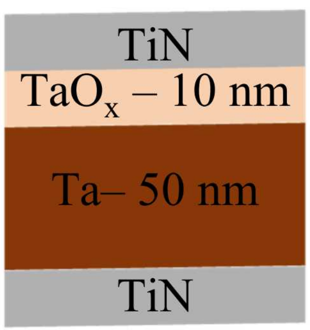
To get the highest 8 bits of information, digital would need to keep a 26 bit intermediate result

Can design an ADC to choose non uniform values to digitize

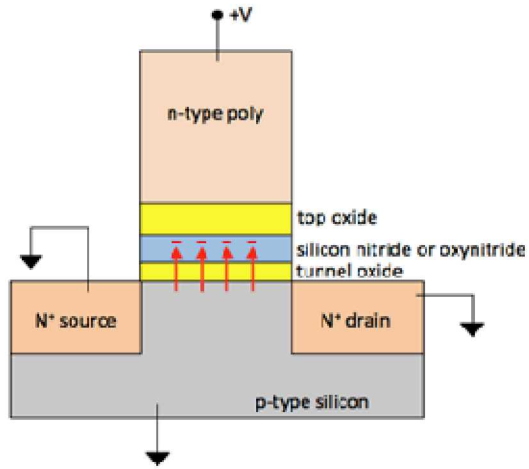


Compare Analog Devices

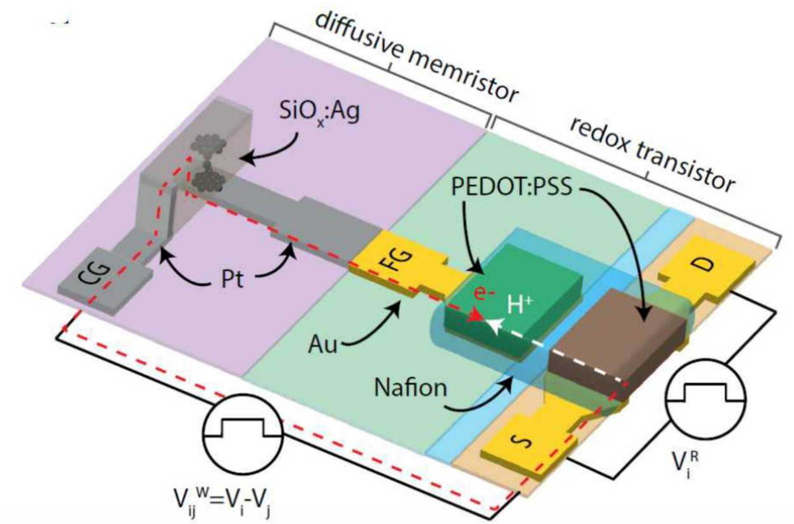
ReRAM



SONOS
Silicon-Oxygen-Nitrogen-Oxygen-Silicon



Ionic Floating-Gate Memory

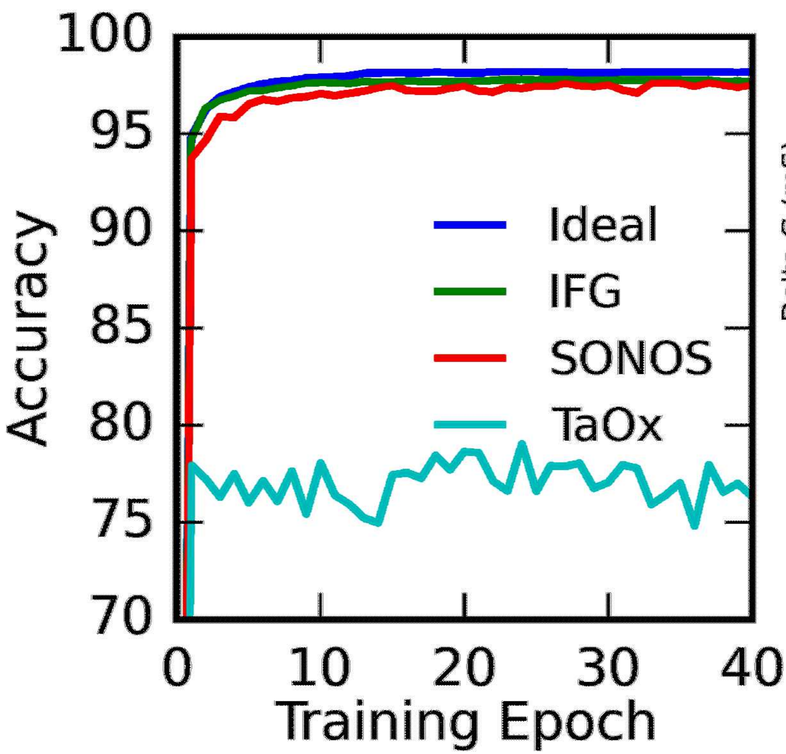


R. B. Jacobs-Gedrim *et al.*, "Impact of Linearity and Write Noise of Analog Resistive Memory Devices in a Neural Algorithm Accelerator," IEEE International Conference on Rebooting Computing (ICRC) Washington, DC, November 2017.

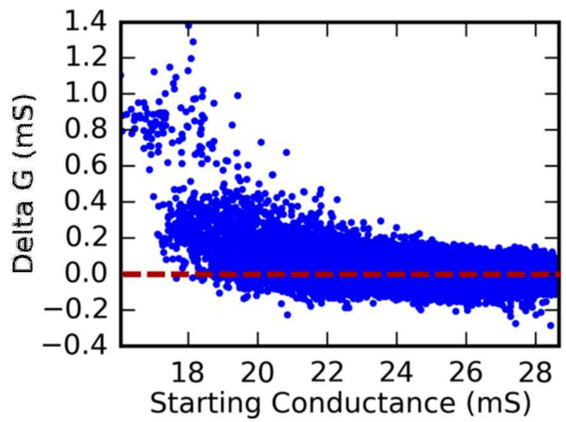
S. Agarwal *et al.*, "Using Floating Gate Memory to Train Ideal Accuracy Neural Networks," *IEEE Journal of Exploratory Solid-State Computational Devices and Circuits,* 2019

E. J. Fuller et al., "Li-Ion Synaptic Transistor for Low Power Analog Computing," *Advanced Materials*, vol. 29, no. 4, p. 1604310, 2017
E. J. Fuller et al., under review

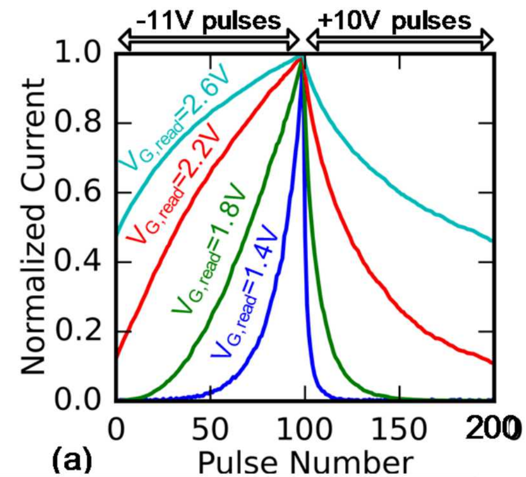
Three Terminal Devices Tend to Have Higher Accuracy

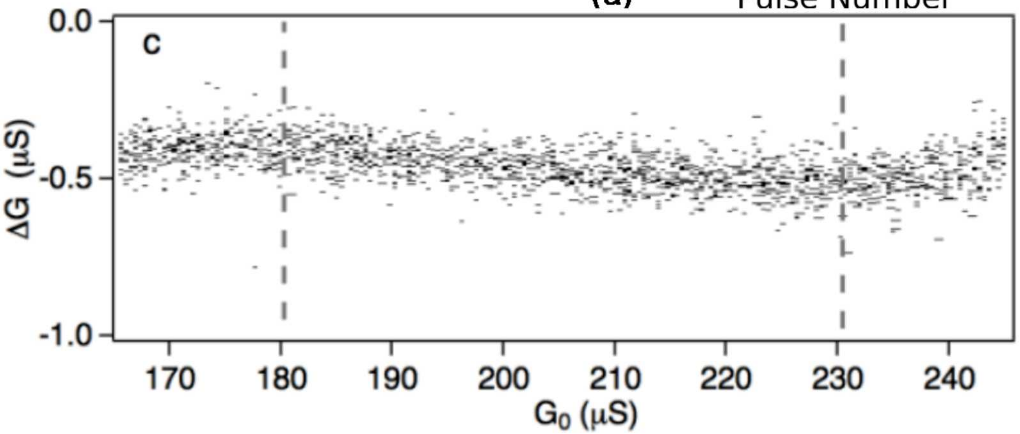


ReRAM



SONOS

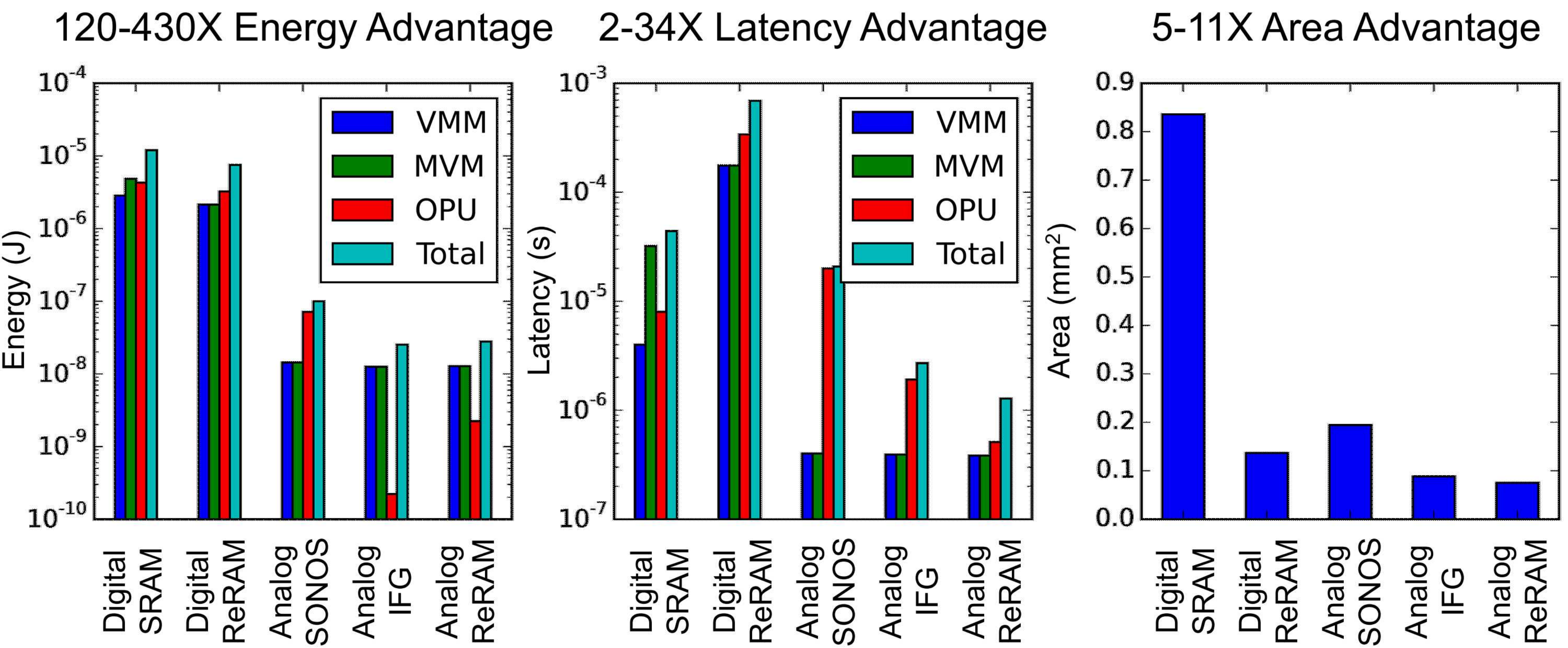




Ionic Floating-Gate

Compare Architectural Advantages

Sandia National Laboratories



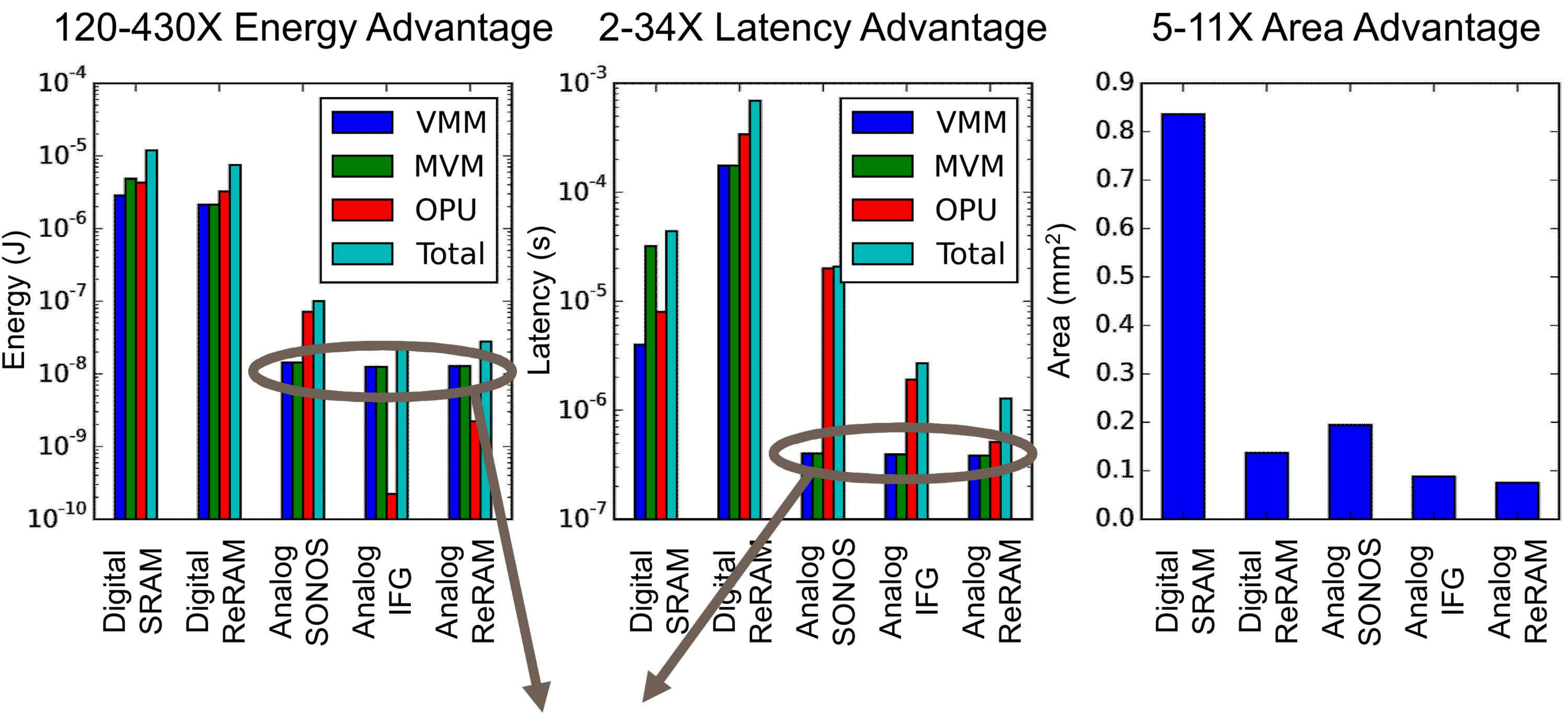
1024 x1024 = 1M array operations, sum over 1 training cycle, 3 operations:

- Vector Matrix Multiply - Matrix Vector Multiply - Outer Product Update

Used a commercial 14/16 nm PDK

***Requires 100 MΩ on state devices

Compare Architectural Advantages: Vector Matrix Multiply

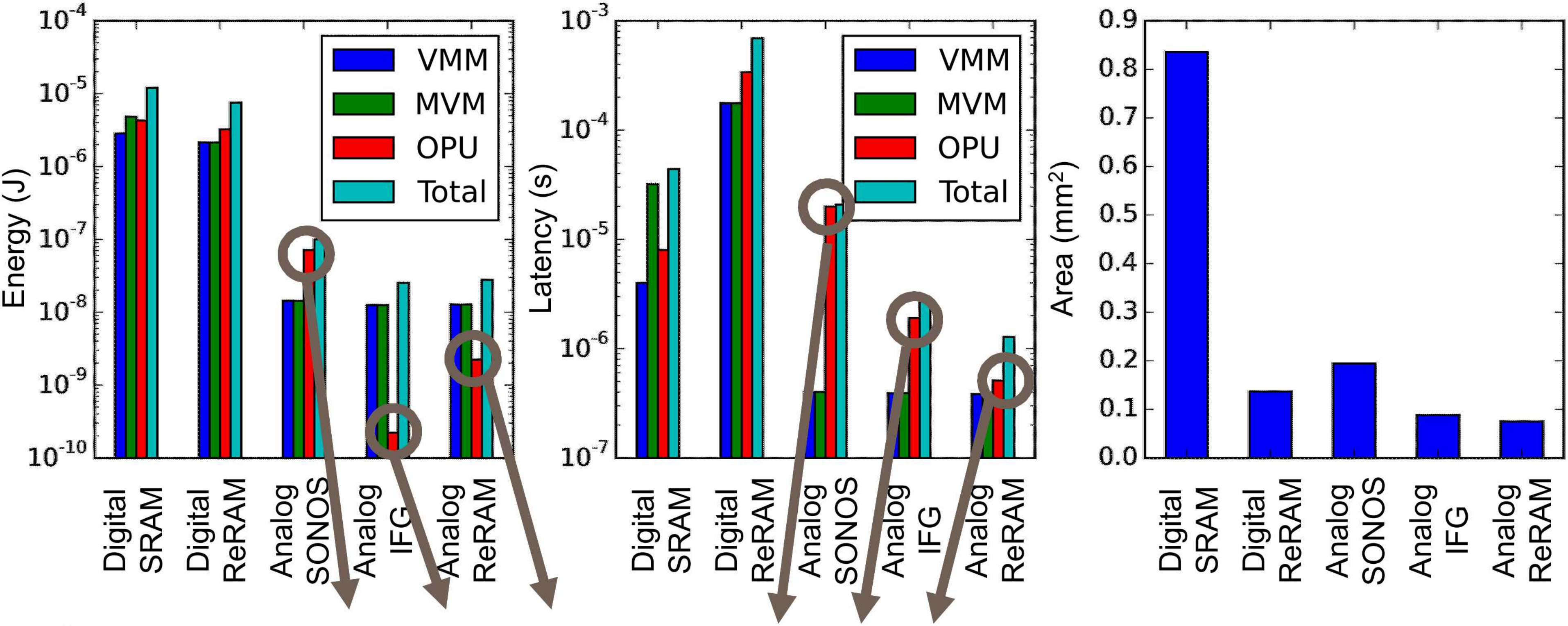


All Analog Vector Matrix Multiply and Matrix Vector Multiply have same energy and latency

- Entirely dominated by ADC, device properties irrelevant

Compare Architectural Advantages: Outer Product Update

120-430X Energy Advantage 2-34X Latency Advantage 5-11X Area Advantage

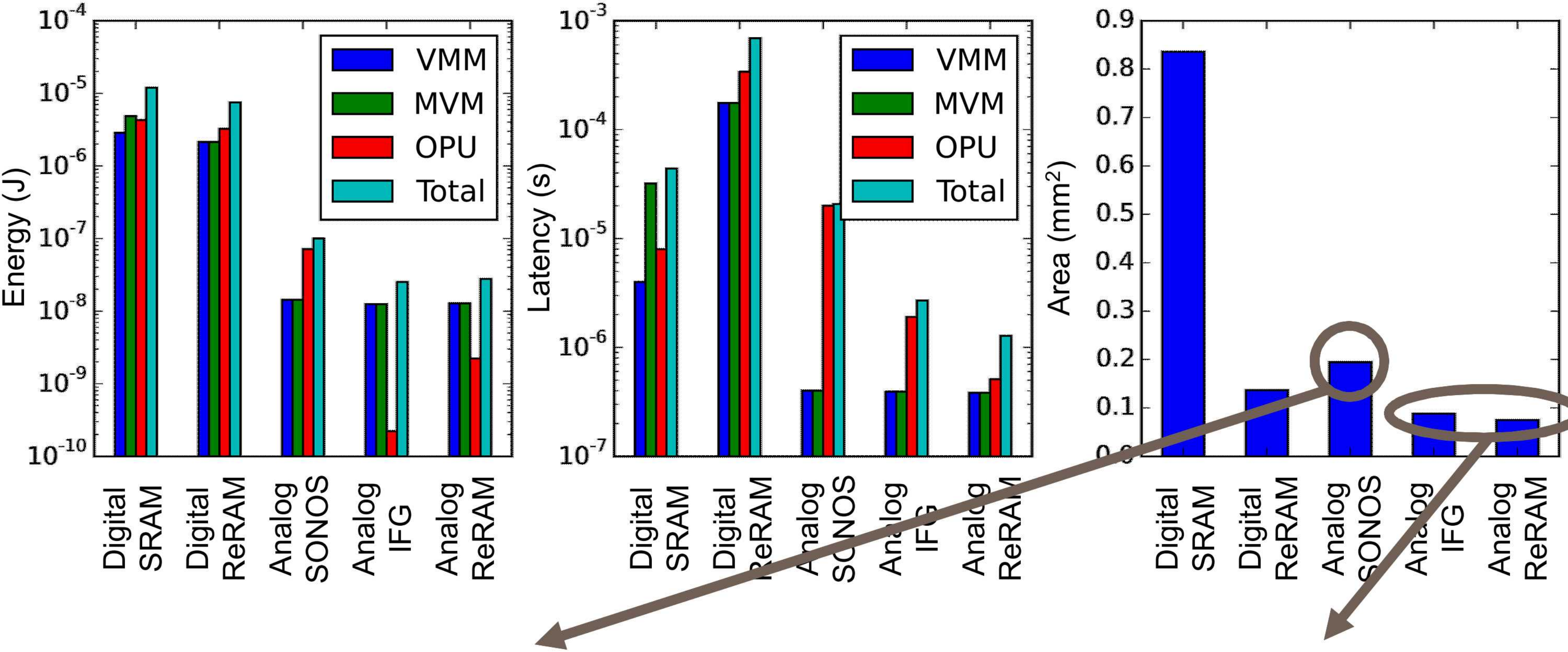


Outer Product Update is device dependent

- SONOS has slow write (~1 ms) and high write voltage (11V)
- IFG and ReRAM write energy negligible compared to VMM
- IFG has extra delay over ReRAM for access device to turn off

Compare Architectural Advantages: Area

120-430X Energy Advantage   2-34X Latency Advantage   5-11X Area Advantage

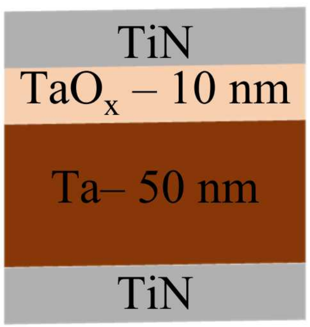


SONOS area cost reasonable, roughly doubles area

IFG and ReRAM go over transistors, area dominated by ADC and DAC

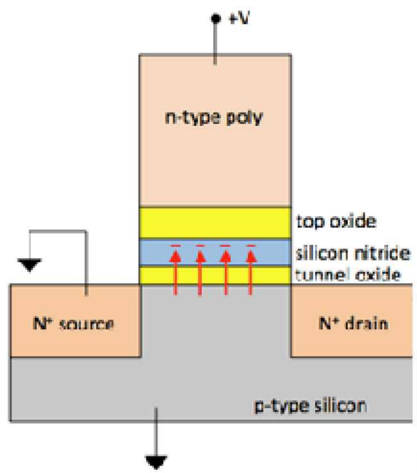
Analog Devices Summary for Training

ReRAM



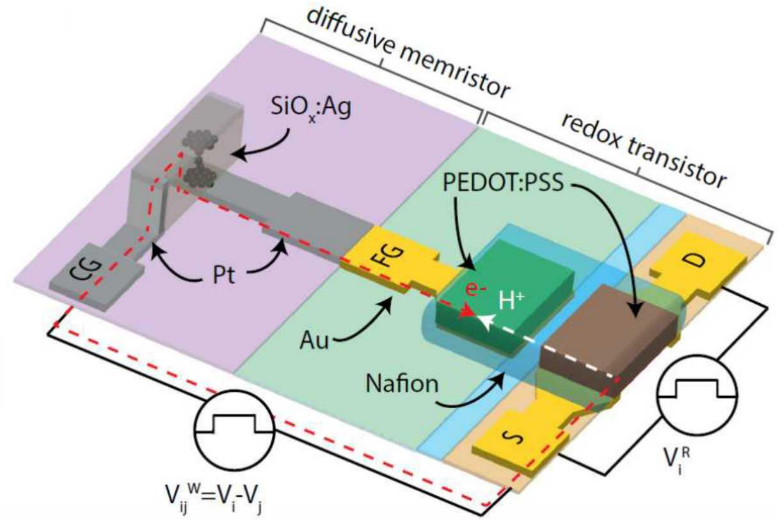
- Large Energy/Area/Latency advantage over digital
- Accuracy not good enough
- Back end of line compatible
- Under commercial development

SONOS
Silicon-Oxygen-Nitrogen-Oxygen-Silicon



- Moderate Energy/Area/Latency advantages over digital
- High Accuracy
- Commercially available
- Need to prove endurance and device to device variability

Ionic
Floating-Gate Memory



- Large Energy/Area/Latency advantages over digital
- High Accuracy
- Not clear how to integrate
- Has retention challenges