RELATIVE EFFICIENCY OF MEMRISTIVE AND DIGITAL NEUROMORPHIC CROSSBARS

Beyond CMOS 2017 Workshop

Christopher Krieger
David Mountain
Mark McLean

Neuromorphic Computation Research Group

The Laboratory for Physical Sciences

Multiply-accumulate is key to today's machine learning

- Bulk of computation is multiply-accumulate (MACC)[†]
  - AlexNet – 2.3 Million weights, 666 million MACCs
  - VGG16 – 14.7 Million weights, 15.3 billion MACCs
- If we can do multiply-accumulate operations efficiently, we can improve machine learning efficiency
- Many efforts to improve this key operation
  - Google TPU
  - Nvidia tensor units
  - Stanford EIE, MIT Eyeriss
  - Memristor crossbar based compute engines

[†] Chen at al. "**Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks." ISSCC 2016.**

Outline

- Memristors as computational elements in neural networks
- Are memristors better than CMOS?
  - Performance / power efficiency experiment
- Memristors in a broader system context

Neurons Using Ohm's Law & Kirchhoff's Current Law

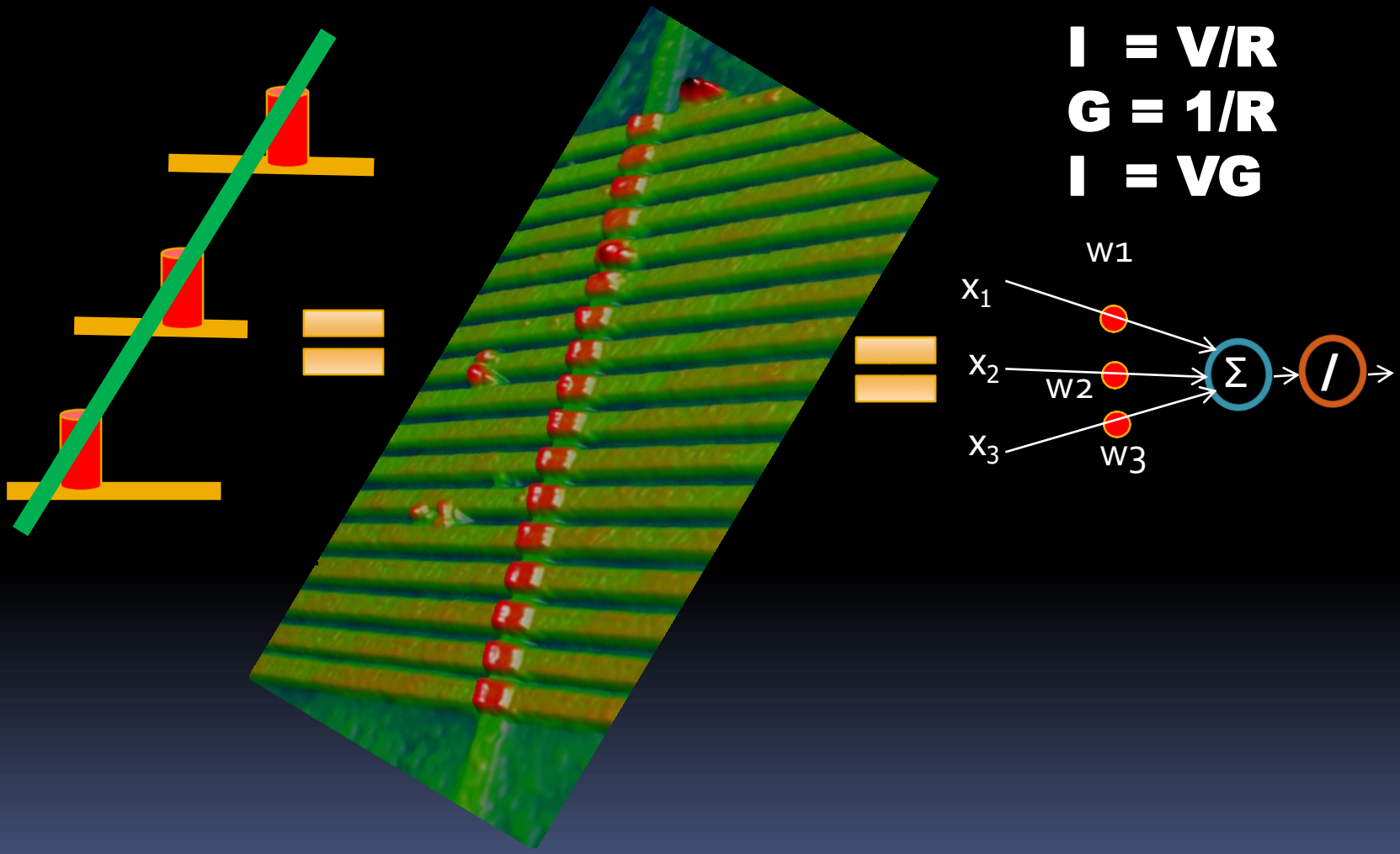
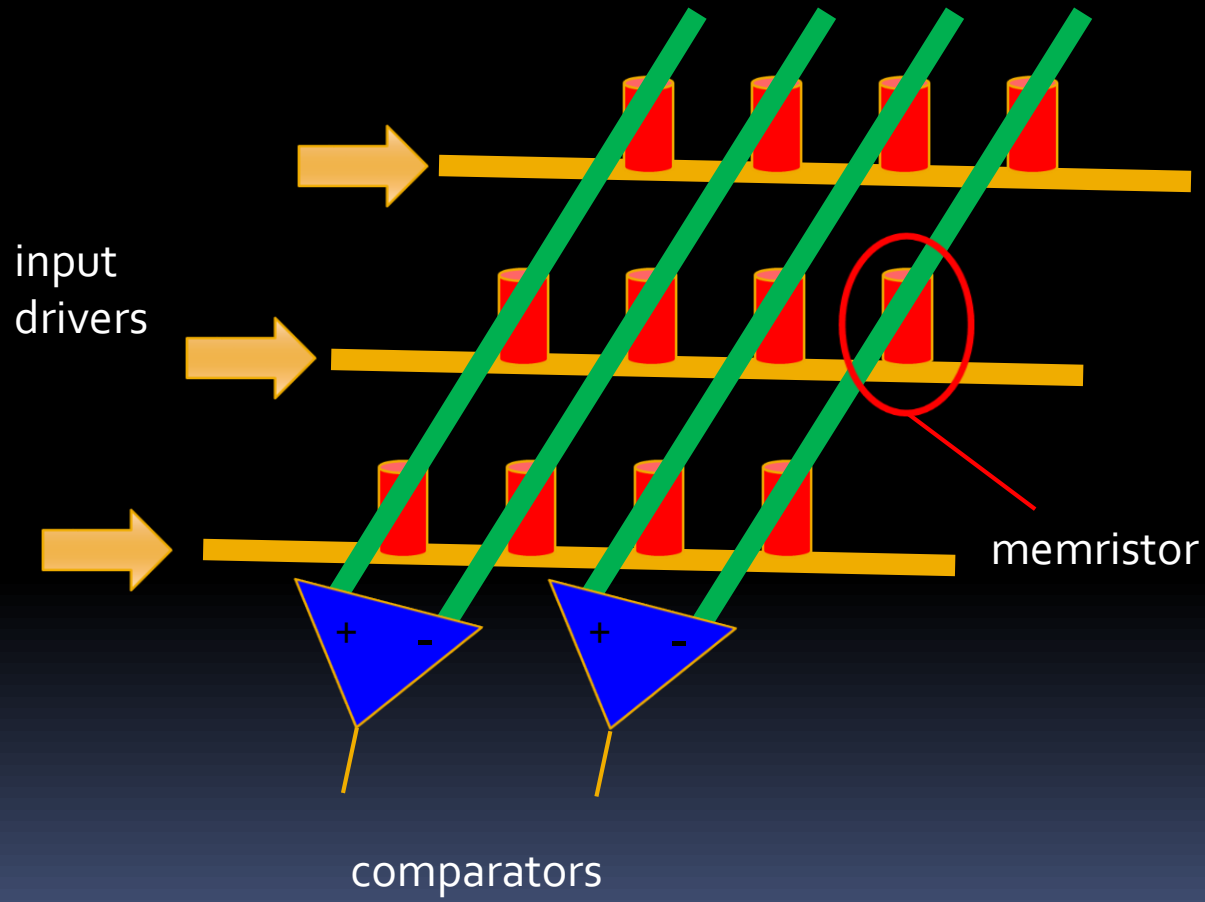


Image: Stan Williams, HP Labs via arstechnica.com

Conceptual 3x4 Crossbar
2 Neurons Shown



Claims about Memristor Systems

- "… can achieve acceleration factors of 30,000x compared to state-of-the-art microprocessors…" – *Frontiers in Neuroscience* 2016
- "… up to 5 orders of magnitude more energy efficiency over GPGPUs for deep neural network processing." – University of Dayton, 2016

Gokman et al, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Frontiers in Neuroscience, vol 10, 2016.*

Hasan and Taha, "A Reconfigurable Low Power High Throughput Architecture for Deep Network Training," ArXiv, March 2016.

Outline

- Memristors as computational elements in neural networks
- **Are memristors better than CMOS?**
  - Performance / power efficiency experiment
- Memristors in a broader system context

Determine Benefit of Memristors

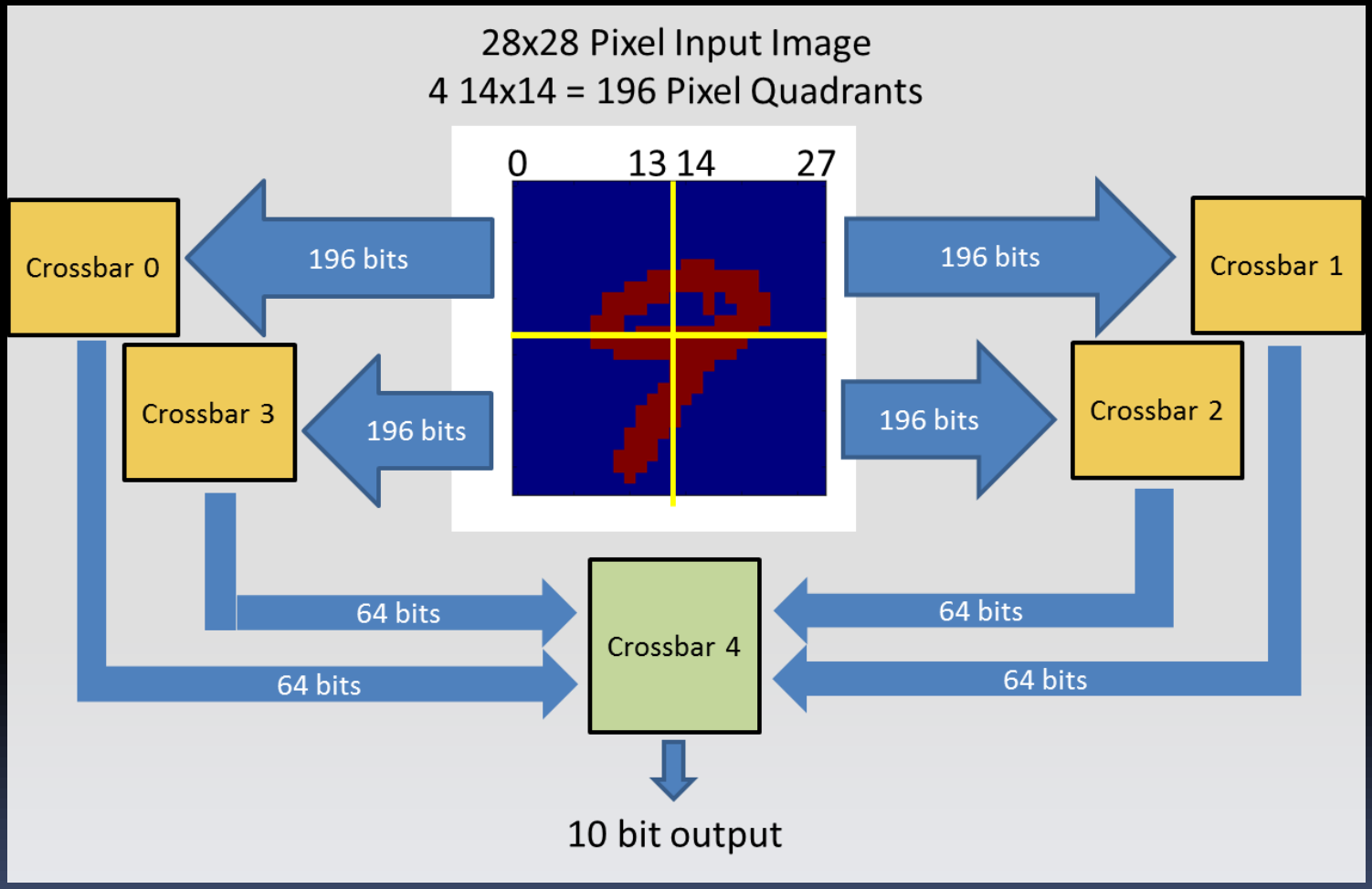
- Goal was to understand *benefit of memristors*
- Isolate the memristors, hold all else constant
- Desired studied would include two systems:
  - System A does computation using memristors
  - System B does computation using digital logic with standard vanilla CMOS transistors

All other design factors identical between two systems

Experiment Requirements

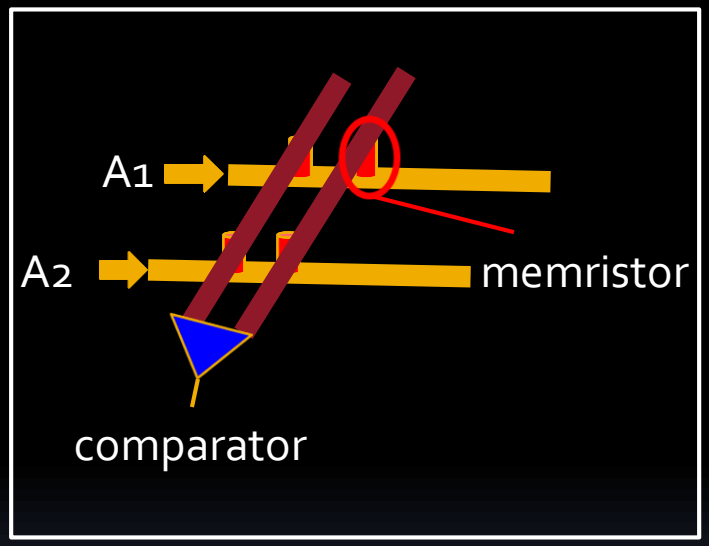
- In-depth simulation
  - Low level SPICE simulation, not spreadsheet
- Not cobbled together from various sources
  - All in a single technology
- Realistic memristors
  - Properties extracted from manufactured memristors
- Running a representative application
  - All system effects of a real task considered

MNIST Neural Network Design

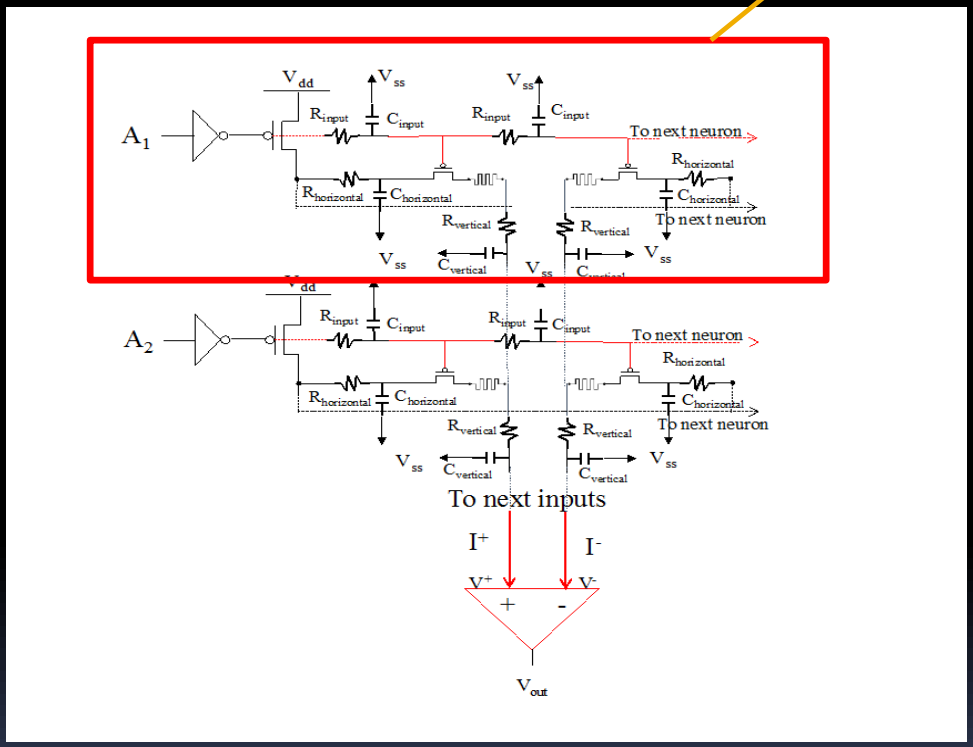


A simple neural network to do handwritten digit recognition (MNIST)

Memristor Design – Single Neuron 256 Input, 64 Output

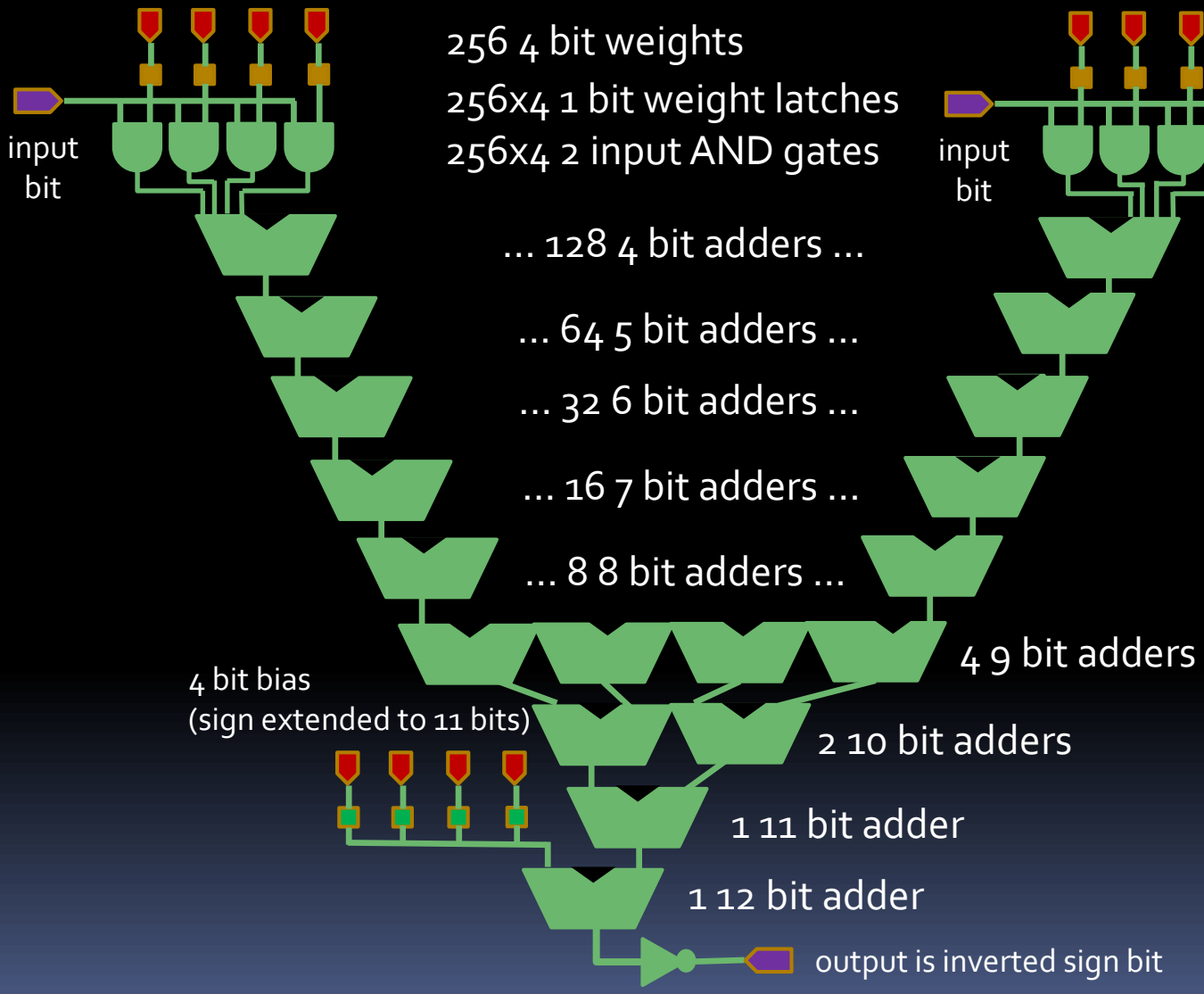


Replicated for every input



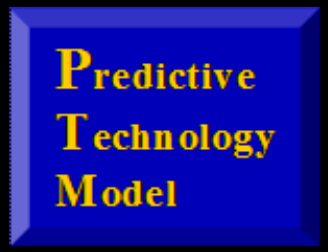
Design Uses ~600 Transistors and 512 Memristors

Digital Design – Single Neuron



About 30,000 transistors per neuron

Experiments



- Detailed simulations of all designs
  - ***Thousands of machine hours of low level SPICE simulations***
    - Used Sandia's Xyce parallel simulator
    - 45 nm CMOS process file from ASU PTM collection
    - Yakopcic-Mountain memristor model
    - Ran on 80 cores across 4 nodes in cluster
    - Longest runtime was 26 days
  - Analog was split into 5 jobs, 1 per crossbar
  - Digital was split into 17 jobs, each a 16 neuron sub-crossbar

First circuit level simulation of a complete application implemented in a memristor based architecture

Results – Memristors Compared with Digital CMOS Circuit

- 40 times smaller
  - 44k vs 1.7M μm^2
- 4x more energy efficient
  - 31.9 pJ/pattern vs 132 pJ/pattern
- 1/4th Performance (Classifications / Second)
  - 6 ns vs 1.5 ns

- 4x Throughput per Watt
- 10x Throughput per Area

No Advantage Greater Than 1 Order of Magnitude

Related Work

- ISAAC
  - A memristive crossbar based system
  - Compared against DaDianNao CMOS accelerator
  - 14.8×, 5.5×, and 7.5× in throughput, energy, and computational density

Shafiee et al. "ISAAC: A Convolutional Neural Network Accelerator With In-Situ Analog Arithmetic in Crossbars," ISCA 2016.

Outline

- Memristors as computational elements in neural networks
- Are memristors better than CMOS?
  - Performance / power efficiency experiment
- **Memristors in a broader system context**

Known Technical Challenges for Memristors

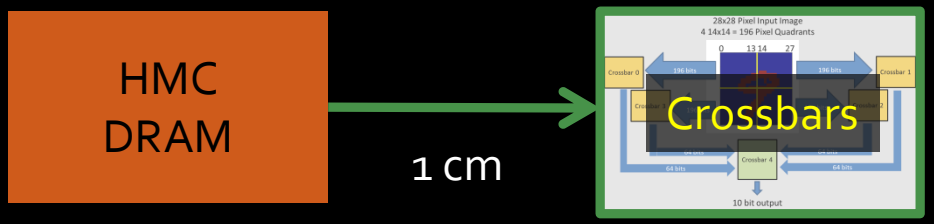
- How will the memristors be programmed?
- Has programming asymmetry been addressed?
- How similar are the characteristics of different memristors in a crossbar?
- What is the memristor yield?
- Are sneak paths controlled?
- How many write cycles before failure?
- Are reads destructive?
- What is the assumed range of resistances?
- Is the power density acceptable?
- Will current density in lower metal layers exceed limits?
- Is the crossbar timing skew tolerable?

Algorithms, materials science/device physics, process technology, analog circuit design, and architecture are all needed

Community is years and millions of dollars away from a mature deployable, manufacturable memristor system

Compute Power: A Fraction of Total Power

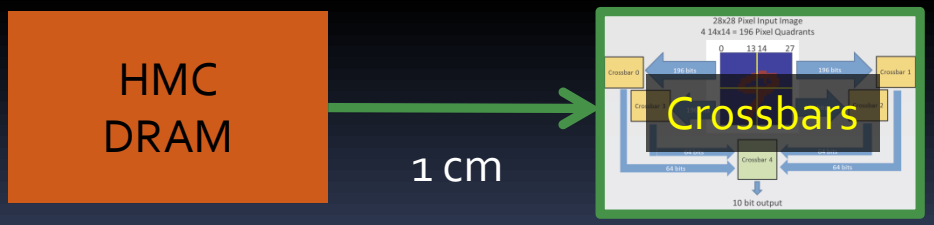
- Assume circuit with Micron HMC memory (12.5 pJ/bit[†]), directly feeding inputs



[†] Jeddeloh and Keeth, **"Hybrid Memory Cube New DRAM Architecture Increases Density and Performance," VLSIT 2012.**

Compute Power: A Fraction of Total Power

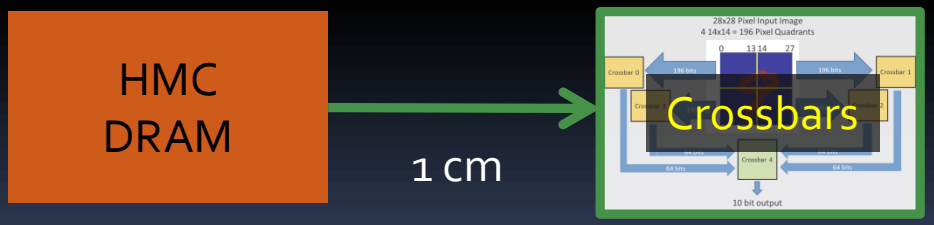
- Assume circuit with Micron HMC memory (12.5 pJ/bit[†]), directly feeding inputs
- Compute power is 1.4% of total power for digital system and 0.3% for memristor system



[†] Jeddeloh and Keeth, **"Hybrid Memory Cube New DRAM Architecture Increases Density and Performance," VLSIT 2012.**

Compute Power: A Fraction of Total Power

- Assume circuit with Micron HMC memory (12.5 pJ/bit[†]), directly feeding inputs
- Compute power is 1.4% of total power for digital system and 0.3% for memristor system
- If we increase computational intensity by 50x with same inputs, memristors cut system power by 3x



[†] Jeddeloh and Keeth, **"Hybrid Memory Cube New DRAM Architecture Increases Density and Performance," VLSIT 2012.**

Conclusion

- Memristors provide a modest energy efficiency and density improvement over CMOS
- When exploring "Beyond CMOS" ideas:
  - Evaluate *early* what benefit the new technology will provide
- Consider the whole system
  - Greatest challenge to next generation ***systems*** is data storage and movement, not compute