SAND2020-7486PE

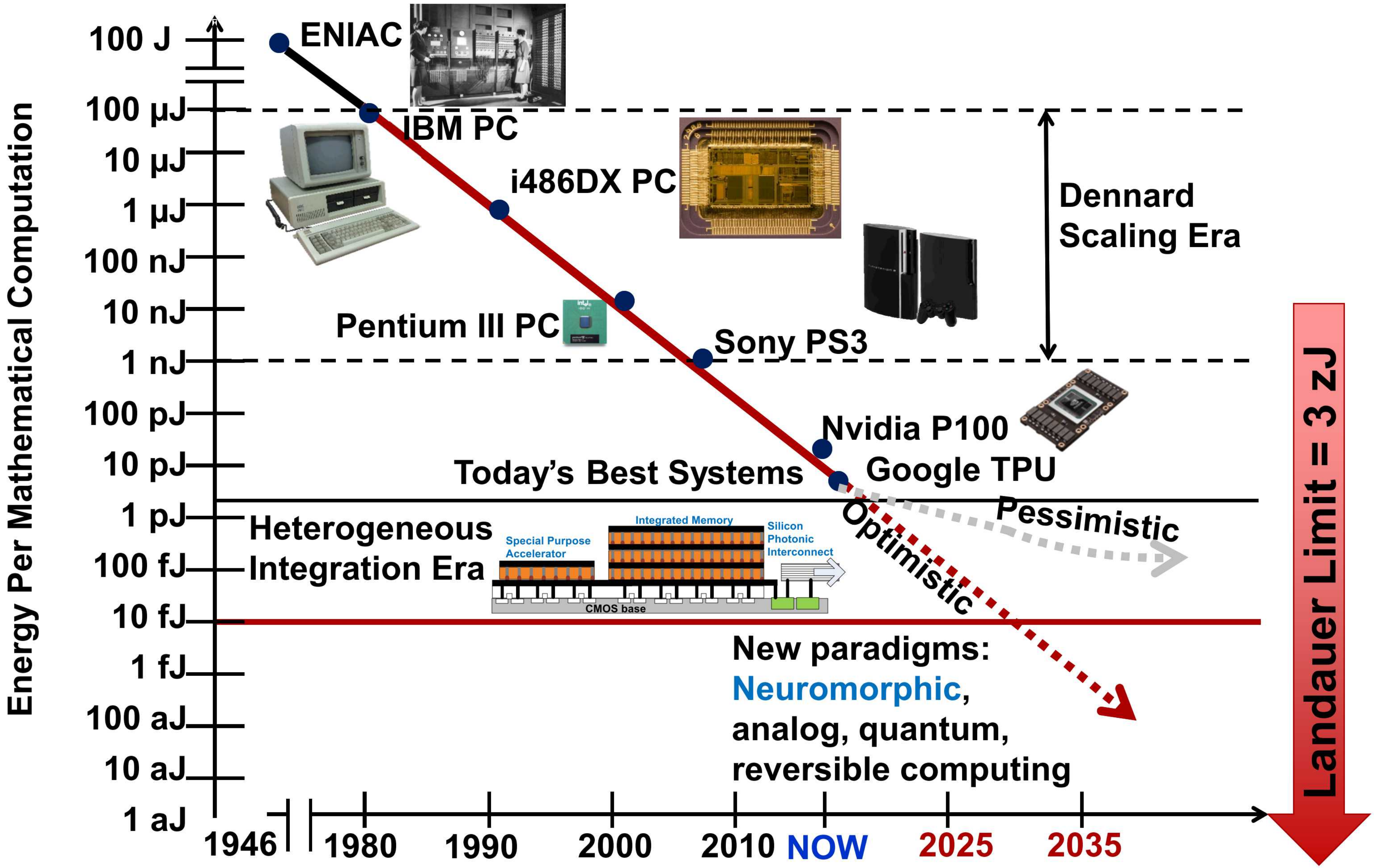
A survey of Emerging Beyond CMOS Devices and Architectures

Sapan Agarwal

Sandia National Laboratories

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.

Evolution of Computing Machinery



Dennard Scaling Era

Heterogeneous Integration Era

New paradigms: Neuromorphic, analog, quantum, reversible computing

Landauer Limit = 3 zJ

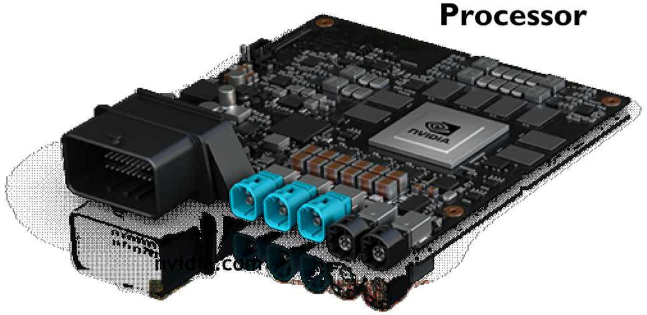
Computing Across Power Envelopes

IoT, Edge, and Mobile Computing

nest.com

Self Driving Cars, Unmanned Arial Vehicles, and Satellite Computing

Nvidia Self-Driving Processor



wikimedia.org

Datacenters, HPC

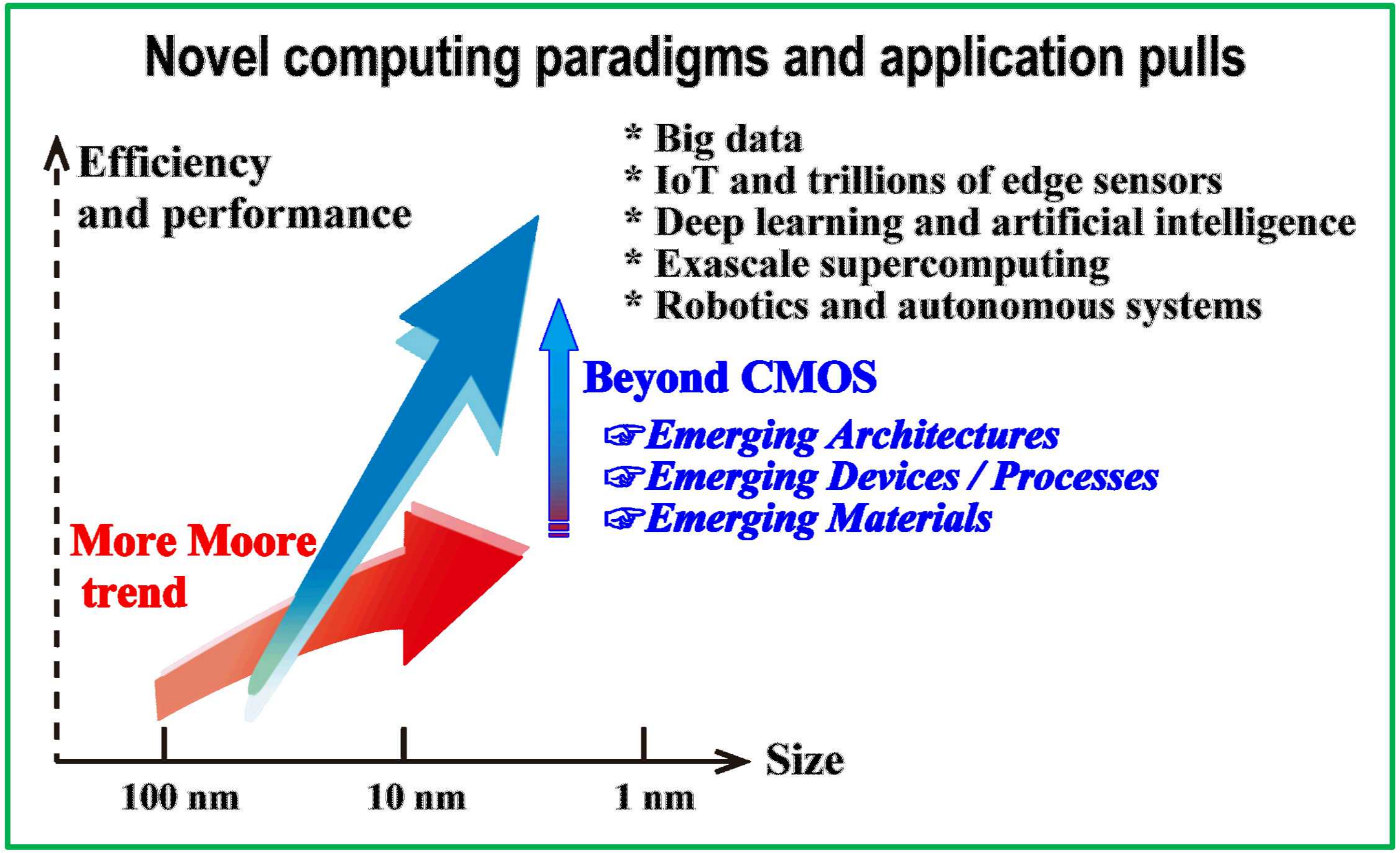
ASCI Red Supercomputer

1W → 10W → 10^2W → 10^3W → 10^4W → 10^5W → 10^6W

IEEE International Roadmap for Devices and Systems

Successor to ITRS (International Technology Roadmap for Semiconductors)

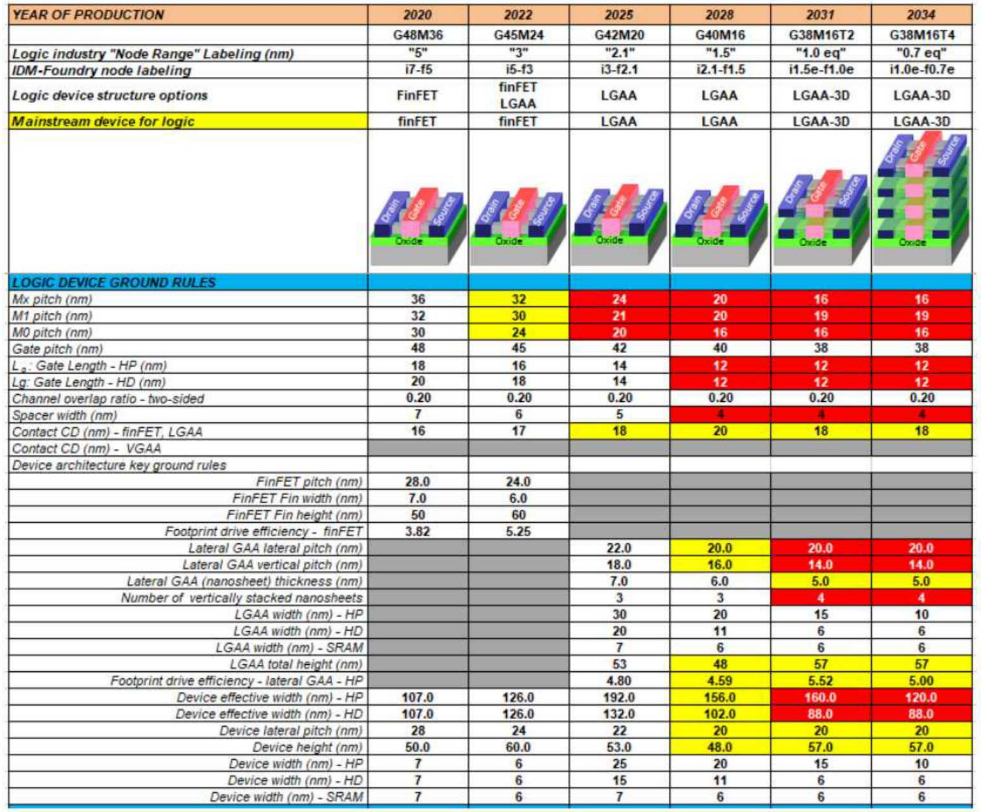
Objective of the Beyond CMOS (BC) Chapter



Novel computing paradigms and application pulls

* Big data
* IoT and trillions of edge sensors
* Deep learning and artificial intelligence
* Exascale supercomputing
* Robotics and autonomous systems

Still road mapping near term semiconductors (Moore Moore)



Drain
Gate
Source
Oxide

YEAR OF PRODUCTION	2020	2022	2025	2028	2031	2034
	G48M36	G45M24	G42M20	G40M16	G38M16T2	G38M16T4
Logic industry "Node Range" Labeling (nm)	"5"	"3"	"2.1"	"1.5"	"1.0 eq"	"0.7 eq"
IDM-Foundry node labeling	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5e-f1.0e	i1.0e-f0.7e
Logic device structure options	FinFET	finFET LGAA	LGAA	LGAA	LGAA-3D	LGAA-3D
Mainstream device for logic	finFET	finFET	LGAA	LGAA	LGAA-3D	LGAA-3D
LOGIC DEVICE GROUND RULES						
Mx pitch (nm)	36	32	24	20	16	16
M1 pitch (nm)	32	30	21	20	19	19
M0 pitch (nm)	30	24	20	16	16	16
Gate pitch (nm)	48	45	42	40	38	38
L_g: Gate Length - HP (nm)	18	16	14	12	12	12
Lg: Gate Length - HD (nm)	20	18	14	12	12	12
Channel overlap ratio - two-sided	0.20	0.20	0.20	0.20	0.20	0.20
Spacer width (nm)	7	6	5	4	4	4
Contact CD (nm) - finFET, LGAA	16	17	18	20	18	18
Contact CD (nm) - VGAA						
Device architecture key ground rules						
FinFET pitch (nm)	28.0	24.0				
FinFET Fin width (nm)	7.0	6.0				
FinFET Fin height (nm)	50	60				
Footprint drive efficiency - finFET	3.82	5.25				
Lateral GAA lateral pitch (nm)			22.0	20.0	20.0	20.0
Lateral GAA vertical pitch (nm)			18.0	16.0	14.0	14.0
Lateral GAA (nanosheet) thickness (nm)			7.0	6.0	5.0	5.0
Number of vertically stacked nanosheets			3	3	4	4
LGAA width (nm) - HP			30	20	15	10
LGAA width (nm) - HD			20	11	6	6
LGAA width (nm) - SRAM			7	6	6	6
LGAA total height (nm)			53	48	57	57
Footprint drive efficiency - lateral GAA - HP			4.80	4.59	5.52	5.00
Device effective width (nm) - HP	107.0	126.0	192.0	156.0	160.0	120.0
Device effective width (nm) - HD	107.0	126.0	132.0	102.0	88.0	88.0
Device lateral pitch (nm)	28	24	22	20	20	20
Device height (nm)	50.0	60.0	53.0	48.0	57.0	57.0
Device width (nm) - HP	7	6	25	20	15	10
Device width (nm) - HD	7	6	15	11	6	6
Device width (nm) - SRAM	7	6	7	6	6	6

2019 Beyond CMOS Team Members and Contributors

Sapan Agarwal
Brad Aimone
Hiro Akinaga
Otitoaleke Akinola
Mustafa Badaroglu
Gennadi Bersuker
Christian Binek
Geoffrey Burr
Leonid Butov
Kerem Camsari
Gert Cauwenberghs
An Chen
Winston Chern
Supriyo Datta
John Dallesasse
Shamik Das
Erik DeBenedictis
Peter Dowben
Tetsuo Endoh
Ben Feinberg
Thomas Ferreira de Lima
Akira Fujiwara
Elliot Fuller
Michael Frank
Paul Franzon
Michael Fuhrer
Mike Garner
Chakku Goplan
Bogdan Govoreanu
Cat Graves
Kohei Hamaya
Masami Hane
Jennifer Hasler
Yoshihiro Hayashi
Toshiro Hiramoto
D. Scott Holmes
Sharon Hu
Francesca Iacopi
Danielle Ilmeni
Jean Anne Incorvia
Engin Ipek
Satoshi Kamiyama
Kiyoshi Kawabata
Asif Khan
Hajime Kobayashi
Suhas Kumar
Ilya Krivorotov
Xiuling Li
Xiang (Shaun) Li
Shy-Jay Lin
Tsu-Jae King Liu
Matthew Marinella
Bicky A. Marque
Rivu Midya
Yoshiyuki Miyamoto
Johannes Muller
Azad Naeemi
Mitchell A. Nahmias
Emre Neftci
Mike Niemier
Dmitri Nikonov
Yutaka Ohno
Chenyun Pan
Ferdinand Peper
Shriram Ramanathan
Mingyi Rao
Shashi Paul
Paul R. Prucnal
Titash Rakshit
Arijit Raychowdhury
Sayeef Salahuddin
Shintaro Sato
Michael Schneider
Bhavin J. Shastri
Xia Sheng
Takahiro Shinada
Urmita Sikder
Greg Snider
John-Paul Strachan
Dimitri Strukov
Naoyuki Sugiyama
Tarek Taha
Alexander N. Tait
Shinichi Takagi
Norikatsu Takaura
Tsutomu Teduka
Yasuhide Tomioka
Wilman Tsai
Tohru Tsuruoka
Zhongrui Wang
R. Stanley Williams
Justin Wong
Dirk Wouters
Patrick Xiao
Kojiro Yagami
J. Joshua Yang
Noboyuki Yoshikawa
Victor Zhirnov

What's the Minimum Energy to Operate a Transistor?

Consider a signal Energy E_{signal}

The probability of an error due to thermal noise is:

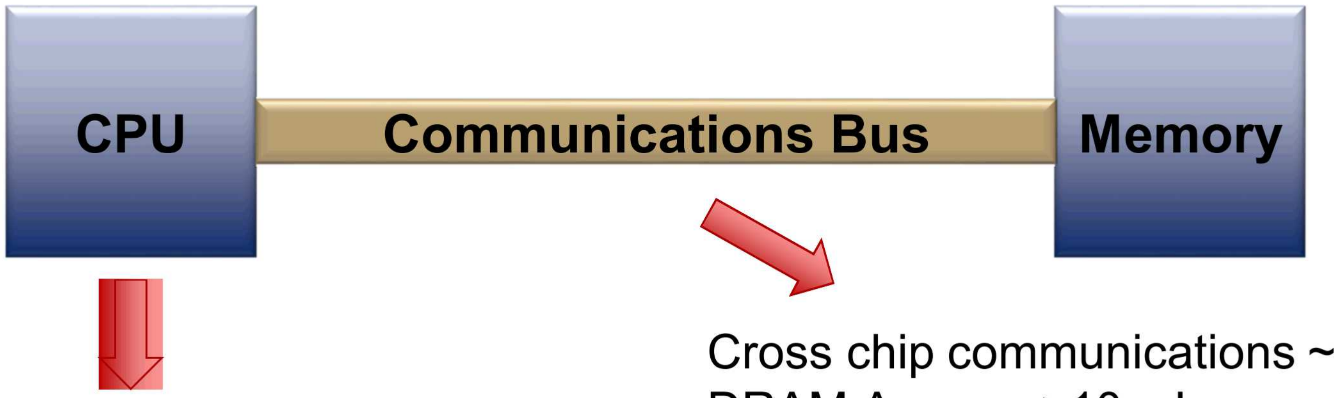
$$P(Error) = e^{-E_{signal}/kT}$$

In order to ensure a full system with billons of transistors is reliable, we need:

$E_{signal} \sim 50$ kT

Landauer – Shannon Limit

How Efficient are Current Systems?



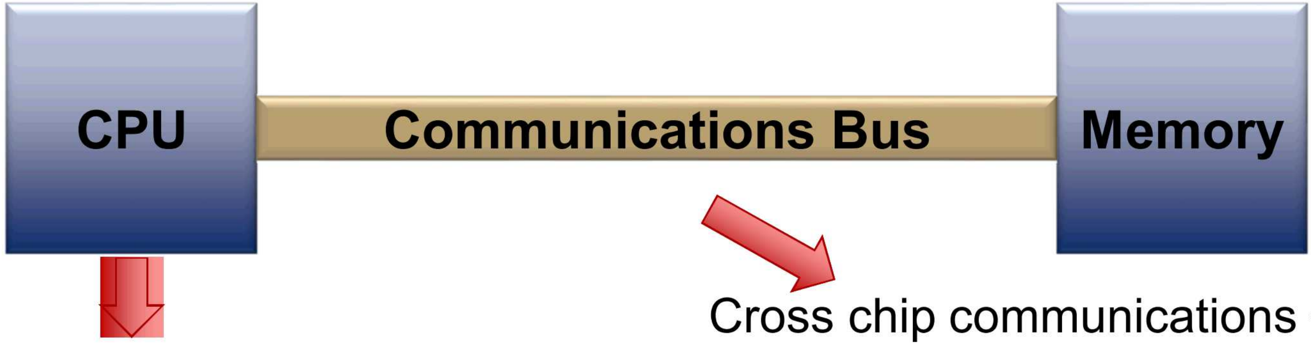
Current Transistors ~ 10 aJ
50kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ

Only 50X Improvement Left for Transistors!

Cross Chip communications are 5 orders of magnitude more energy!

Beyond Moore Technologies



Current Transistors ~ 10 aJ
50kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ

Extending Von Neumann

- Low Voltage or Novel Transistors
- Optical Communications
- Reduced Data Movement
  - New On-Chip Memory
  - Processing near Memory

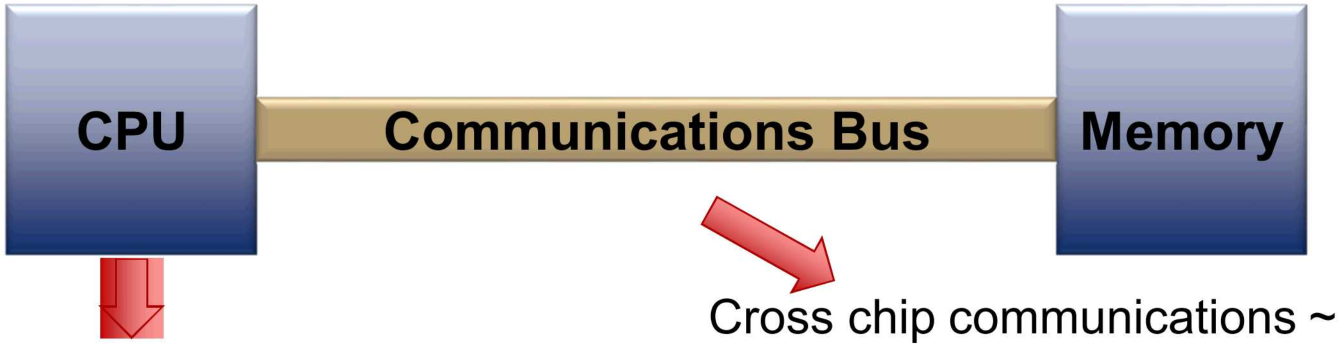
Alternate Computing Paradigms

- Neuromorphic
- Analog
  - Computing with memory devices
- Quantum
- Stochastic
- Approximate

Going Below 50 kT

- Error Correction
- Reversible Computing
  - Adiabatic Computing / Energy Recycling
- Superconducting

Beyond Moore Technologies



Current Transistors ~ 10 aJ
50kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ

Extending Von Neumann

- Low Voltage or Novel Transistors
- Optical Communications
- Reduced Data Movement
  - New On-Chip Memory
  - Processing near Memory

Alternate Computing Paradigms

- Neuromorphic
- Analog
  - Computing with memory devices
- Quantum
- Stochastic
- Approximate

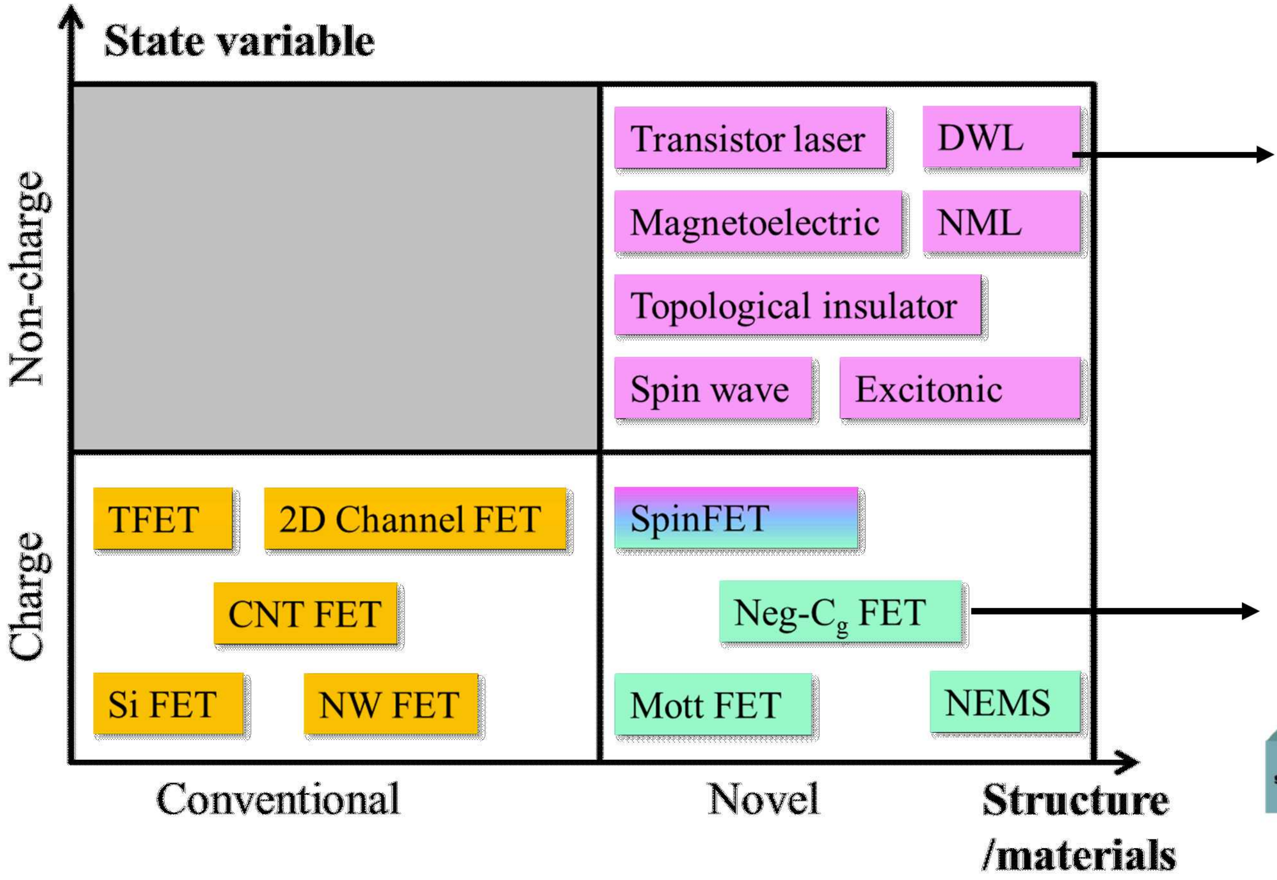
Going Below 50 kT

- Error Correction
- Reversible Computing
  - Adiabatic Computing / Energy Recycling
- Superconducting

Extending Von Neumann – New Transistors

Lowering voltage lowers CV² energy of communications

Logic and Information Processing Devices



State variable

Non-charge

Charge

Transistor laser

DWL

Magnetoelectric

NML

Topological insulator

Spin wave

Excitonic

TFET

2D Channel FET

CNT FET

Si FET

NW FET

SpinFET

Neg-C_g FET

Mott FET

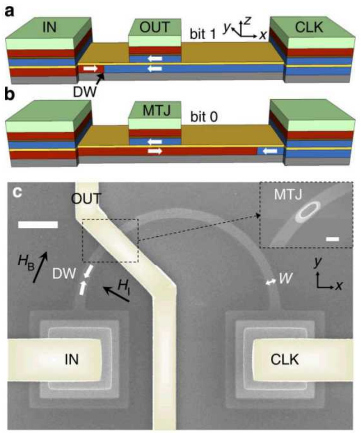
NEMS

Conventional

Novel

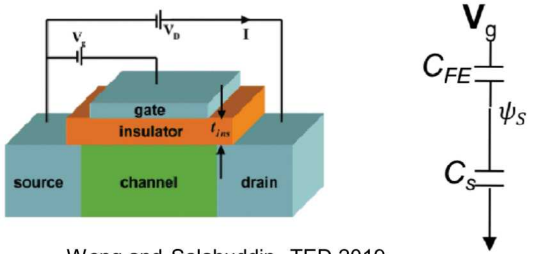
Structure /materials

Domain Wall Logic



J.A. Incorvia et al, Nature Comm 7, 2016

Negative Capacitance FET

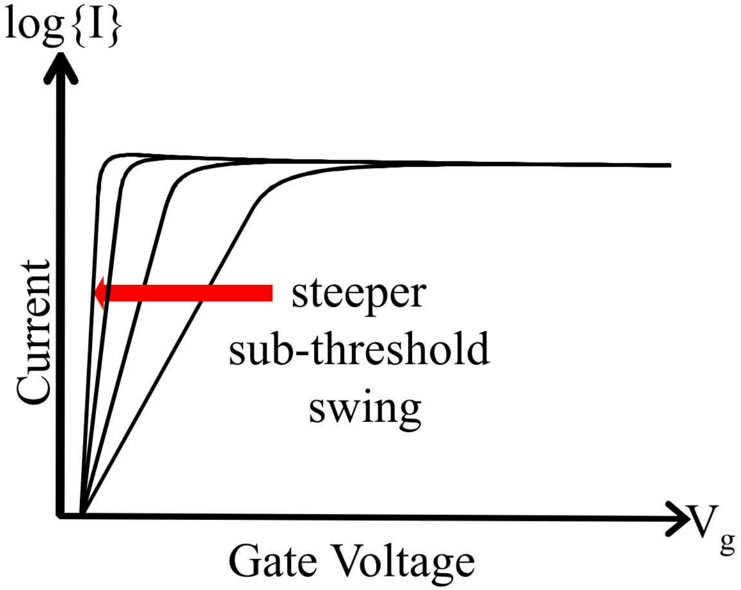


Wong and Salahuddin, TED 2019

A New Switch has to Satisfy Three Specifications

Low Active Power (CV^2 energy)

- Steepness (or sensitivity)
  - o switches with only a few milli-volts
  - o 60mV/decade ⇒ **1mV/decade**

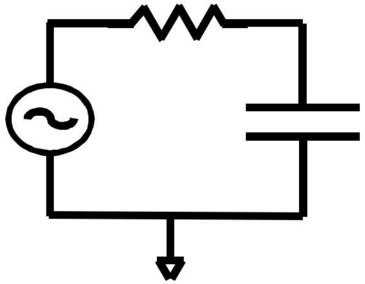


Low Leakage Power

- On/Off ratio: **10^5 : 1**

High Speed (RC delay)

- High Conductance Density
  - o **1 milli-Siemen/micron**



Need Restoring Logic for Novel State Variables

- Need to be able to drive multiple output transistors across multiple stages of logic

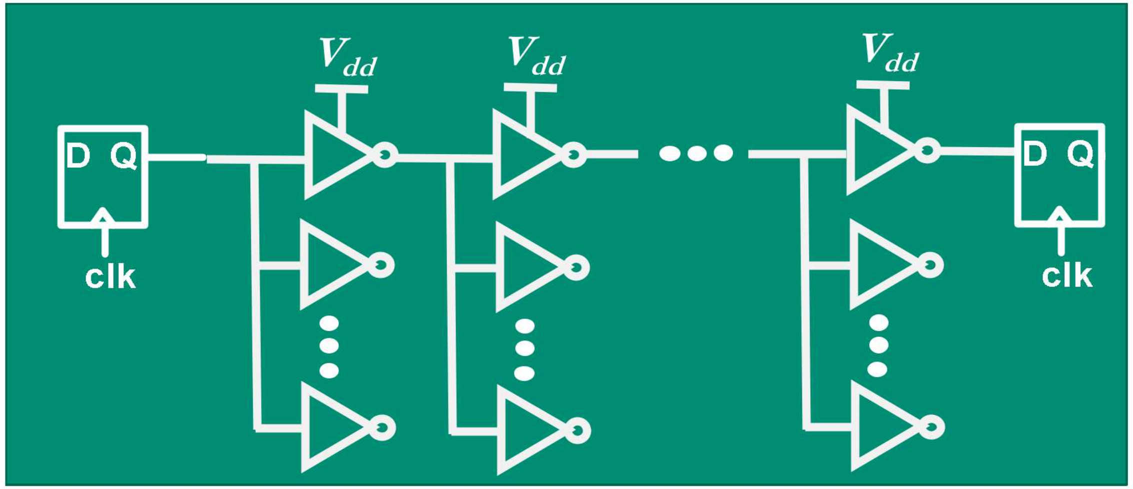
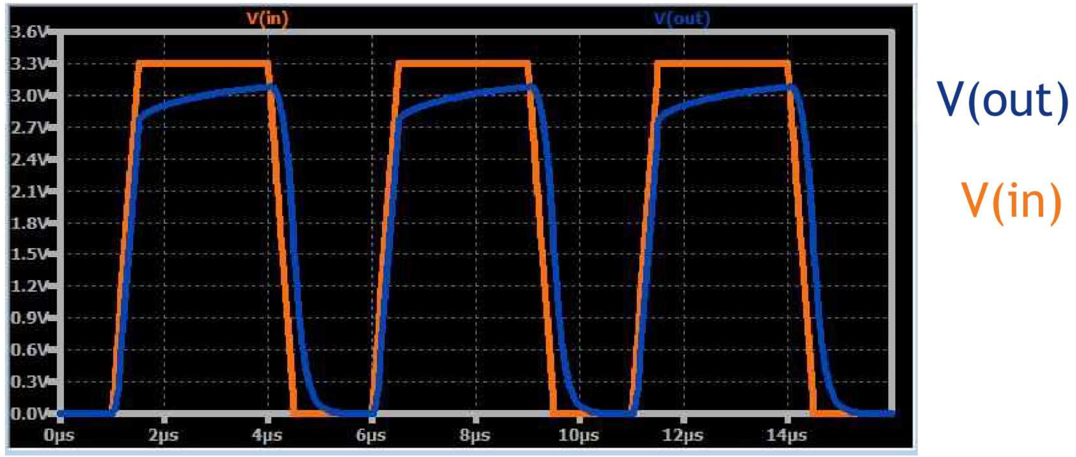
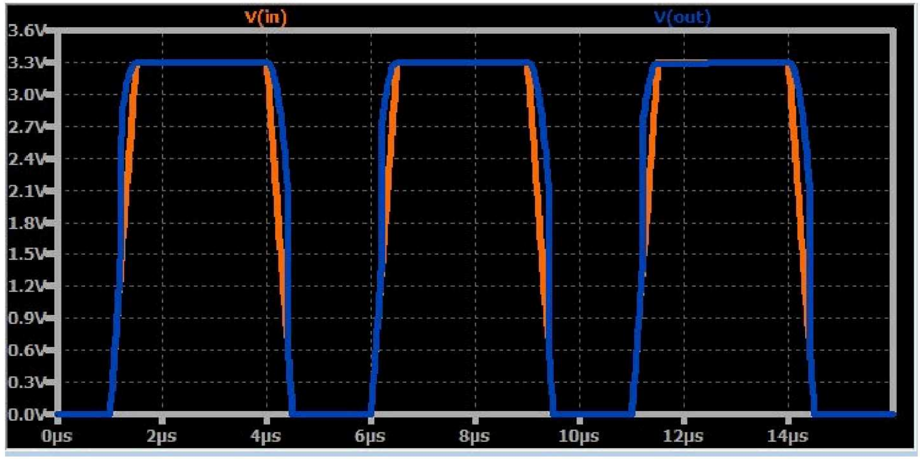


Figure from Elad Elon, 2010 E3S Retreat

Not Restoring



Restoring

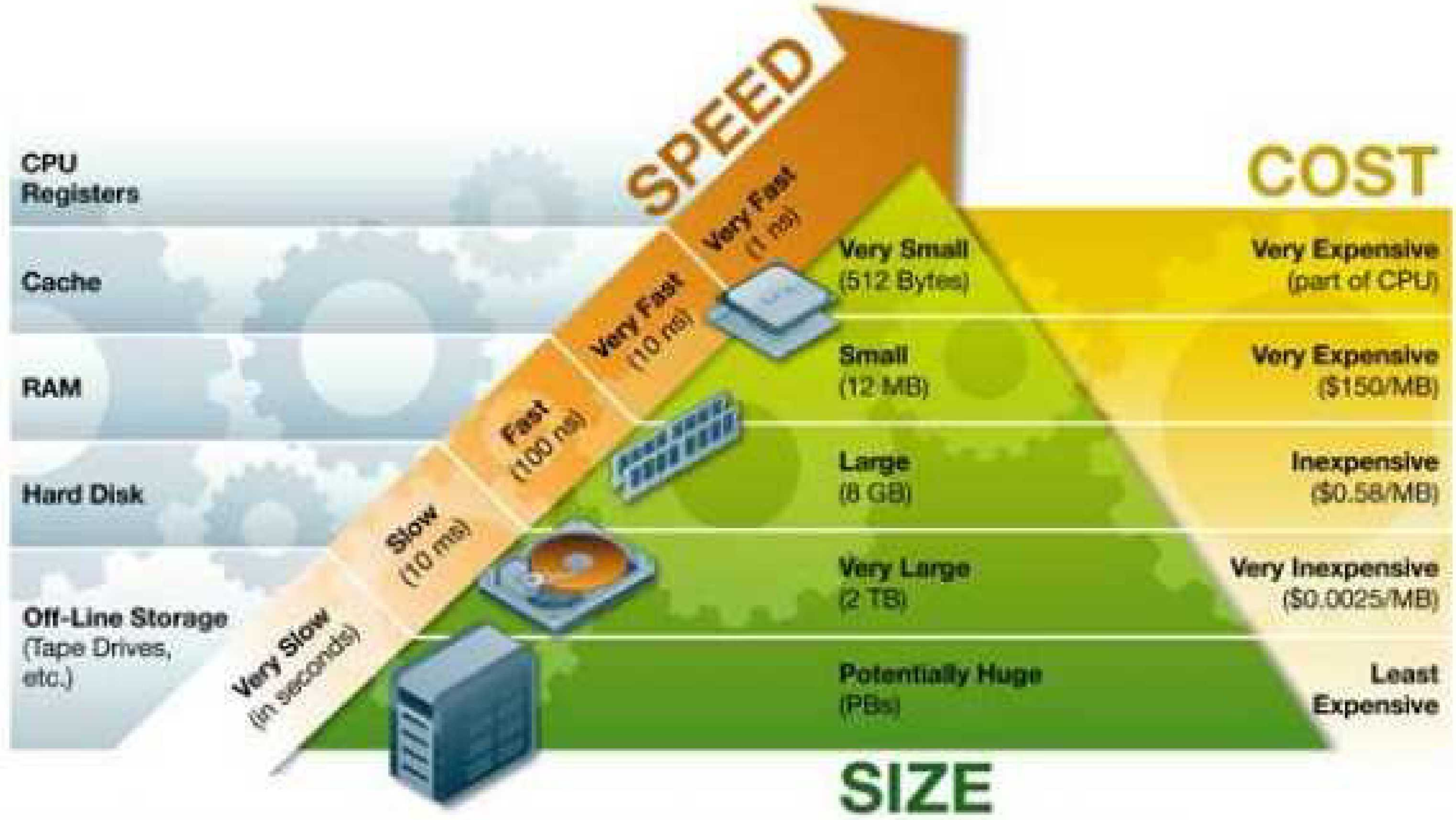


Inputs slightly below V_{dd} restored to V_{dd} and inputs slightly above 0 are driven to zero

https://www.allaboutcircuits.com/technical-articles/restoring-digital-signals-in-pass-transistor-logic/

The Memory Hierarchy

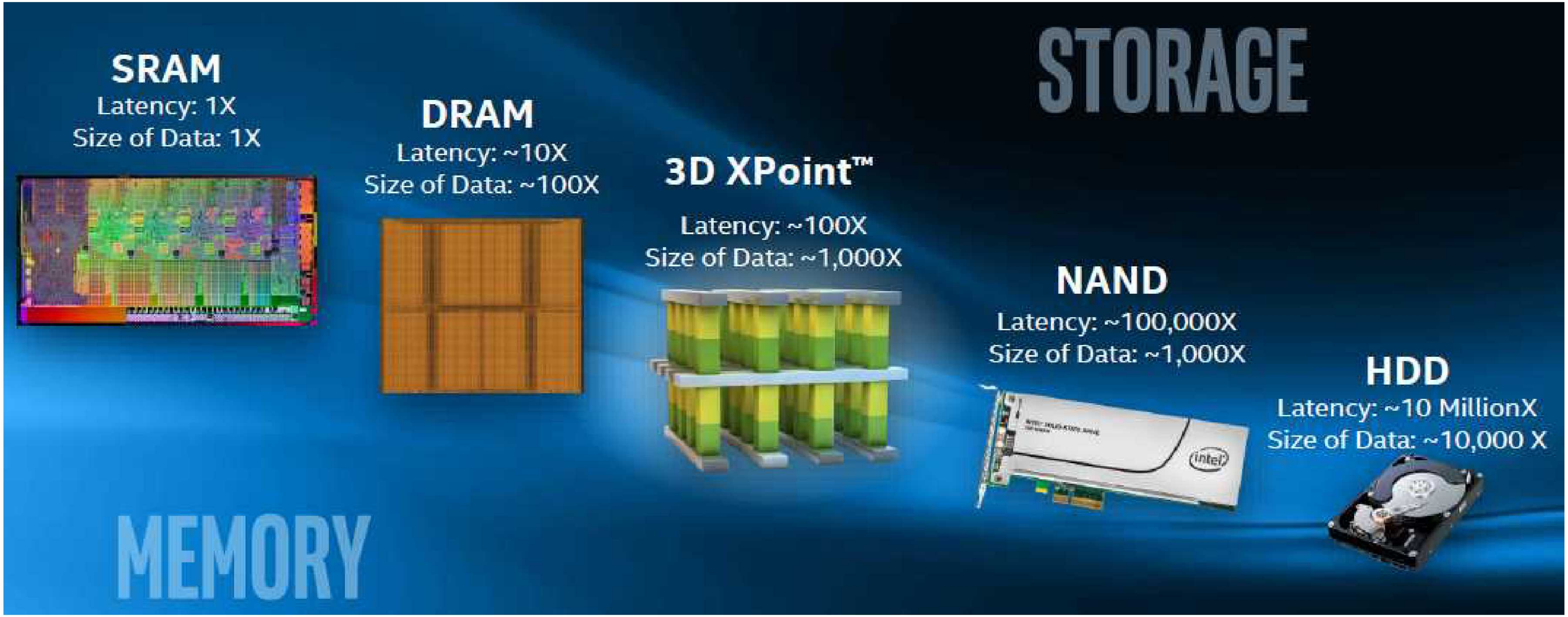
10^7 difference in speed from HDD to Registers



Source: http://www.ts.avnet.com/uk/products_and_solutions/storage/hierarchy.html

10^{10} difference in density from HDD to Registers

Storage Class Memory - Intel/Micron 3D XPoint

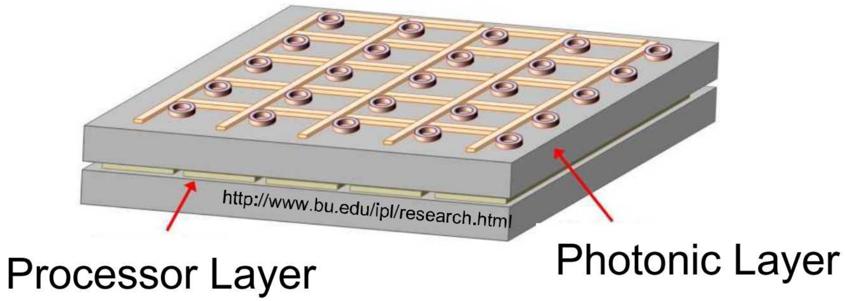


Also Samsung Z-NAND:

- Re-optimize flash for speed rather than density (single level per cell)
- Comparable to Intel Optane (3D XPoint) products

Extending Von Neumann – Reduced Data Movement

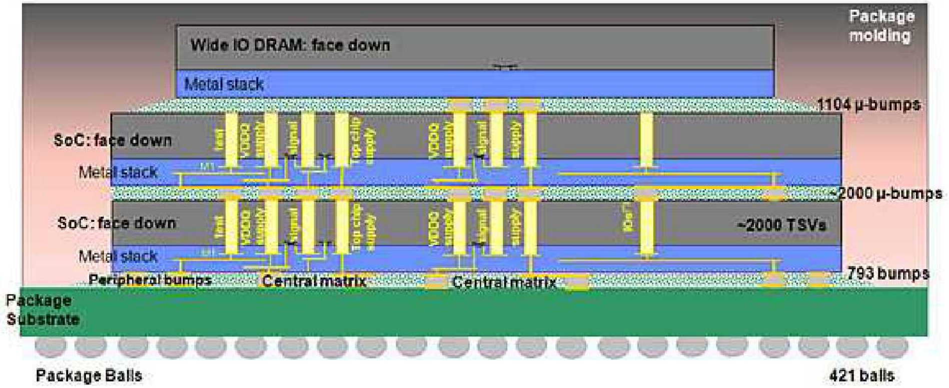
Optical Interconnects



http://www.bu.edu/ipl/research.html

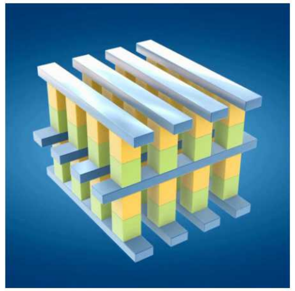
100 fJ to 1 pJ

2.5D & 3D Integration



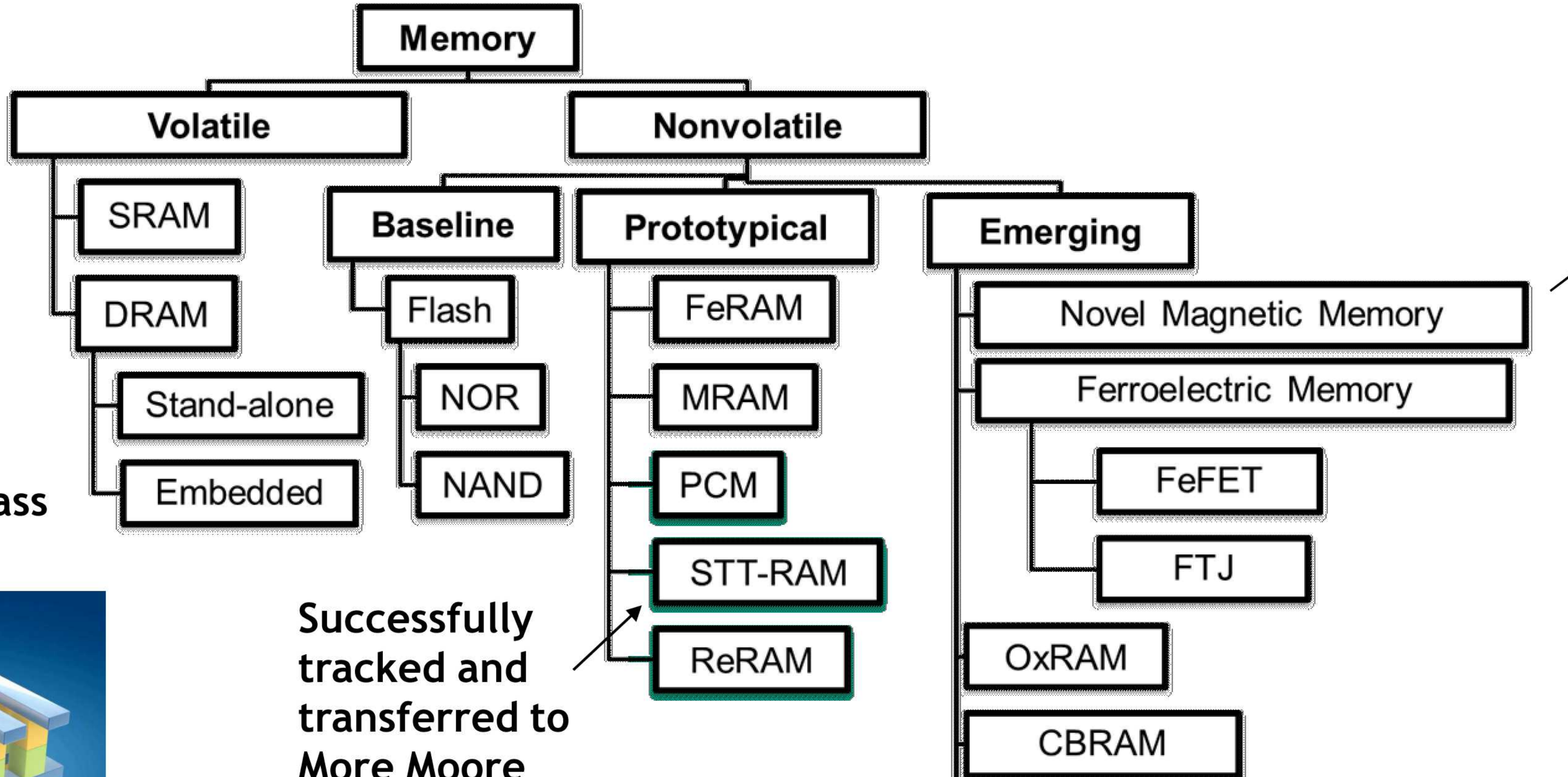
Richard Goering, "Three Die Stack -- A Big Step "Up" for 3D-ICs with TSVs" Cadence blog

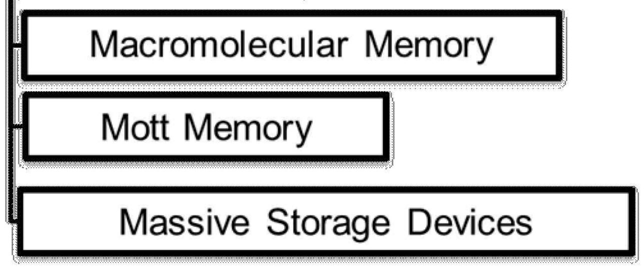
Embedded Memory



- Embed new denser nonvolatile memories in the processor
- Add simple processing to DRAM or disk drive controllers "Processing in Memory"

Emerging Memory Devices

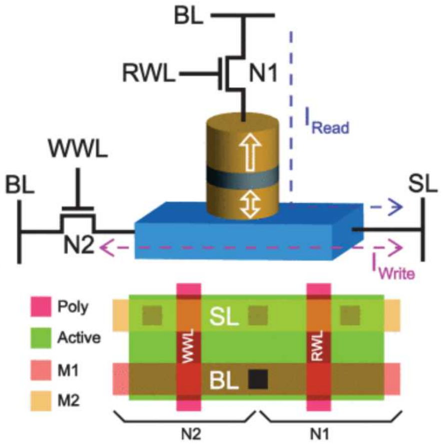




- Memory
  - Volatile
    - SRAM
    - DRAM
      - Stand-alone
      - Embedded
  - Nonvolatile
    - Baseline
      - Flash
        - NOR
        - NAND
    - Prototypical
      - FeRAM
      - MRAM
      - PCM
      - STT-RAM
      - ReRAM
    - Emerging
      - Novel Magnetic Memory
      - Ferroelectric Memory
        - FeFET
        - FTJ
      - OxRAM
      - CBRAM
      - Macromolecular Memory
      - Mott Memory
      - Massive Storage Devices

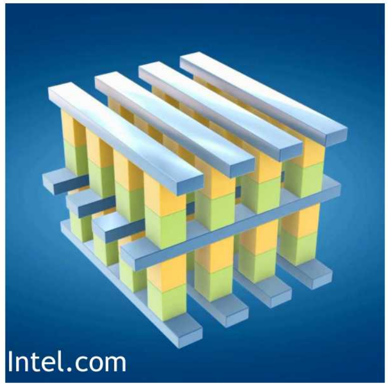
Successfully tracked and transferred to More Moore

SOT Memory



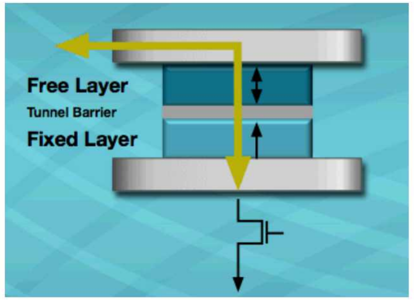
Z. Wang EDL 39, 2018

Storage Class Memory

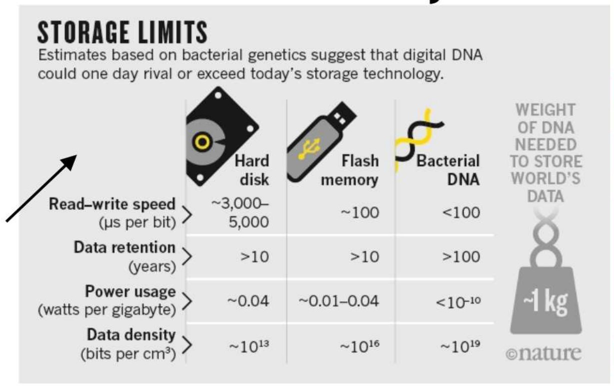


Intel.com

STT-MRAM



DNA Memory



STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	Hard disk	Flash memory	Bacterial DNA
Read–write speed (μs per bit)	~3,000–5,000	~100	<100
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01–0.04	$<10^{-10}$
Data density (bits per cm^3)	$\sim10^{13}$	$\sim10^{16}$	$\sim10^{19}$

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA

~1 kg

©nature

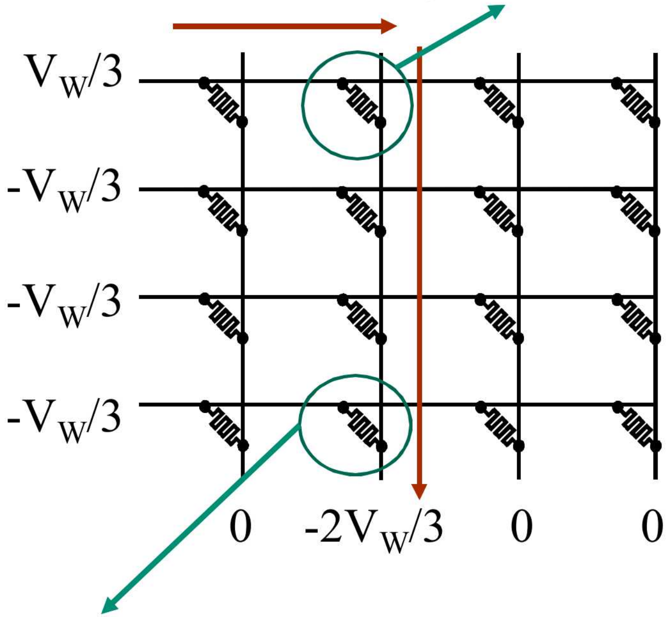
Advantages of Emerging Memories

- Back end of line integration
  - Can integrate in the metal layers directly on top of logic
- 3D stackable
- Higher endurance (10^9 to 10^{12}) relative to flash (10^4), but not as good as DRAM ($>10^{16}$)
- Nonvolatile: >10 year retention, no standby leakage
- O(10ns) read and write times

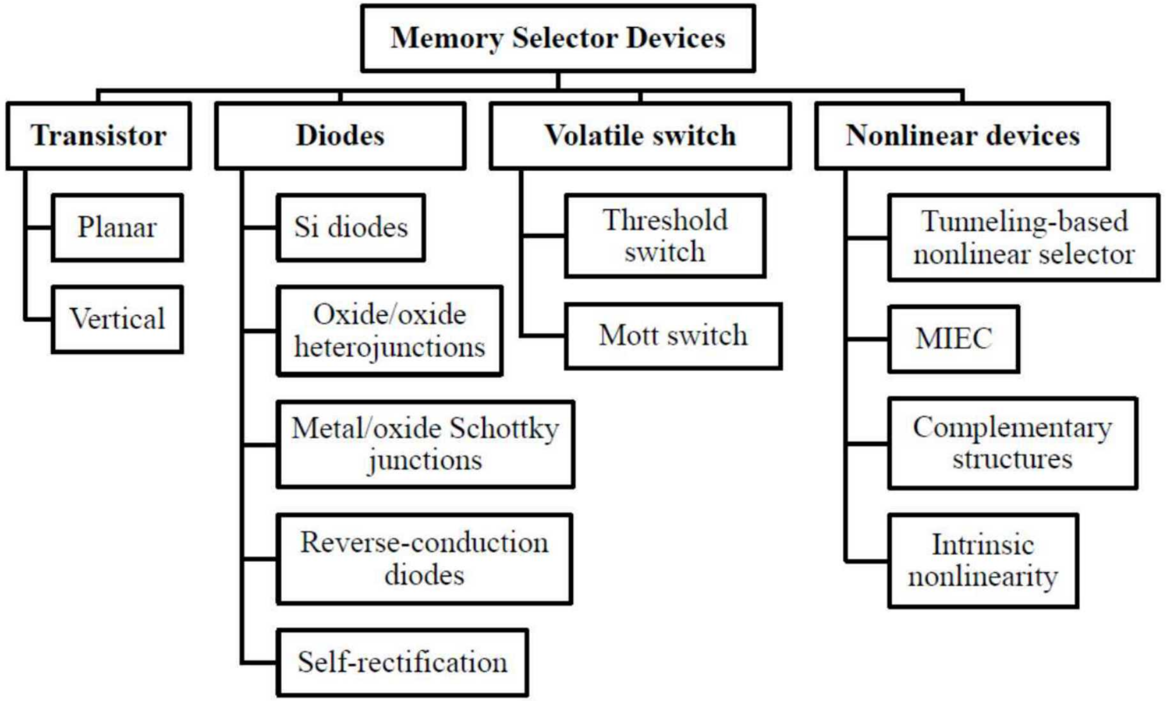
The Need for a Select Device for Resistive Memories

V/3 write scheme

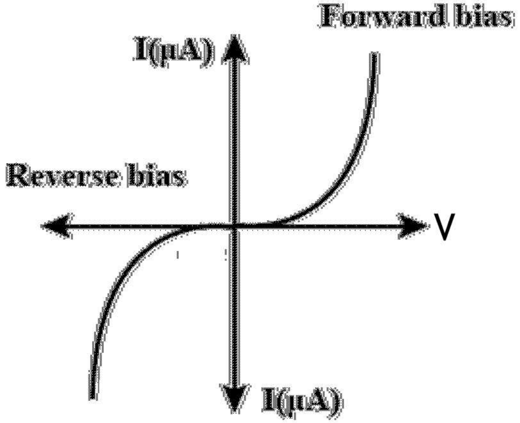
Target device sees W_{write}



Half selected device sees $V_{write}/3$

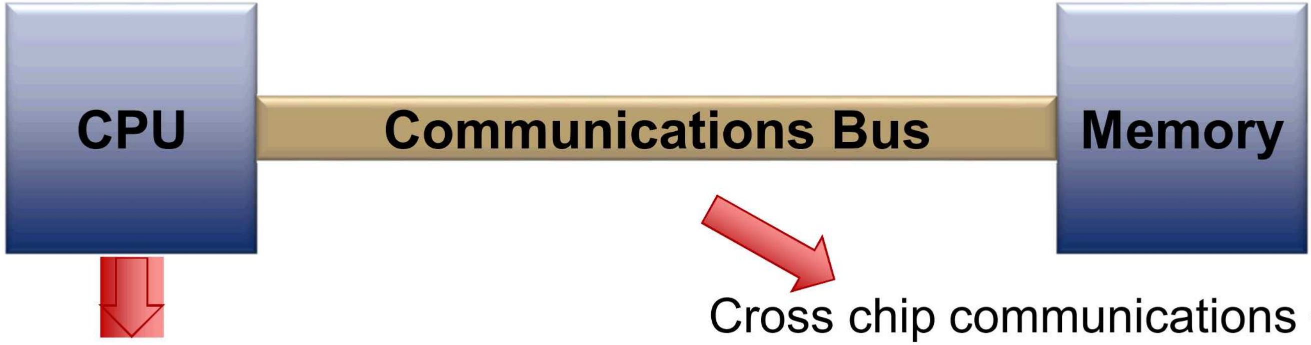


- Memory Selector Devices
  - Transistor
    - Planar
    - Vertical
  - Diodes
    - Si diodes
    - Oxide/oxide heterojunctions
    - Metal/oxide Schottky junctions
    - Reverse-conduction diodes
    - Self-rectification
  - Volatile switch
    - Threshold switch
    - Mott switch
  - Nonlinear devices
    - Tunneling-based nonlinear selector
    - MIEC
    - Complementary structures
    - Intrinsic nonlinearity



- Need a nonlinear 2 terminal back end of line compatible device to block current from half selected and unselected devices
- The larger the on/off ratio, the larger the possible array and therefore density
- Typically needs to be bi-directional

Beyond Moore Technologies



Current Transistors ~ 10 aJ
50kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ

Extending Von Neumann

- Low Voltage or Novel Transistors
- Optical Communications
- Reduced Data Movement
  - New On-Chip Memory
  - Processing near Memory

Alternate Computing Paradigms

- Neuromorphic
- Analog
  - Computing with memory devices
- Quantum
- Stochastic
- Approximate

Going Below 50 kT

- Error Correction
- Reversible Computing
  - Adiabatic Computing / Energy Recycling
- Superconducting

Alternate Computing Paradigms

Crossbar Based Computing Architectures

- Vector Matrix Multiplication
- Outer Product Update
- Crossbar Based Matrix Solvers
- Ternary Content Addressable Memory

Neuro-inspired Computing

- Hyperdimensional Computing
- Local Learning Rules
- Spiking Neural Networks

Probabilistic and Stochastic Circuits

Computing With Dynamical Systems

- Simulated Annealing
- Coupled Oscillator based energy minimization

Alternate Computing Paradigms

Crossbar Based Computing Architectures

- Vector Matrix Multiplication
- Outer Product Update
- Crossbar Based Matrix Solvers
- Ternary Content Addressable Memory

Neural Network
Training Accelerators

Neuro-inspired Computing

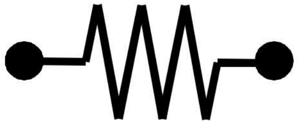
- Hyperdimensional Computing
- Local Learning Rules
- Spiking Neural Networks

Probabilistic and Stochastic Circuits

Computing With Dynamical Systems

- Simulated Annealing
- Coupled Oscillator based energy minimization

Use Resistive Memories for Local Computation

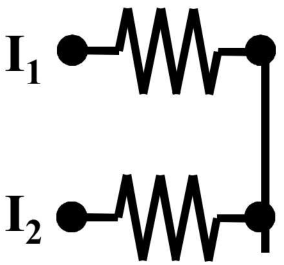


$$V = I \times R$$

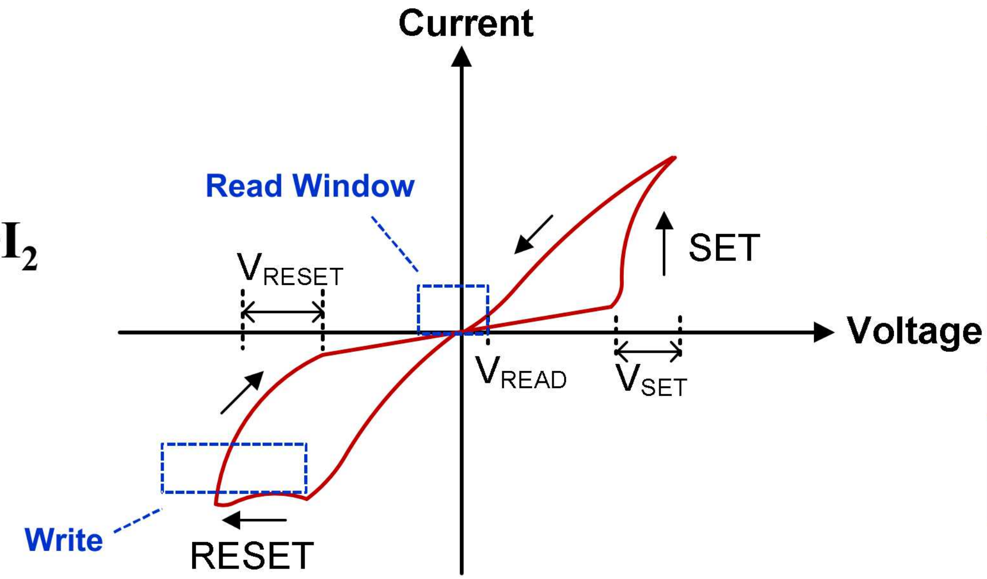
$$I = G \times V$$

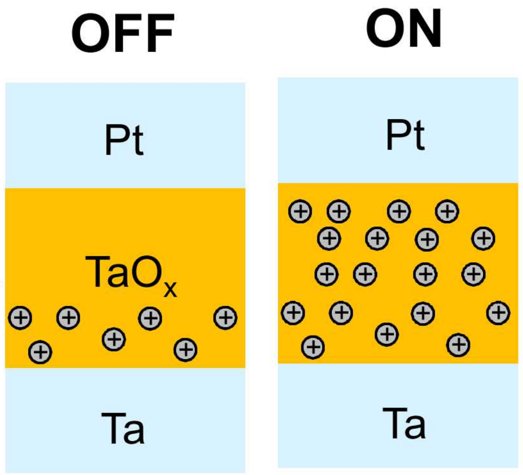
multiplication

- A resistive memory or ReRAM is a programmable resistor
  - Apply small voltages allows the conductance to be read: I = G × V
  - Apply large voltages to change the resistance



Addition: $I=I_1+I_2$





Vector Matrix Multiply: Directly Process in the Memory Itself

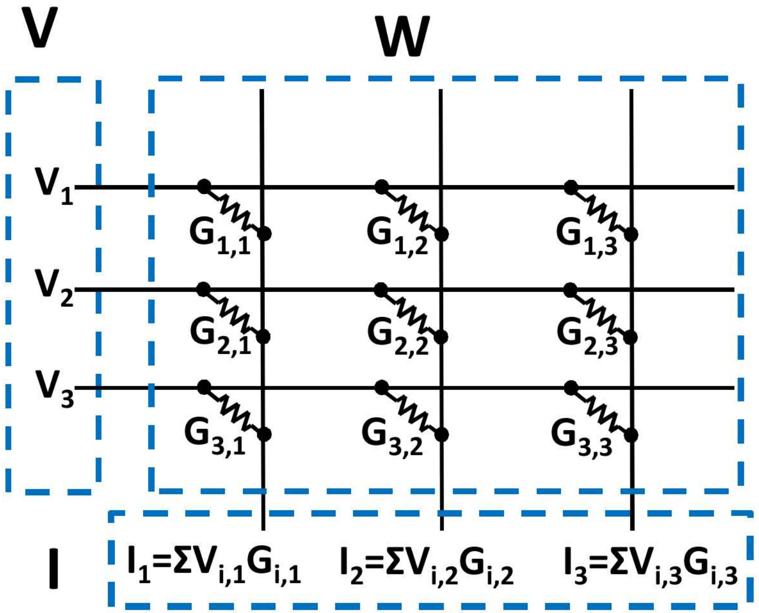
Mathematical

$$V^T W = I$$

$$\begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} =$$

$$\begin{bmatrix} I_1 = \Sigma V_{i,1} W_{i,1} & I_2 = \Sigma V_{i,2} W_{i,2} & I_3 = \Sigma V_{i,3} W_{i,3} \end{bmatrix}$$

Electrical



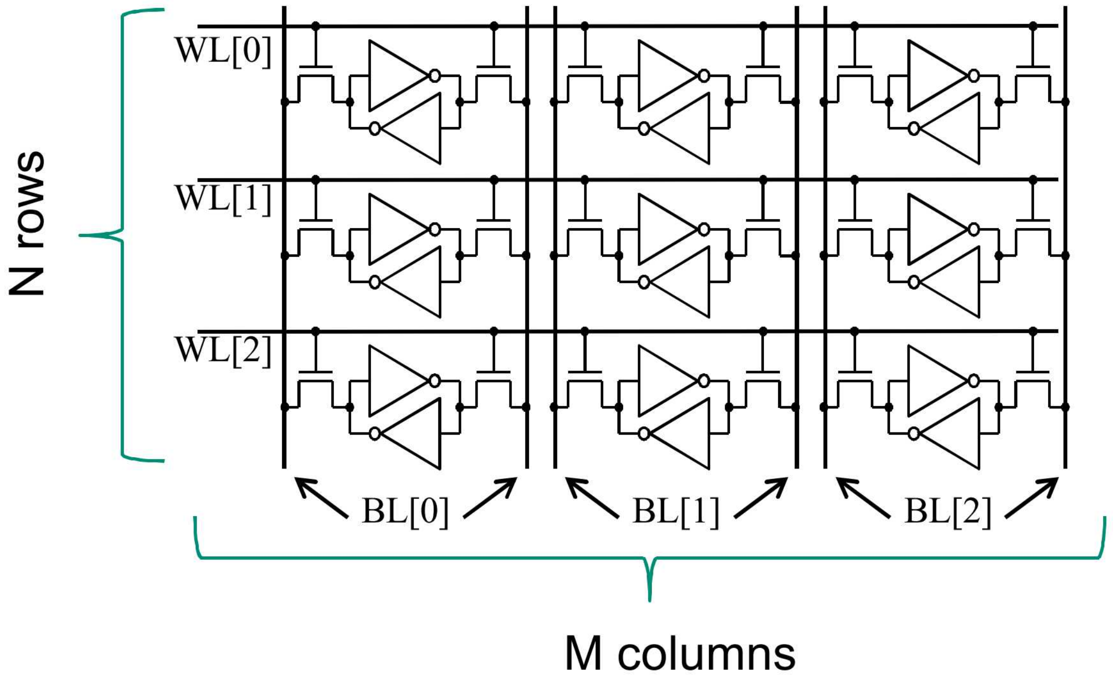
<10 fJ MAC
<10 fJ Update
>100 TOPS/W

Analog is efficiently and naturally able to combine computation and data access

Large-scale processing in memory with a multiplier and adder at each real-valued memory location

- Energy to charge the crossbar is CV^2
- $E \propto C \propto$ number of RRAMs $\propto N \times M$

SRAM Arrays Require Charging Columns Multiple Times



SRAMs must be read one row at a time, charging M columns

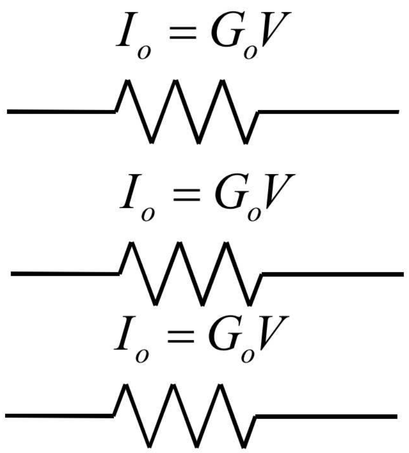
Each column wire length is O(N).

Energy = N Rows × M Columns × O(N) wire length

Energy ~ $O(N^2 \times M)$

O(N) times worse than a crossbar!

The Noise Limited Energy to Read a Crossbar Column is Independent of Crossbar Size



Measure N resistors and determine the total output current with some signal to noise ratio (SNR)*

What is the minimum energy?

$$Energy = \underbrace{V^2 G_O \times N}_{\text{Power in each resistor × number of resistors}} \times \frac{1}{\Delta f}$$

Determined by noise and SNR

$$\text{Thermal Noise} = \langle \Delta I^2 \rangle$$
$$= N \times (4k_bT \times G_o \times \Delta f)$$

$$SNR^2 = \frac{(NI_o)^2}{\langle \Delta I^2 \rangle}$$

$$\frac{1}{\Delta f} = 4k_bT \times SNR^2 \times \frac{1}{V^2 G_o \times N}$$

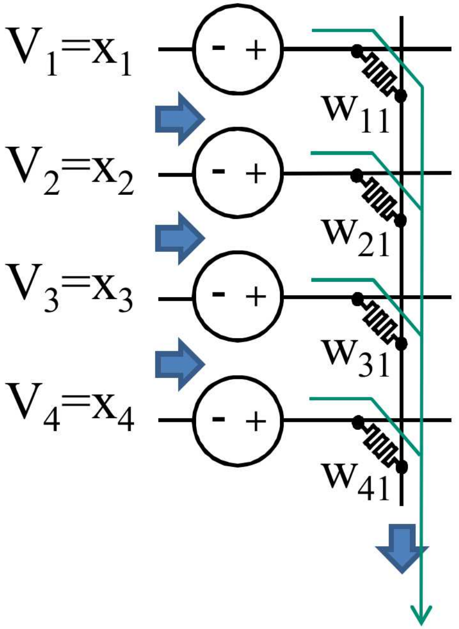
If we double the number of resistors, we can double the speed to get the same energy and SNR.

This is because the noise scales as sqrt(N) while the signal scales as N

$$Energy = 4k_bT \times SNR^2$$

*we are assuming we need some fixed precision on the output, and don't need full floating point accuracy

Need to Use Analog to Efficiently Discard Precision



Sum 1024 8 bit weights X 8 bit inputs:

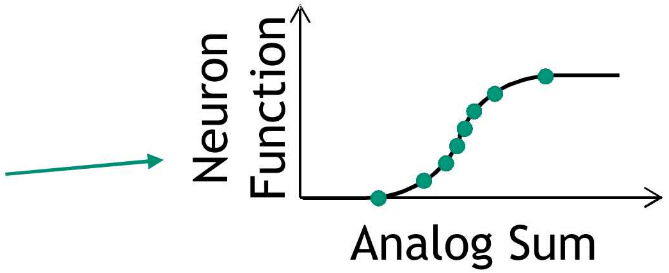
- Result has 26 bits of information!
- A 26 bit ADC would eliminate any analog advantage!

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

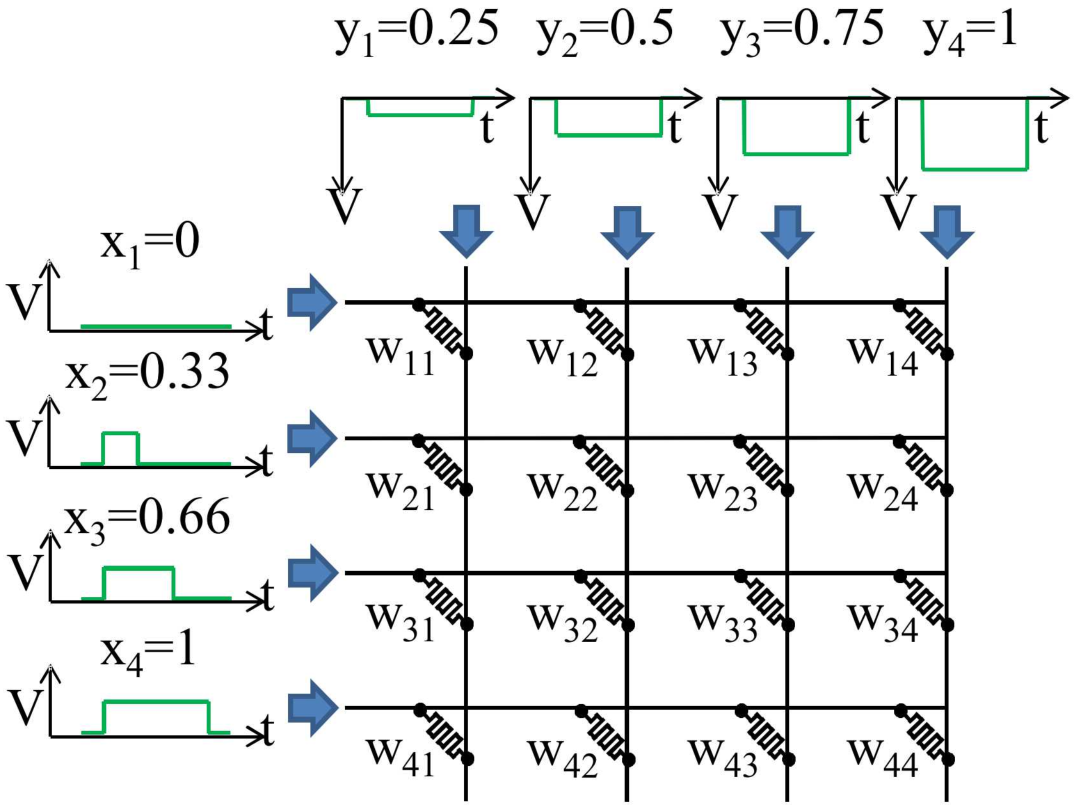
- i.e. digitize only 8 bits or fewer

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



Outer Product Update: Parallel Write



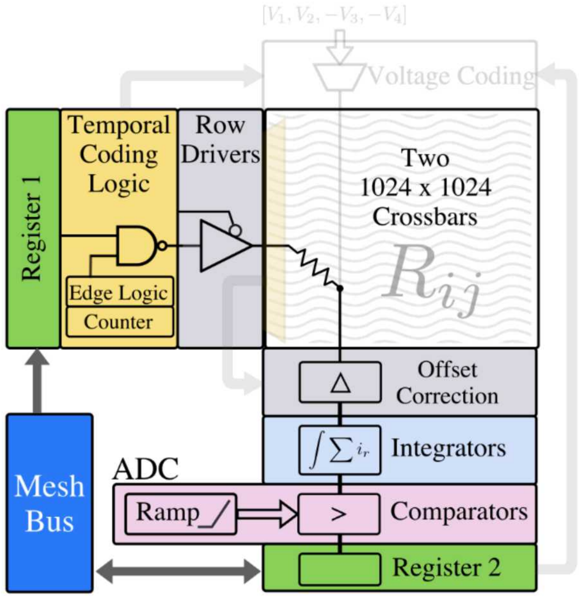
Energy to charge the crossbar is CV^2

$E \propto C \propto$ number of RRAMs $\propto N\times M$

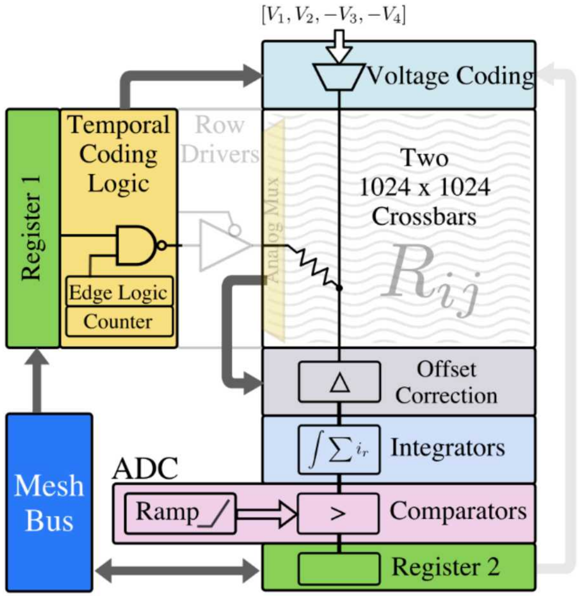
$$E \sim O(N\times M)$$

Example: Design an Neural Network Training Accelerator

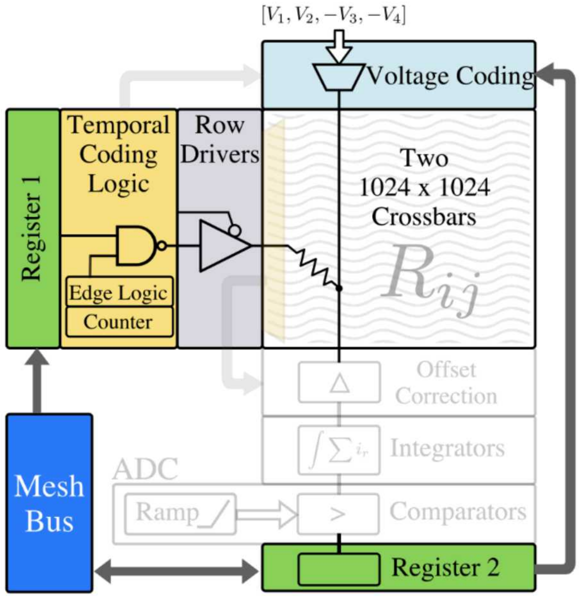
Vector Matrix Multiply



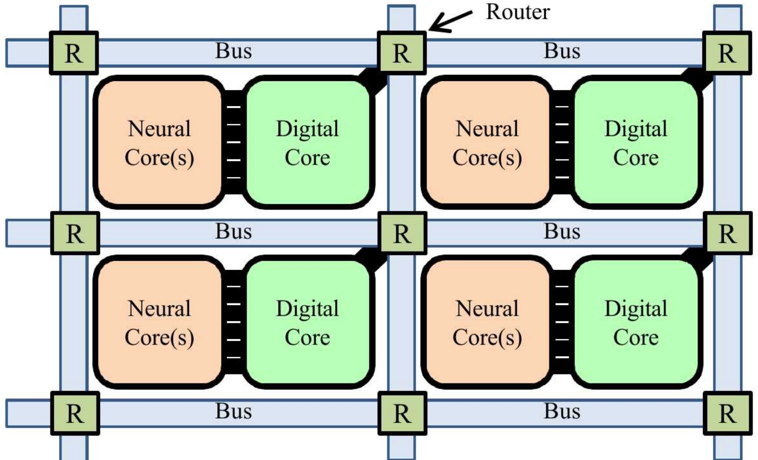
Matrix Vector Multiply



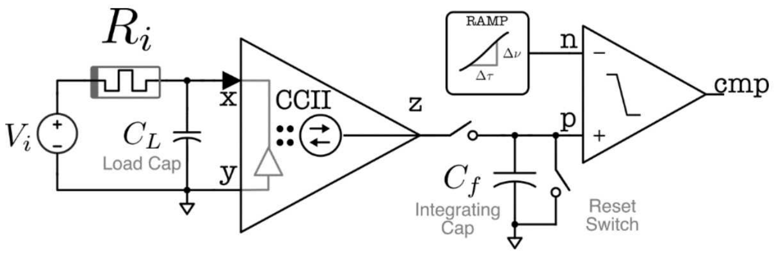
Outer product Update



Chip Architecture



Neuron Circuitry



Energy Area and Latency Advantages of an Analog Accelerator

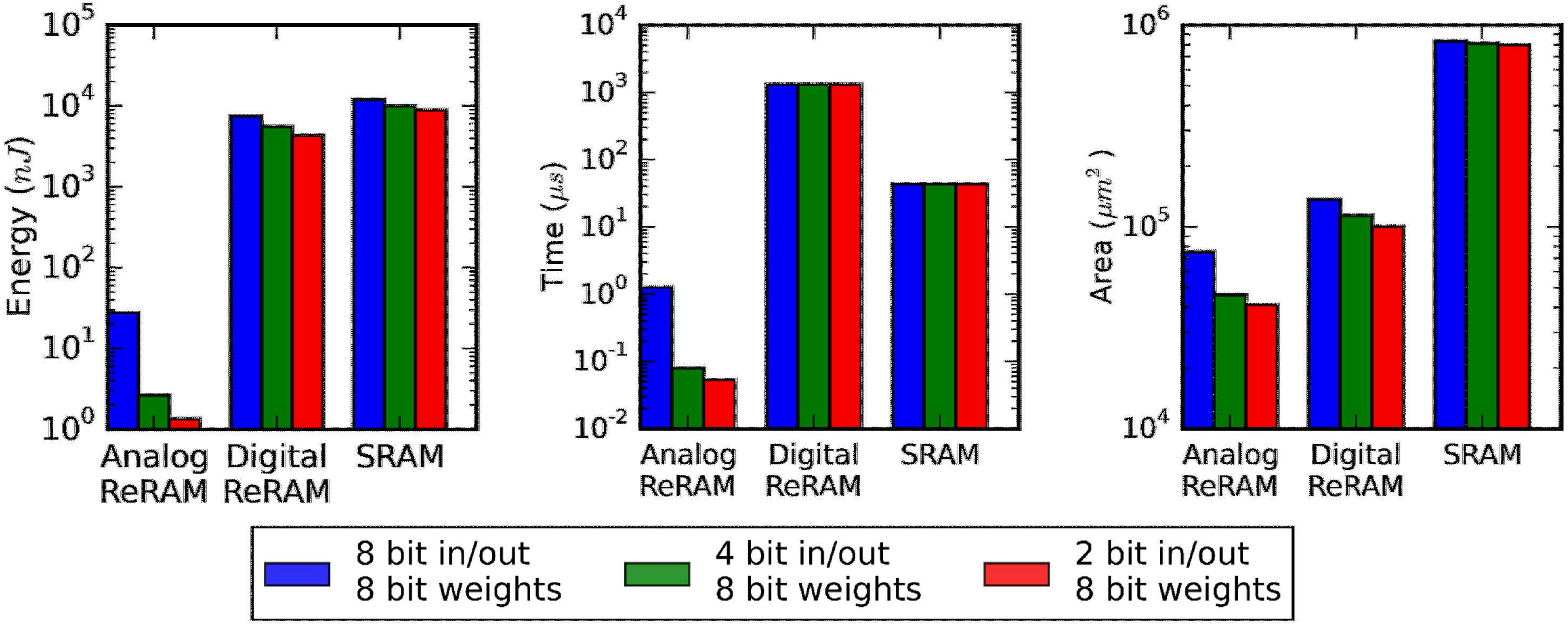
1024 x1024 = 1M array operations, sum over 1 training cycle, 3 operations:

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

Energy
430 – 6,900X over SRAM

Latency
35 – 800X over SRAM

Area
11 – 20X over SRAM

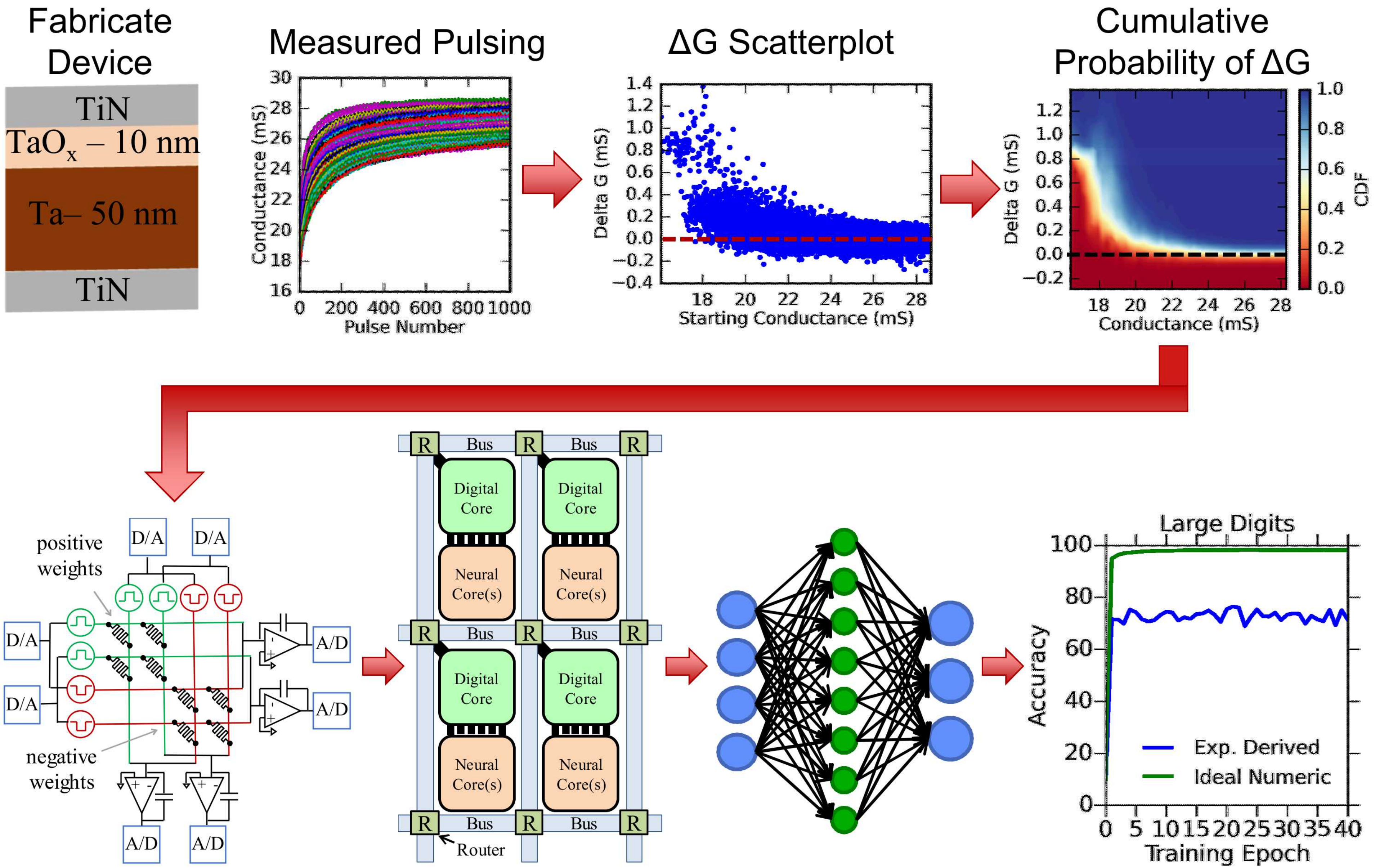


Used a commercial 14/16 nm PDK

***Requires 100 MΩ on state devices

M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, "Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* 2018.

Go from Measurement to Accuracy



Fabricate Device

Measured Pulsing

ΔG Scatterplot

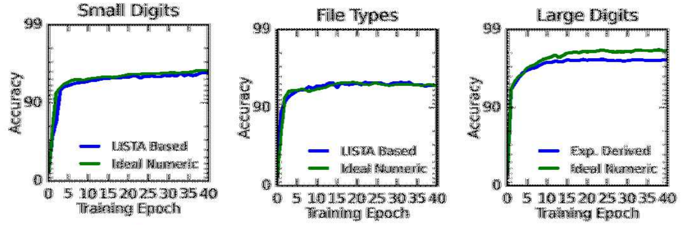
Cumulative Probability of ΔG

Large Digits

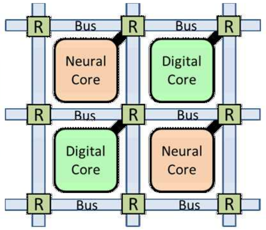
Multiscale Model of a Neural Training Accelerator

Target Algorithms
- Deep Learning
- Sparse Coding
- Liquid State Machines

Algorithms



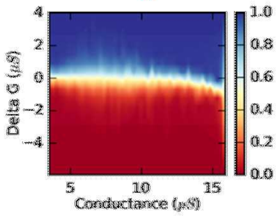
Architecture

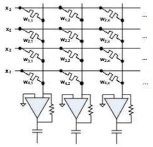


Modified McPAT/CACTI: Model performance and energy requirements

CROSS SIM

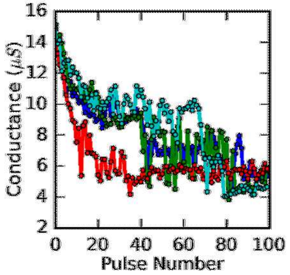
Sandia Cross-Sim: Translates device measurements and crossbar circuits to algorithm-level performance





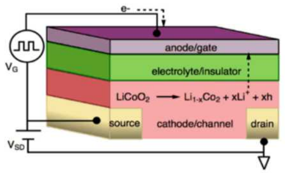
Circuits

Sandia's Xyce Circuit Sim: Simulate crossbar circuits based on our devices

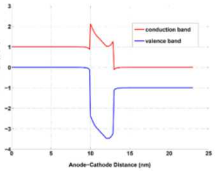


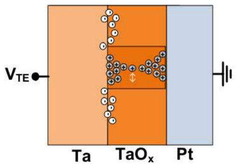
Memristor fabrication and measurements in MESAFab

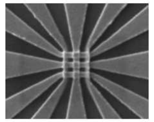
Devices

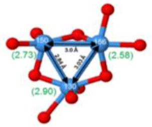


Drift-diffusion model of ReRAM band diagram & transport (REOS, Charon)





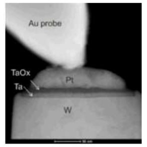


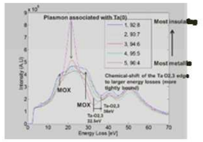


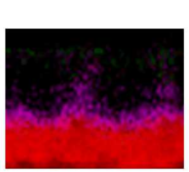
DFT of model of oxide physics, bands

Materials

In situ TEM of filament switching: Use DFT model to interpret EELS signature

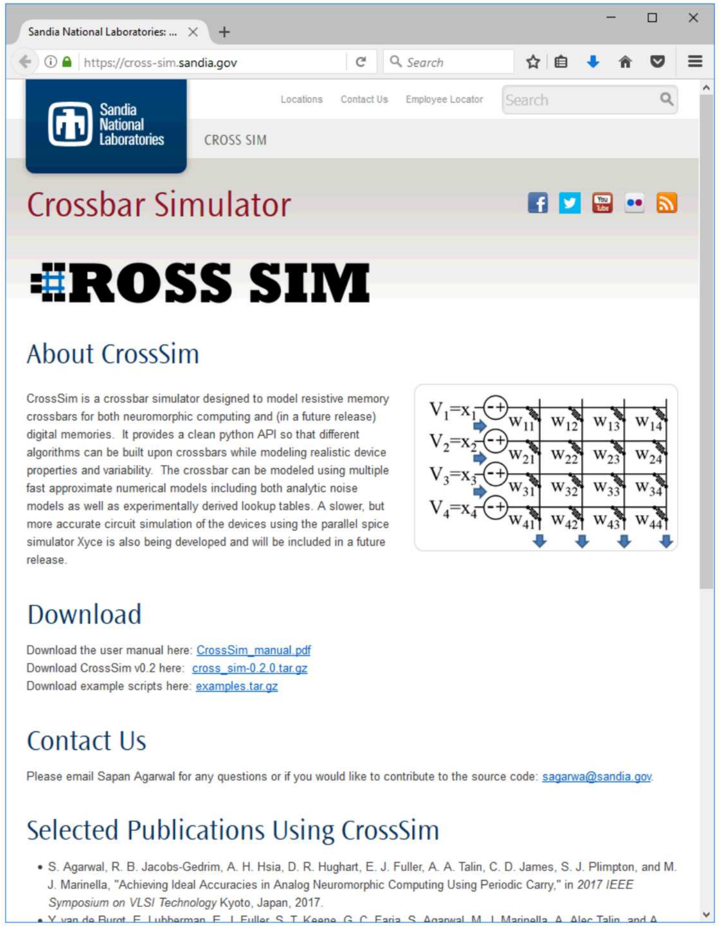






CROSS SIM

https://cross-sim.sandia.gov


Sandia National Laboratories: ...
https://cross-sim.sandia.gov
Search
Locations
Contact Us
Employee Locator
Search
Sandia National Laboratories
CROSS SIM

Crossbar Simulator

CROSS SIM

About CrossSim

CrossSim is a crossbar simulator designed to model resistive memory crossbars for both neuromorphic computing and (in a future release) digital memories. It provides a clean python API so that different algorithms can be built upon crossbars while modeling realistic device properties and variability. The crossbar can be modeled using multiple fast approximate numerical models including both analytic noise models as well as experimentally derived lookup tables. A slower, but more accurate circuit simulation of the devices using the parallel spice simulator Xyce is also being developed and will be included in a future release.

$V_1=x_1$
$V_2=x_2$
$V_3=x_3$
$V_4=x_4$
W_{11} W_{12} W_{13} W_{14}
W_{21} W_{22} W_{23} W_{24}
W_{31} W_{32} W_{33} W_{34}
W_{41} W_{42} W_{43} W_{44}

Download

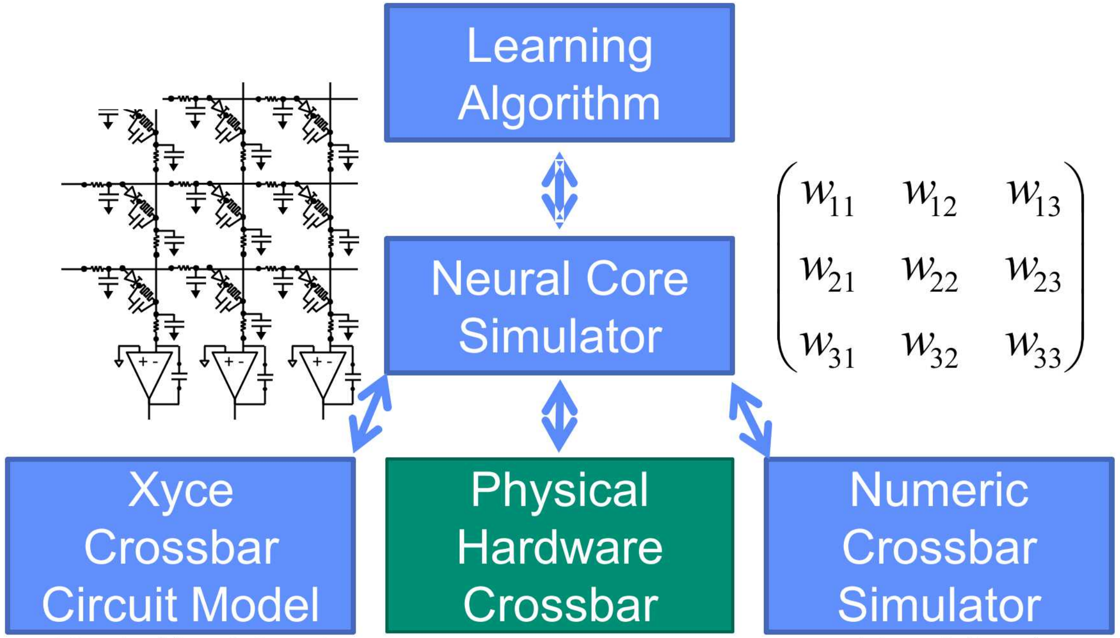
Download the user manual here: CrossSim_manual.pdf
Download CrossSim v0.2 here: cross_sim-0.2.0.tar.gz
Download example scripts here: examples.tar.gz

Contact Us

Please email Sapan Agarwal for any questions or if you would like to contribute to the source code: sagarwa@sandia.gov.

Selected Publications Using CrossSim

- S. Agarwal, R. B. Jacobs-Gedrim, A. H. Hsia, D. R. Hughart, E. J. Fuller, A. A. Talin, C. D. James, S. J. Plimpton, and M. J. Marinella, "Achieving Ideal Accuracies in Analog Neuromorphic Computing Using Periodic Carry," in *2017 IEEE Symposium on VLSI Technology* Kyoto, Japan, 2017.
- Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. Alec Talin, and A.

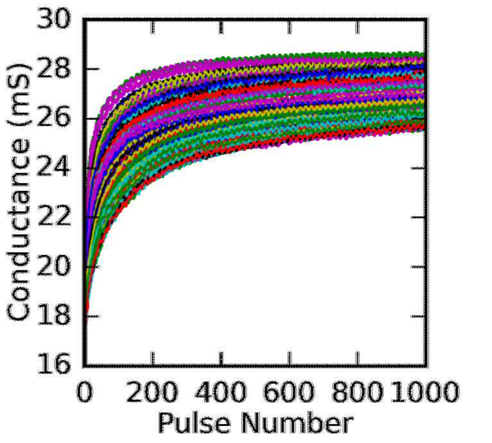

Learning Algorithm
Neural Core Simulator
Xyce Crossbar Circuit Model
Physical Hardware Crossbar
Numeric Crossbar Simulator

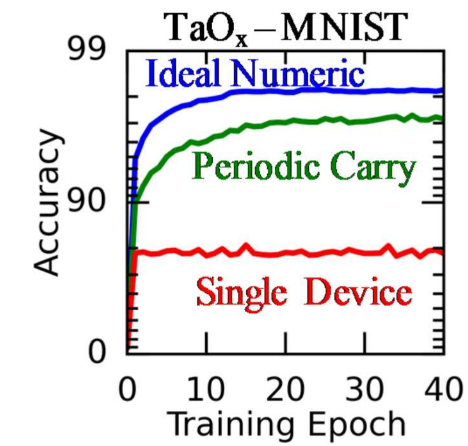
$$\begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix}$$

Detailed but slow

Fast but approximate

Measured Devices ↔ **Algorithmic Performance**



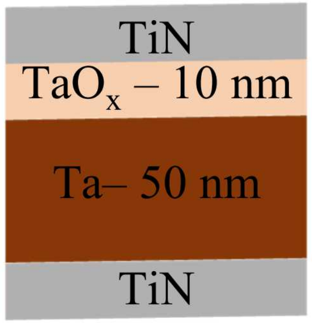


Simple Python API:

```
# Do a matrix vector multiplication
result = neural_core.run_xbar_mvm(vector)
```

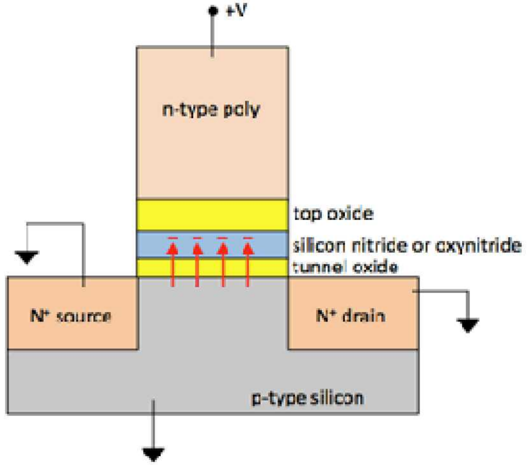
Compare Analog Devices

ReRAM



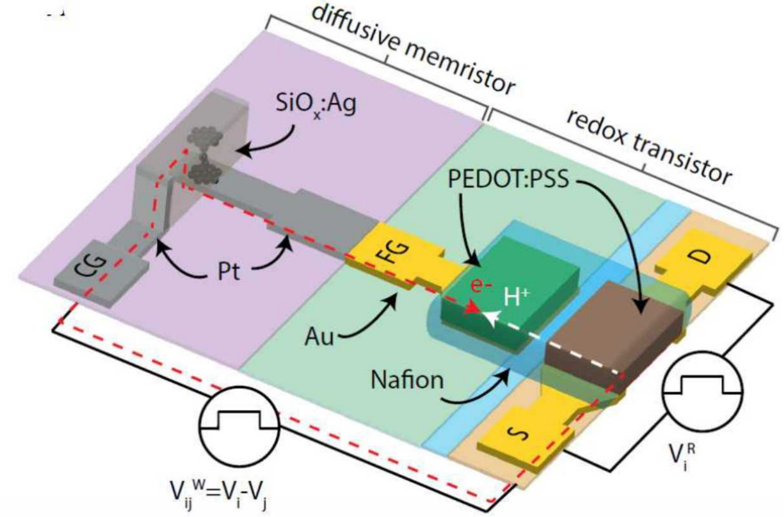
R. B. Jacobs-Gedrim *et al.*, "Impact of Linearity and Write Noise of Analog Resistive Memory Devices in a Neural Algorithm Accelerator," IEEE International Conference on Rebooting Computing (ICRC) Washington, DC, November 2017.

SONOS Silicon-Oxygen-Nitrogen-Oxygen-Silicon



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

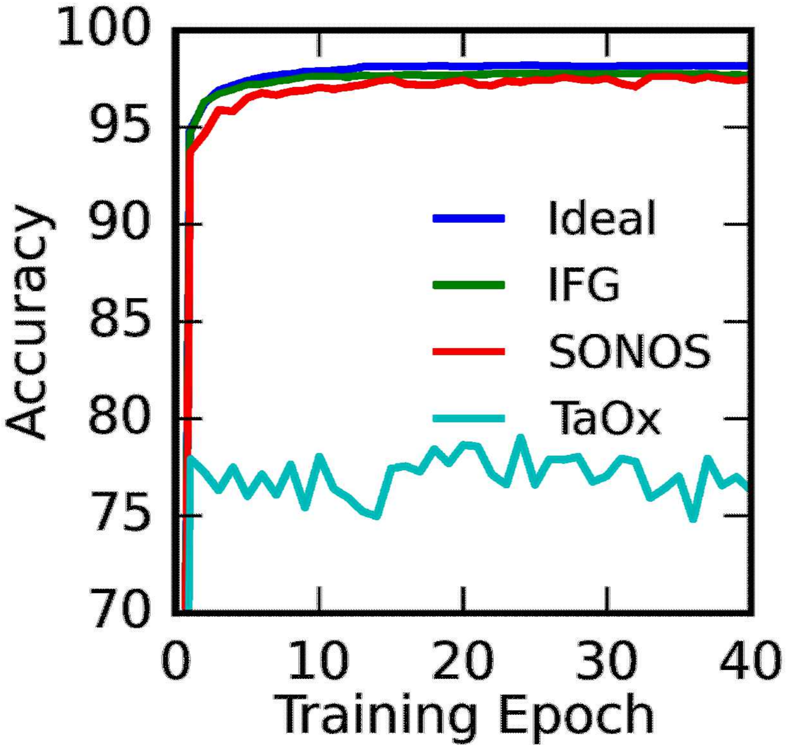
Ionic Floating-Gate Memory



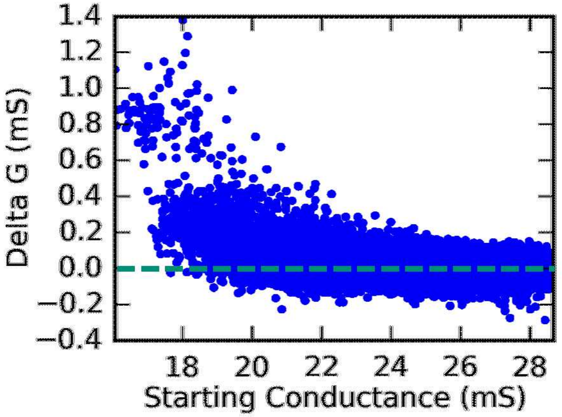
E. J. Fuller *et al.*, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science,* vol. 364, no. 6440, pp. 570-574, 2019.

Three Terminal Devices Tend to Have Higher Accuracy

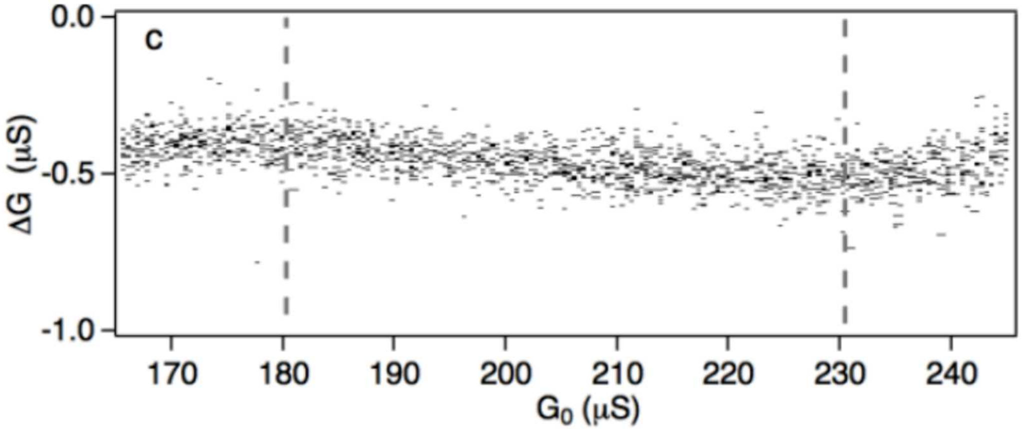
MNIST Handwritten Digits



ReRAM

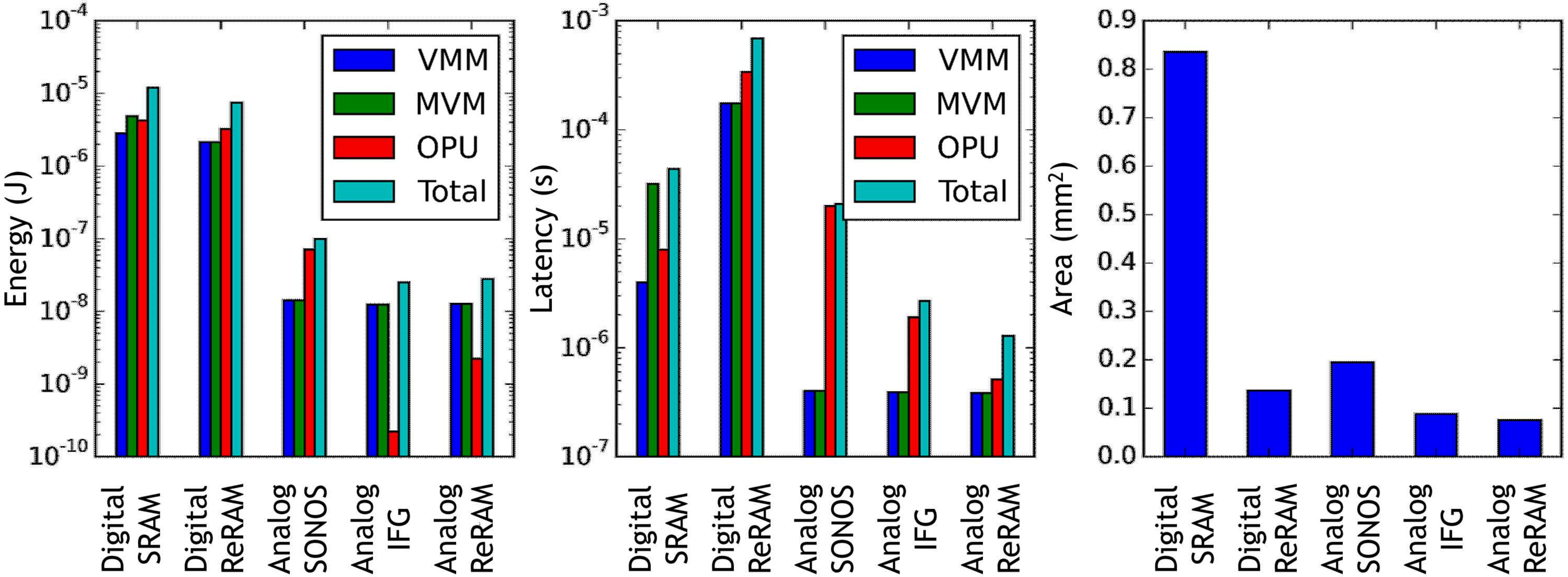


- Lower write noise
- Lower write asymmetry



Ionic Floating-Gate

Compare Architectural Advantages



120-430X Energy Advantage

2-34X Latency Advantage

5-11X Area Advantage

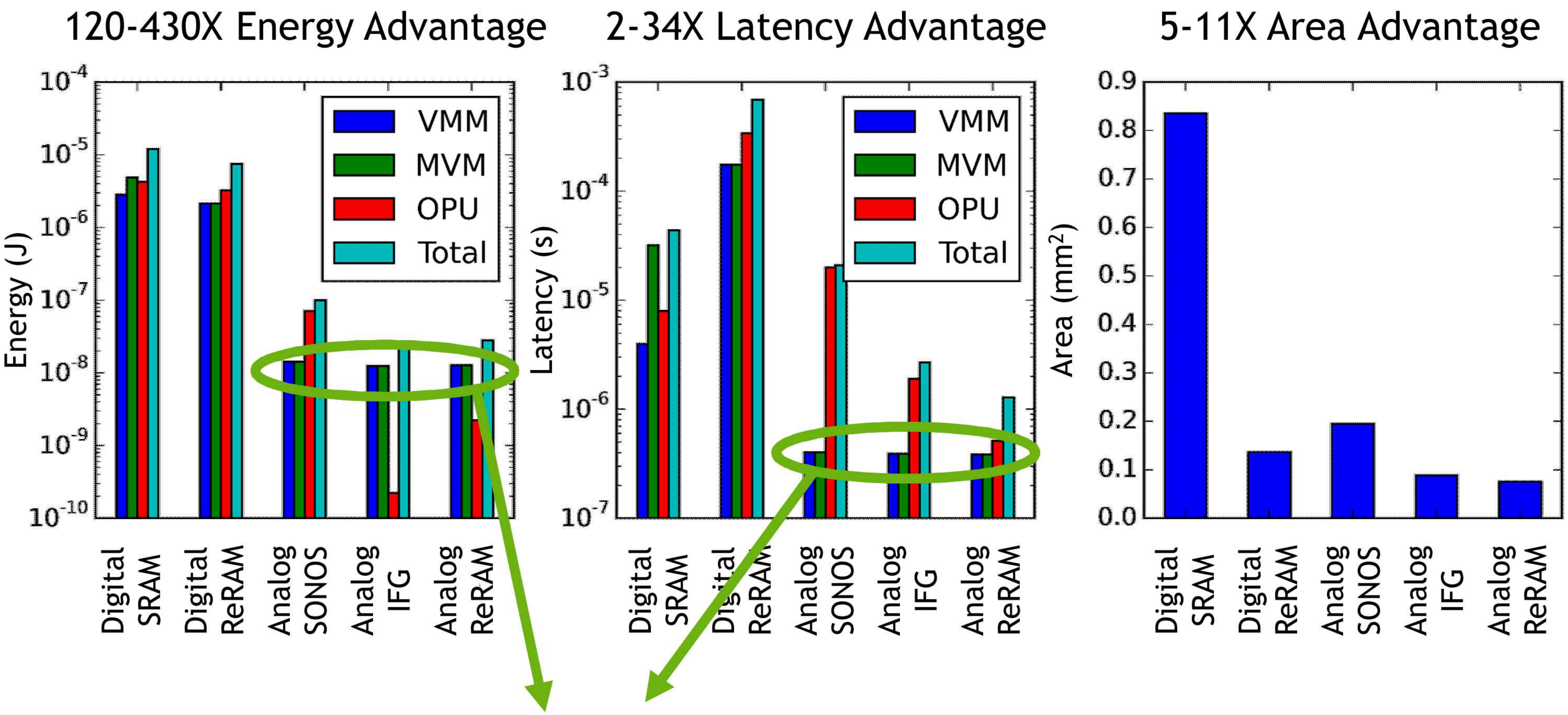
1024 x1024 = 1M array operations, sum over 1 training cycle, 3 operations:

- 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



120-430X Energy Advantage

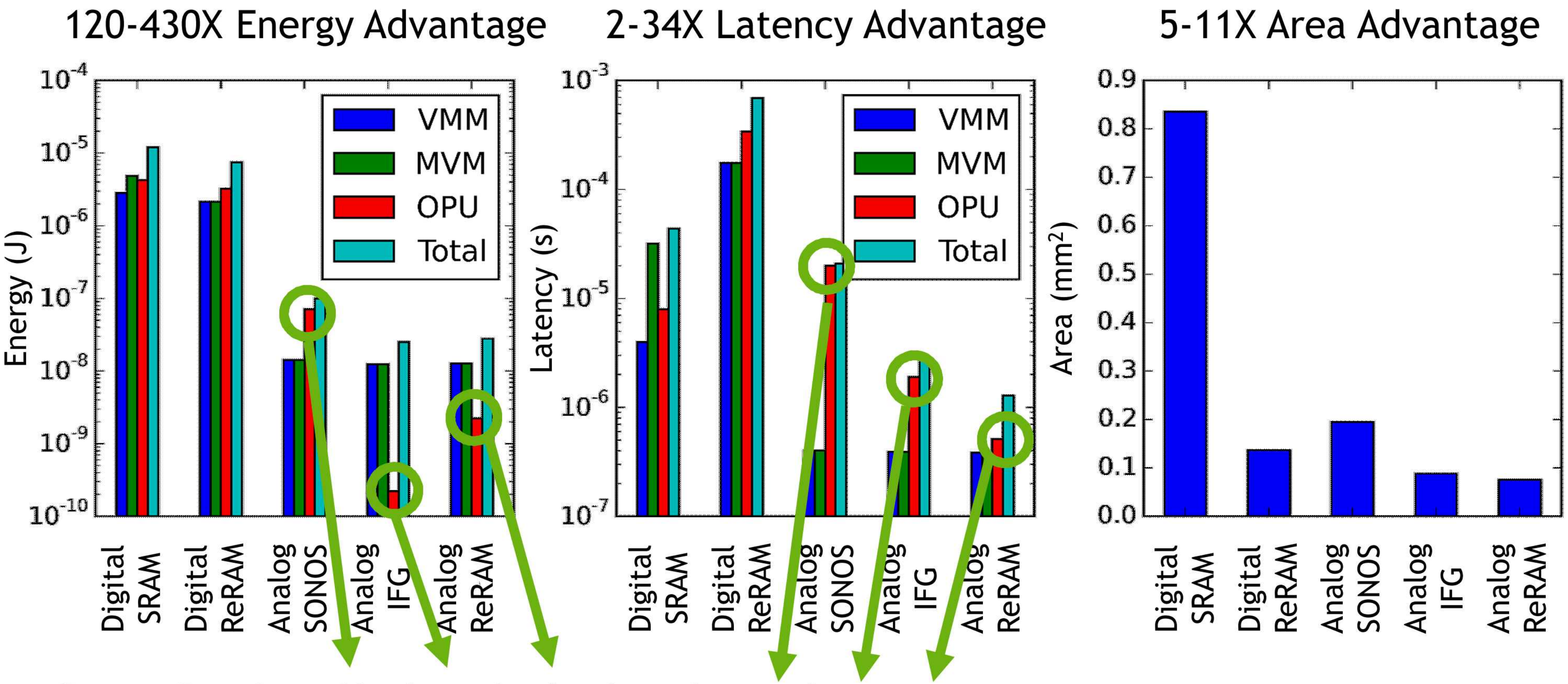
2-34X Latency Advantage

5-11X Area Advantage

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

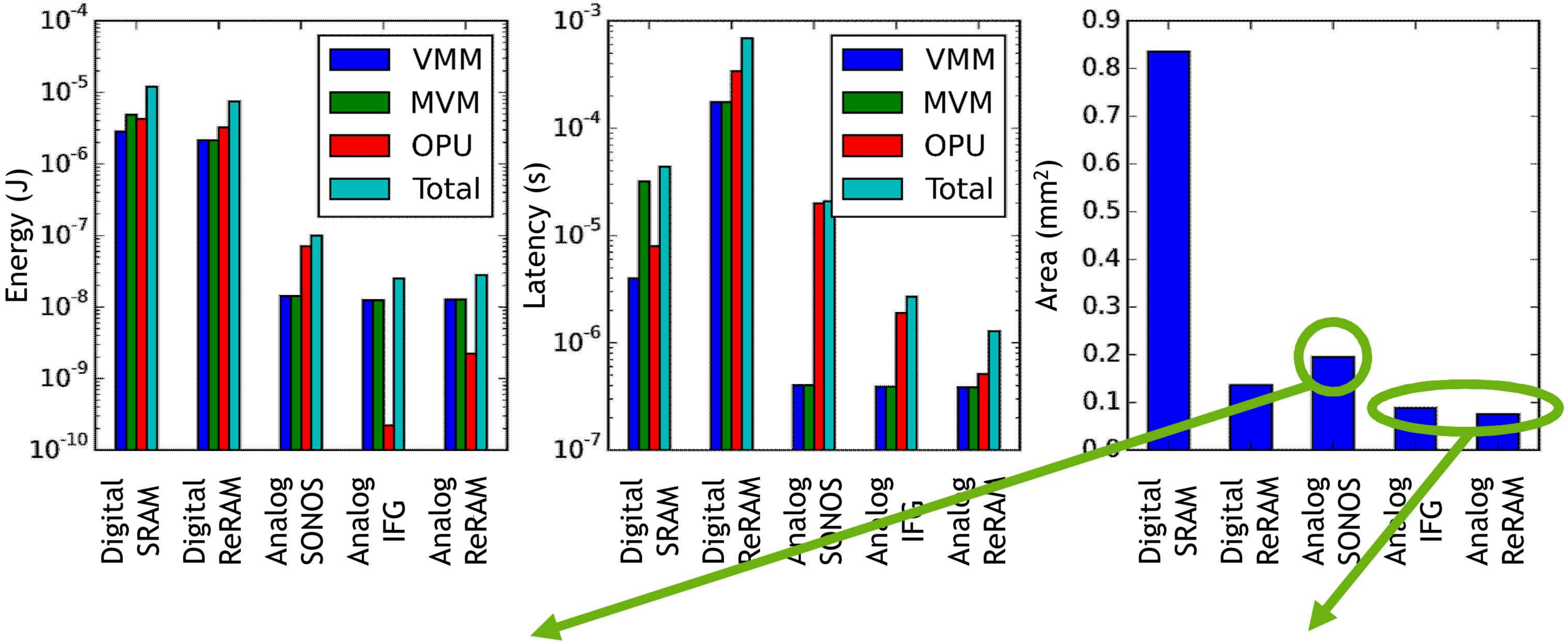
2-34X Latency Advantage

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

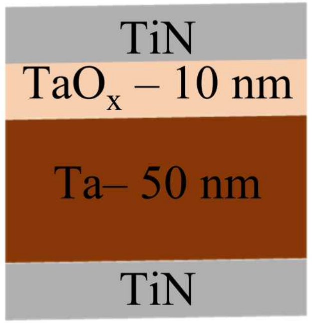
5-11X Area Advantage

SONOS area cost reasonable, roughly doubles area

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

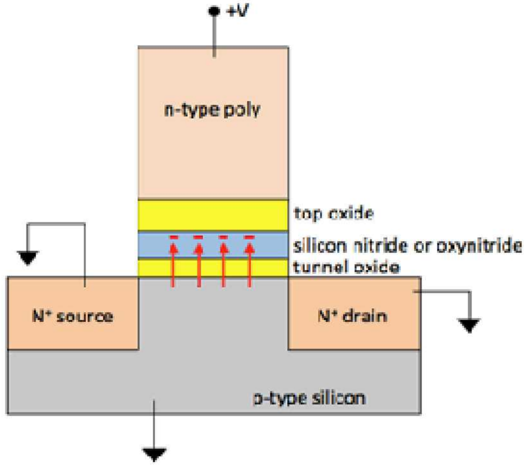
Analog Devices Summary for Training

ReRAM



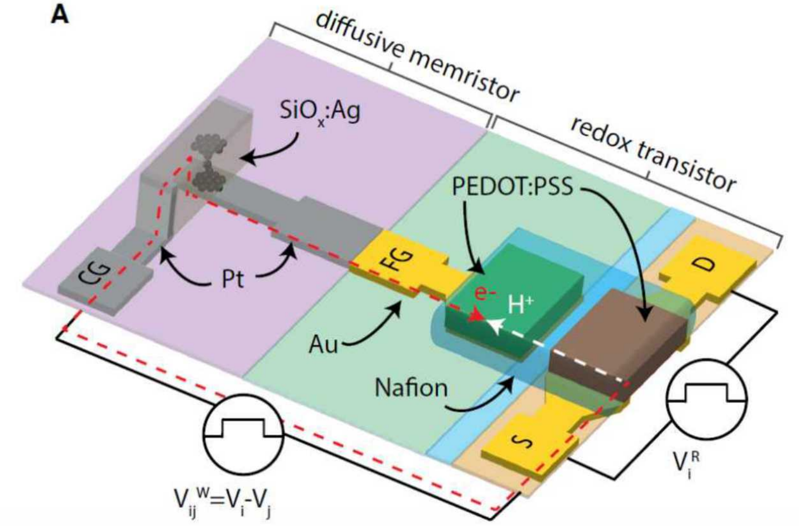
- Large Energy/Area/Latency advantage over digital
- Accuracy not good enough
- Back end of line compatible
- Under commercial development

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

Alternate Computing Paradigms

Crossbar Based Computing Architectures

- Vector Matrix Multiplication
- Outer Product Update
- Crossbar Based Matrix Solvers
- Ternary Content Addressable Memory

Neuro-inspired Computing

- Hyperdimensional Computing
- Local Learning Rules
- Spiking Neural Networks

Probabilistic and Stochastic Circuits

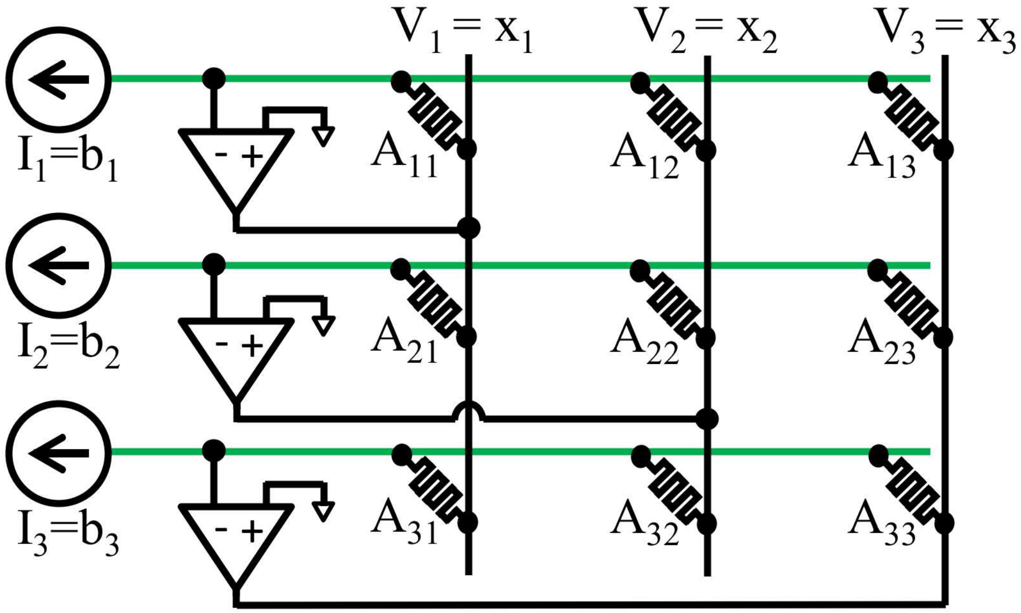
Computing With Dynamical Systems

- Simulated Annealing
- Coupled Oscillator based energy minimization

Analog Matrix Inversion

Analog matrix inversion can perform a dense approximate matrix solve

$$b = Ax \Rightarrow I_{row} = \sum_{col} G_{row,col}(V_{col} - \cancelto{0}{V_{row}})$$

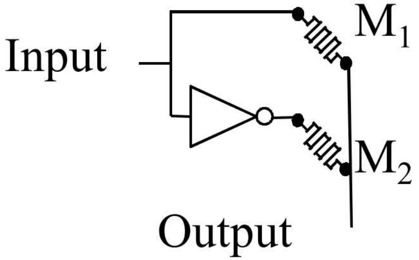


Challenges:

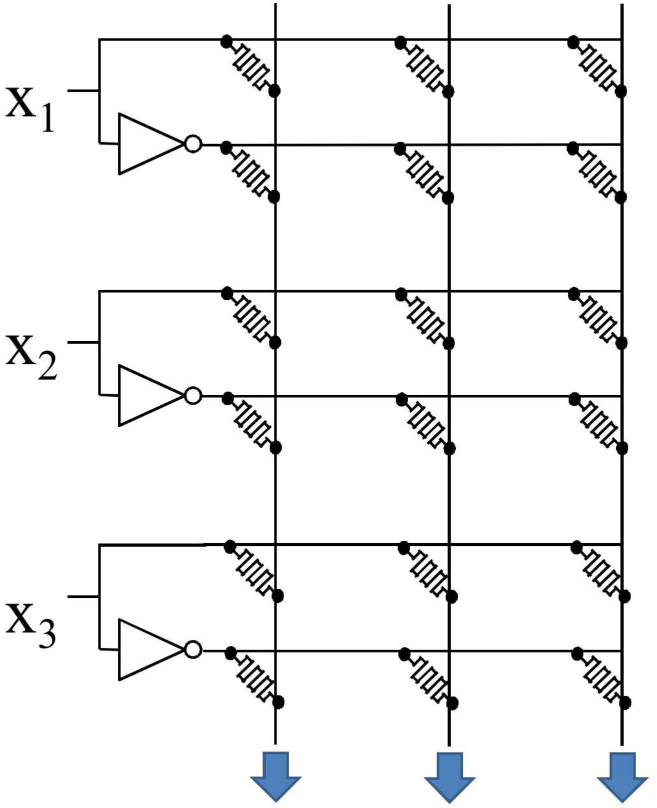
- Matrix inversion is non-linear, limiting how the computation can be split for large matrices
- Analog non-idealities can cause significant errors

Ternary Content Addressable Memory (TCAM)

- Can do very efficient fast pattern matching to search stored data
  - Data analytics, k-nearest neighbors machine learning
  - Sparse matrix multiplication
  - Associative Computing
- Crossbars can implement extremely efficient TCAMs



Input	Stored State	M_1 State	M_1 Output	M_2 State	M_2 Output	Total Output
0	0	0	0	1	1	1
1	0	0	0	1	0	0
0	1	1	0	0	0	0
1	1	1	1	0	0	1
0	x	1	0	1	1	1
1	x	1	1	1	0	1



Alternate Computing Paradigms

Crossbar Based Computing Architectures

- Vector Matrix Multiplication
- Outer Product Update
- Crossbar Based Matrix Solvers
- Ternary Content Addressable Memory

Neuro-inspired Computing

- Hyperdimensional Computing
- Local Learning Rules
- Spiking Neural Networks

Probabilistic and Stochastic Circuits

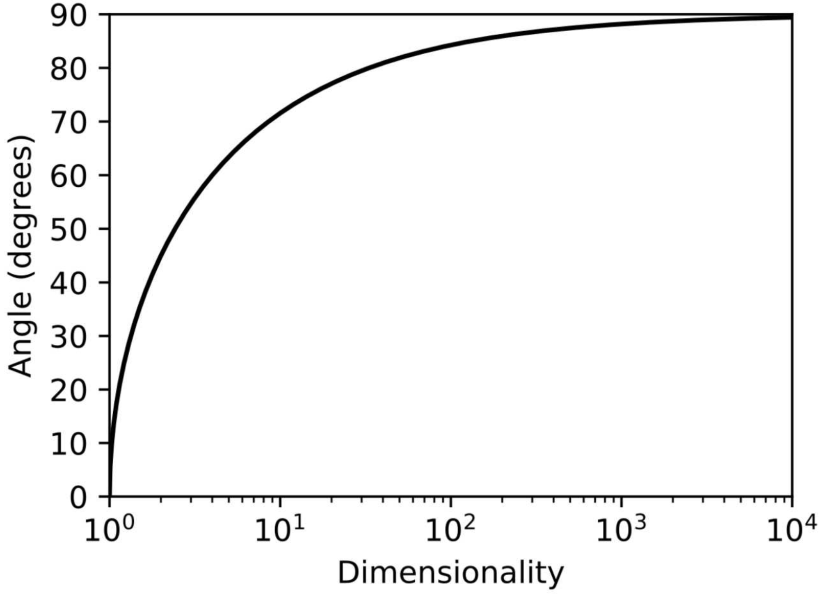
Computing With Dynamical Systems

- Simulated Annealing
- Coupled Oscillator based energy minimization

Hyper-Dimensional Computing

Store data in redundant high dimensional vectors
1 0 0 1 0 1 1 0 1 1 0 1 1 0 0 1 0 1 0 1 1 1 0 0 0 1 1 0 0 1 0

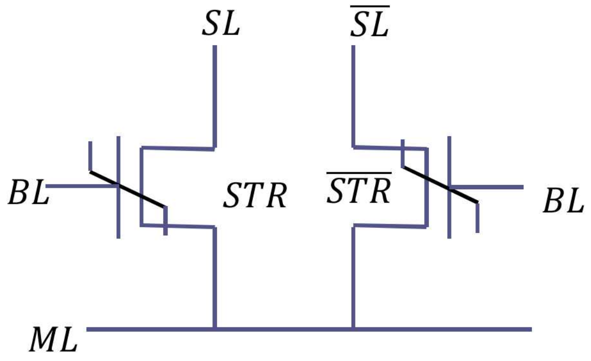
As dimensionality increases, two random vectors are nearly orthogonal



Justin C. Wong. "Negative capacitance and hyperdimensional computing for unconventional low-power computing." PhD thesis, EECS Department, University of California, Berkeley, Dec 2018.

Encode data by combining vectors such that the more similar the data is, the smaller the angle between the vectors

Use ferroelectric content addressable memory to enable pattern matching



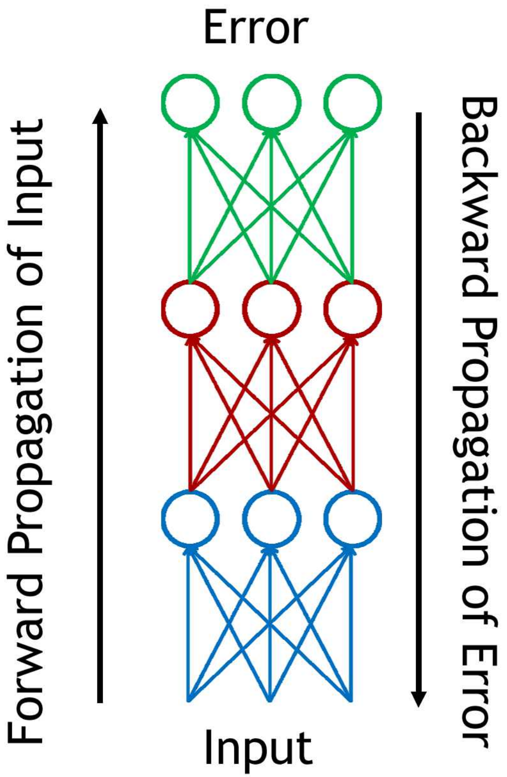
New hardware allows for processing large scale vectors and therefore new algorithms that would otherwise be computationally inefficient

Figures from Justin Wong and Sayeef Salahuddin

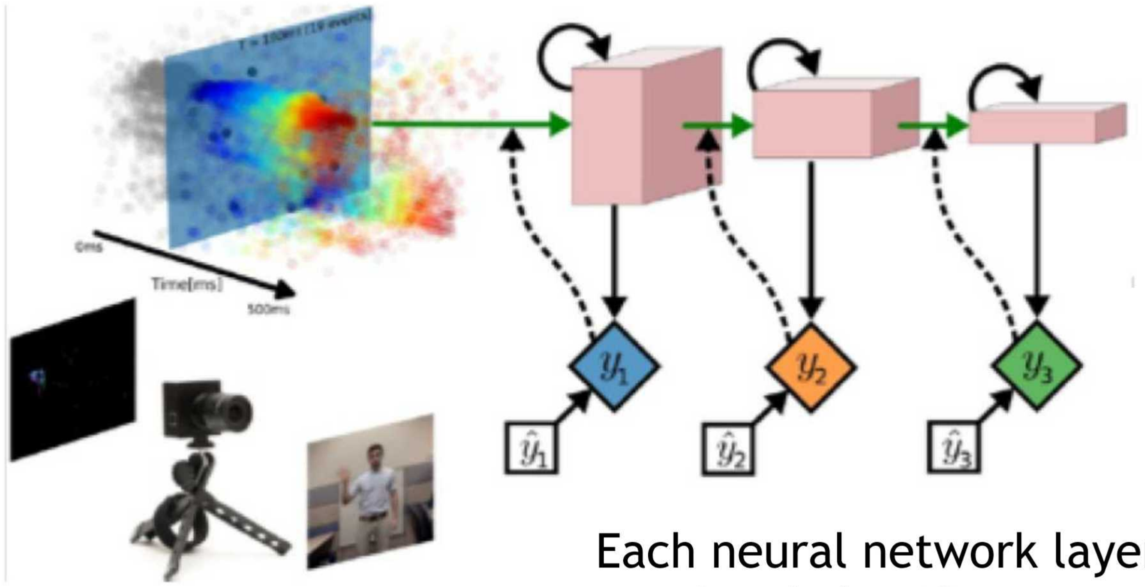
Local Learning Rules

Minimizing Data Movement Directly Minimizes Energy and Latency

Training neural networks requires backpropagating information across all layers resulting in long range communication and storage of intermediate values



Want learning rules that can train using only local information that is present at a given moment in time



Each neural network layer has its own local classifier

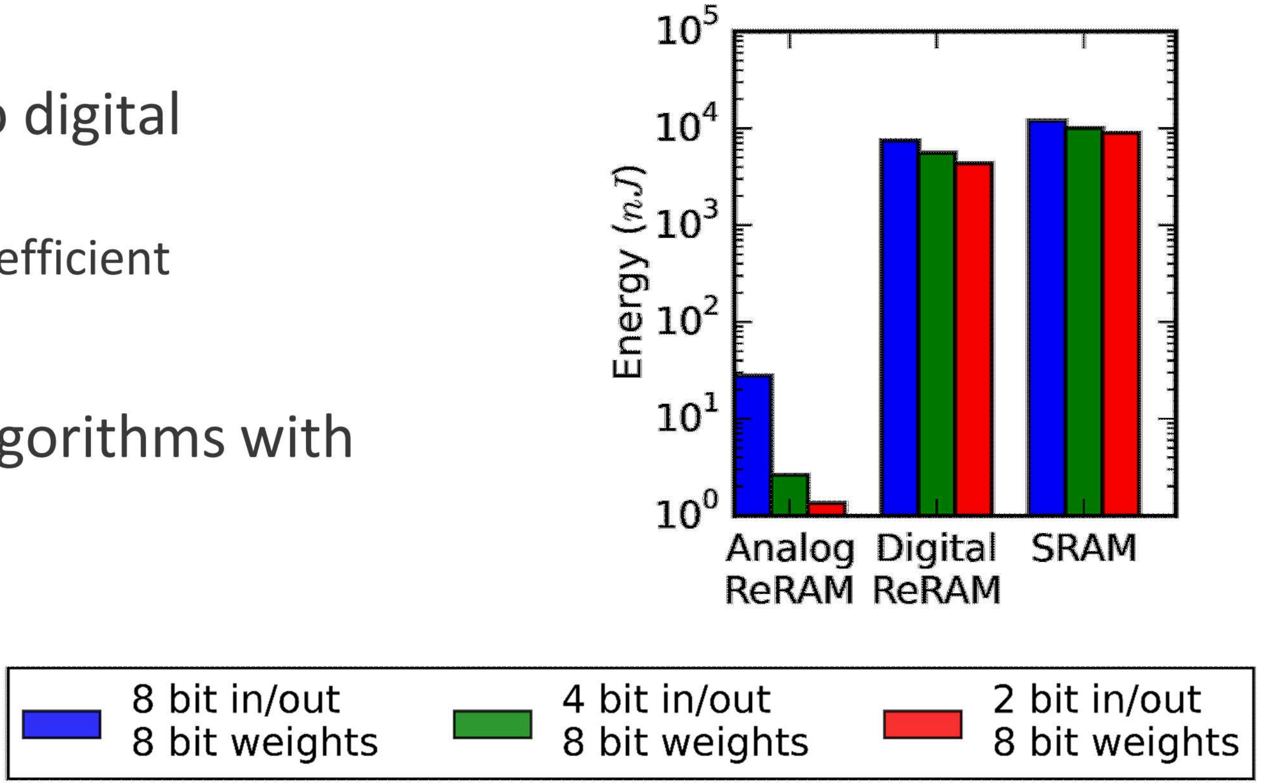
Kaiser, Jacques, Hesham Mostafa, and Emre Neftci. "Synaptic Plasticity Dynamics for Deep Continuous Local Learning (DECOLLE)." *Frontiers in Neuroscience* 14 (2020): 424.

Spiking Neural Networks

Minimizing Data Movement Directly Minimizes Energy and Latency

- For sparse data, communicating only non-zero values is more energy efficient than communicating all data
  - Need to account for overhead of including an address in flexible routing based networks
- Analog system energy is limited by analog to digital conversion
  - Binary outputs from an analog system are far more efficient
- Key challenge is developing high accuracy algorithms with binary inputs and outputs

Analog Neural Network Training Accelerator



Alternate Computing Paradigms

Crossbar Based Computing Architectures

- Vector Matrix Multiplication
- Outer Product Update
- Crossbar Based Matrix Solvers
- Ternary Content Addressable Memory

Neuro-inspired Computing

- Hyperdimensional Computing
- Local Learning Rules
- Spiking Neural Networks

Probabilistic and Stochastic Circuits

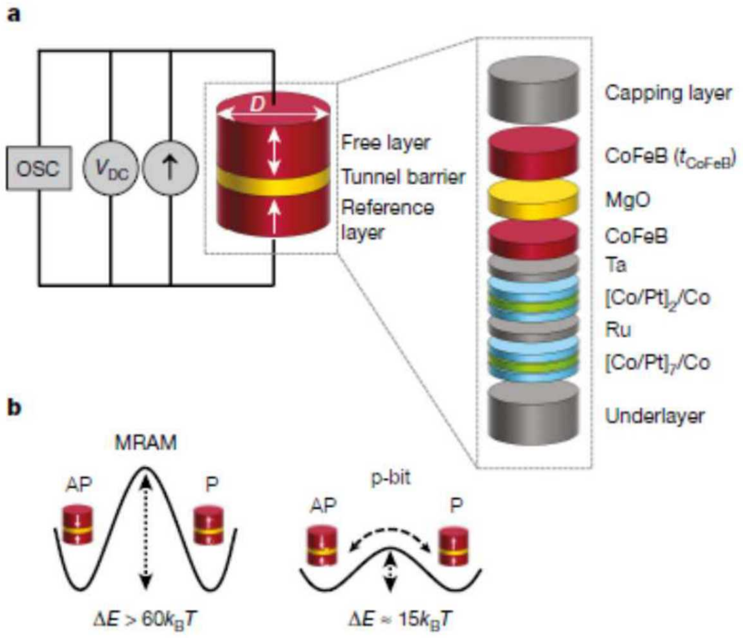
Computing With Dynamical Systems

- Simulated Annealing
- Coupled Oscillator based energy minimization

Probabilistic and Stochastic Circuits

- Generating good random numbers is very computationally intensive
- Compact devices that provide true randomness with tunable probabilities enable new stochastic computing paradigms

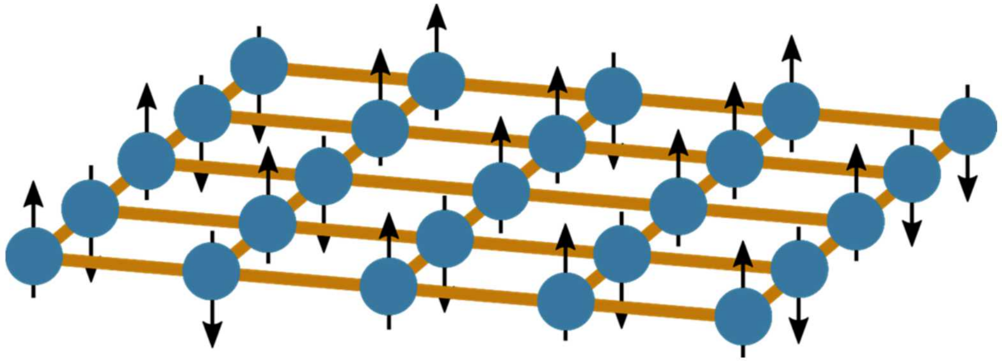
Magnetic Tunnel Junction



Borders, William A., et al. "Integer factorization using stochastic magnetic tunnel junctions." *Nature* 573.7774 (2019): 390-393

- Single Electron Bipolar Avalanche Transistor
  - Avalanche breakdown is stochastic
- ReRAM
  - The intrinsic variability of memristive switching provides a source of randomness
- Contact-Resistive RAM
- CMOS - ring oscillator jitter
- Stochastic Josephson Junctions

Can map optimization problems to a set of connected stochastic bits: The Ising problem



minimize: $$H = -\sum_{i,j} J_{ij} S_i S_j$$

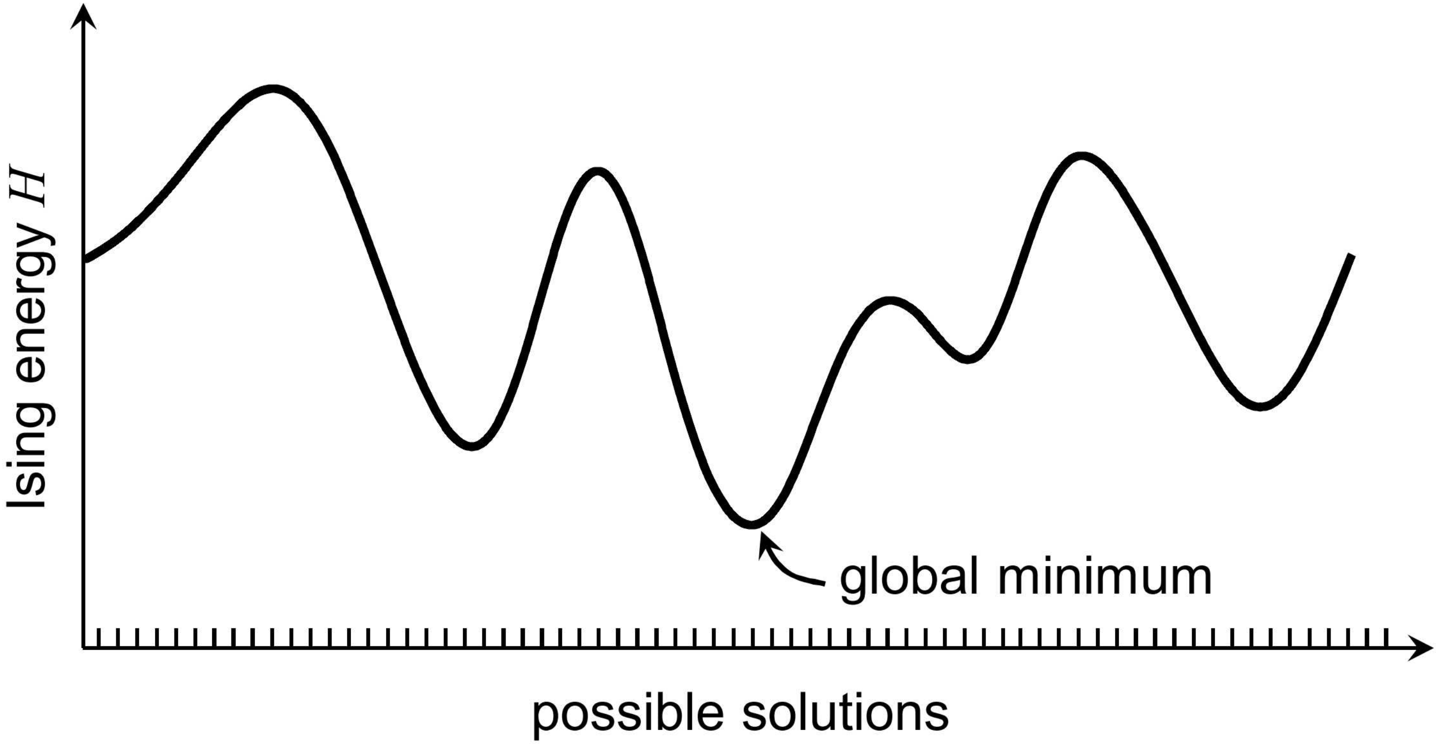
coupling

spins $\{-1, +1\}$

A. Lucas, "**Ising formulations of many NP problems**," 2014

Slide from Tianyao P. Xiao and Eli Yablonovitch

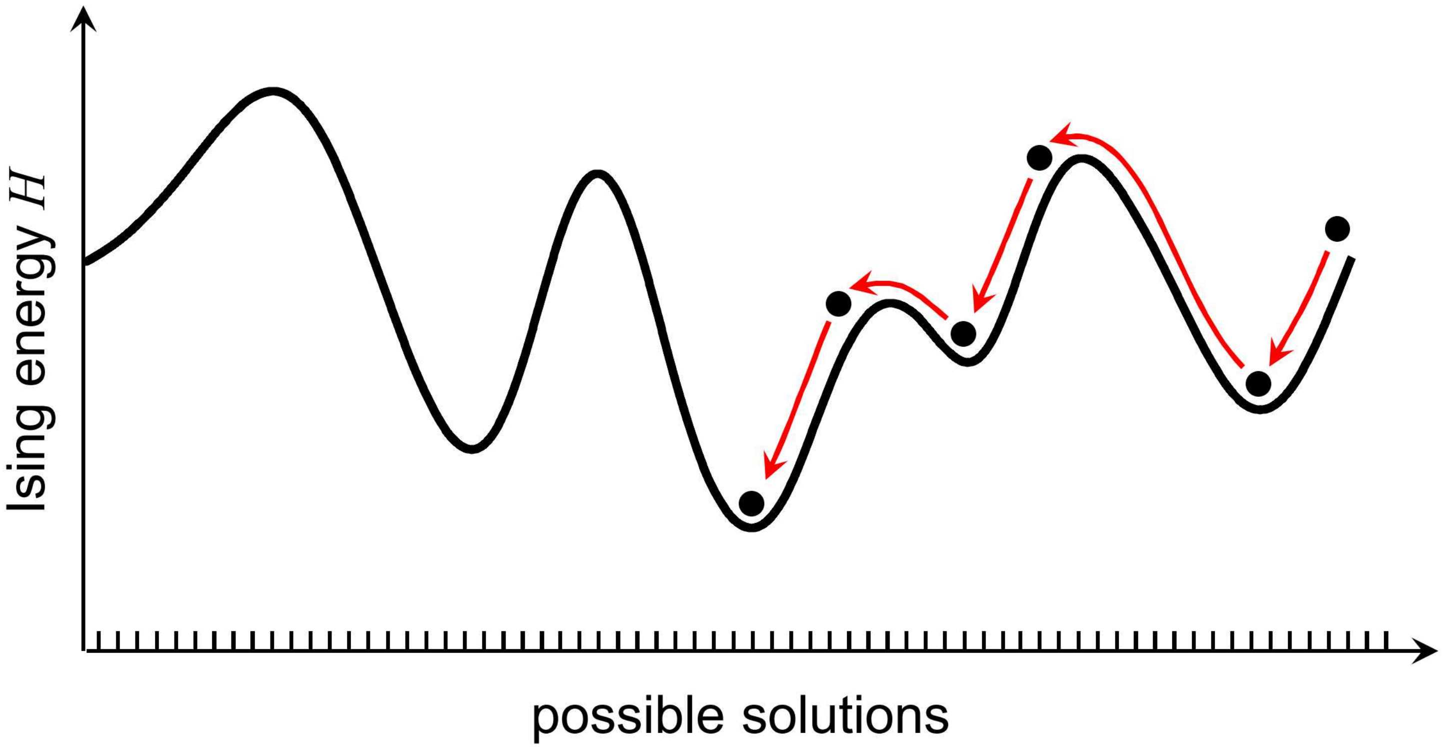
Landscape of the Ising problem



Slide from Tianyao P. Xiao and Eli Yablonovitch

Method 1: **Simulated annealing**

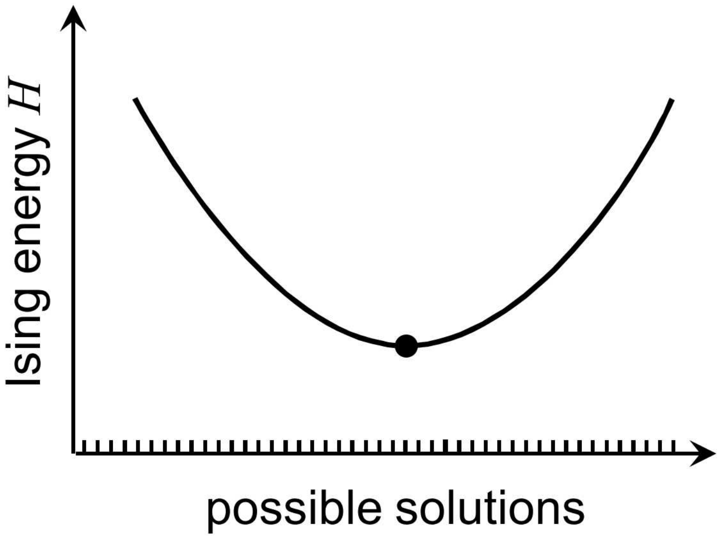
(and other digital heuristics)



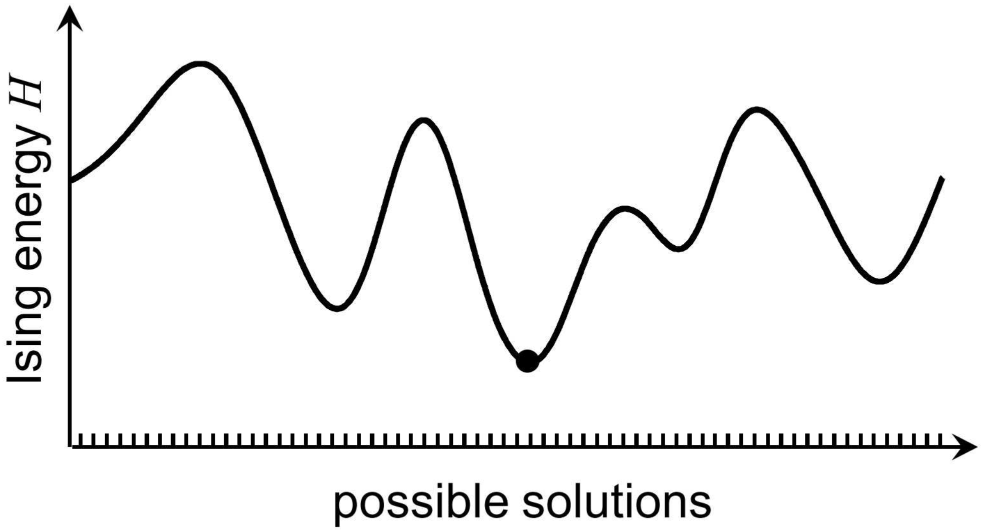
Slide from Tianyao P. Xiao and Eli Yablonovitch

Method 2: **Adiabatic quantum optimization**

Prepare the ground state of a simple problem



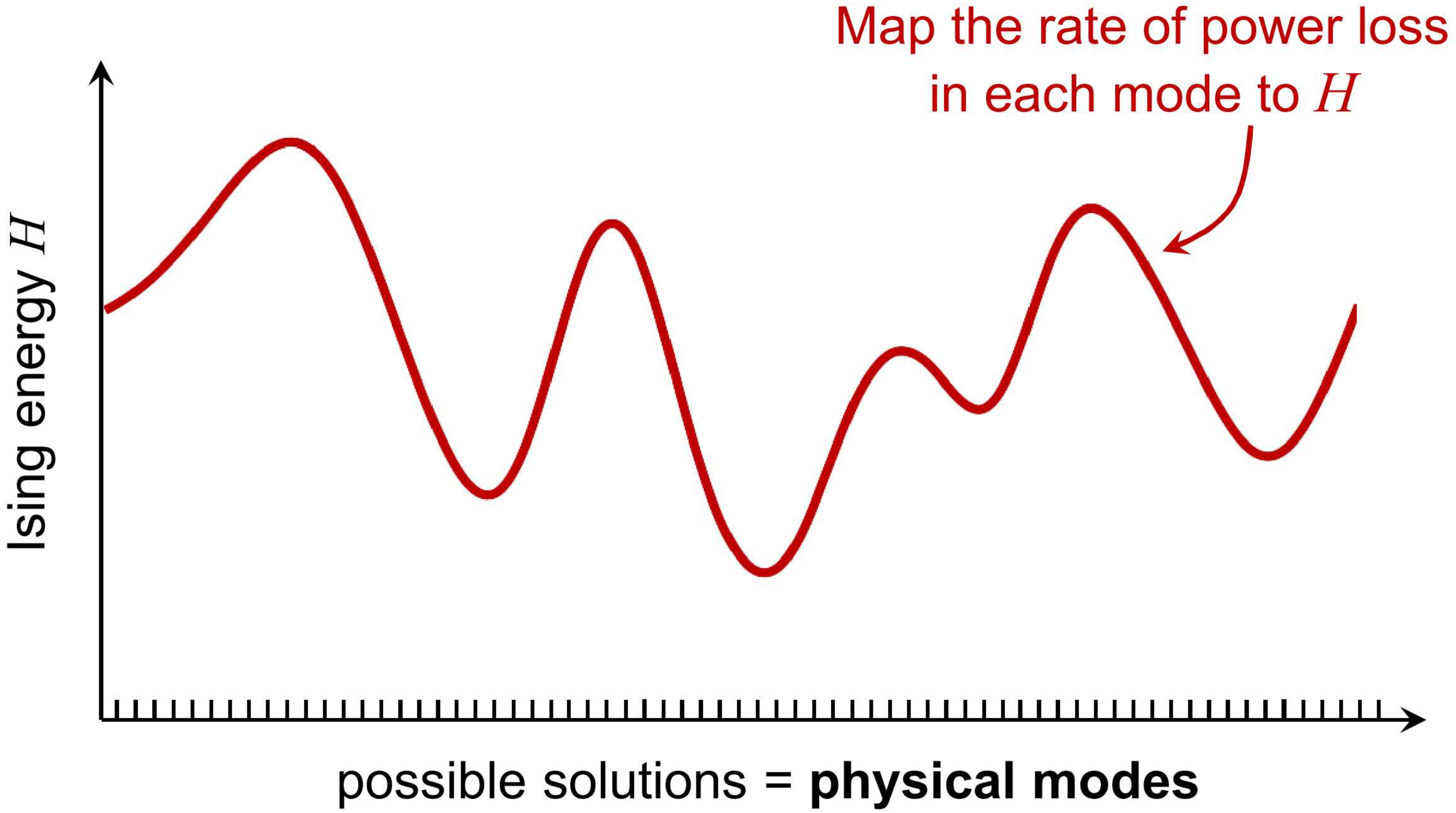
Transform the system into the desired problem



If done *slowly enough*, the system is guaranteed to remain in the ground state during the full evolution

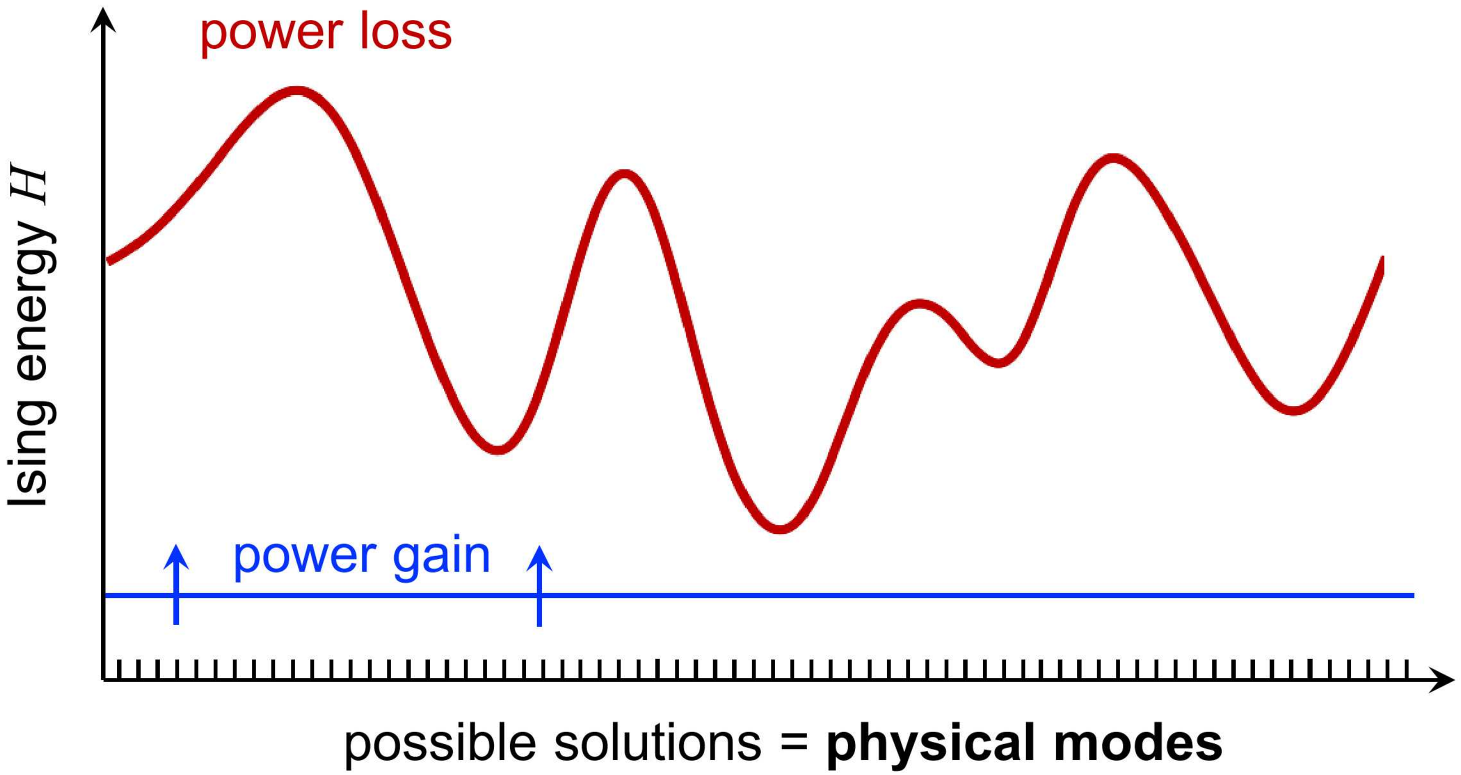
Slide from Tianyao P. Xiao and Eli Yablonovitch

Method 3: **First to threshold**


Map the rate of power loss
in each mode to H
Ising energy H
possible solutions = **physical modes**

Slide from Tianyao P. Xiao and Eli Yablonovitch

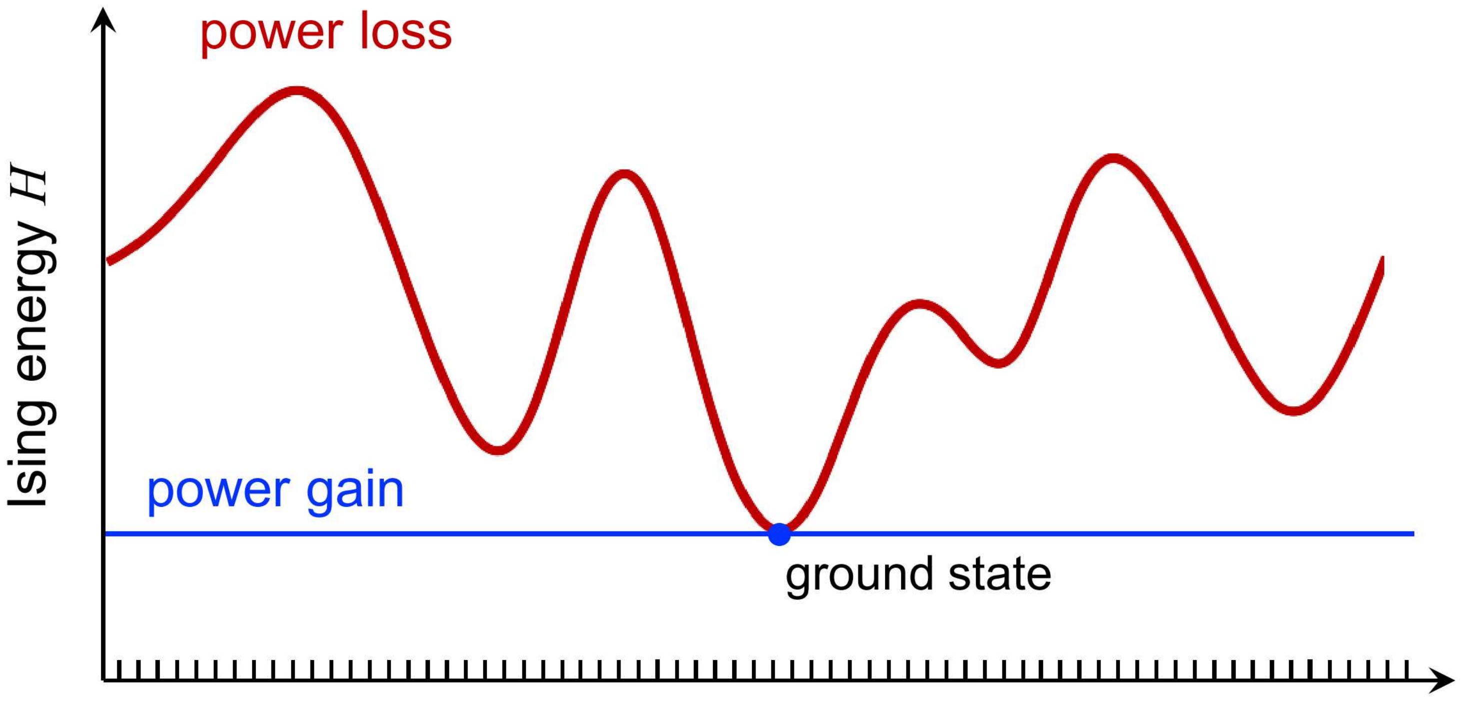
Method 3: **First to threshold**



No mode is stable – noise dominates in circuit

Slide from Tianyao P. Xiao and Eli Yablonovitch

Method 3: **First to threshold**



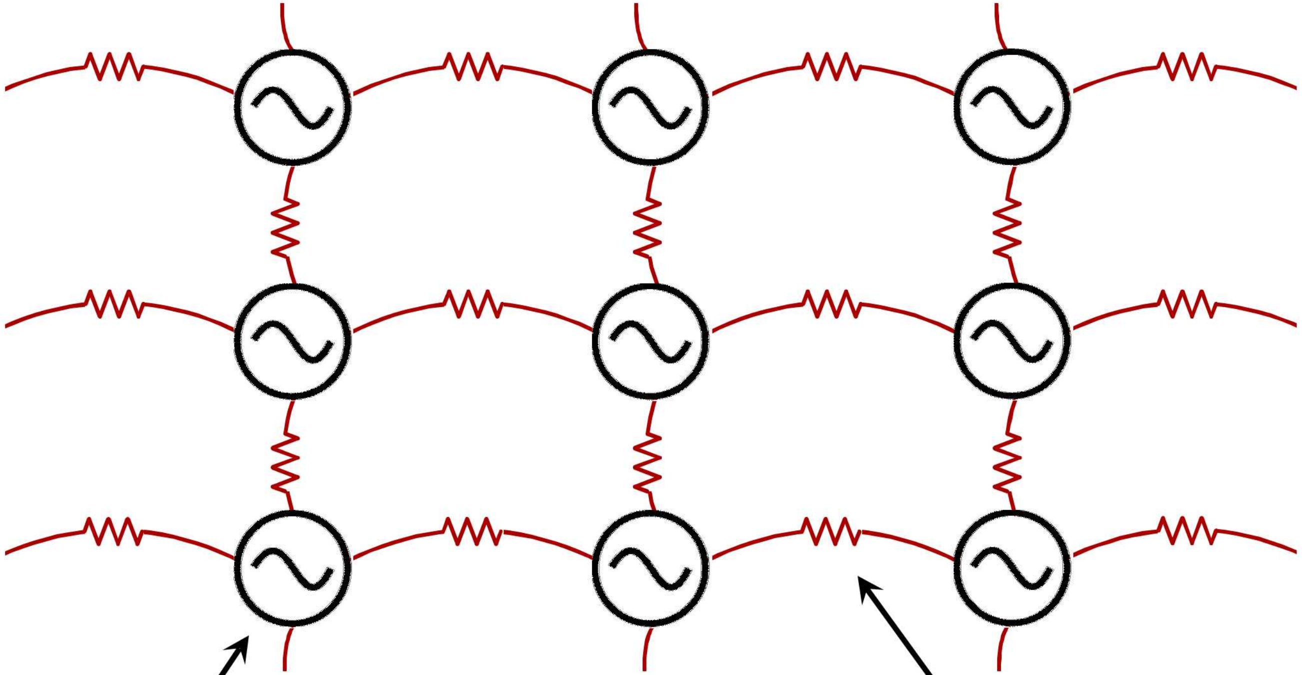
possible solutions = **physical modes**

A stable mode emerges, representing the ground state!

Slide from Tianyao P. Xiao and Eli Yablonovitch

Analog electronic Ising machine

(high-level view)



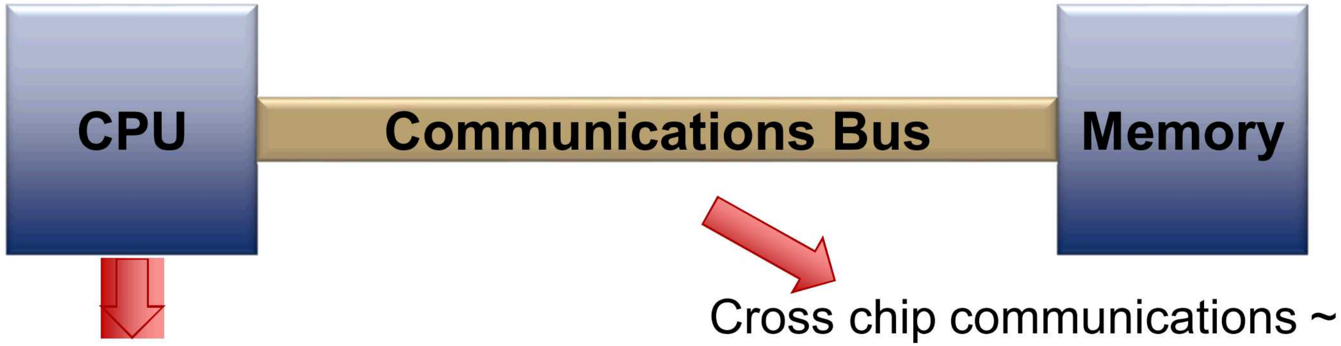
bistable LC oscillators

resistive coupling elements

Slide from Tianyao P. Xiao and Eli Yablonovitch

T. P. Xiao, “Optoelectronics for refrigeration and analog circuits for combinatorial optimization,” Ph.D. dissertation, Chapter 6, University of California, Berkeley, 2019

Beyond Moore Technologies



Current Transistors ~ 10 aJ
50kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ

Extending Von Neumann

- Low Voltage or Novel Transistors
- Optical Communications
- Reduced Data Movement
  - New On-Chip Memory
  - Processing in Memory

Alternate Computing Paradigms

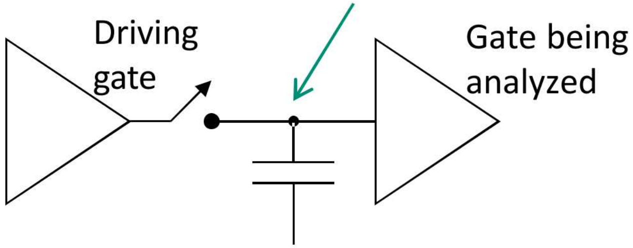
- Neuromorphic
- Analog
  - Computing with memory devices
- Quantum
- Stochastic
- Approximate

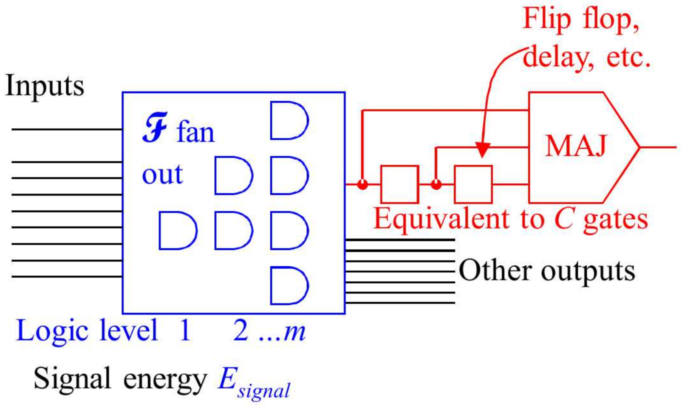
Going Below 50 kT

- Error Correction
- Reversible Computing
  - Adiabatic Computing / Energy Recycling
- Superconducting

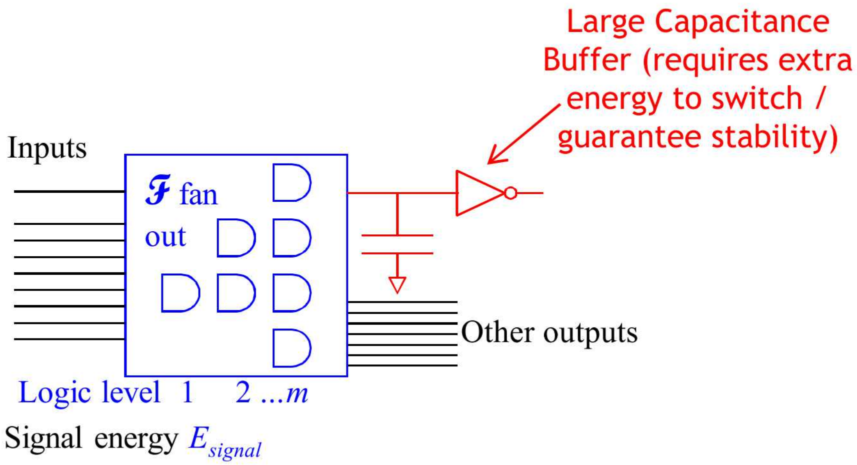
Temporal Error Correction

$$E_{signal} = CV^2 \longrightarrow P(Error) = e^{-E_{signal}/kT}$$





Use multiple samples and take the majority of the result

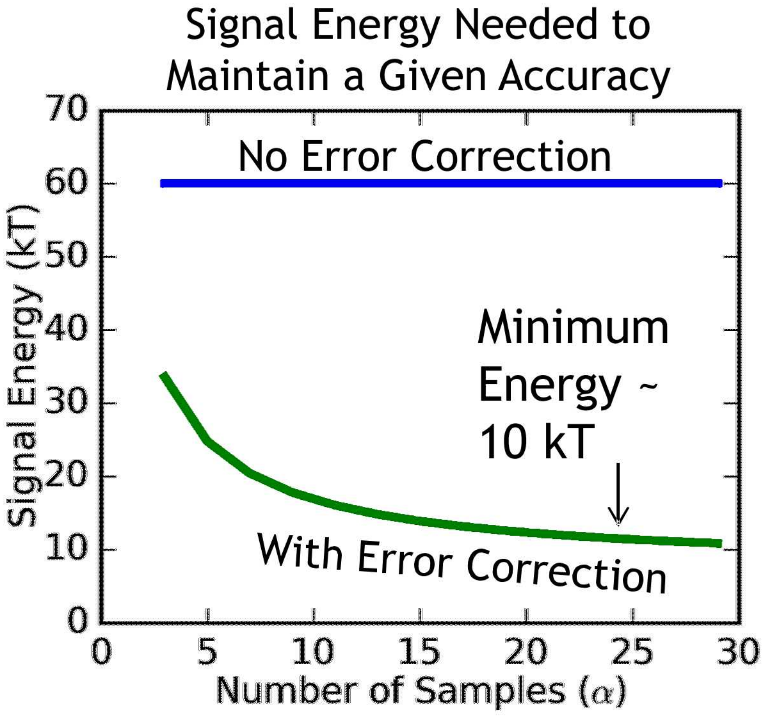


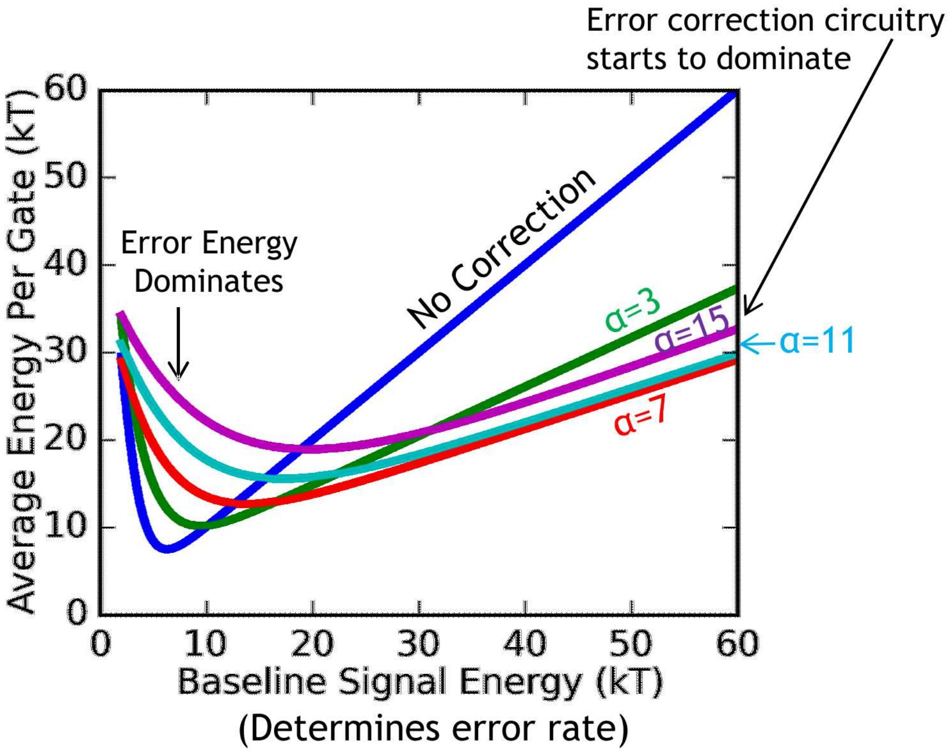
Large Capacitance Buffer (requires extra energy to switch / guarantee stability)

Use high capacitance stage / low pass filter to integrate out transient thermal errors

How Much Energy is Saved by a Majority Gate?

Take the majority of α samples





Consider a 16x16 multiplier

- 48 levels of logic depth
- 32 inputs/outputs (any input can affect any output)

Can get around a 2X reduction in energy

Reversible Computing

Consider a signal Energy E_{signal}

The probability of an error due to thermal noise is:

$$P(Error) = e^{-E_{signal}/kT}$$

In order to ensure a full system with billons of transistors is reliable, we need:

$$E_{signal} \sim 50\ \mathrm{kT}$$

Landauer – Shannon Limit

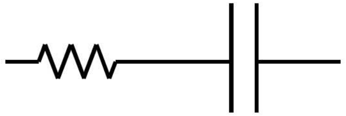
What is E_{signal}?

It could be the energy on a single irreversible gate

It could also be the energy in a reversible system that computes a complex logic function the comprises many logical functions

In both cases the signal energy is the same!

Adiabatic Computing



To switch, need Q= C x V_{dd}

Conventional

- supply charge through a resistor, R, with voltage V_{dd} across it.
- The time it takes is RC

$$E = \frac{V_{dd}{}^2}{R} \times RC = CV_{dd}{}^2 = V_{dd} \times Q$$

Adiabatic

- Reduce the power burned in the resistor, by minimizing the voltage across it
  - Charge the circuit with a lower current, I_{low}
- This takes a longer time, τ, to get the required charge: $Q = C \times V_{dd} = I_{low} \times \tau$

$$\begin{aligned} E &= I_{low}{}^2 \times R \times \tau \\ &= I_{low} \times R \times \left(I_{low} \times \tau\right) \\ &= I_{low} \times R \times Q \\ &= V_{low} \times Q \end{aligned}$$

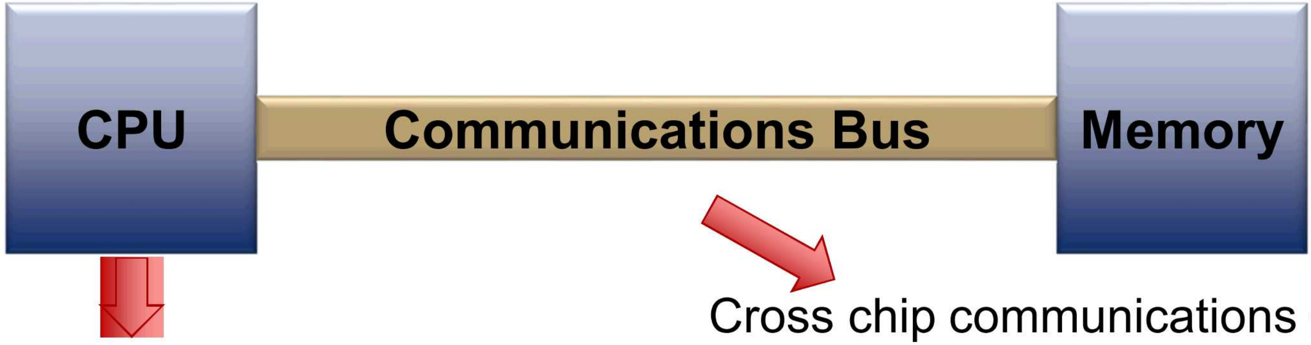
Energy Reduced by:

$$\frac{E_{adiabatic}}{E_{classical}} = \frac{V_{low}}{Vdd} = \frac{I_{low}R}{V_{dd}} = \frac{\frac{CV_{dd}}{\tau}R}{V_{dd}} = \frac{RC}{\tau}$$

Delay Increased By: $\frac{\tau}{RC}$

(ignoring factors of 2 for simplicity)

Summary: There are many ways to extend Moore’s Law!



Current Transistors ~ 10 aJ
50kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ

Extending Von Neumann

- Low Voltage or Novel Transistors
- Optical Communications
- Reduced Data Movement
  - New On-Chip Memory
  - Processing near Memory

Alternate Computing Paradigms

- Neuromorphic
- Analog
  - Computing with memory devices
- Quantum
- Stochastic
- Approximate

Going Below 50 kT

- Error Correction
- Reversible Computing
  - Adiabatic Computing / Energy Recycling
- Superconducting

Sandia is Hiring! sandia.gov/careers

~12,500 people in Albuquerque, NM

~1,800 people in Livermore, CA

Fulfilling Our National Security Mission

Nuclear Deterrence

Defense Nuclear Nonproliferation

National Security Programs

Energy & Homeland Security

Advanced Science & Technology