Device Technologies for the N3XT 1,000X Improvement in Computing Performance

H.-S. Philip Wong
Stanford University

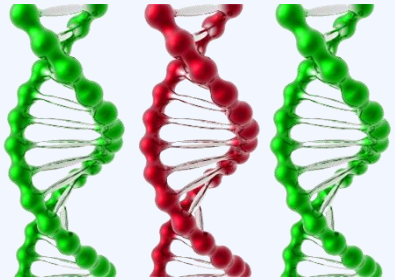
21st century workload:

Large amounts of loosely-structured data

- Streaming video/audio
- Natural languages
- Real-time sensor
- Contextual environment

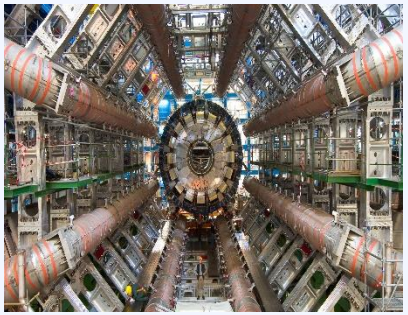
Abundant-Data Applications

Genomics



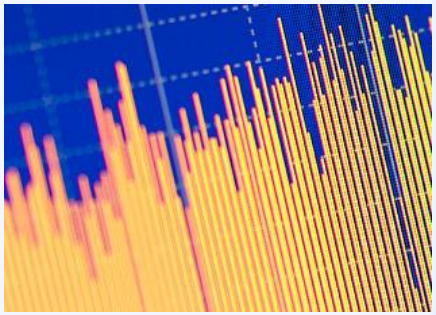
Smart Cities

Science



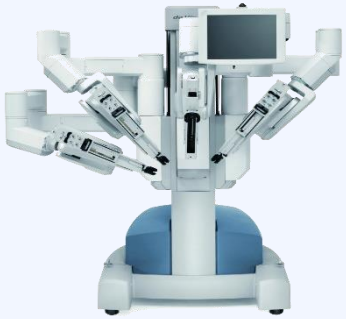
Retail

Finance

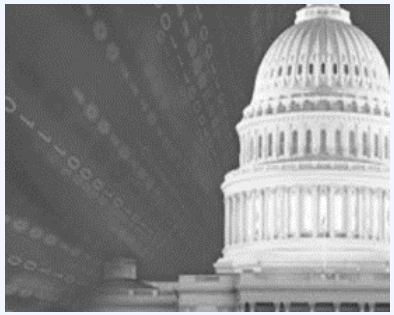


Security

Health Care

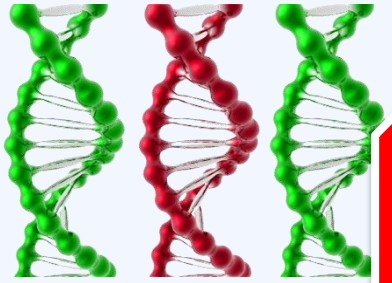


Government



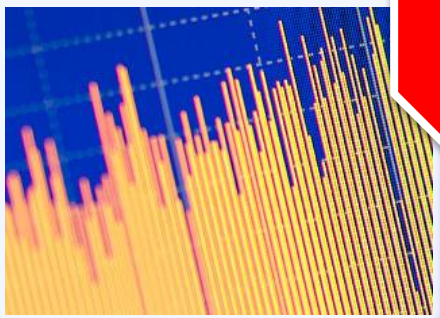
Abundant-Data Applications

Genomics

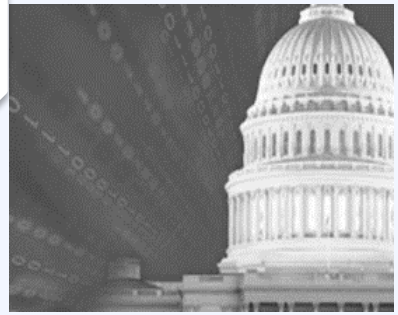


Retail

Finance



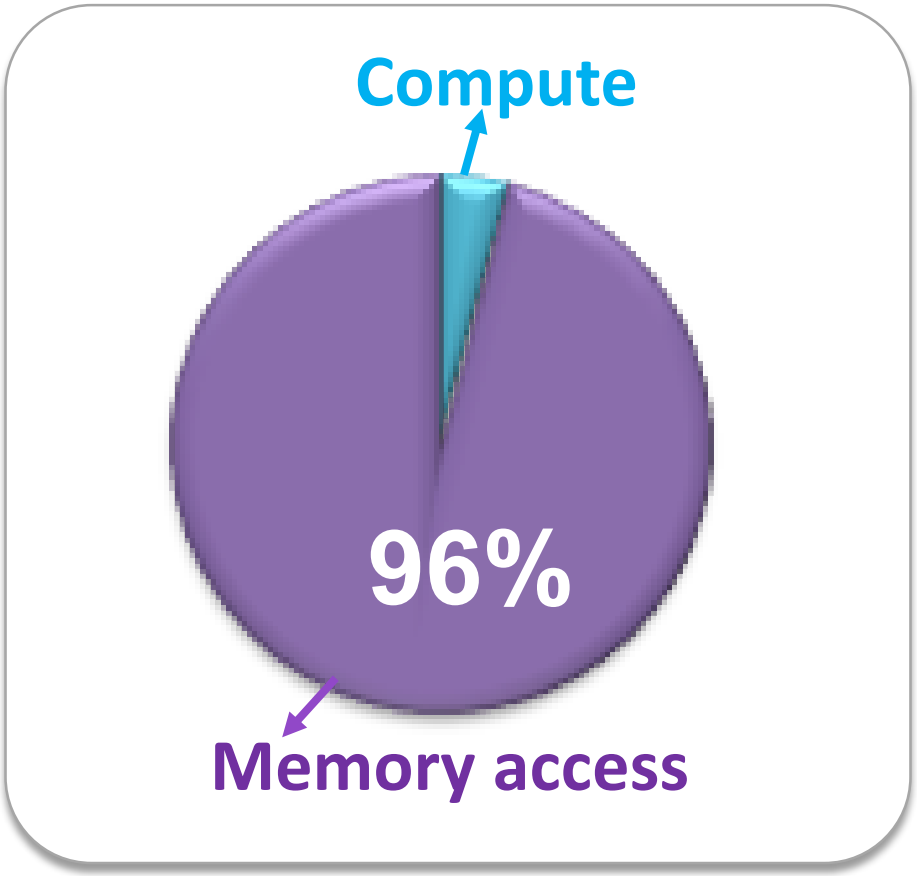
Government



Computational demands *exceed* Processing capability

Abundant-Data Applications

Huge memory wall



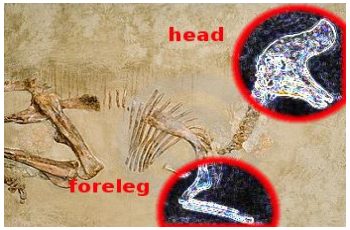
Application execution time

Source: S. Mitra (Stanford)

Hardware-Software Co-Evolution

Commodity hardware

Today



Advanced image analysis

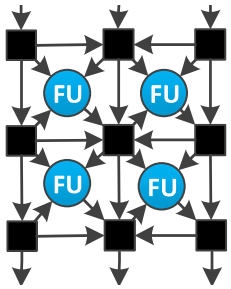
Hardware

Software

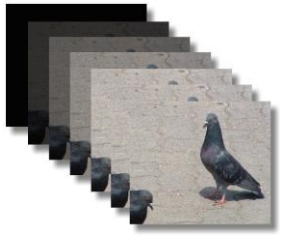
Source: Subhasish Mitra, Chris Ré

Hardware-Software Co-Evolution

CGRAs



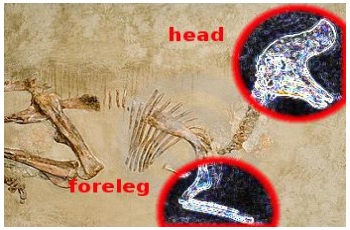
Tomorrow



Automatic video annotation

Commodity hardware

Today



Advanced image analysis

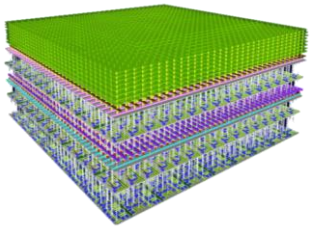
Hardware

Software

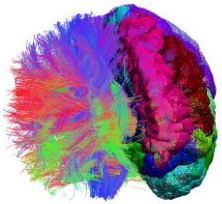
Source: Subhasish Mitra, Chris Ré

Hardware-Software Co-Evolution

N3XT

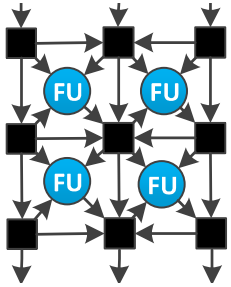


The Day After Tomorrow



Brain network

CGRAs

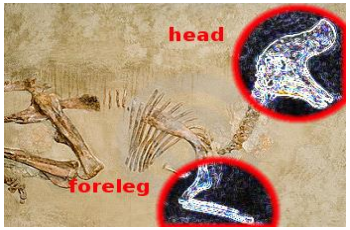


Tomorrow

Automatic video annotation

Commodity hardware

Today



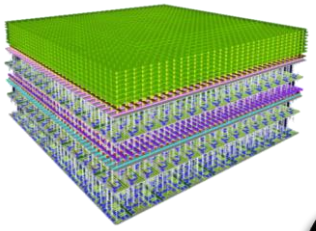
Advanced image analysis

Hardware

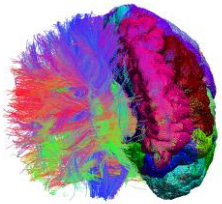
Software

Source: Subhasish Mitra, Chris Ré

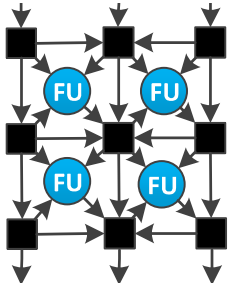
SUPERCELL



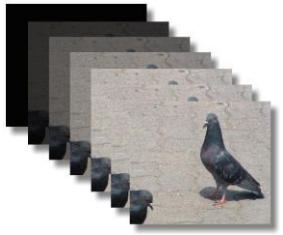
N3XT



Brain network

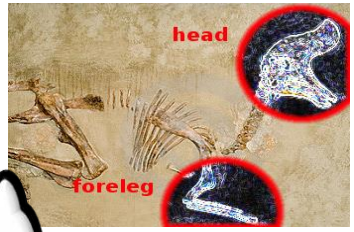


CGRAs



Automatic video annotation

Commodity hardware



Advanced image analysis

The Day After Tomorrow

Tomorrow

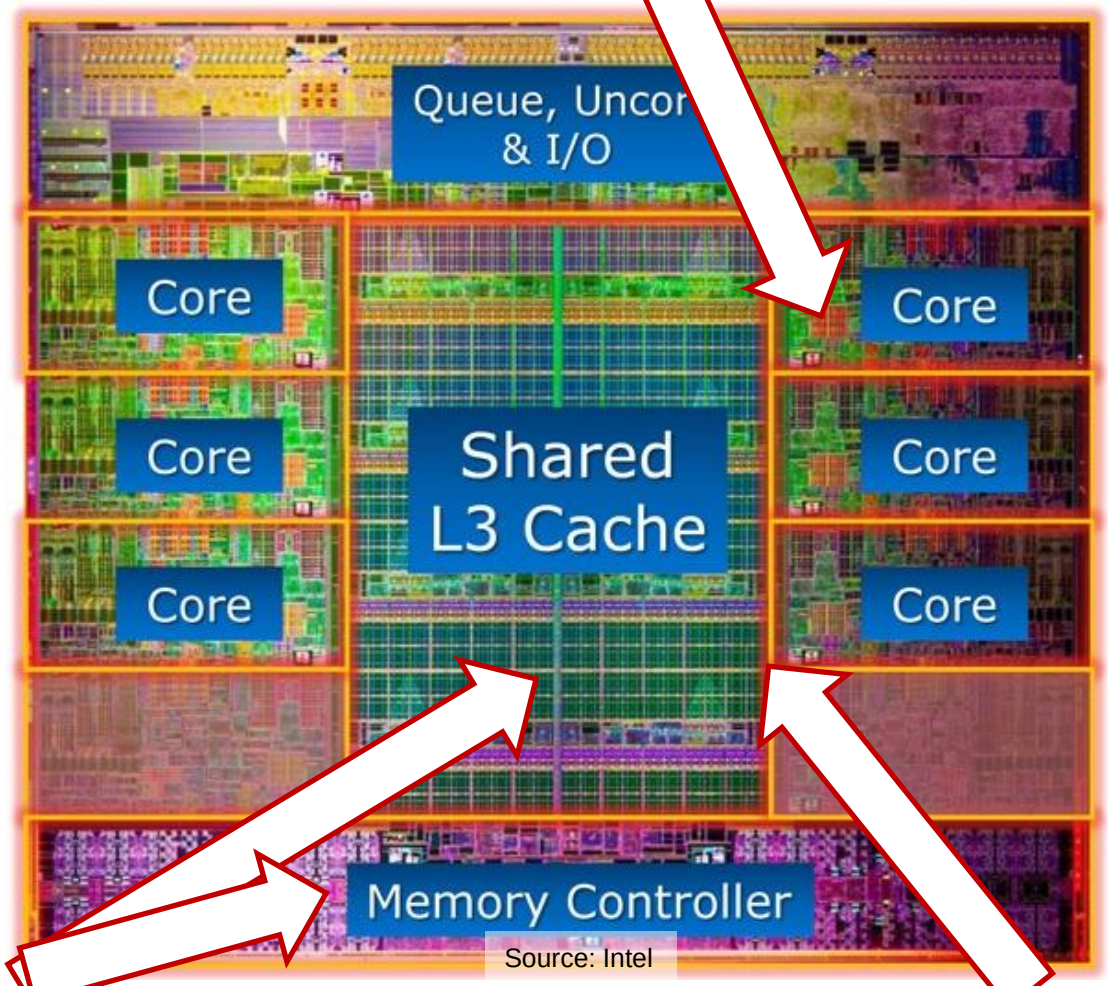
Today

Hardware

Software

Source: Subhasish Mitra, Chris Ré

Logic



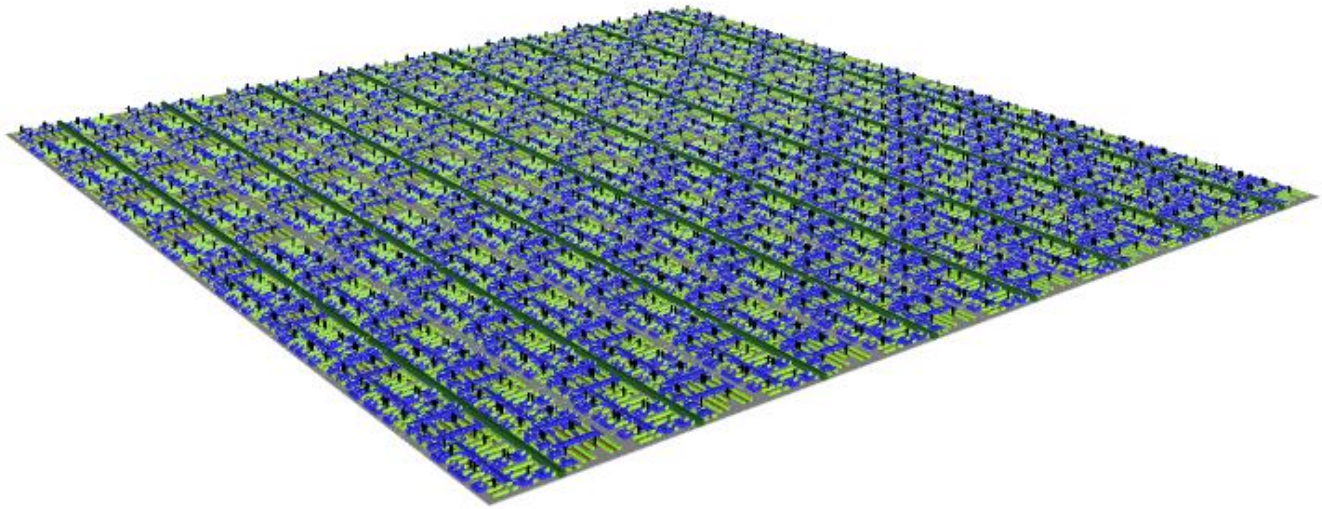
Source: Intel

Memory

Wire

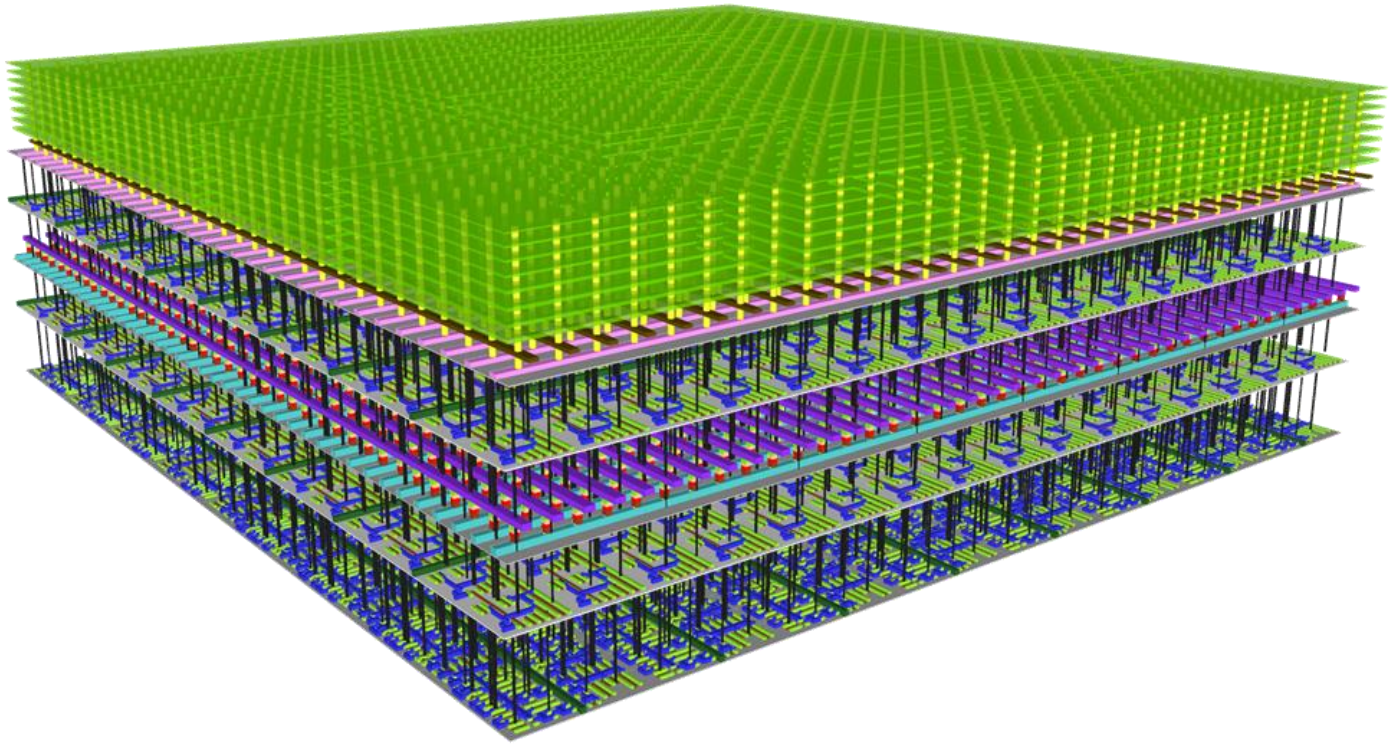
N3XT Nanosystems

Computation immersed in memory



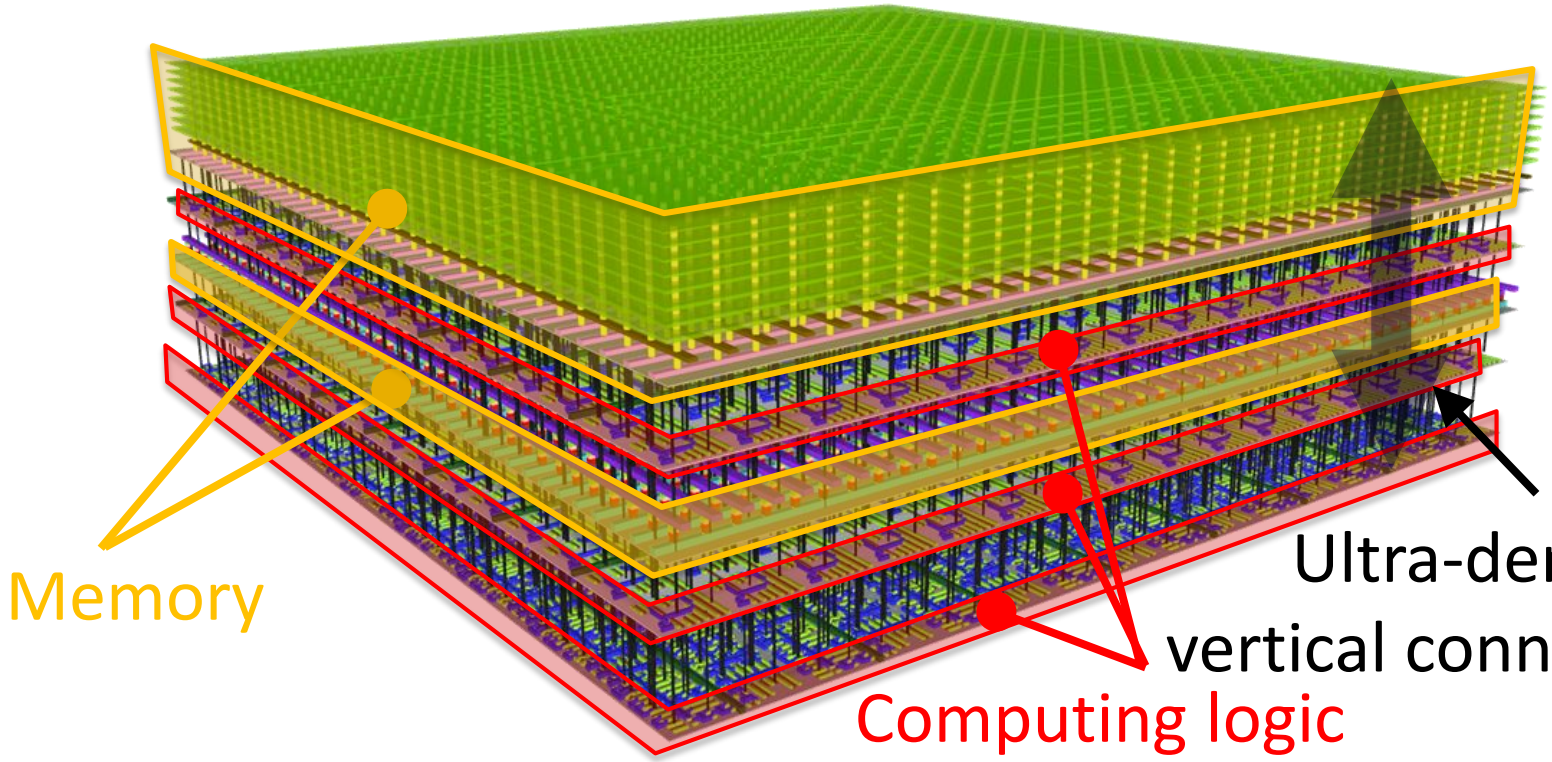
N3XT Nanosystems

Computation immersed in memory



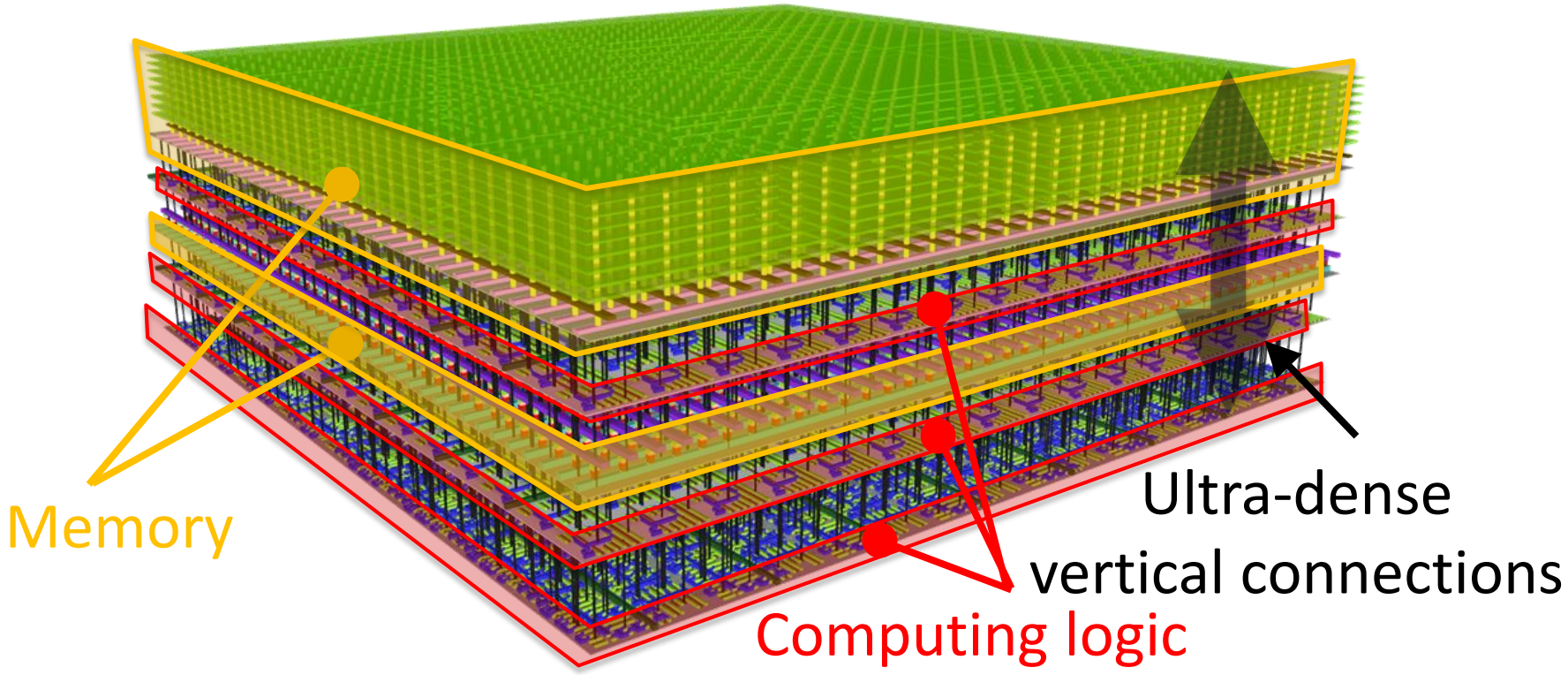
N3XT Nanosystems

Computation immersed in memory



N3XT Nanosystems

Computation immersed in memory



Impossible with today's technologies

A Nanotechnology-Inspired Grand Challenge for Future Computing

OCTOBER 20, 2015 AT 6:00 AM ET BY LLOYD WHITMAN, RANDY BRYANT, AND TOM KALIL

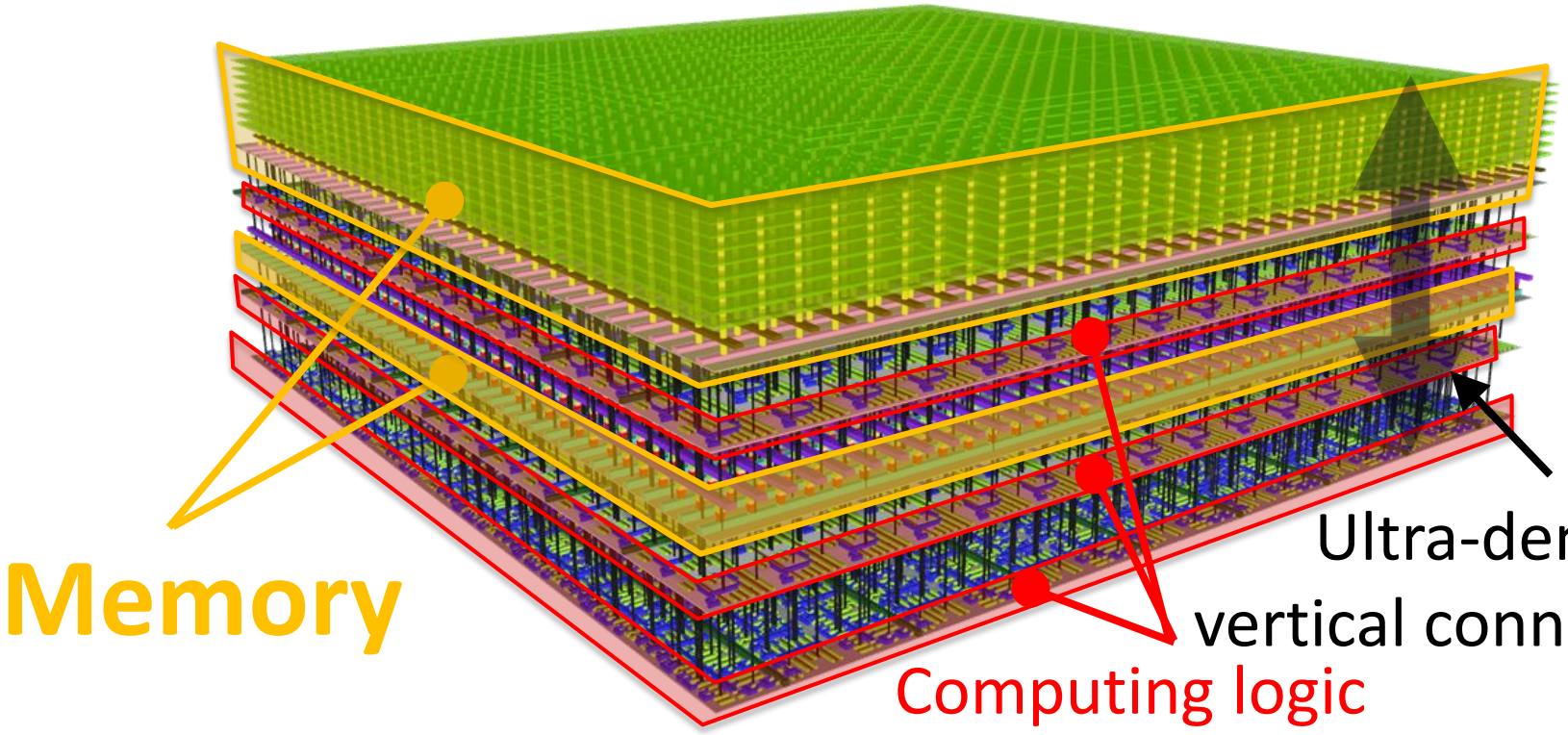
Summary: Today, the White House is announcing a grand challenge to develop transformational computing capabilities by combining innovations in multiple scientific disciplines.

In June, the Office of Science and Technology Policy issued a Request for Information seeking suggestions for *Nanotechnology-Inspired Grand Challenges for the Next Decade.* After considering over 100 respo
three Adminis
Computing Ini

Many of these breakthroughs will require new kinds of nanoscale devices and materials integrated into **three-dimensional systems** and may take a decade or more to achieve.

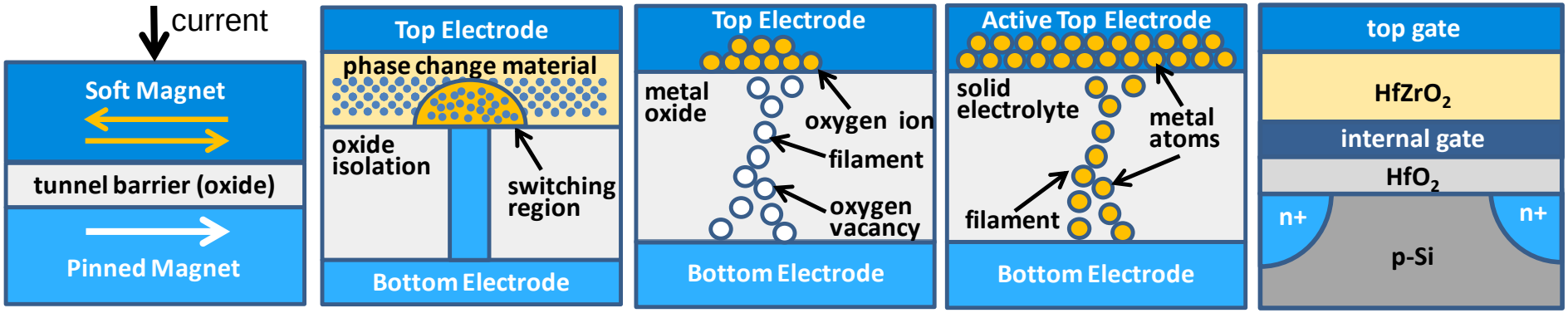
N3XT Nanosystems

Computation immersed in memory



“New” Memories

Random access, non-volatile, no erase before write, on-chip integration



STT-MRAM

Spin torque transfer magnetic random access memory

PCM

Phase change memory

RRAM

Resistive switching random access memory

CBRAM

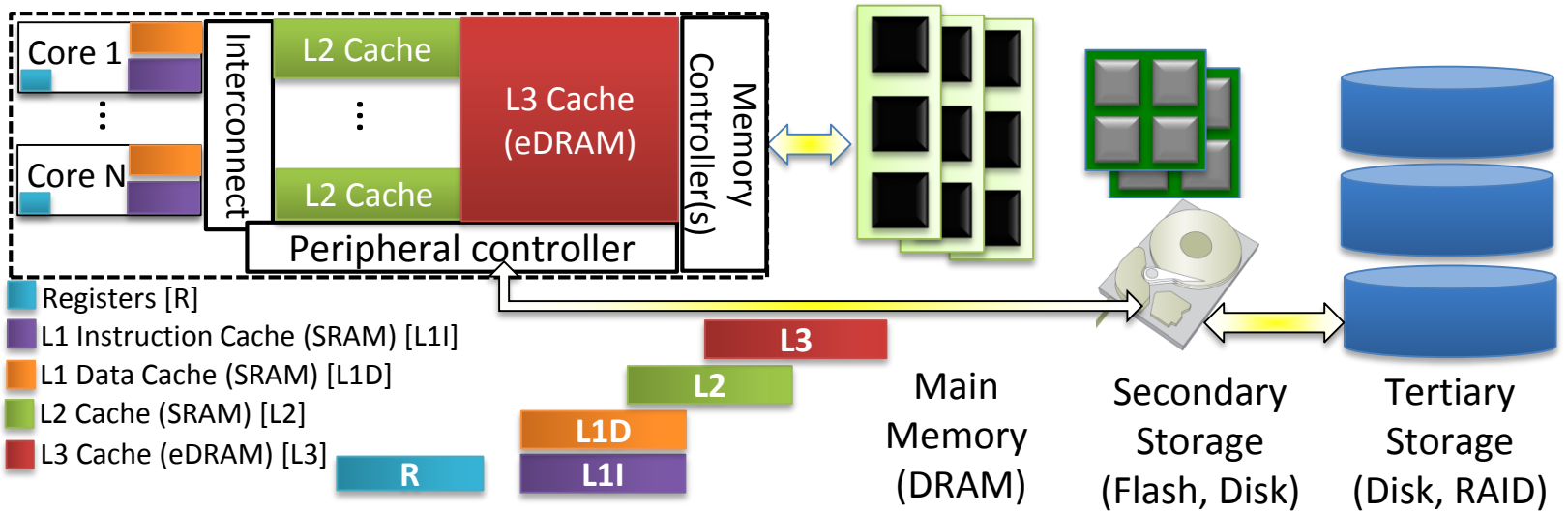
Conductive bridge random access memory

FERAM

Ferro-electric random access memory

Memory Hierarchy

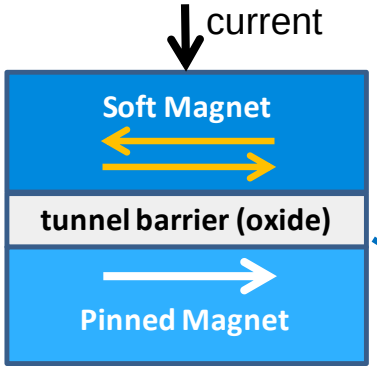
Must change drastically in the coming decade



Speed	0.1's ns		1 - 10's ns	10's -100's ns	0.1-10's ms	> 100's ms
CPU cycles	1		10 - 100	>100	10^6 - 10^8	$>10^9$
Size (bytes)	100	K	M	G	T	P

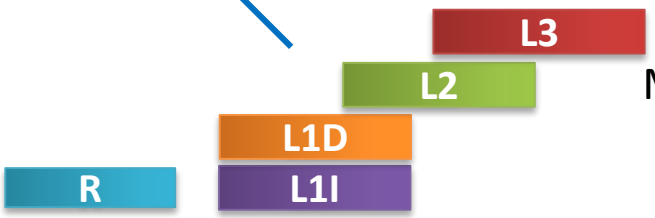
H.-S. P. Wong, S. Salahuddin, *Nature Nanotech* (2015)

STT-MRAM



You already know:

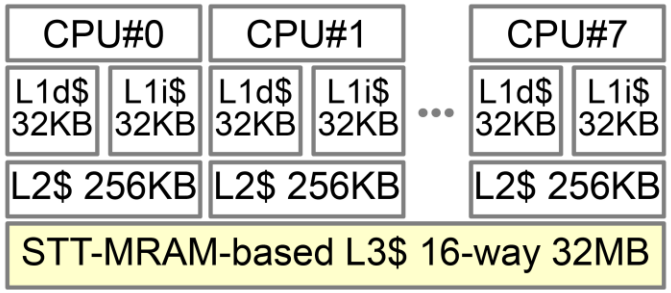
- $V_{PROG} \leq 0.5V$, $V_{DD} = 1.2 – 1.5$ V
- A few – 10's of ns read/write
- Write current ~ 10's µA (F<45 nm)
- "infinite" endurance
- 1T1MTJ, > $6F^2$
- Scalable to 20 nm (expt.)

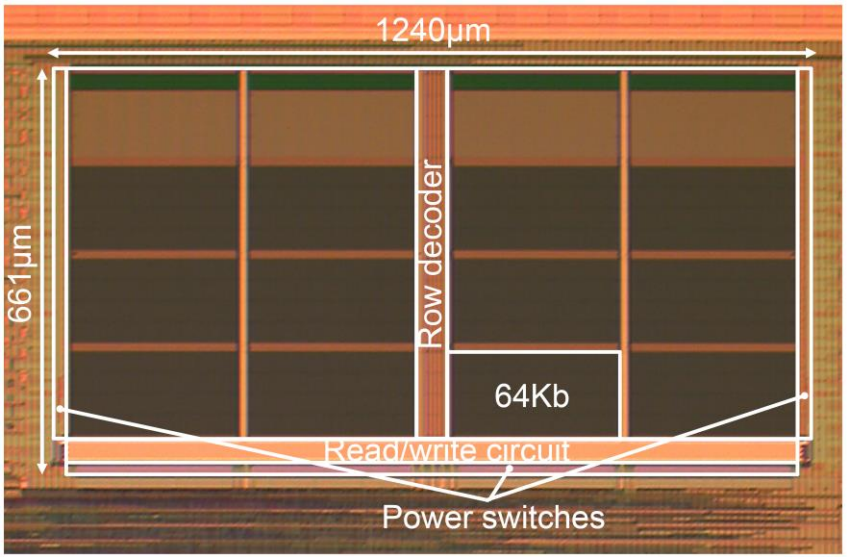


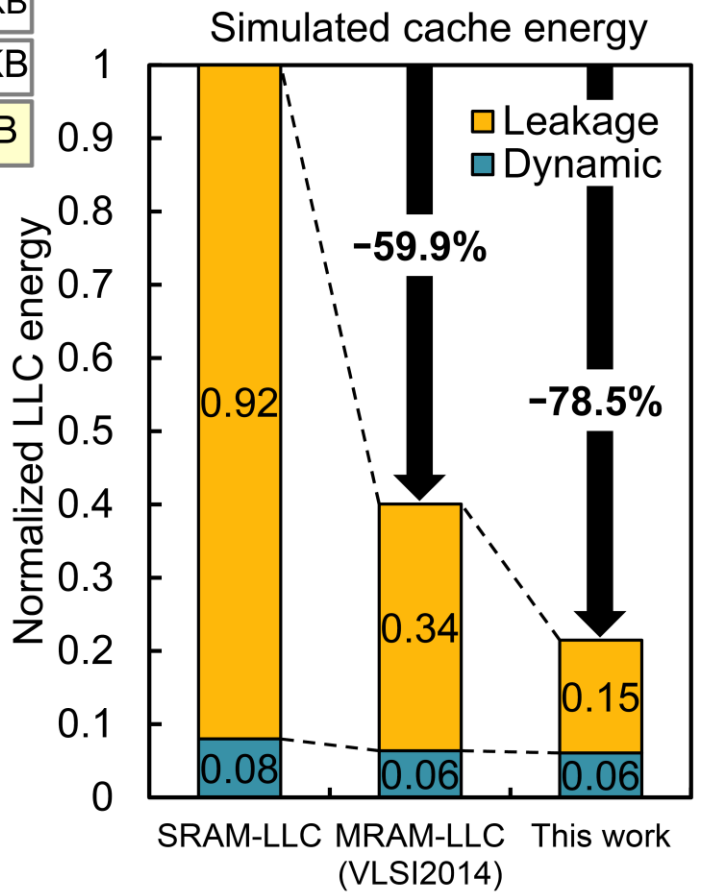
			Main Memory (DRAM)	Secondary Storage (Flash, Disk)	Tertiary Storage (Disk, RAID)
Speed	0.1's ns	1 - 10's ns	10's -100's ns	0.1-10's ms	> 100's ms
CPU cycles	1	10 - 100	>100	10^6 - 10^8	$>10^9$
Size (bytes)	100	K M	G	T	P

Embedded STT-MRAM Cache

Technology	65-nm CMOS 47-nm MTJ
Macro size	0.8196mm^2
Capacity	1Mb
I/O width	256bits
Read speed	3.3ns
Write speed	3.0ns
Read energy	71.2µJ/MHz
Write energy	166.2µJ/MHz

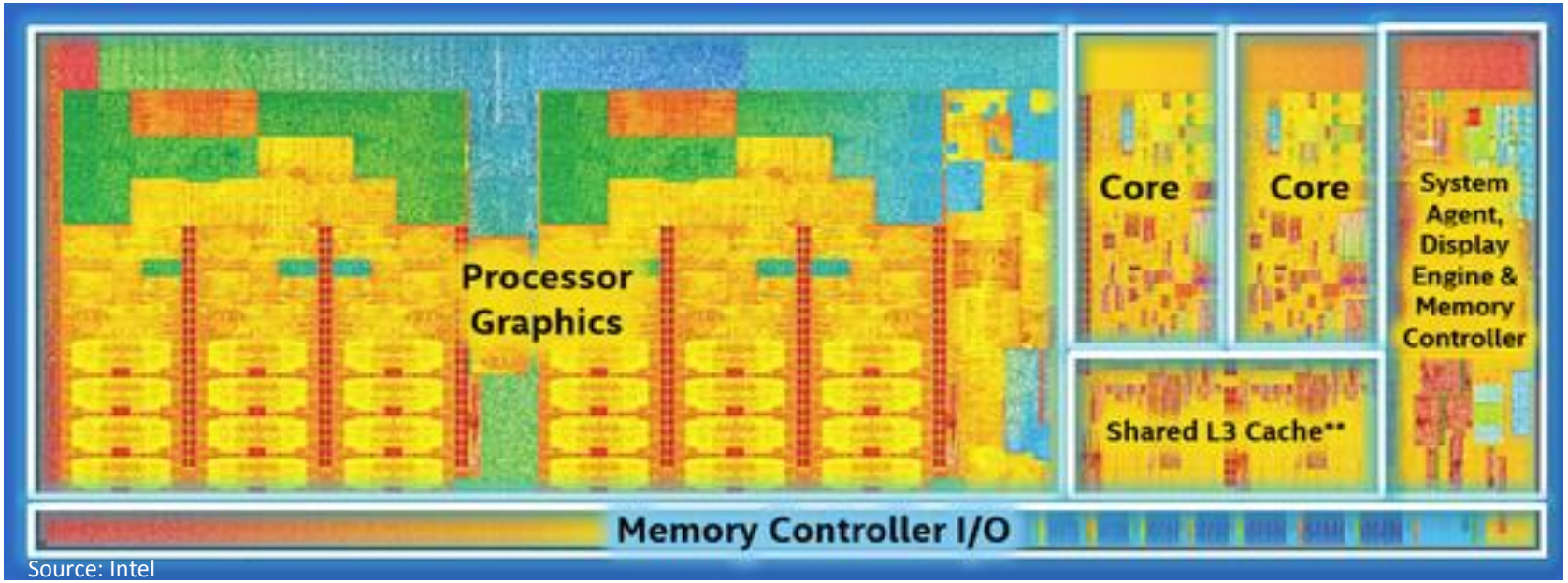






H. Noguchi et al., "A 3.3ns-Access-Time 71.2µW/MHz 1Mb Embedded STT-MRAM Using Physically Eliminated Read-Disturb Scheme and Normally-Off Memory Architecture," *ISSCC*, pp. 136 – 137 (2015) [Toshiba]

Why has STT-MRAM not arrived?



Source: Intel

Maybe I will be proven wrong soon …

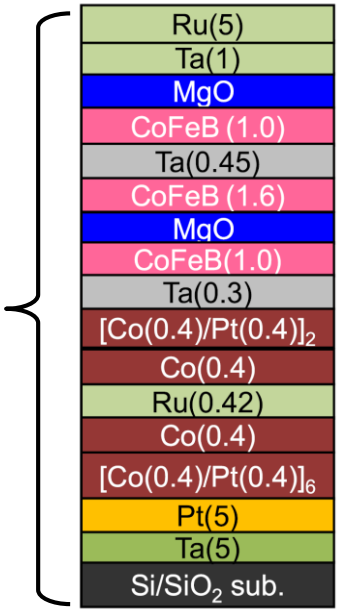
Homework for STT-MRAM Folks

Device technology

- Current (10's μA) too high
- Spin Hall effect promises to reduce current by 10X
- Scalability demonstrated to 20 nm so far

Manufacturing

- Not ready at the same time as logic (cf. SRAM)
- MTJ – stacks of 16 layers , 4Å thick
- Deposited and ion milled across 300 mm wafer
- Does not survive 400°C BEOL fab temperature



(# = nm)

STT-MRAM Design Opportunity

Retention time, $\tau = \tau_o \exp(\Delta)$

- 10-year retention: $\Delta = 40$

Thermal stability, $\Delta = E_B / k_B T$

- Large array $\Delta = 60$

Write current, $I_{co} \propto E_B$

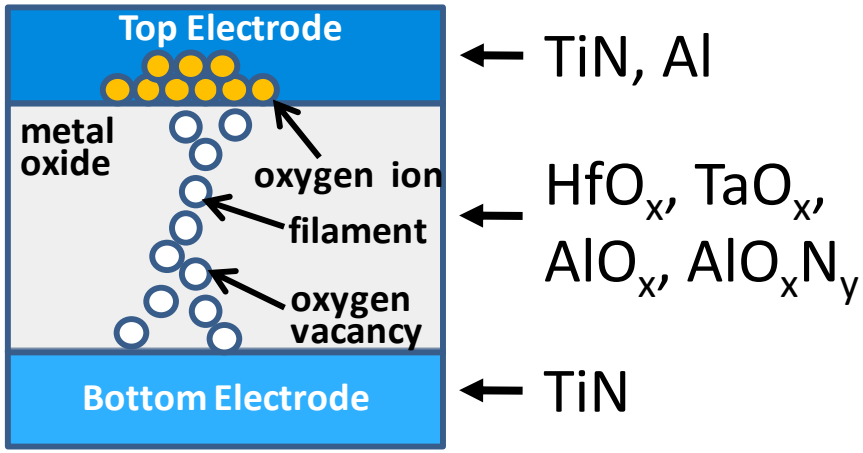
But we don't need 10 year retention for cache and main memory!

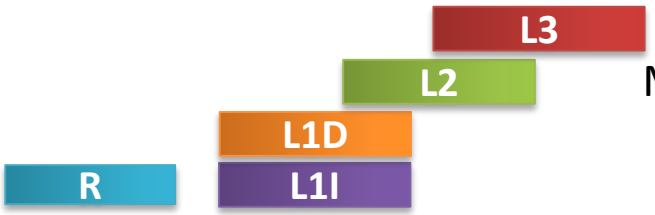
Write current can be significantly reduced

RRAM – Fab Friendly, "Easy" to Make

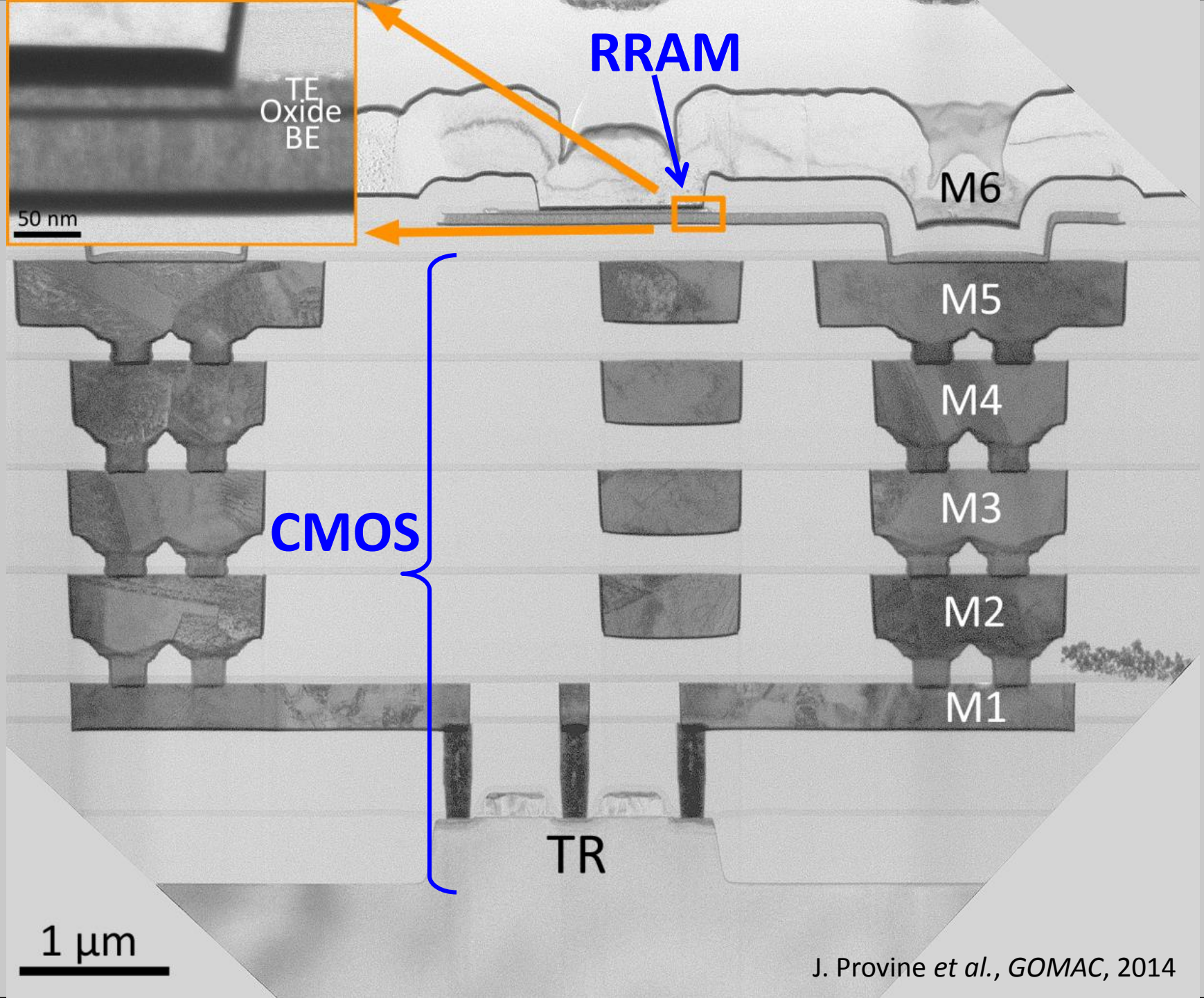
You already know:

- 10 ns read/write, V_{PROG} ~1 – 2 V
- Write current ~ nA – 10's µA
- 1E12 cycles endurance @ device
- 1T1R, ~ $6F^2$
- 3D RRAM (similar to 3D NAND)
- Scalable to < 5 nm & smaller
- Fab-friendly, easy to embed





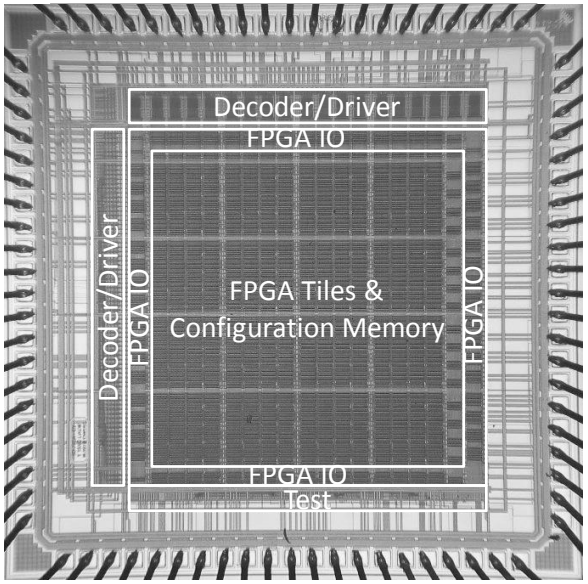
	Main Memory (DRAM)	Secondary Storage (Flash, Disk)	Tertiary Storage (Disk, RAID)
Speed	10's -100's ns	0.1-10's ms	> 100's ms
CPU cycles	>100	10^6 - 10^8	>10^9
Size (bytes)	G	T	P



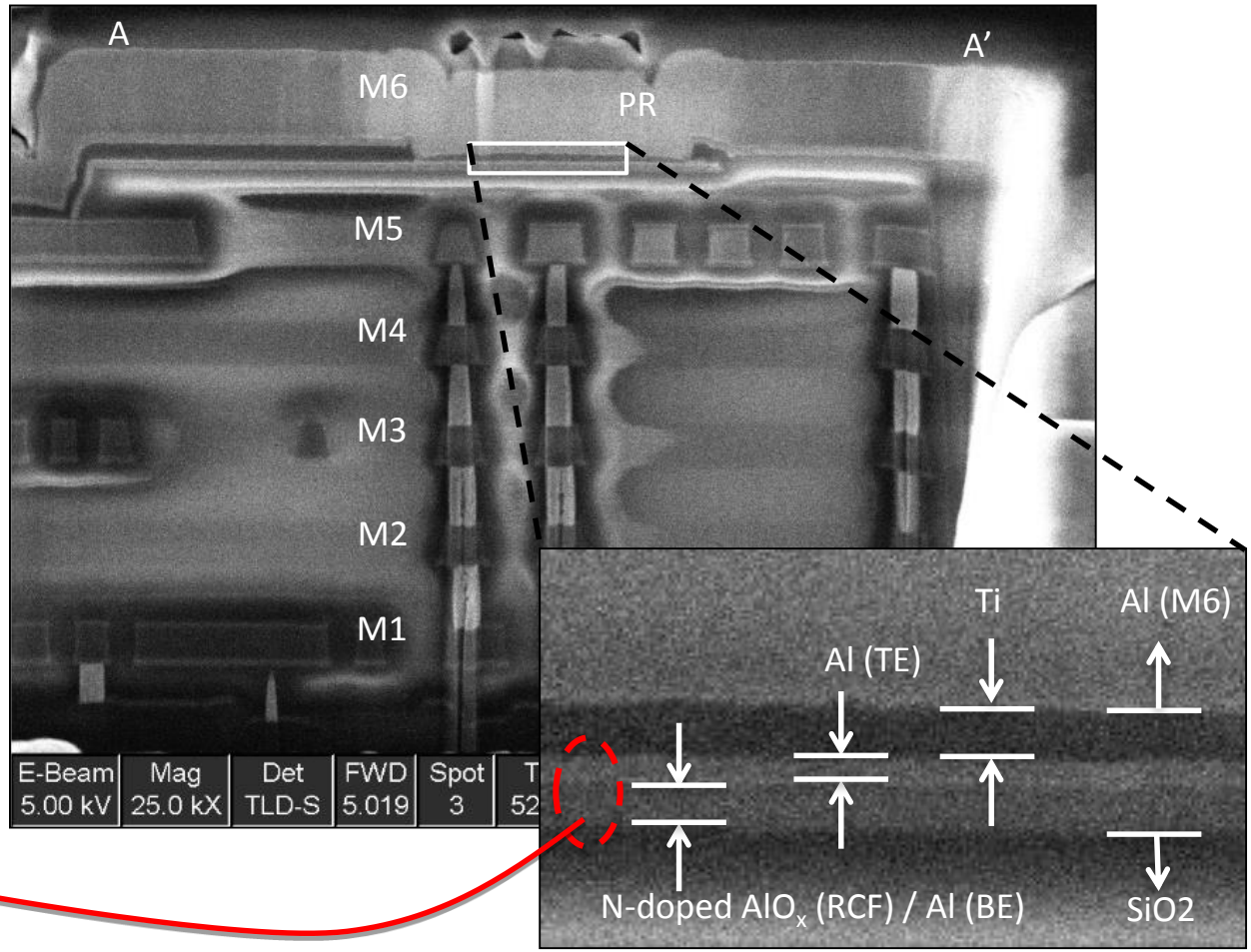
J. Provine *et al.*, *GOMAC*, 2014

RRAM FPGA Integration

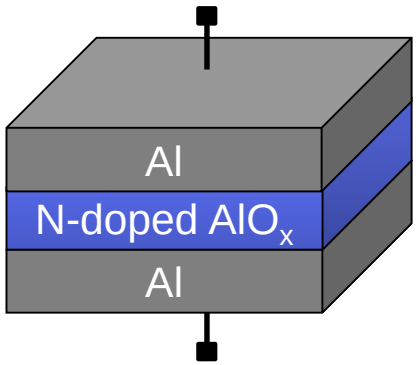
FPGA Chip



Cross-section View



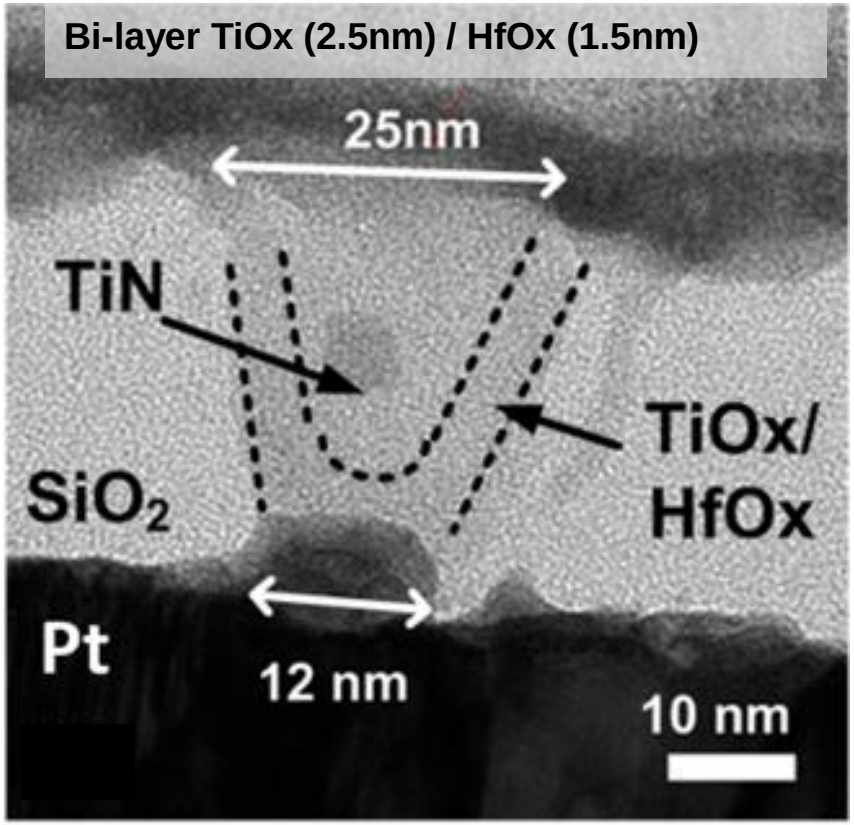
Programmable Resistor

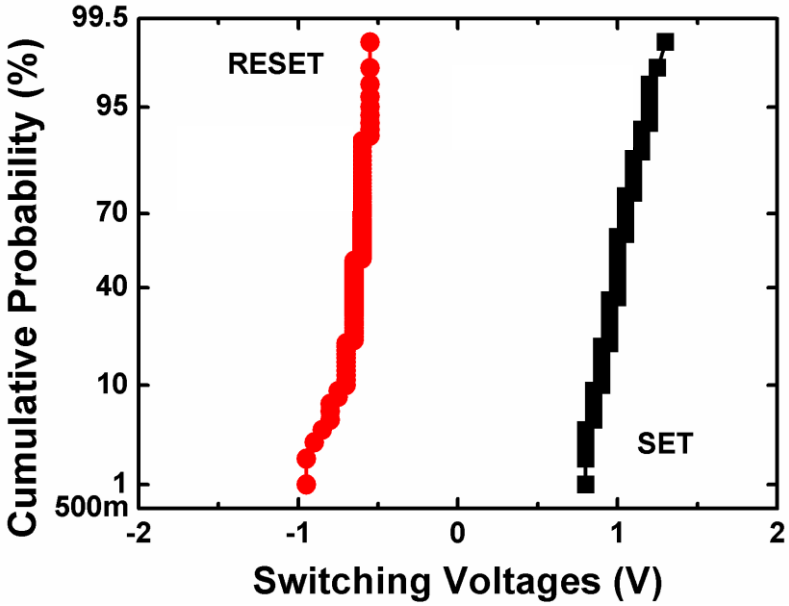


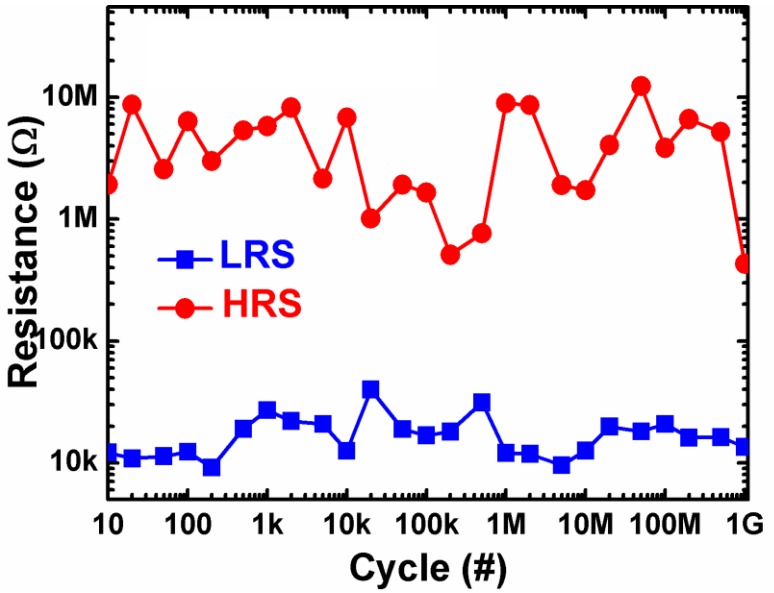
Y. Liauw et al., *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012.*

RRAM Is Scalable

Scalable: 12 nm



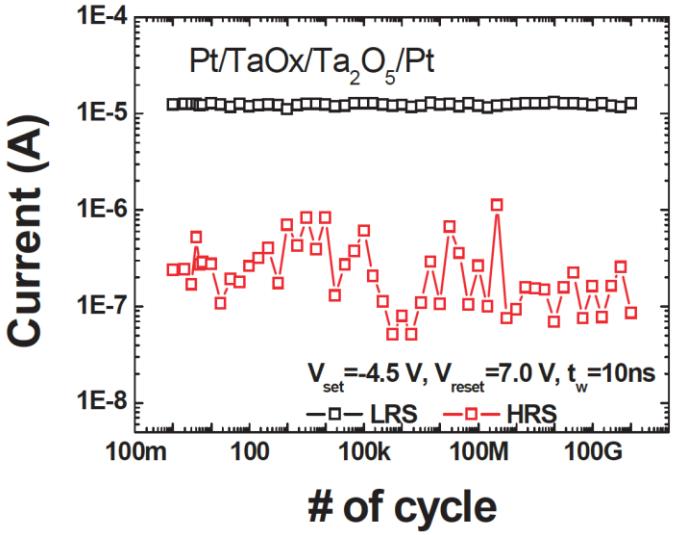




Y. Wu et al., *IEDM* 2013.

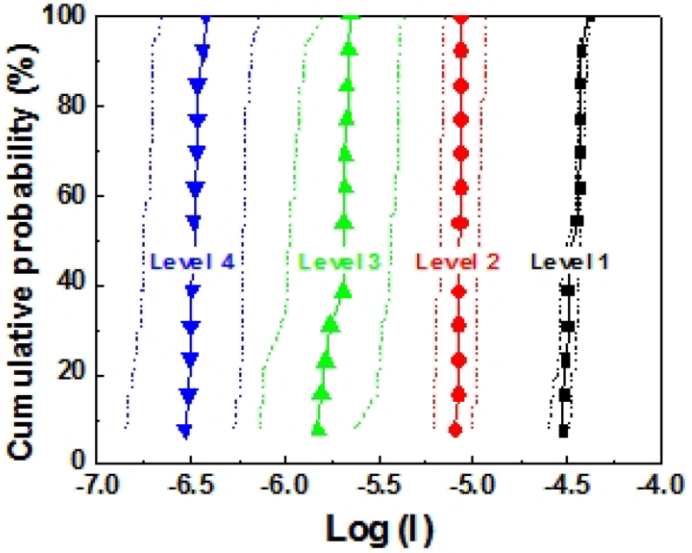
RRAM Promises

10^{12} cycles

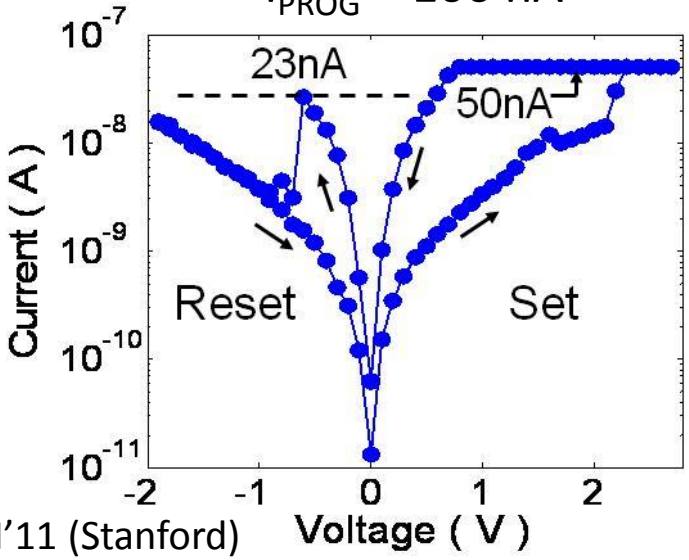


VLSI'11 (Samsung)
VLSI'12 (Samsung)

Multi-bit

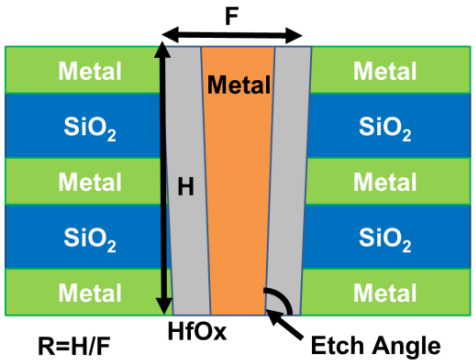


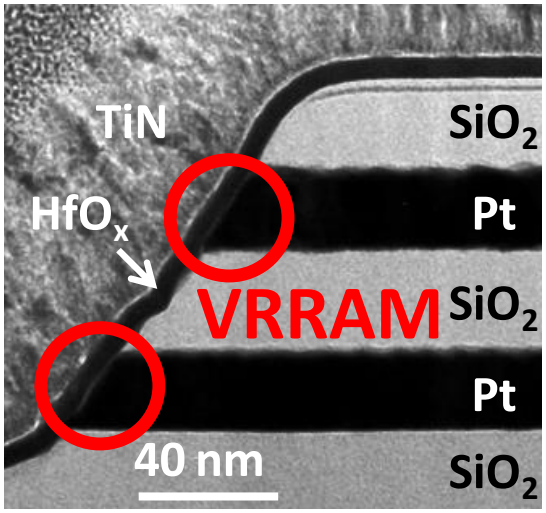
I_{PROG} < 100 nA



VLSI'11 (Stanford)

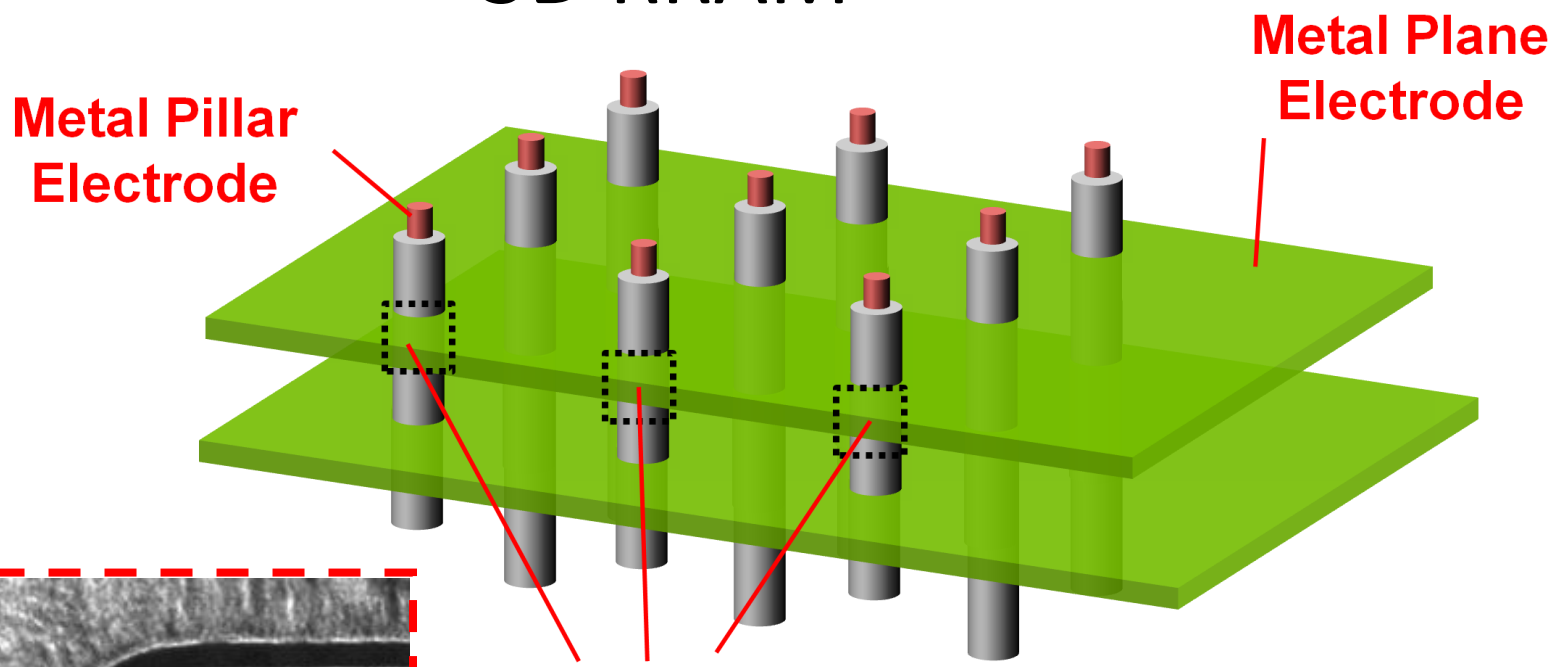
Bit-Cost Scalable

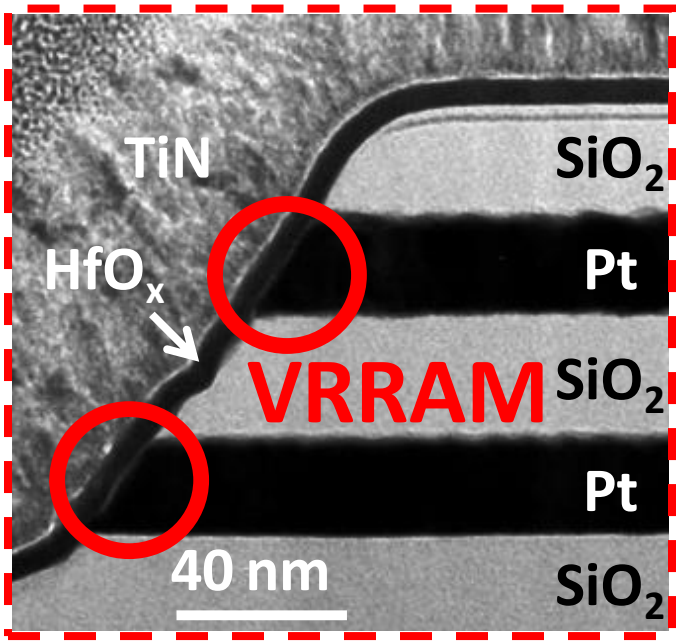


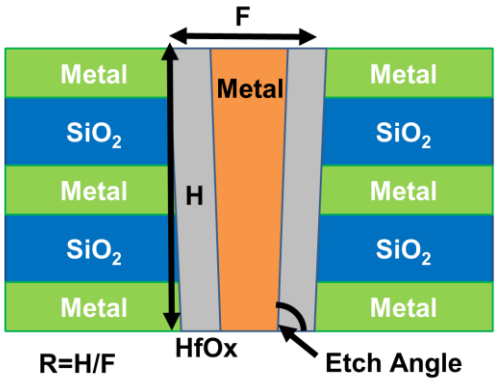


IEDM'12 (Stanford)

3D RRAM

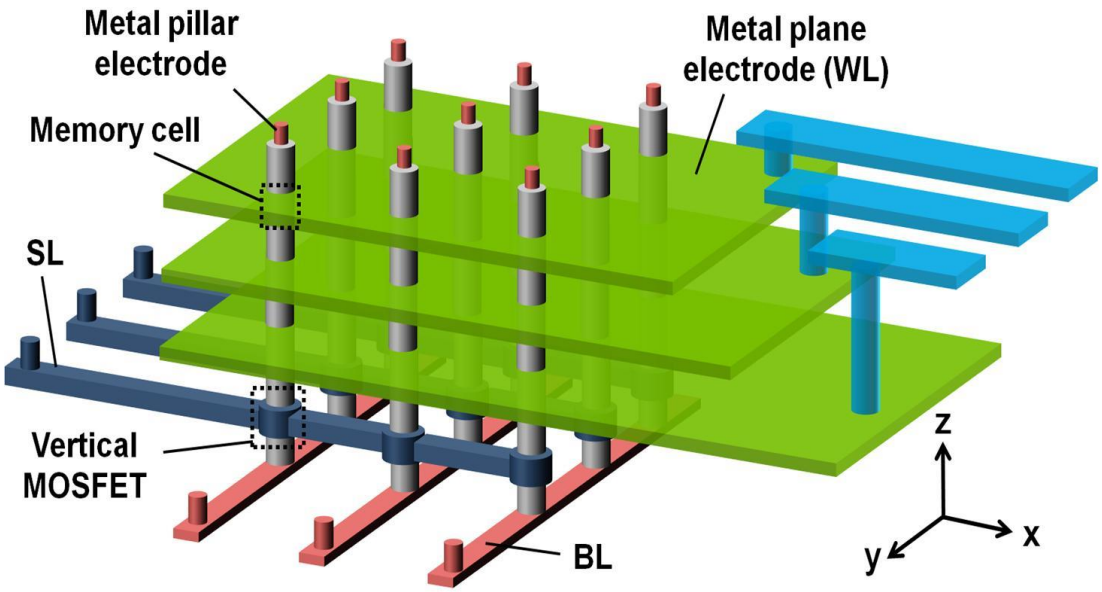






S. Yu, H.-Y. Chen *et al.*, *Symp. VLSI Tech*. 2013

High Density 3D Memory



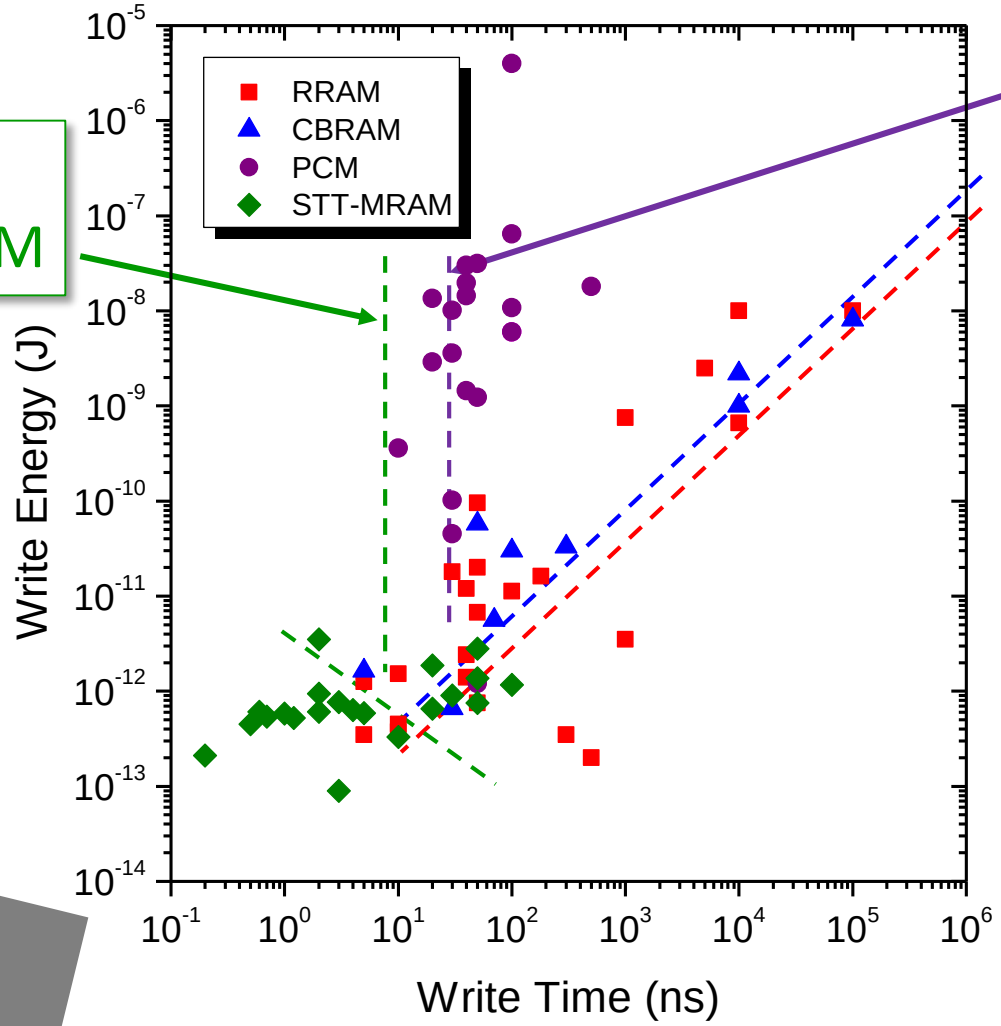
- < 1 µA
- 1 – 2 V
- 5 ns
- > 1G cycles
- F = 5 nm
- 128 layers
- 64 Tb per chip

IEDM ’12, ’13, ’14
VLSI ’13, ’14, ’16
DATE ’15, Nature Comm ’15

Energy vs Speed @ Device Level

Nothing faster than STT-MRAM

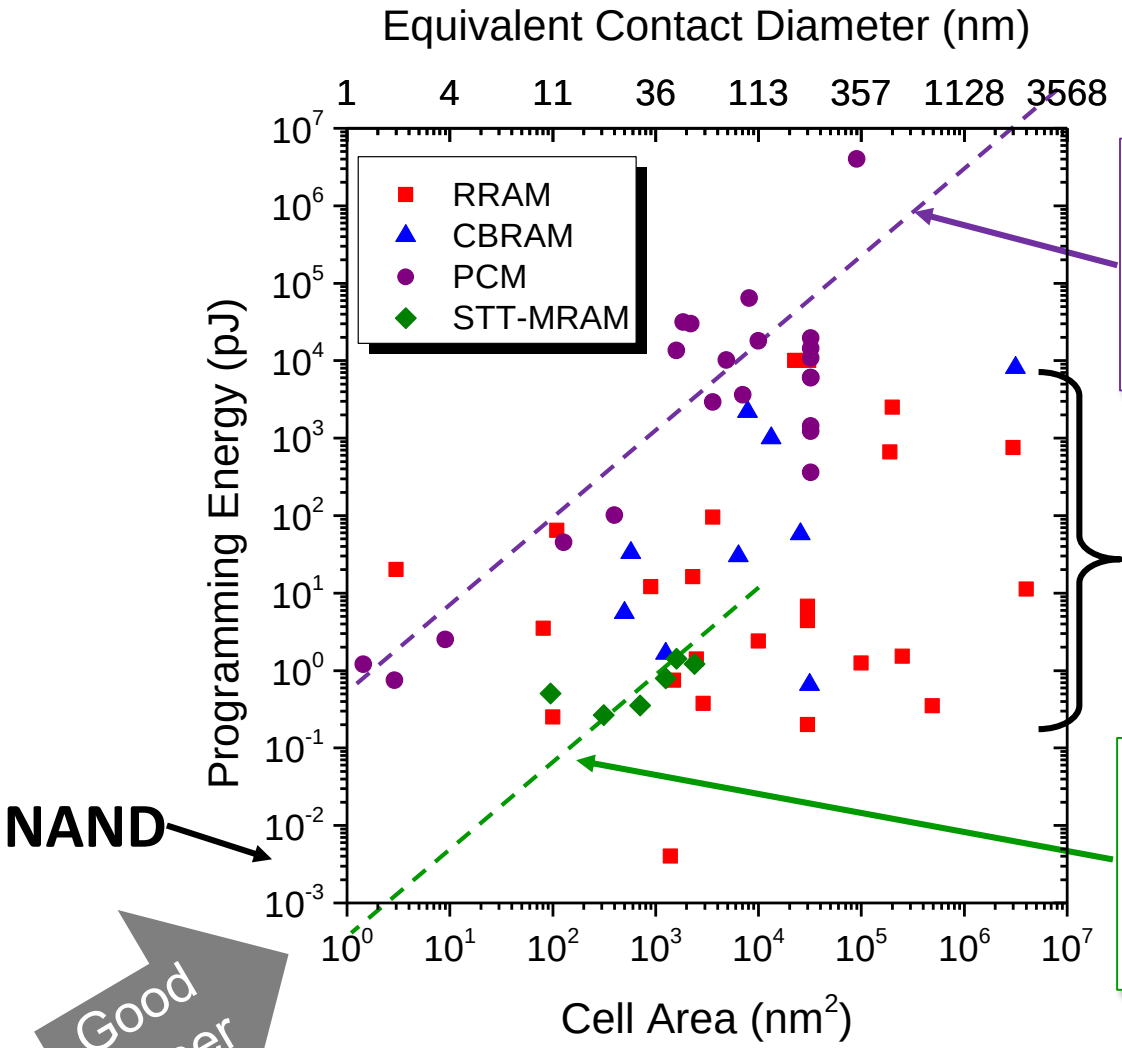
Speed limited by physics



Good corner

https://nano.stanford.edu/stanford-memory-trends

Write Energy Scaling @ Device Level



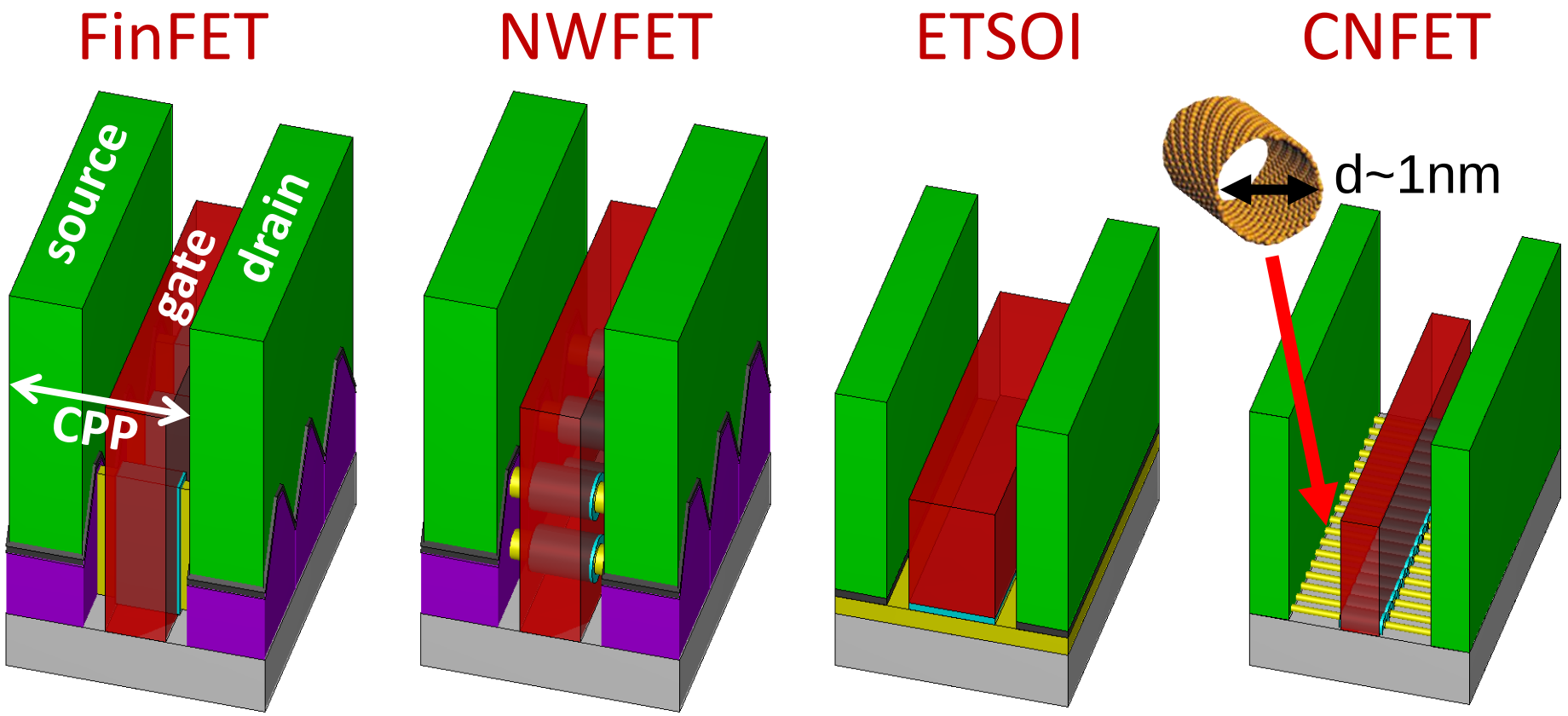
PCM: Proportional to area

RRAM & CBRAM: Material and resistance values

STT-MRAM Proportional to area

https://nano.stanford.edu/stanford-memory-trends

Logic Switches for 10, 7, 5, 3, 1 … nm



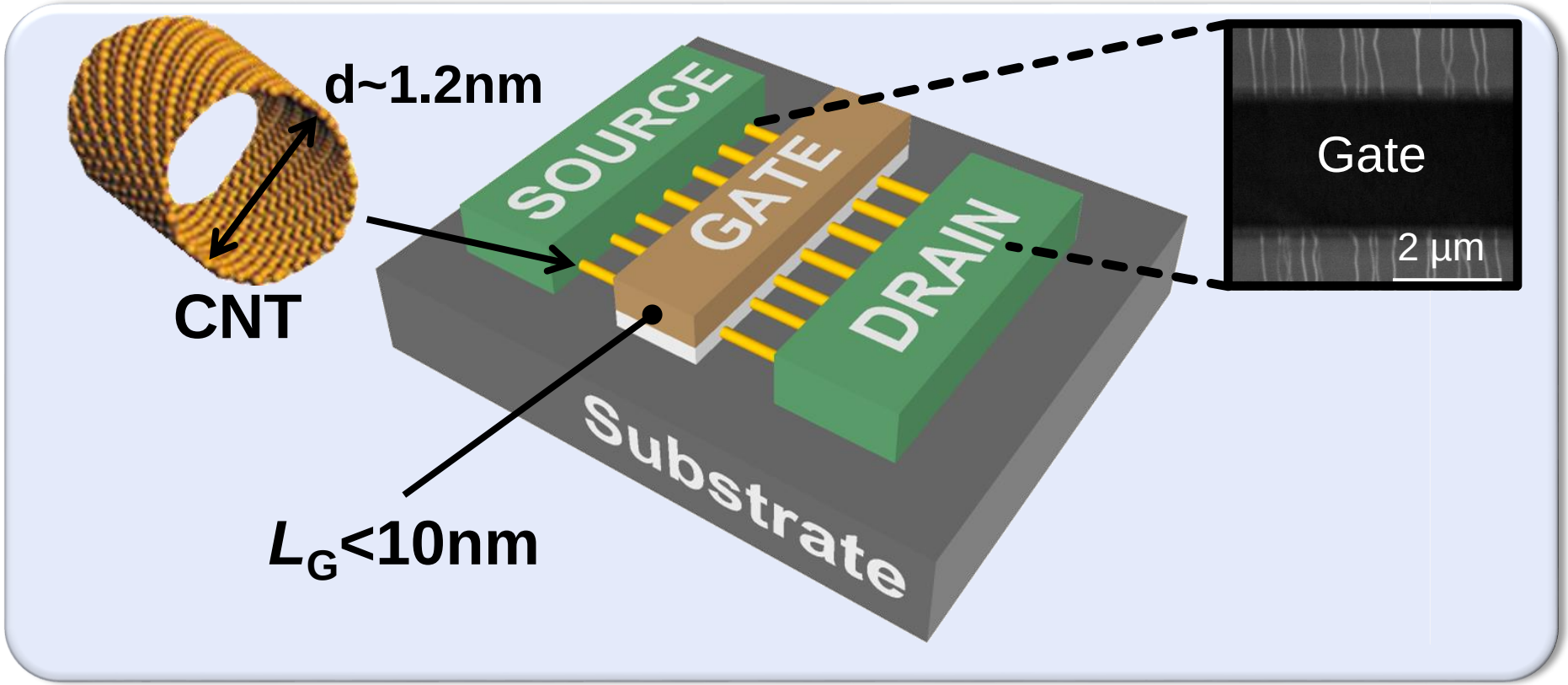
- Thin channel body (electrostatics control)
- Minimize parasitic capacitance and resistance

NWFET = “nanowire FET”, ETSOI = “extremely-thin silicon-on-insulator”, CNFET = “carbon nanotube FET”

Source: G. Hills, S. Mitra (Stanford)

Carbon Nanotube FET (CNFET)

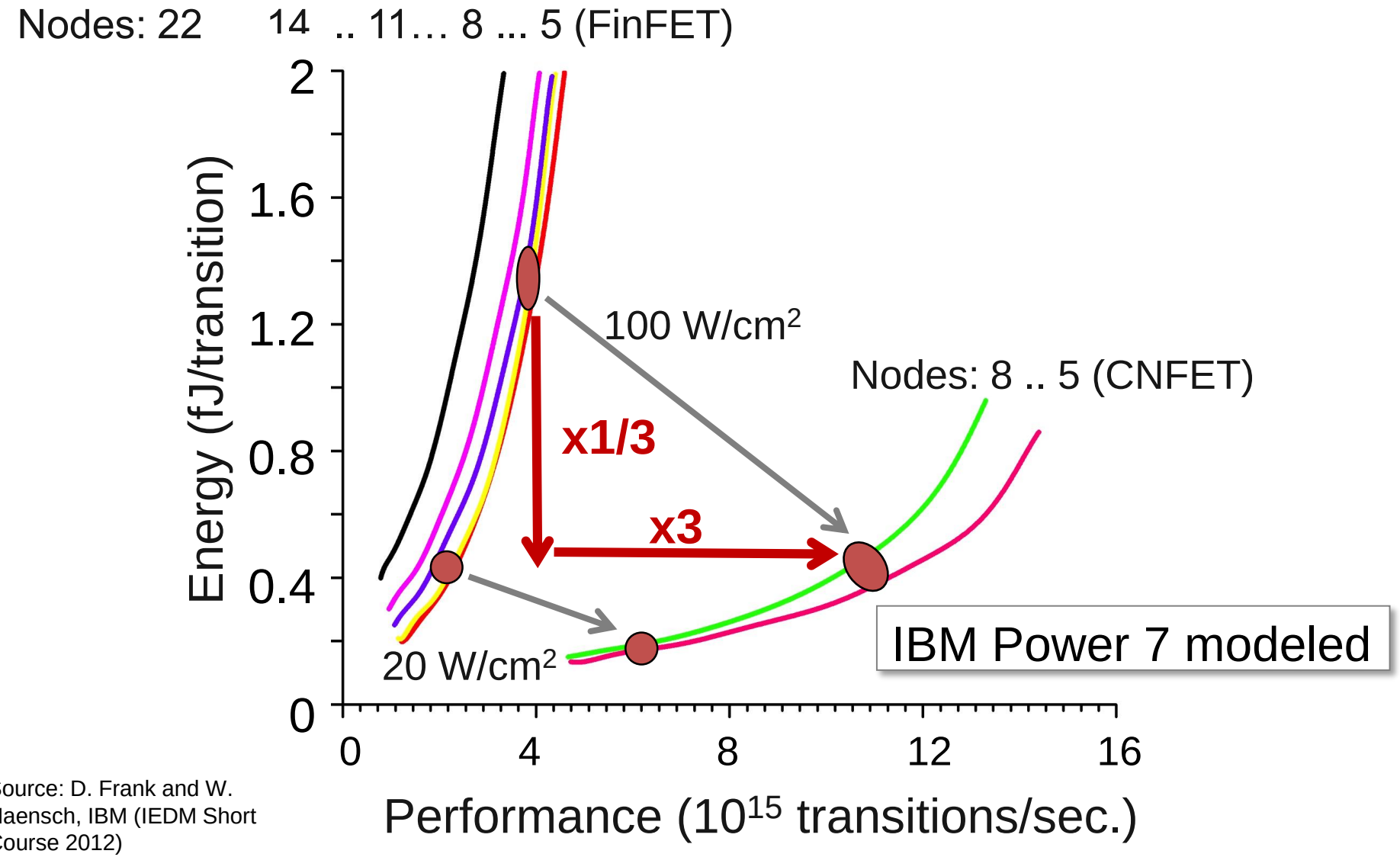
- **10X EDP Benefit** vs. silicon
  - Sub-10 nm node VLSI circuits



EDP: "energy-delay product"

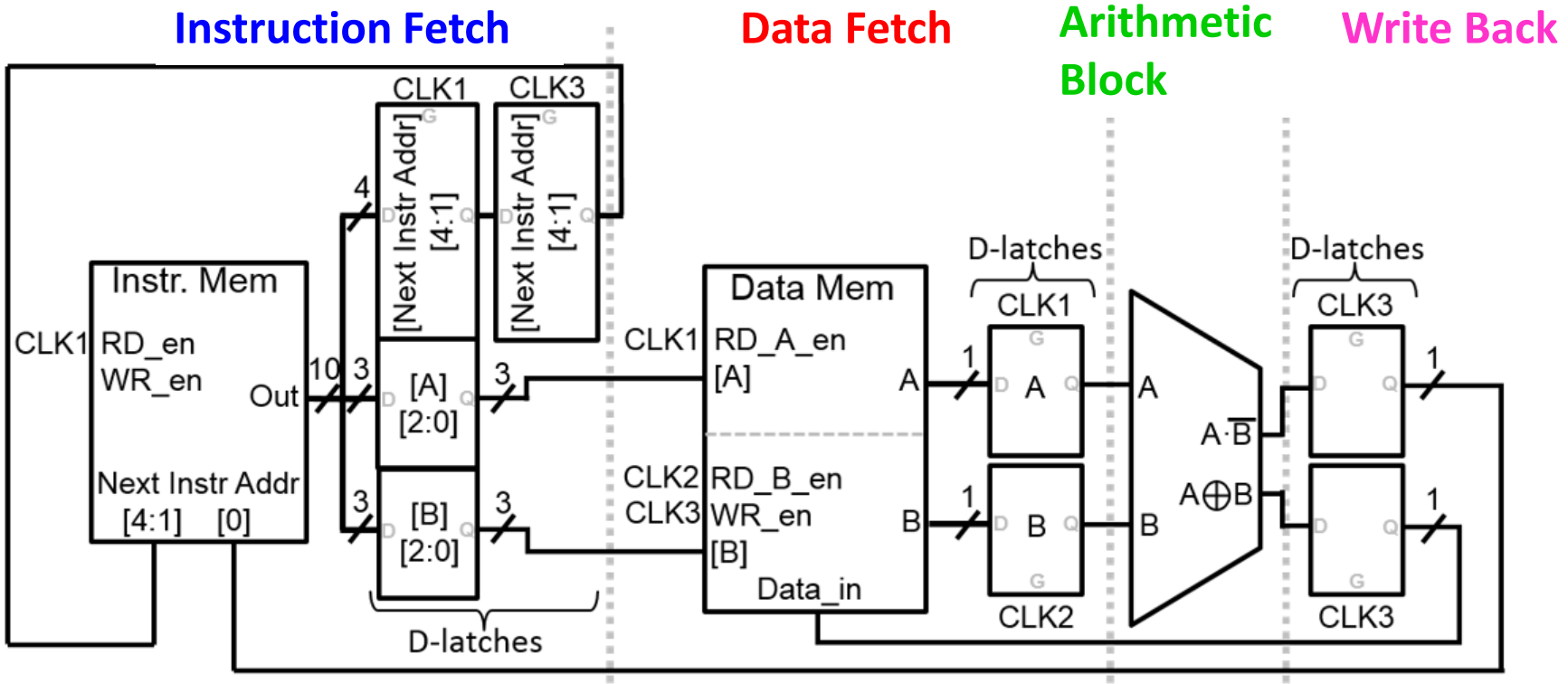
Source: G. Hills, S. Mitra (Stanford)

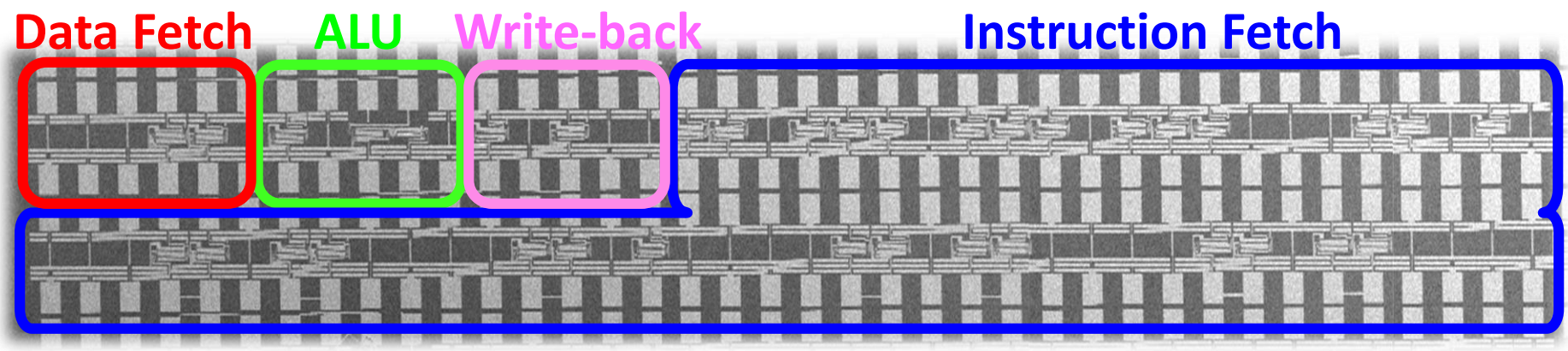
CNFET: ~10X EDP vs. Si-CMOS



Source: D. Frank and W. Haensch, IBM (IEDM Short Course 2012)

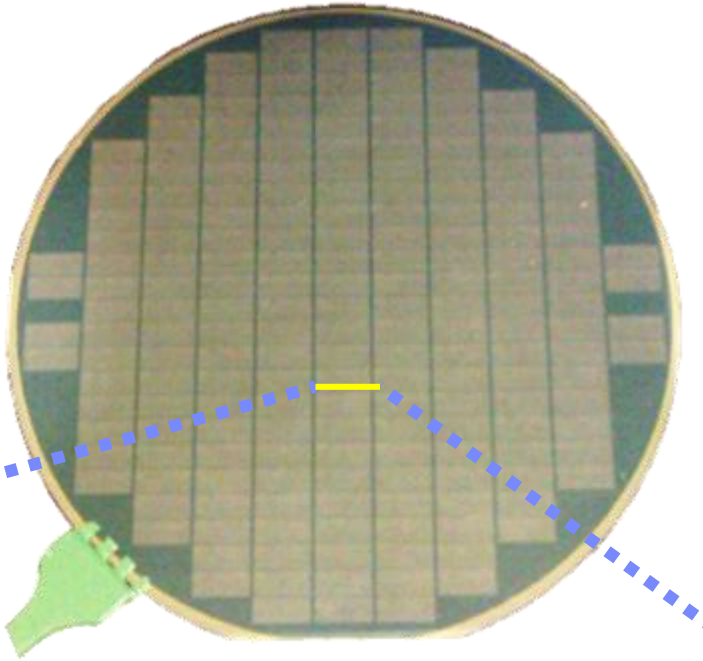
CNT Computer

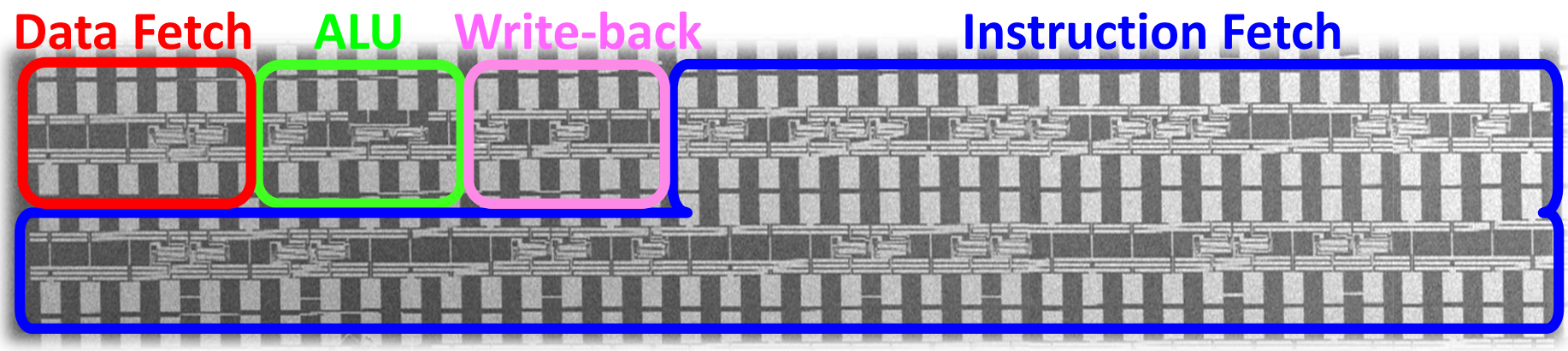




M. Shulaker et al., Nature 13

CNT Computer



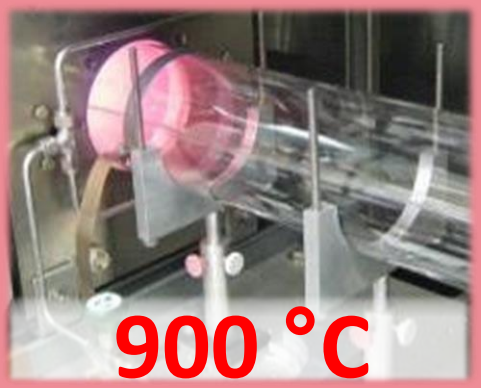


M. Shulaker et al., Nature 13

CNFET Enables Monolithic 3D

CNT transfer decouples high temperature growth

High-temperature CNT growth



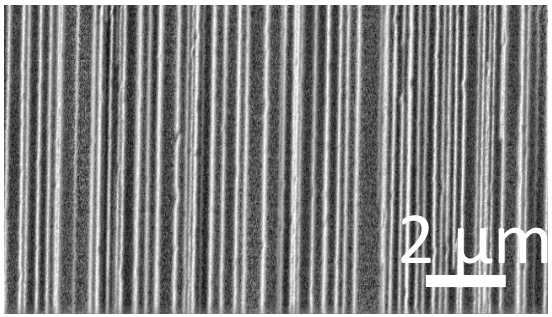
CNT transfer

120 °C

Low-temperature circuit fabrication

Growth

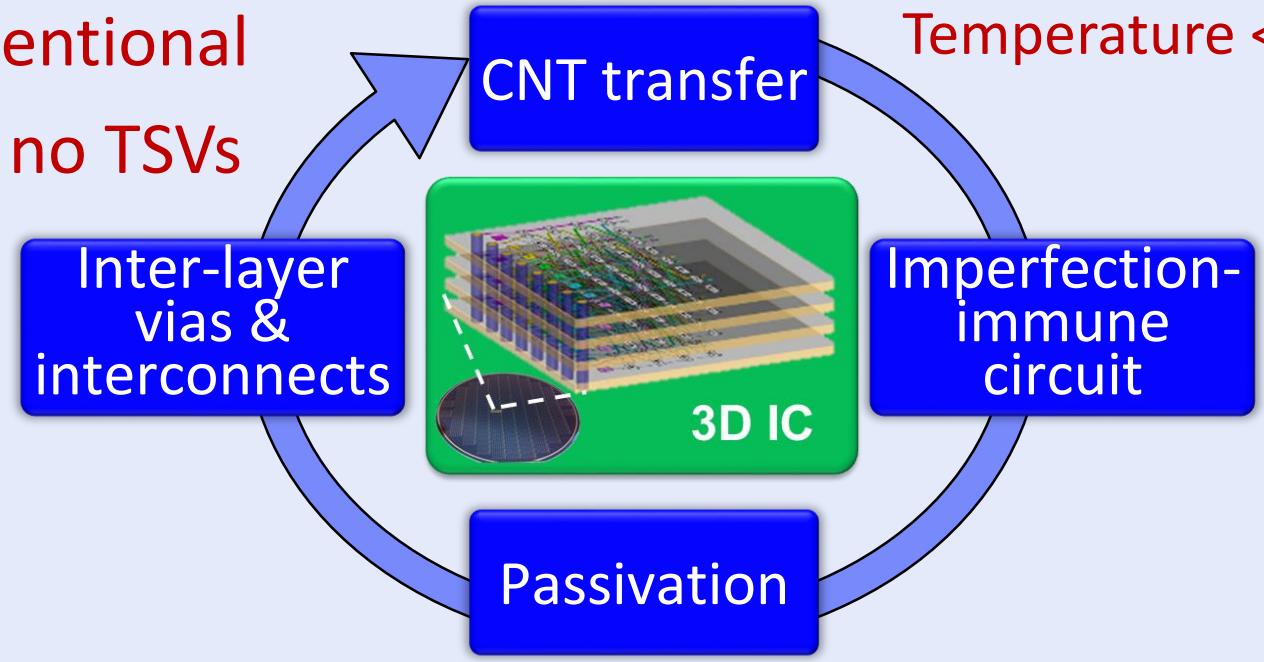
Post-transfer



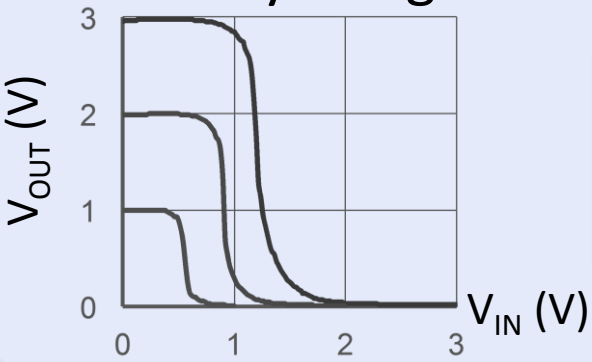
CNFET: Monolithic 3D Fabrication Flow

Conventional vias, no TSVs

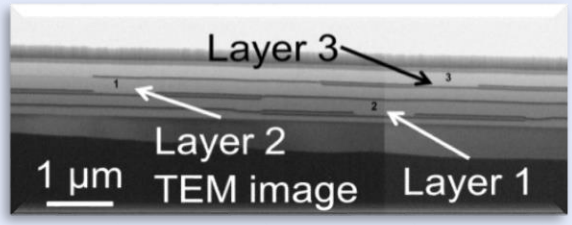
Temperature < 250 °C



Inter-layer logic



3-Layer integration



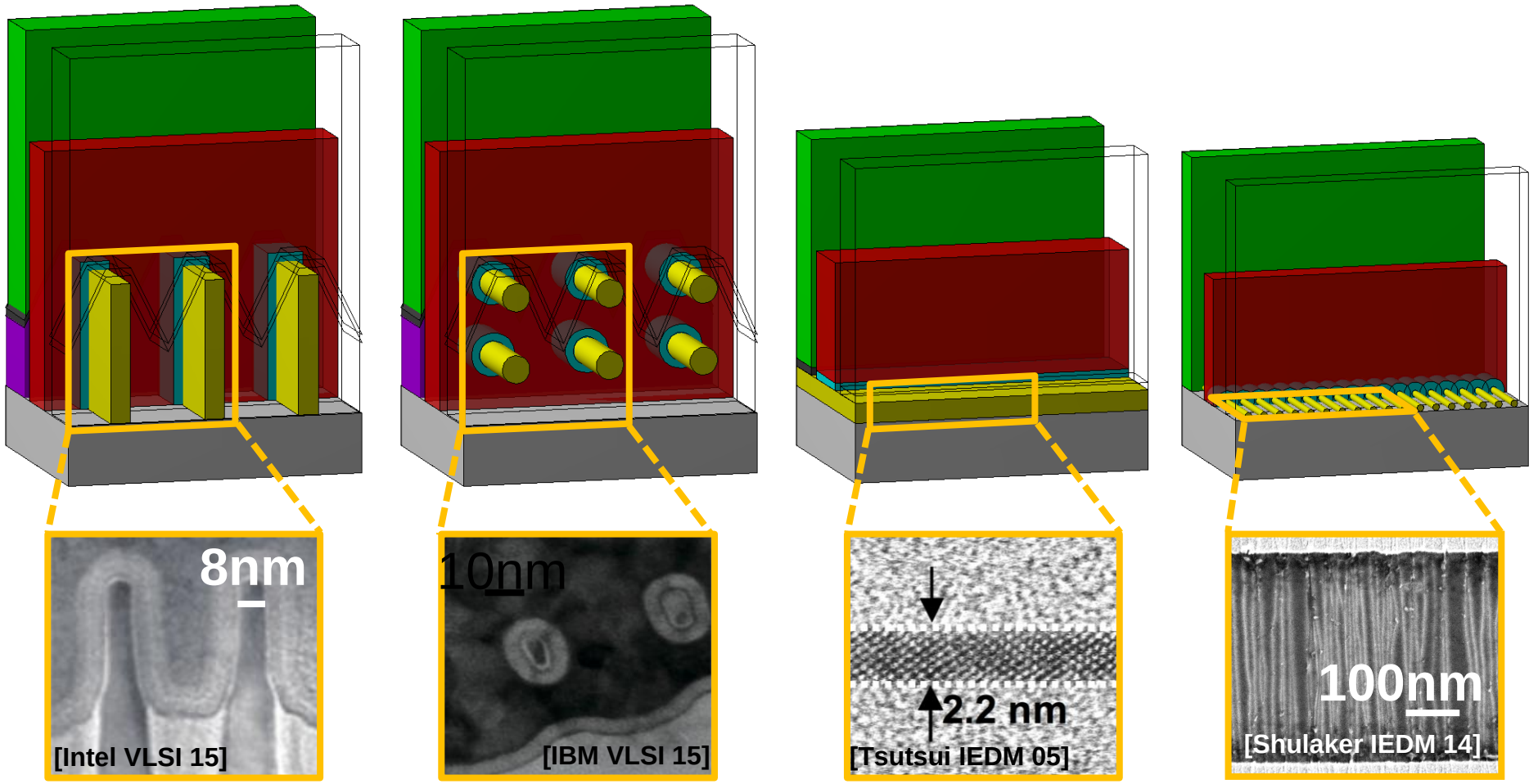
Memory on logic

Layer 2 Layer 1

H. Wei, IEDM 09, IEDM 13

2D → 3D → 2D Structure, 2D → 3D Integration

FinFET NWFET ETSOI CNFET



Source: G. Hills, S. Mitra (Stanford)

Computation Immersed in Memory

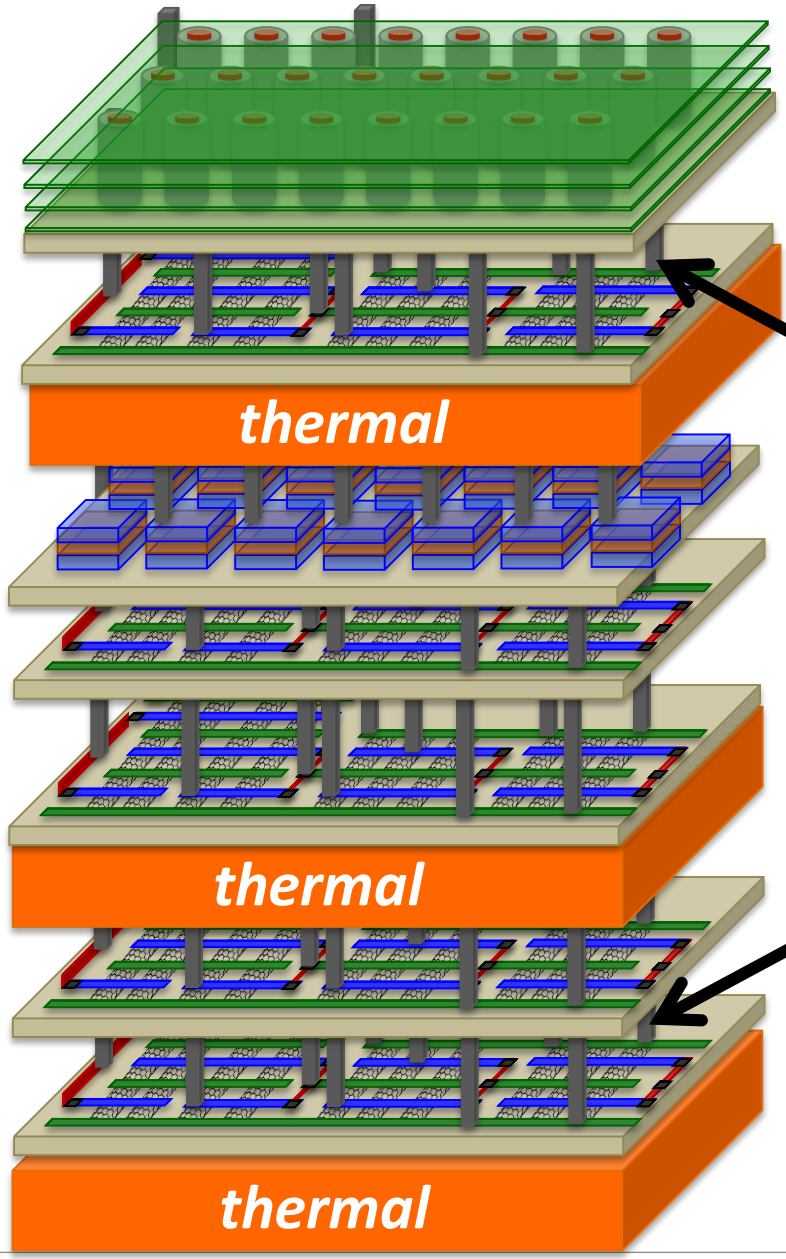
3D Resistive RAM
Massive storage

1D CNFET, 2D FET
Compute, RAM access

MRAM
Quick access

1D CNFET, 2D FET
Compute, RAM access

1D CNFET, 2D FET
Compute, Power, Clock



Not TSV

Ultra-dense, fine-grained vias

Silicon compatible

Energy-Efficient Abundant-Data Computing: The N3XT 1,000×

Aly et al., *IEEE Computer*, 2015

Mohamed M. Sabry Aly, Mingyu Gao, Gage Hills, Chi-Shuen Lee, Greg Pitner, M Tony F. Wu, and Mehdi Asheghi, Stanford University

Jeff Bokor, University of California, Berkeley

Franz Franchetti, Carnegie Mellon University

Kenneth E. Goodson and Christos Kozyrakis, Stanford University

Igor Markov, University of Michigan, Ann Arbor

Kunle Olukotun, Stanford University

Larry Pileggi, Carnegie Mellon University

Eric Pop, Stanford University

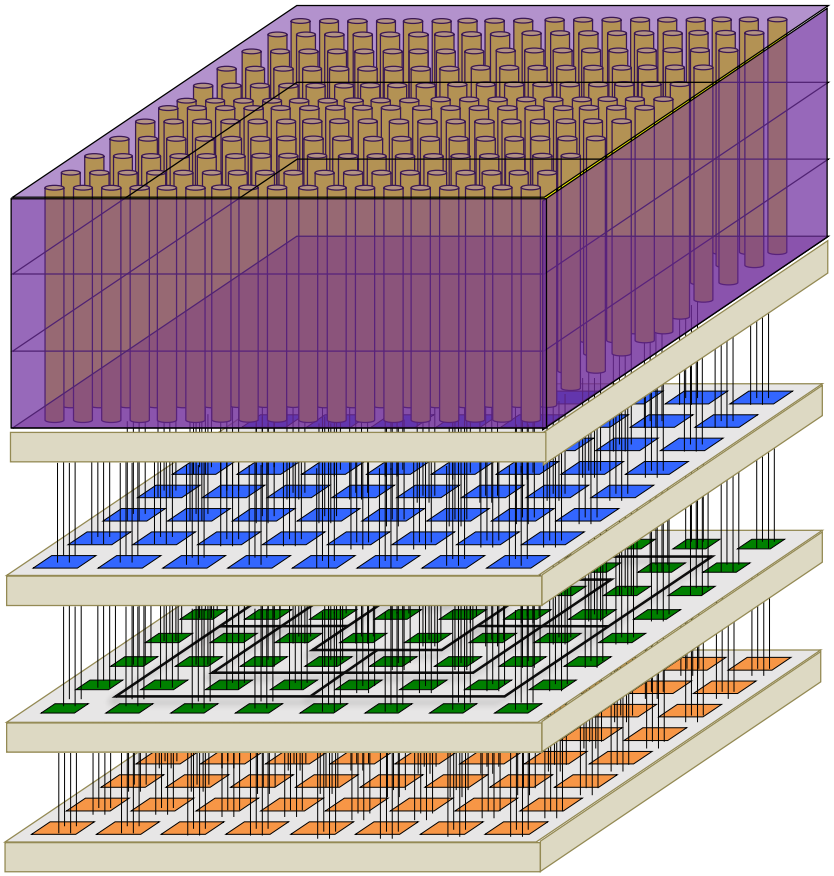
Jan Rabaey, University of California, Berkeley

Christopher Ré, H.-S. Philip Wong, and Subhasish Mitra, Stanford University

Next-generation information technologies will process unprecedented amounts of loosely structured data that overwhelm existing computing systems. N3XT improves the energy efficiency of abundant-data applications 1,000-fold by using new logic and memory technologies, 3D integration with fine-grained connectivity, and new architectures for computation immersed in memory.

N3XT Architecture

- Full physical design
  - Standard tool flow



Memory: RRAM
CNFET access

Memory control:
CNFET

Cache: STTRAM
CNFET access

Cores: CNFET

2 Million ILVs

Source: M. Sabry Aly, S. Mitra (Stanford)

N3XT: 850× EDP Benefit

DeepDive

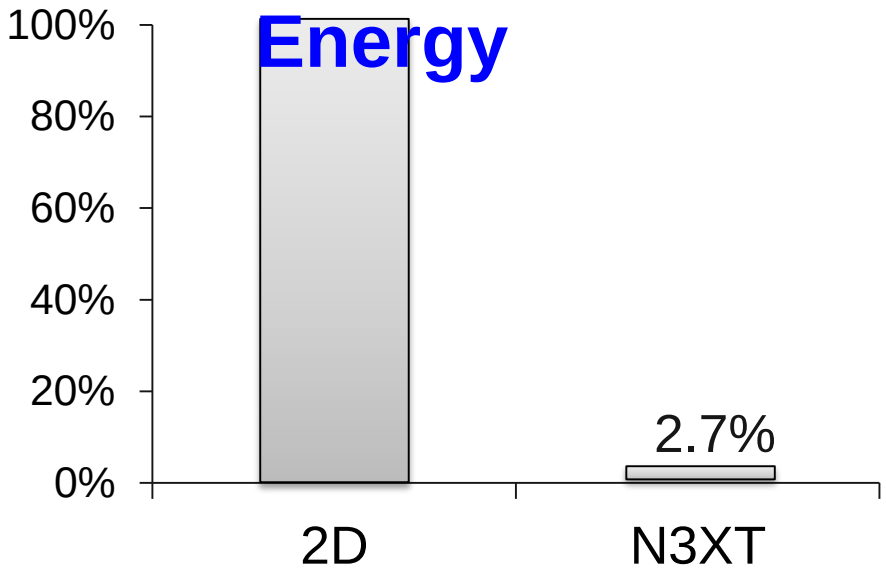
intel

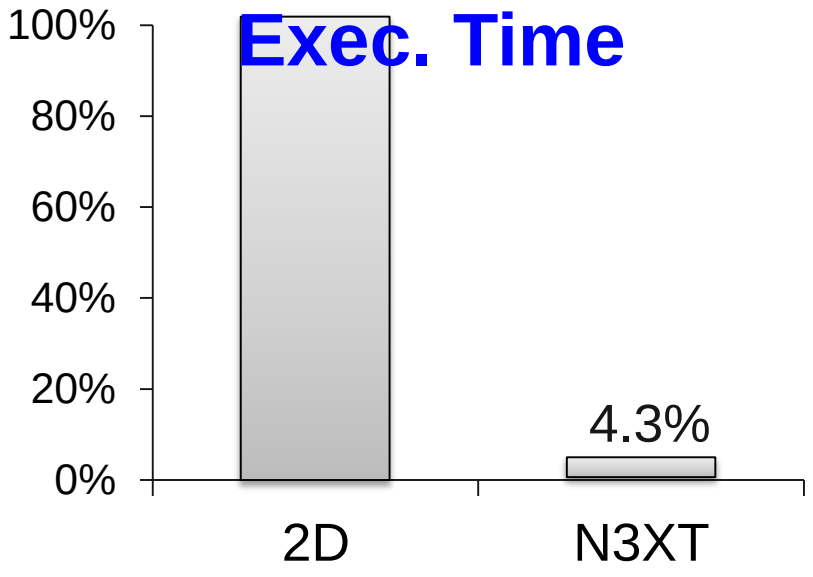
BigDataBench

Data-intensive computing

IBM graph analytics

Same power density (67 W/cm^2) and peak temperature 63°C





PageRank app.

M. Aly…S. Mitra, *IEEE Computer* (2015)

N3XT: 850× EDP Benefit

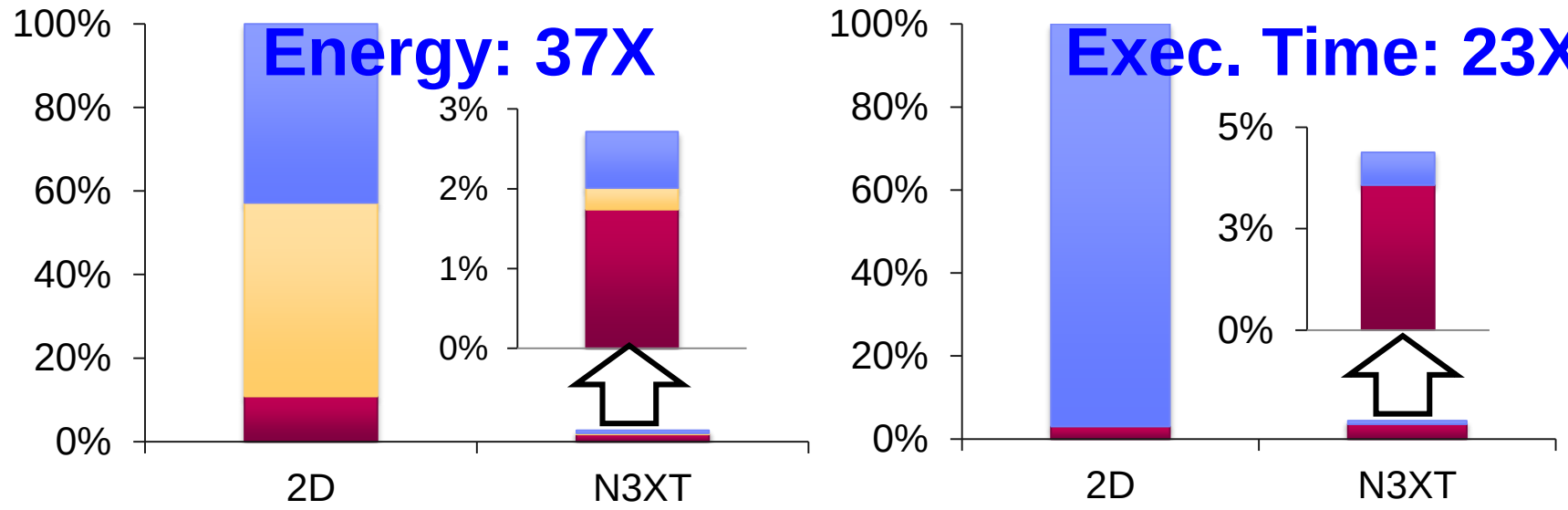
DeepDive

intel

BigDataBench

Data-intensive computing

IBM graph analytics

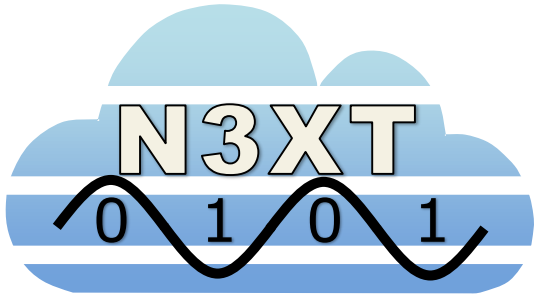


PageRank app.

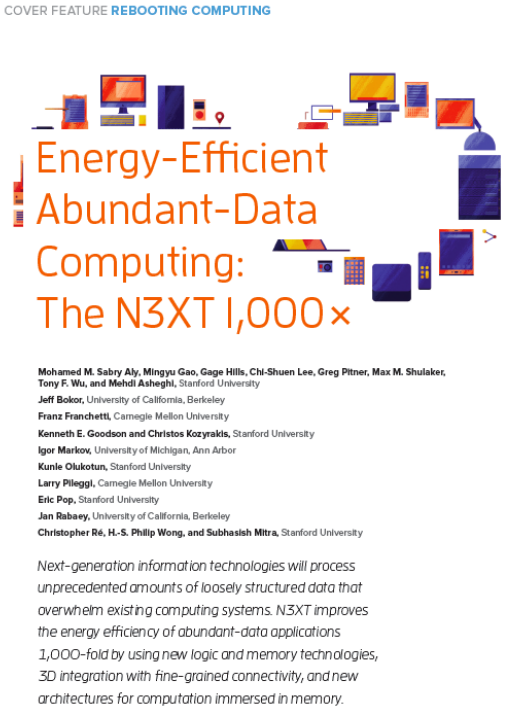
M. Aly…S. Mitra, *IEEE Computer* (2015)

Nanosystems

Nano-Engineered Computing Systems Technology



End-to-end approach

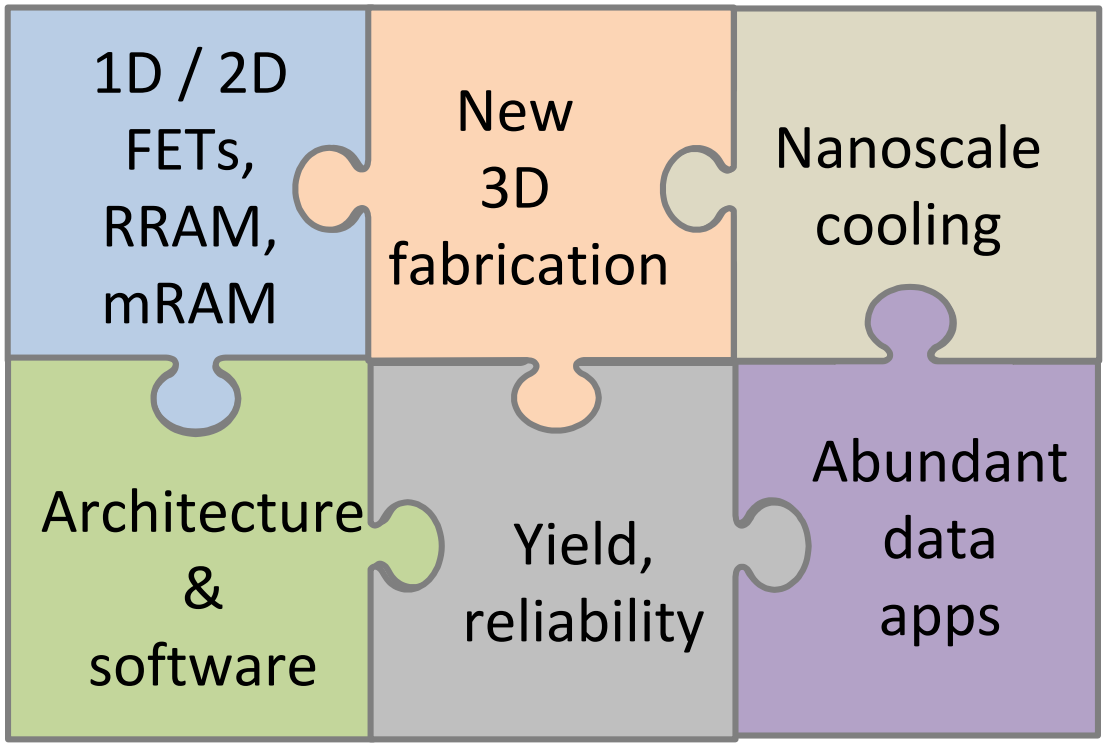

COVER FEATURE REBOOTING COMPUTING

Energy-Efficient Abundant-Data Computing: The N3XT 1,000×

Mohamed M. Sabry Aly, Mingyu Gao, Gage Hills, Chi-Shuen Lee, Greg Pitner, Max M. Shulaker, Tony F. Wu, and Mehdi Asheghi, Stanford University
Jeff Bokor, University of California, Berkeley
Franz Franchetti, Carnegie Mellon University
Kenneth E. Goodson and Christos Kozyrakis, Stanford University
Igor Markov, University of Michigan, Ann Arbor
Kunle Olukotun, Stanford University
Larry Pileggi, Carnegie Mellon University
Eric Pop, Stanford University
Jan Rabaey, University of California, Berkeley
Christopher Ré, H.-S. Philip Wong, and Subhasish Mitra, Stanford University

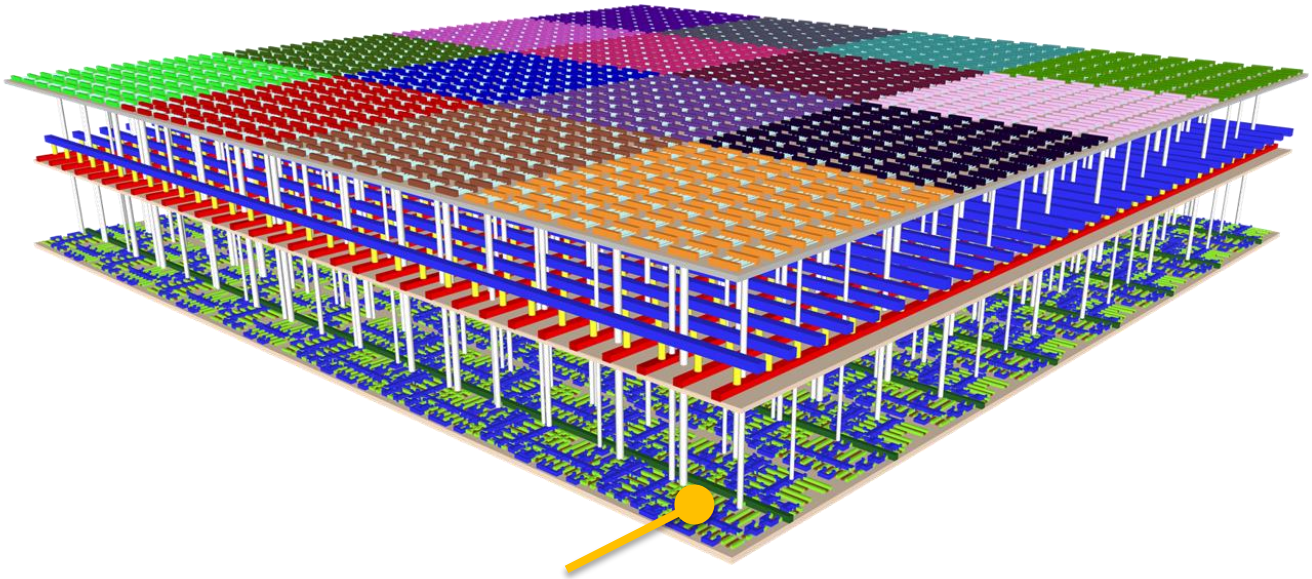
Next-generation information technologies will process unprecedented amounts of loosely structured data that overwhelm existing computing systems. N3XT improves the energy efficiency of abundant-data applications 1,000-fold by using new logic and memory technologies, 3D integration with fine-grained connectivity, and new architectures for computation immersed in memory.

[M. Aly et al., IEEE Computer '15]

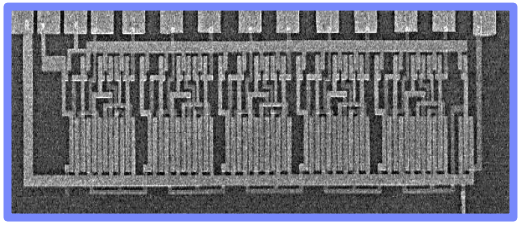


Can Build Nano*systems* Today

>2 Million carbon nanotube FETs, 1 Mbit Resistive RAM

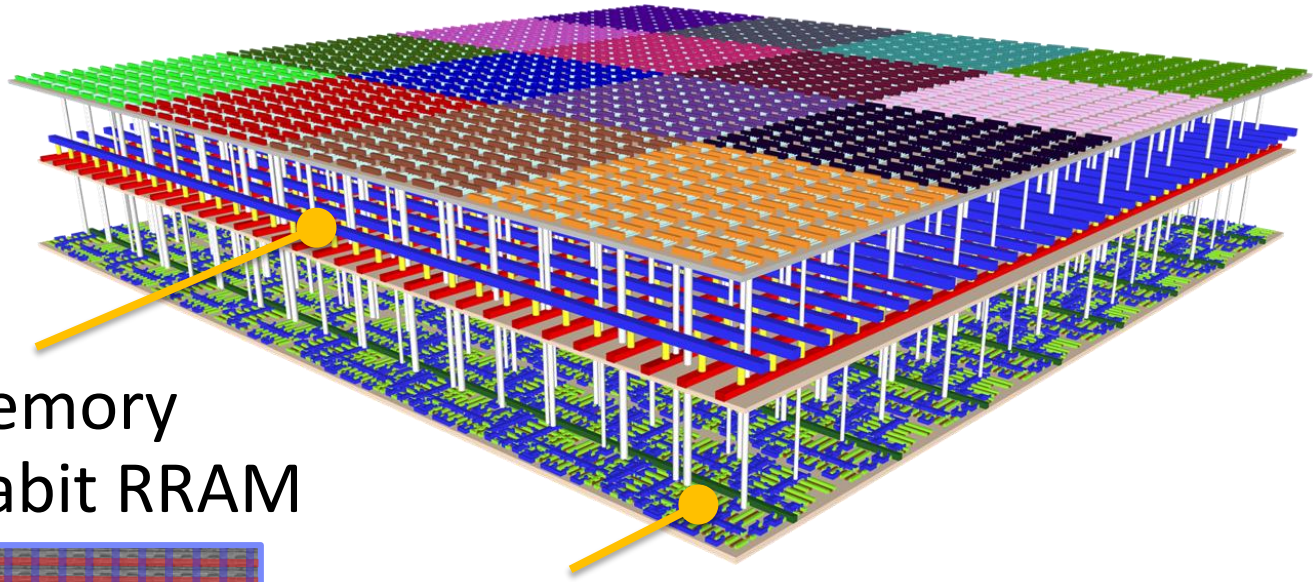


CNT computing logic
Classification accelerator

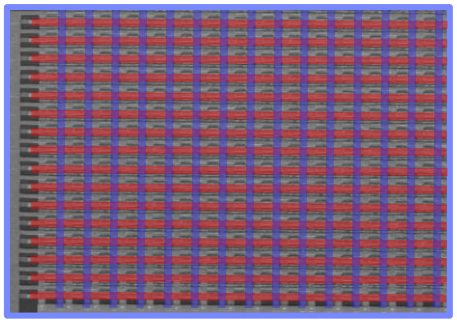


[M. Shulaker et al., submitted '16]

Can Build Nano*systems* Today



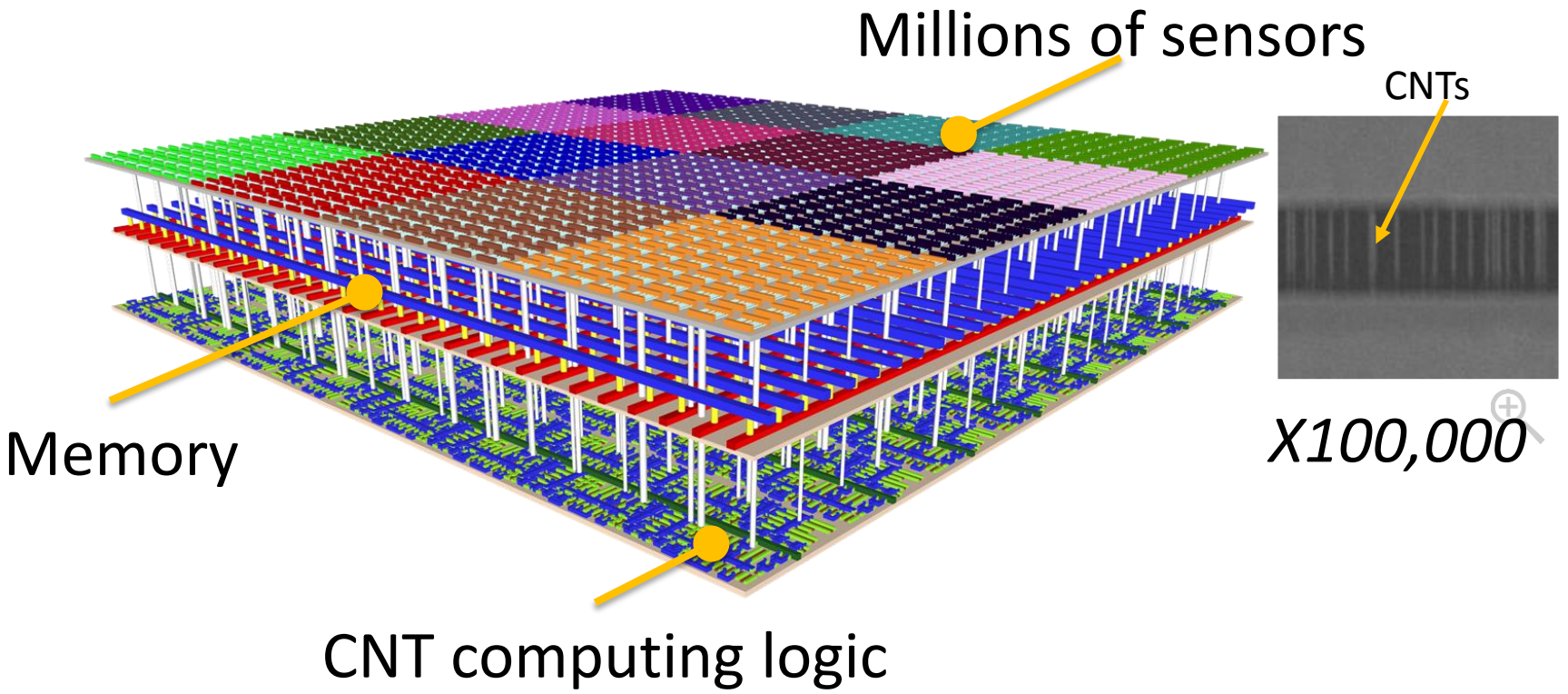
Memory
1 Megabit RRAM



CNT computing logic

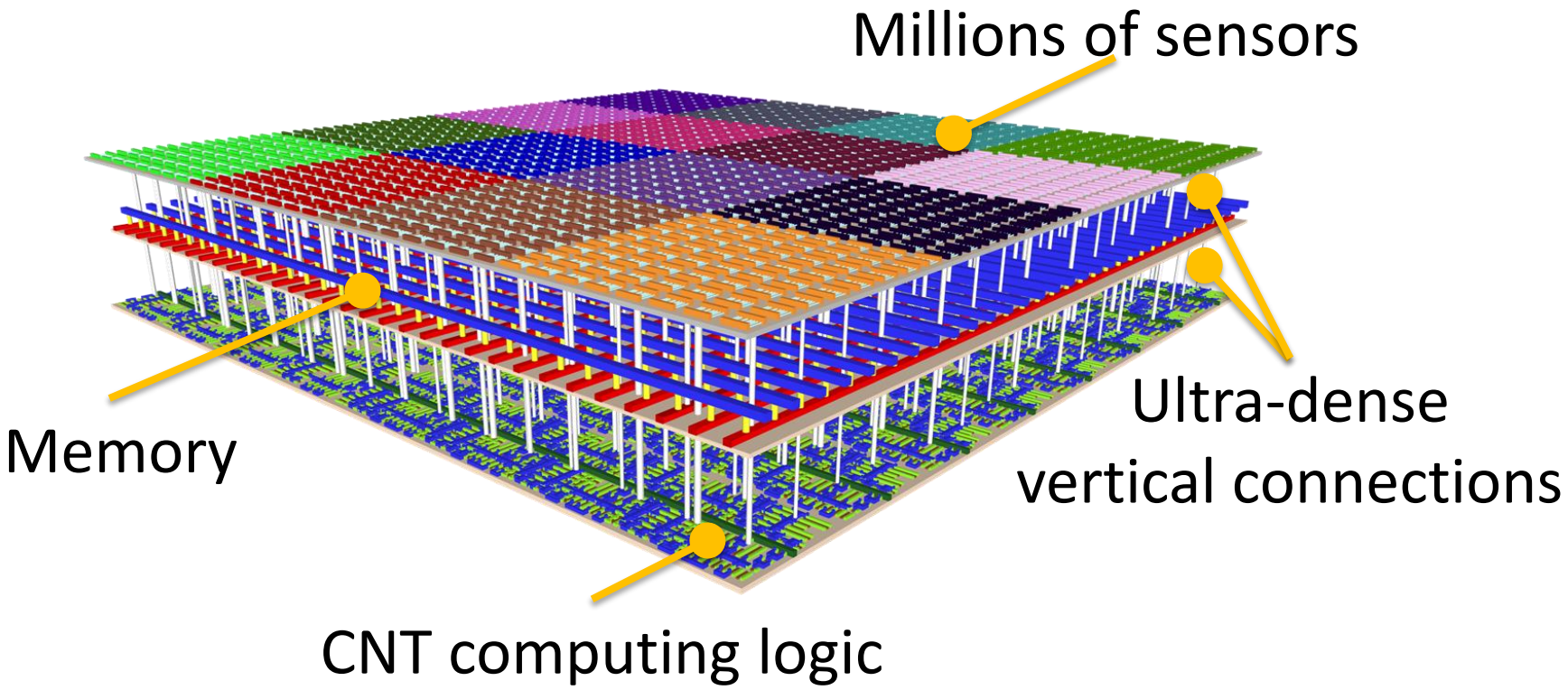
[M. Shulaker et al., submitted '16]

Can Build Nano*systems* Today



[M. Shulaker et al., submitted '16]

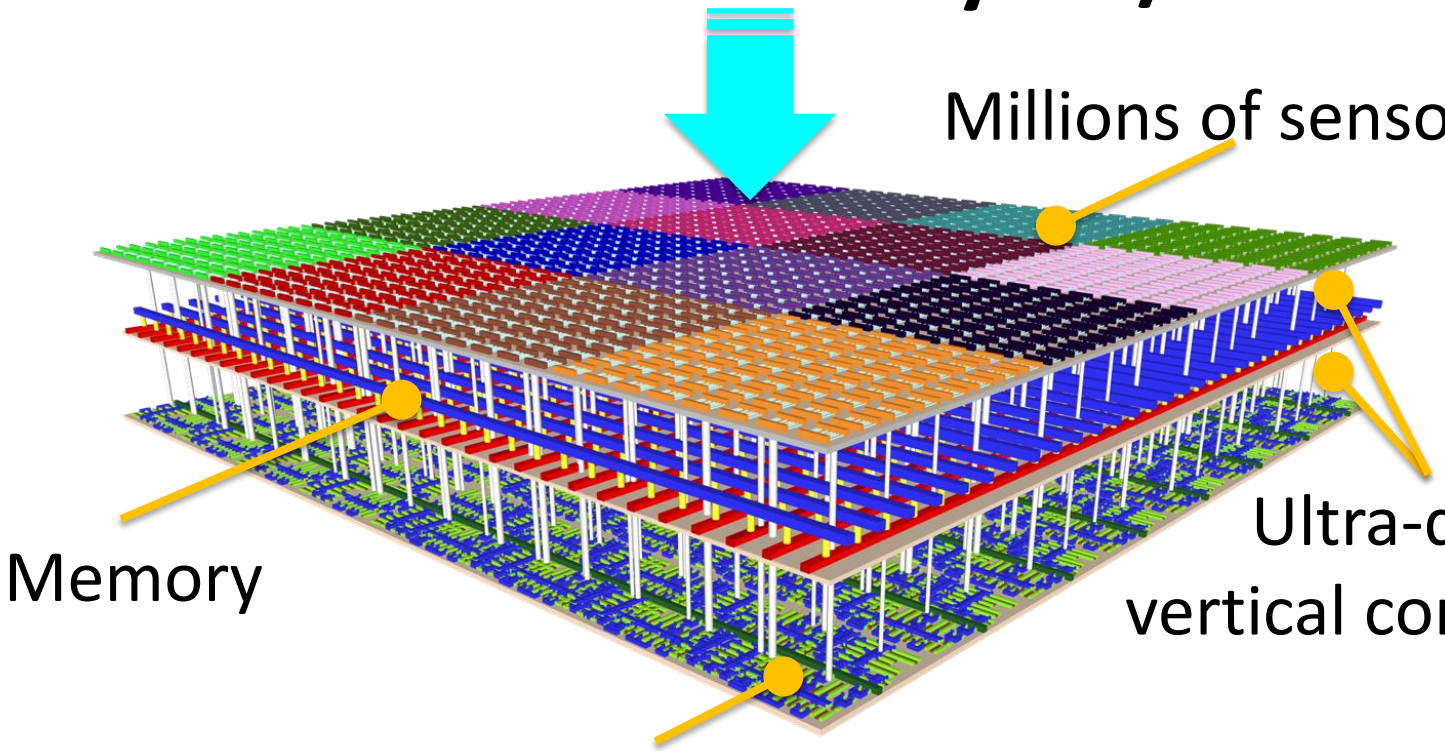
Can Build Nano*systems* Today



[M. Shulaker et al., submitted '16]

Can Build Nano*systems* Today

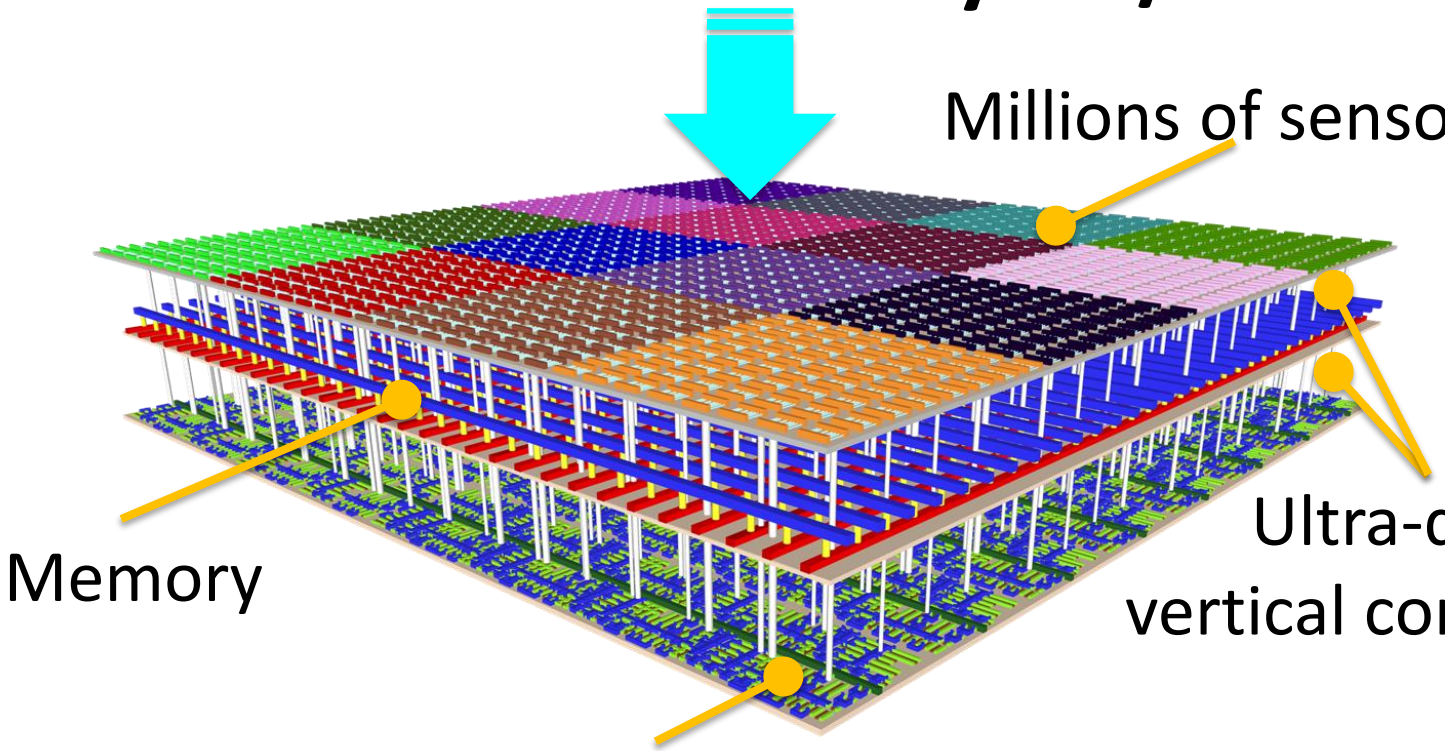
Abundant data: Terabytes / second



[M. Shulaker et al., submitted '16]

Can Build Nano*systems* Today

Abundant data: Terabytes / second

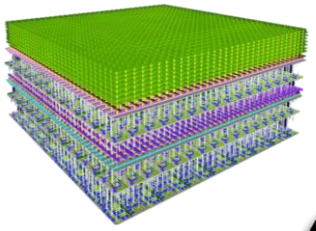


In-situ classification:
Extensive, accurate

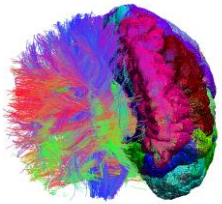
[M. Shulaker et al., submitted '16]

Source: Subhasish Mitra, Chris Ré

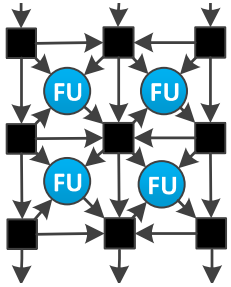
SUPERCELL



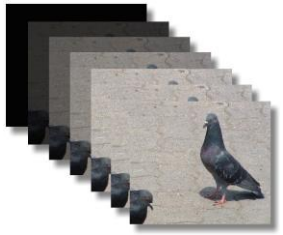
N3XT



Brain network

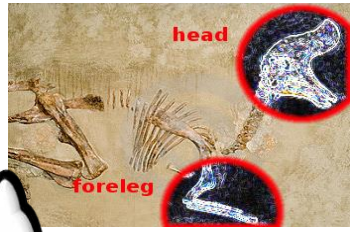


CGRAs



Automatic video annotation

Commodity hardware



Advanced image analysis

The Day After Tomorrow

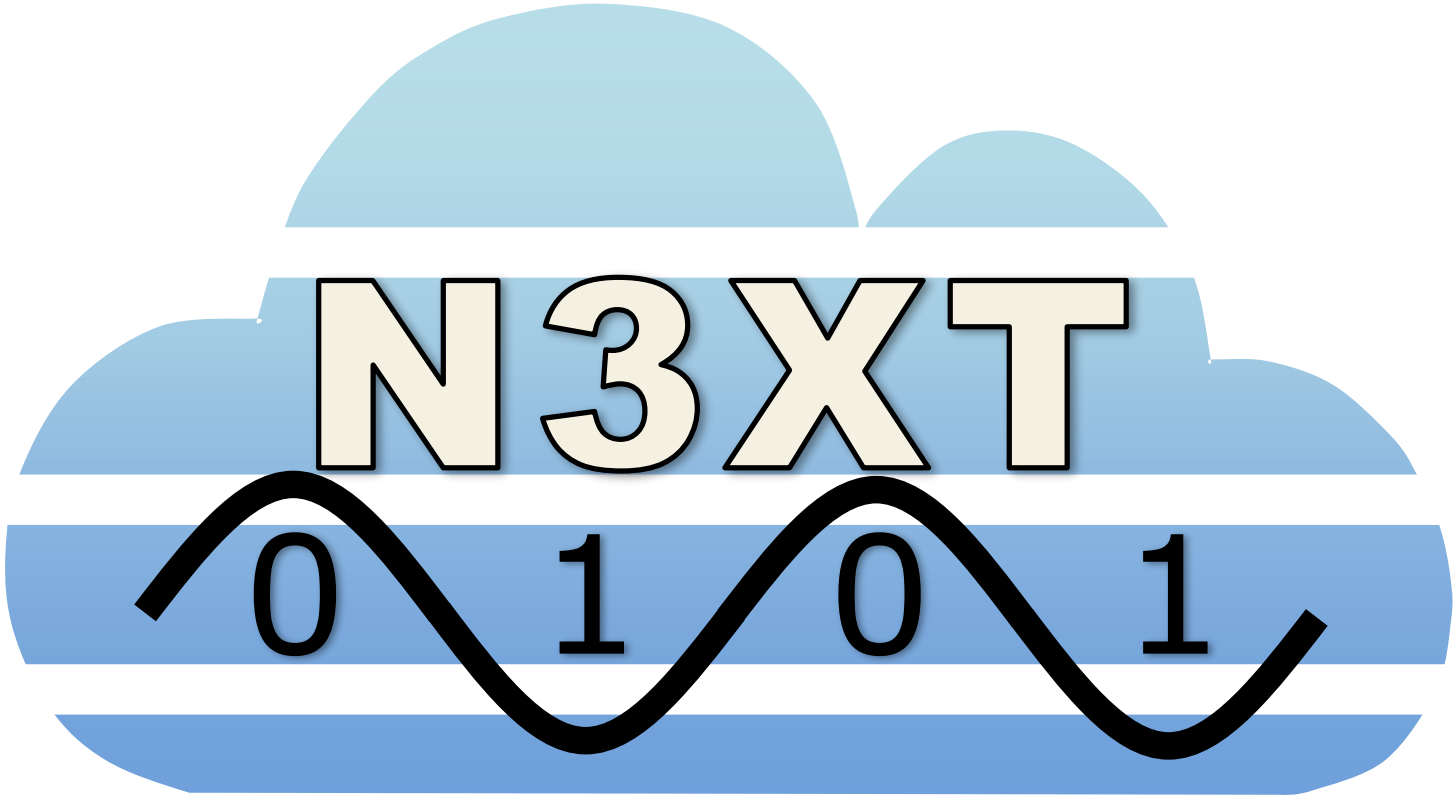
Tomorrow

Today

Hardware

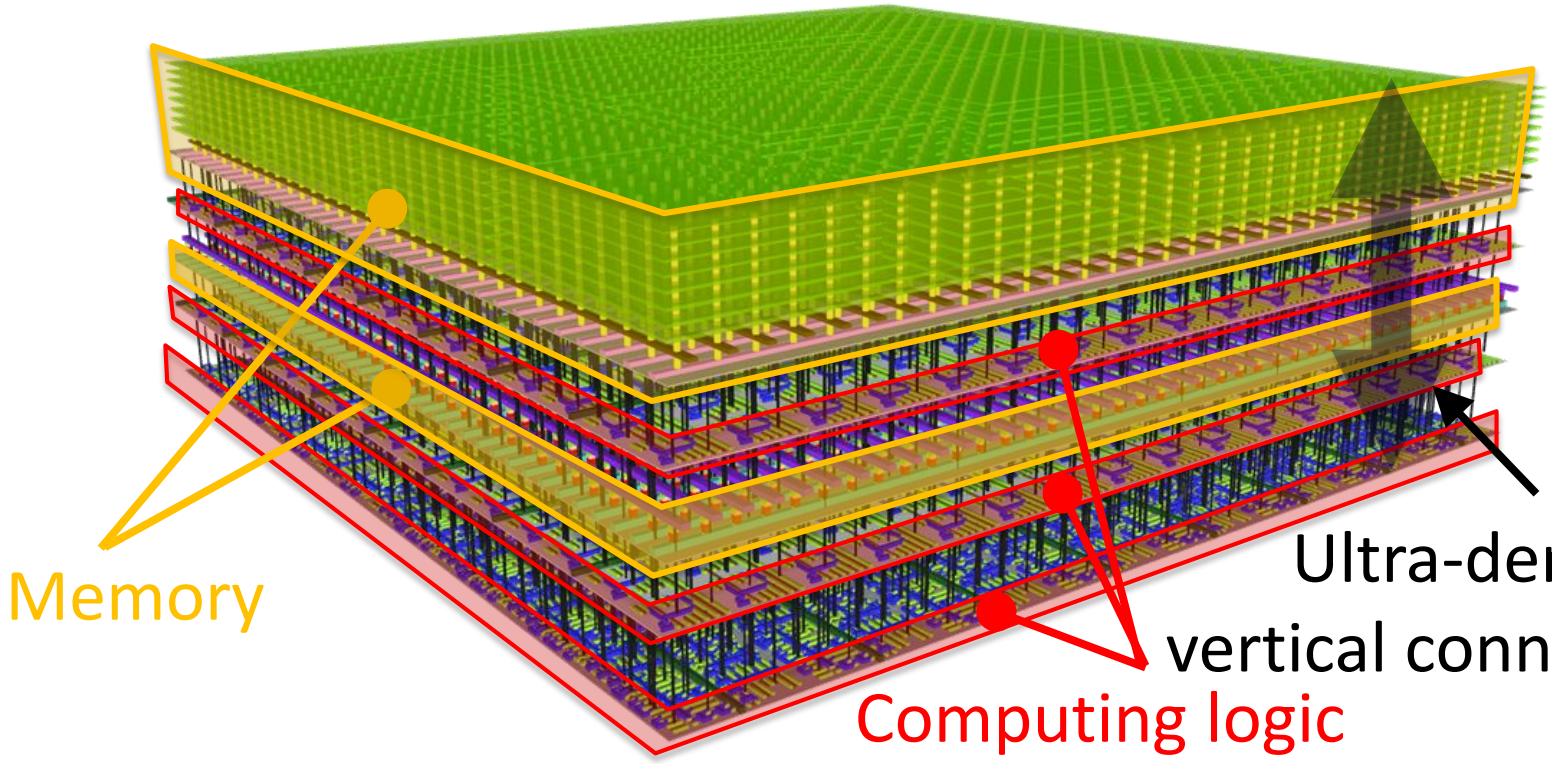
Software

Source: Subhasish Mitra, Chris Ré


N3XT
0 1 0 1

N3XT Nanosystems

Computation immersed in memory



A Nanotechnology-Inspired Grand Challenge for Future Computing

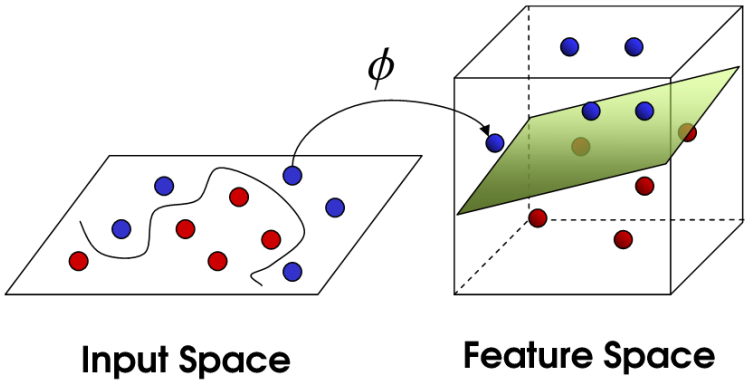
OCTOBER 20, 2015 AT 6:00 AM ET BY LLOYD WHITMAN, RANDY BRYANT, AND TOM KALIL

Summary: Today, the White House is announcing a grand challenge to develop transformational computing capabilities by combining innovations in multiple scientific disciplines.

In June,
suggest
over 10
three A
Compu

These nanotechnology innovations will have to be developed in close coordination with new computer architectures, and will likely be informed by our growing understanding of the brain—a remarkable, fault-tolerant system that consumes less power than an incandescent light bulb.

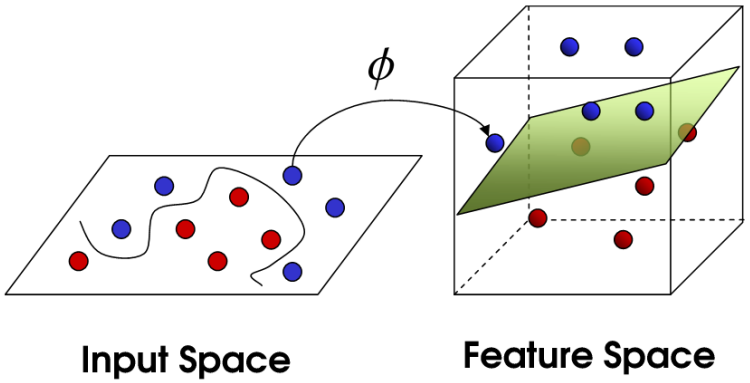
Goal: Energy Efficiency of the Brain



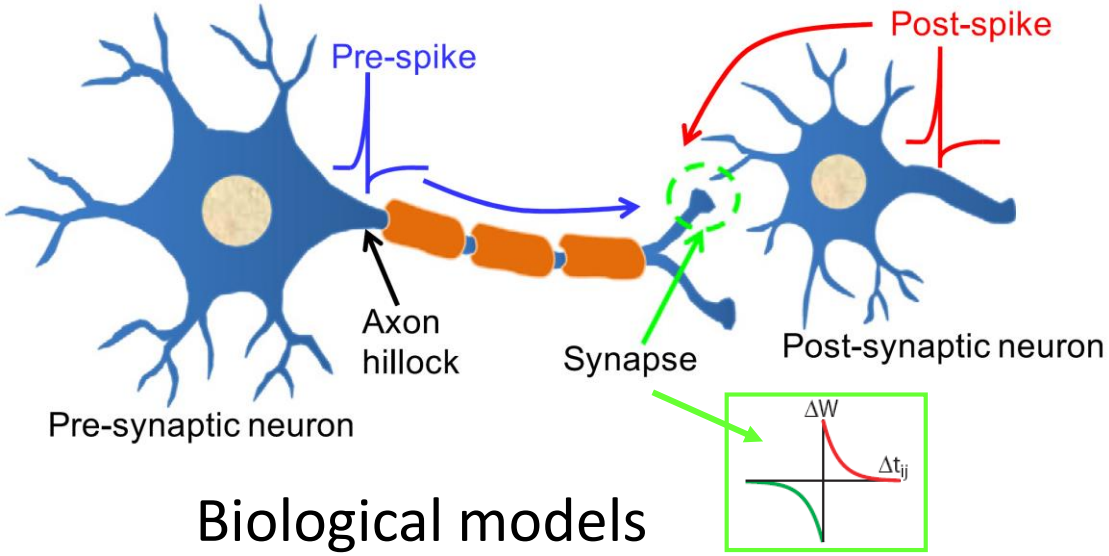
Learning algorithms

Image sources: stackoverflow

Goal: Energy Efficiency of the Brain



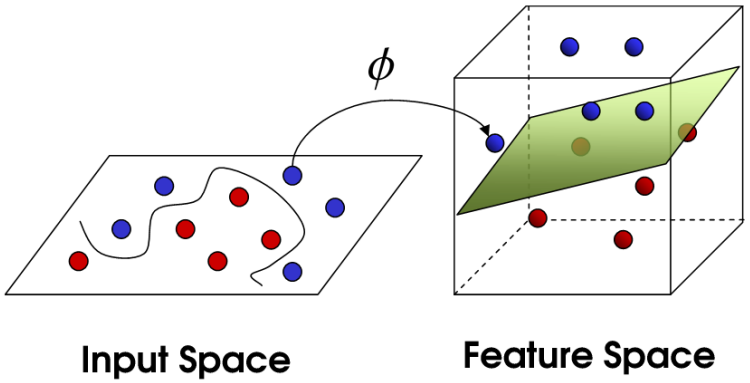
Learning algorithms



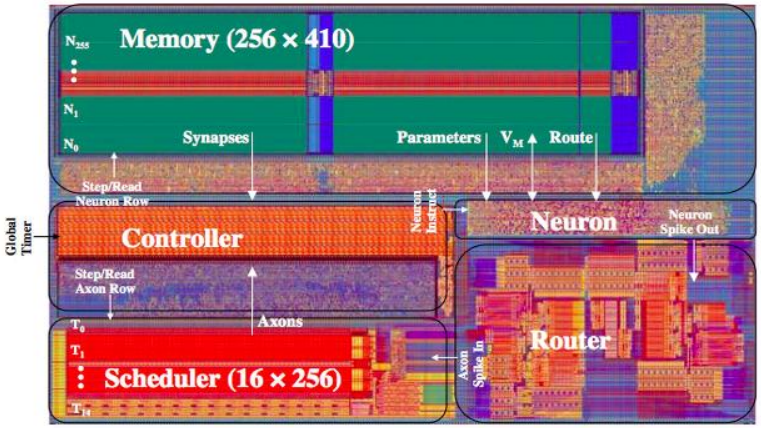
Biological models

Image sources: stackoverflow, Stanford

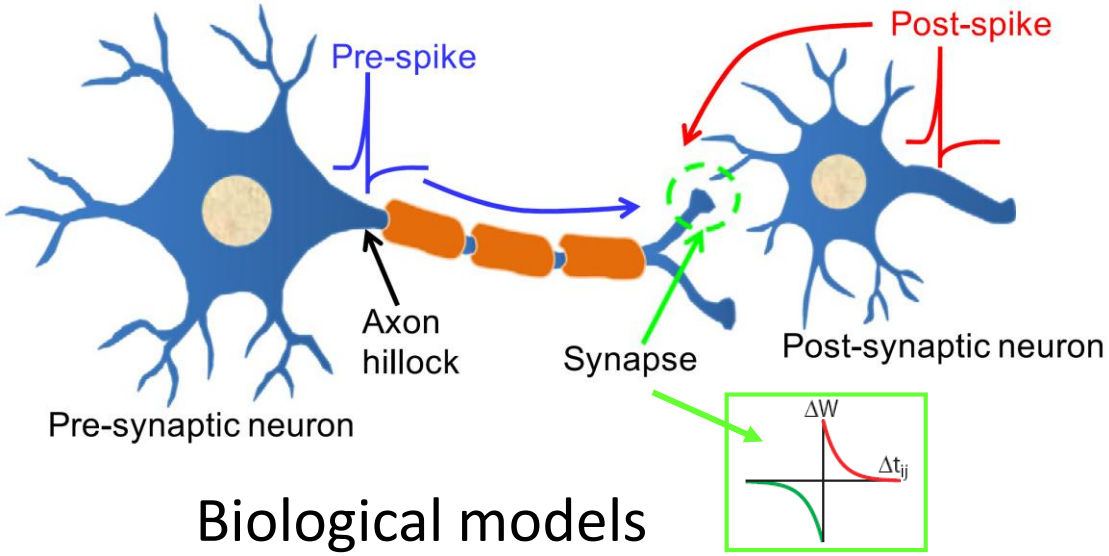
Goal: Energy Efficiency of the Brain



Learning algorithms



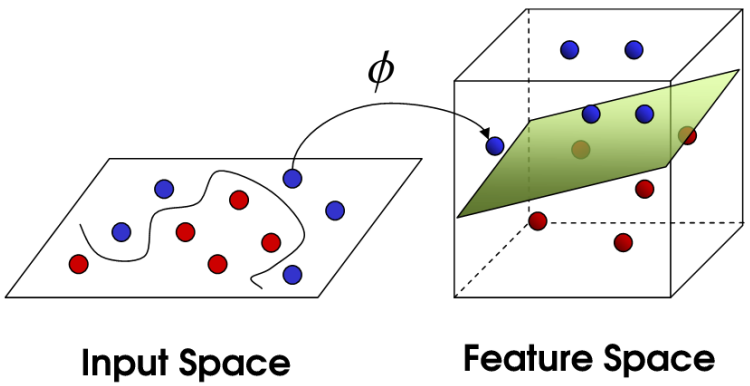
Electronic implementation



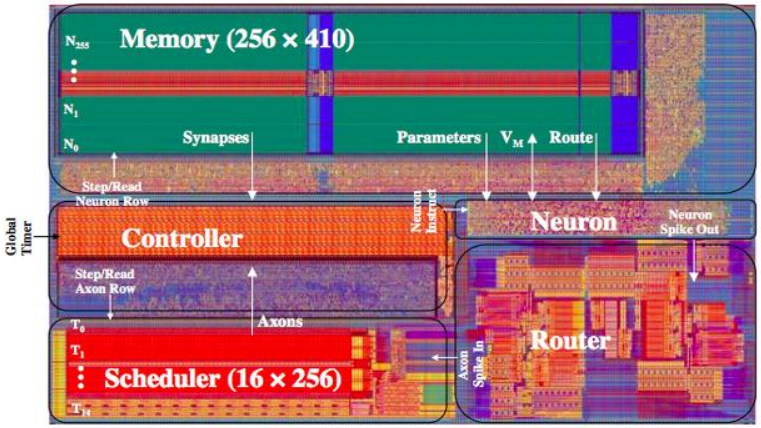
Biological models

Image sources: stackoverflow, Stanford, IBM

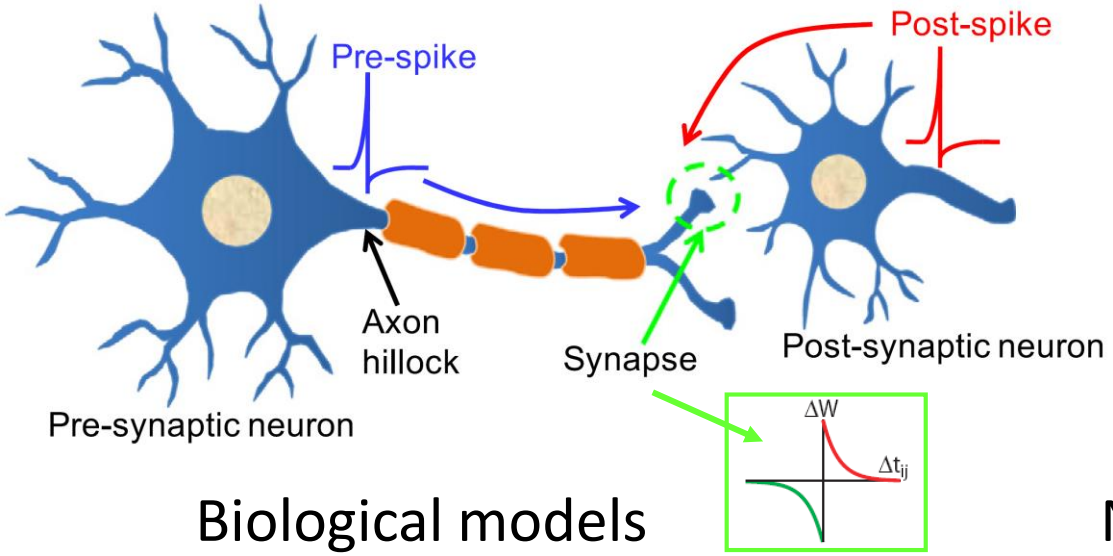
Goal: Energy Efficiency of the Brain



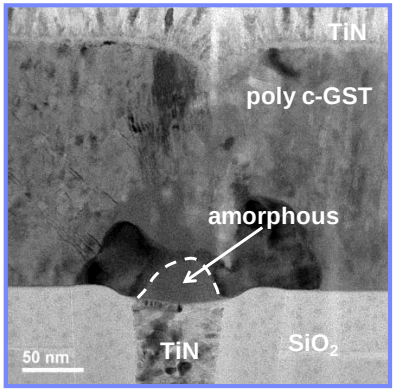
Learning algorithms



Electronic implementation



Biological models



Nanoscale electronic synapse

Image sources: stackoverflow, Stanford, IBM

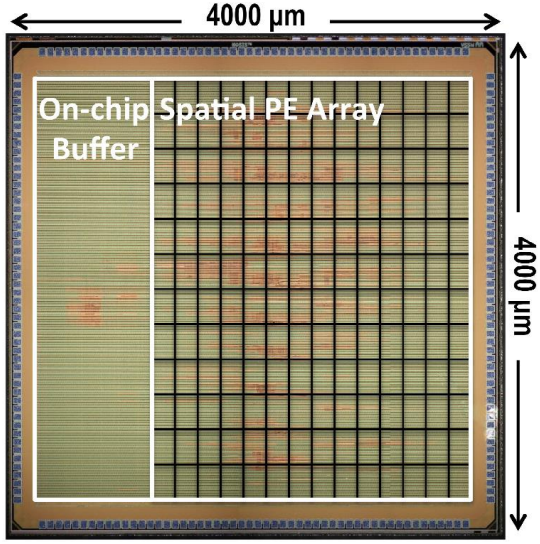
Scale Up Requires Energy Efficiency

	Application	Hardware used	Estimated power consumption
Large scale	**Emulating 4.5% of human brain**:10^{13} synapses, 10^9 neurons	Blue Gene/P: 36,864 nodes, 147,456 cores	**2.9 MW** (LINPACK)
	Deep sparse autoencoder: 10^9 synapses, 10M images	1,000 CPUs (16,000 cores)	**~100 kW** (cores only)
Small to moderate scale	**Convolutional neural net** with 60M synapses, 650K neurons	2 GPUs	**1,200 W**
	Restricted Boltzmann Machine: 28M synapses; 69,888 neurons	GPU	**550 W**
		CPU	**65 W**
	Processing 1 s of speech using **deep neural network**	GPU	**238 W**
		CPU (4 cores)	**80 W**

S. B. Eryilmaz et al., IEDM 2015

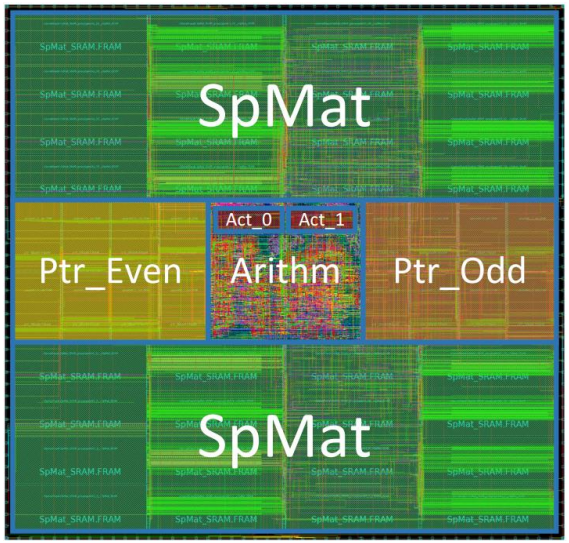
Custom Hardware for Energy Efficient AI

Y.-H. Chen et al, ISSCC 2016



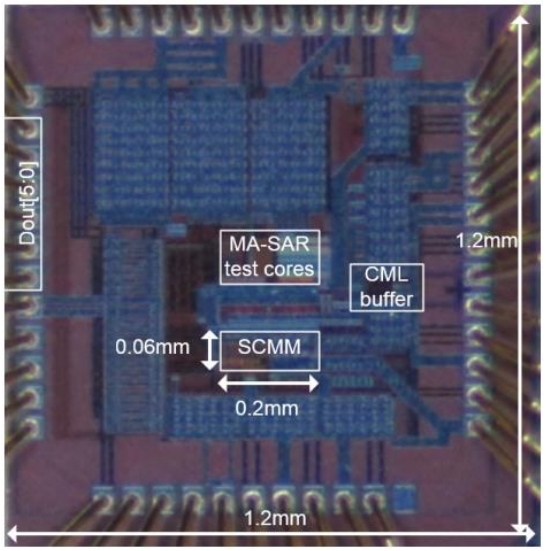
Inference in CNN hardware
280 mW for AlexNet
[MIT]

S. Han et al., arXiv 2016



Inference in deep compressed networks
600 mW for AlexNet
[Nvidia/Stanford]

E. Lee et al, ISSCC 2016



Matrix-vector multiplication with switch caps
8 TOPS/W @ 2.5 GHz
[Stanford]

Approaches of Neuromorphic Hardware

Approaches of Neuromorphic Hardware

Biology-based models / algorithms

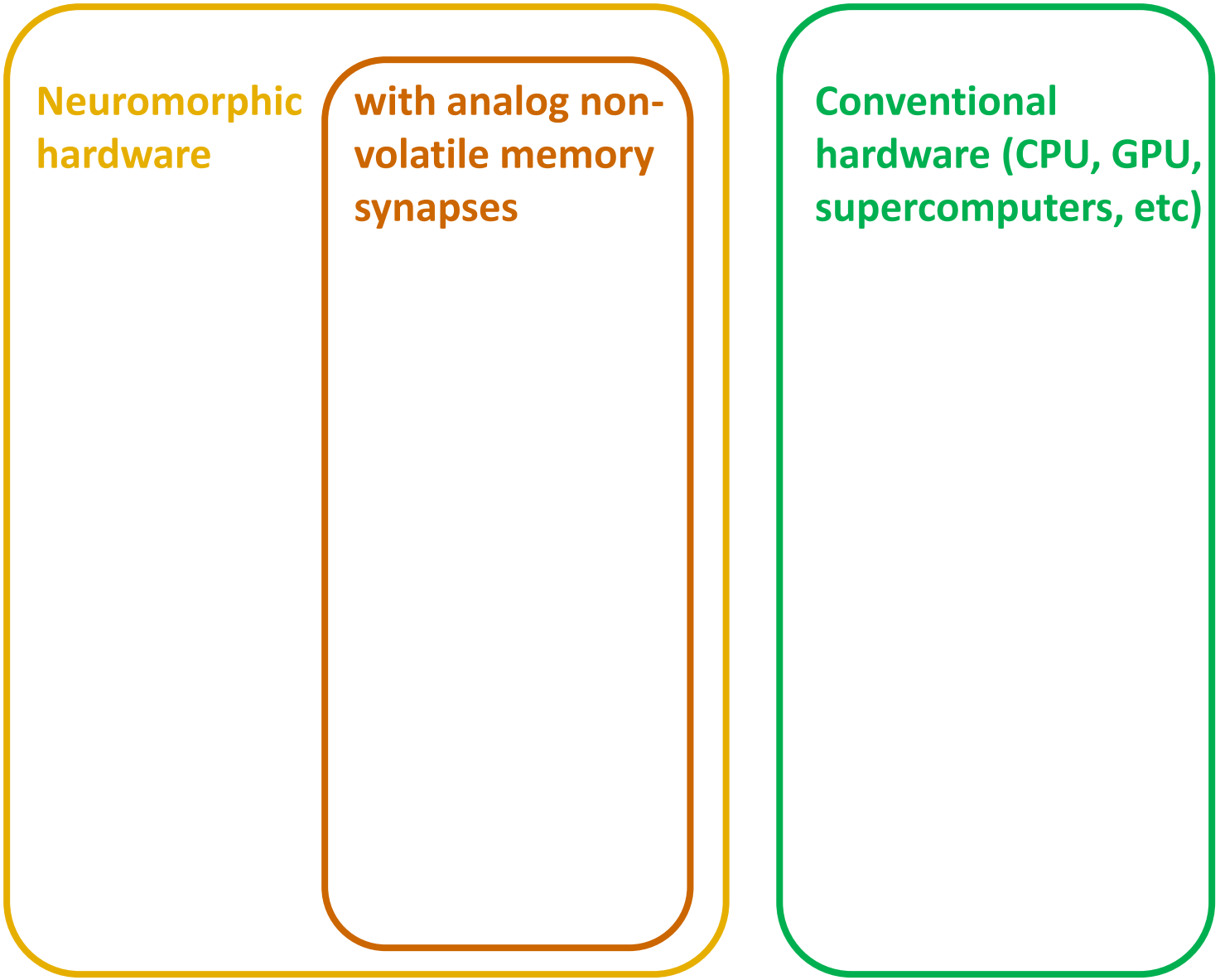
Conventional ML algorithms

Approaches of Neuromorphic Hardware

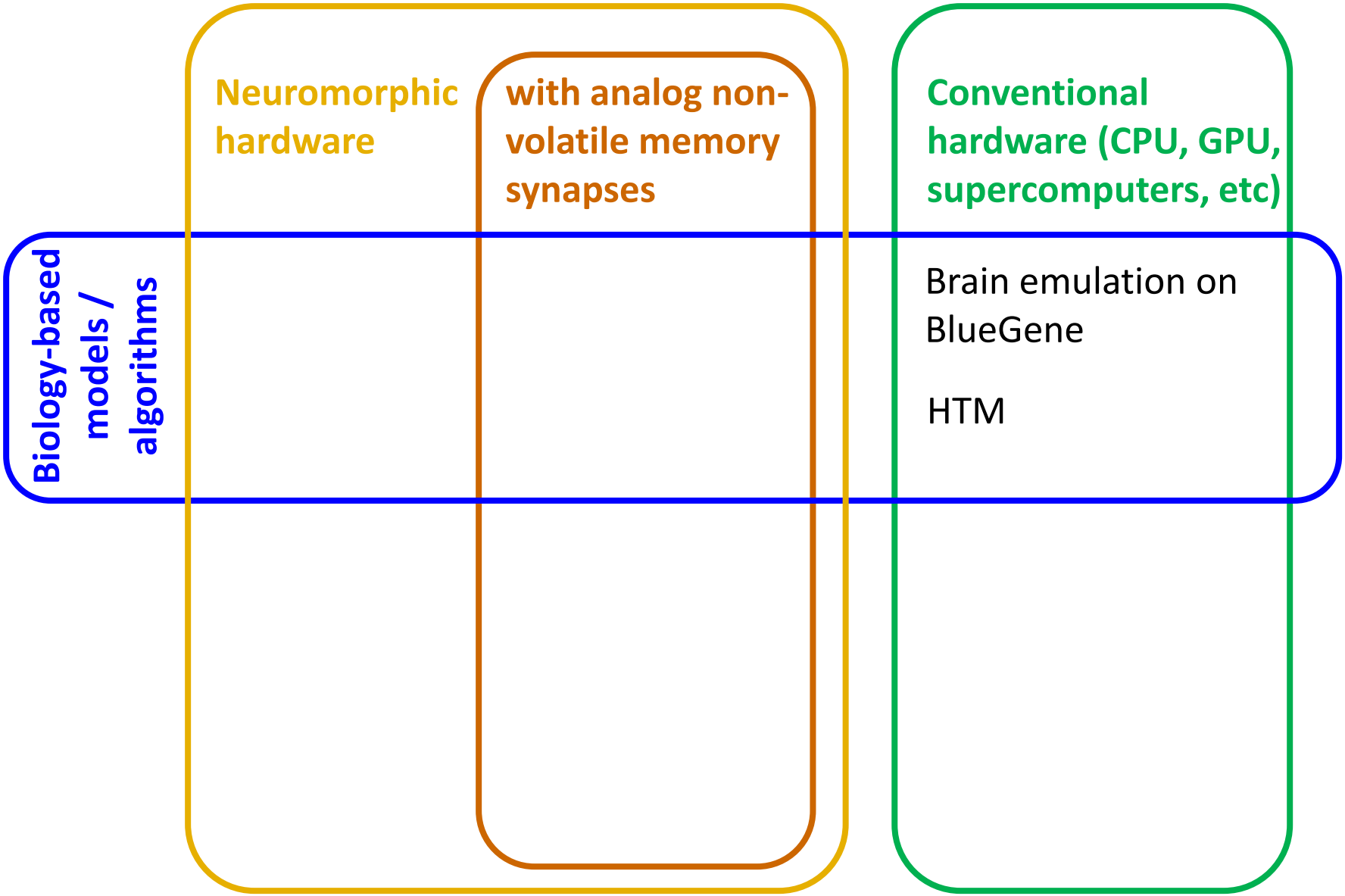
Neuromorphic hardware

Conventional hardware (CPU, GPU, supercomputers, etc)

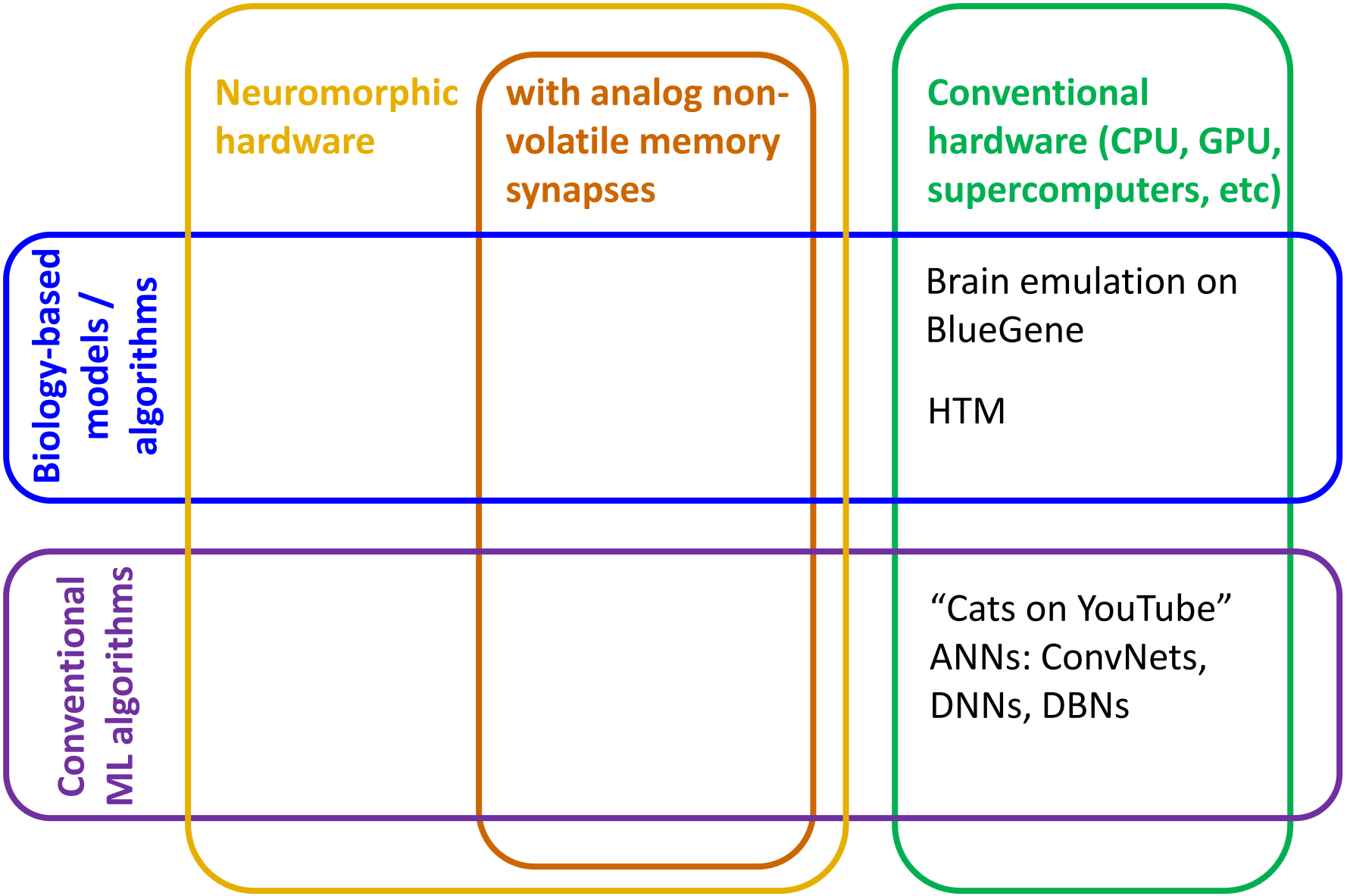
Approaches of Neuromorphic Hardware



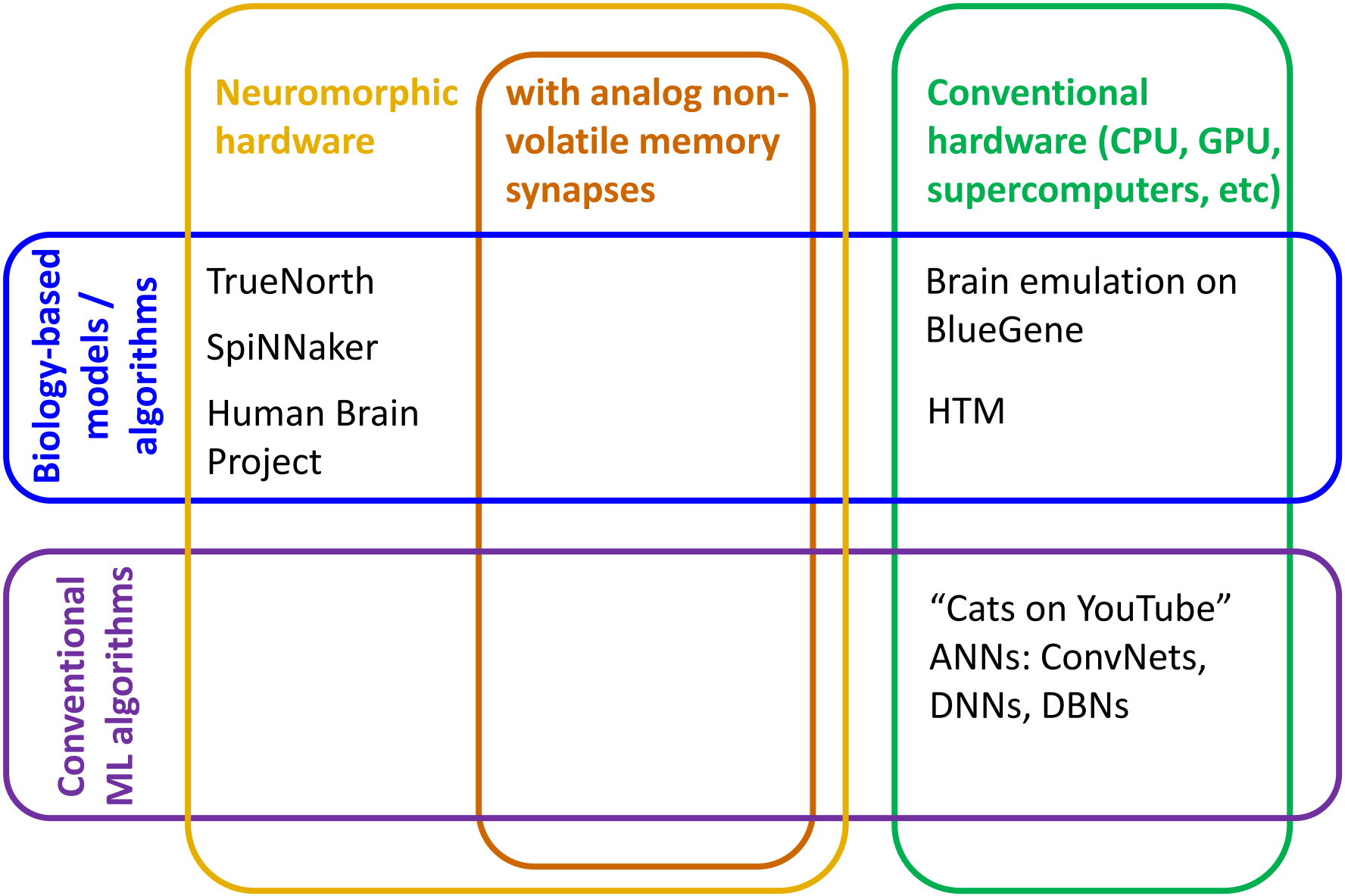
Approaches of Neuromorphic Hardware



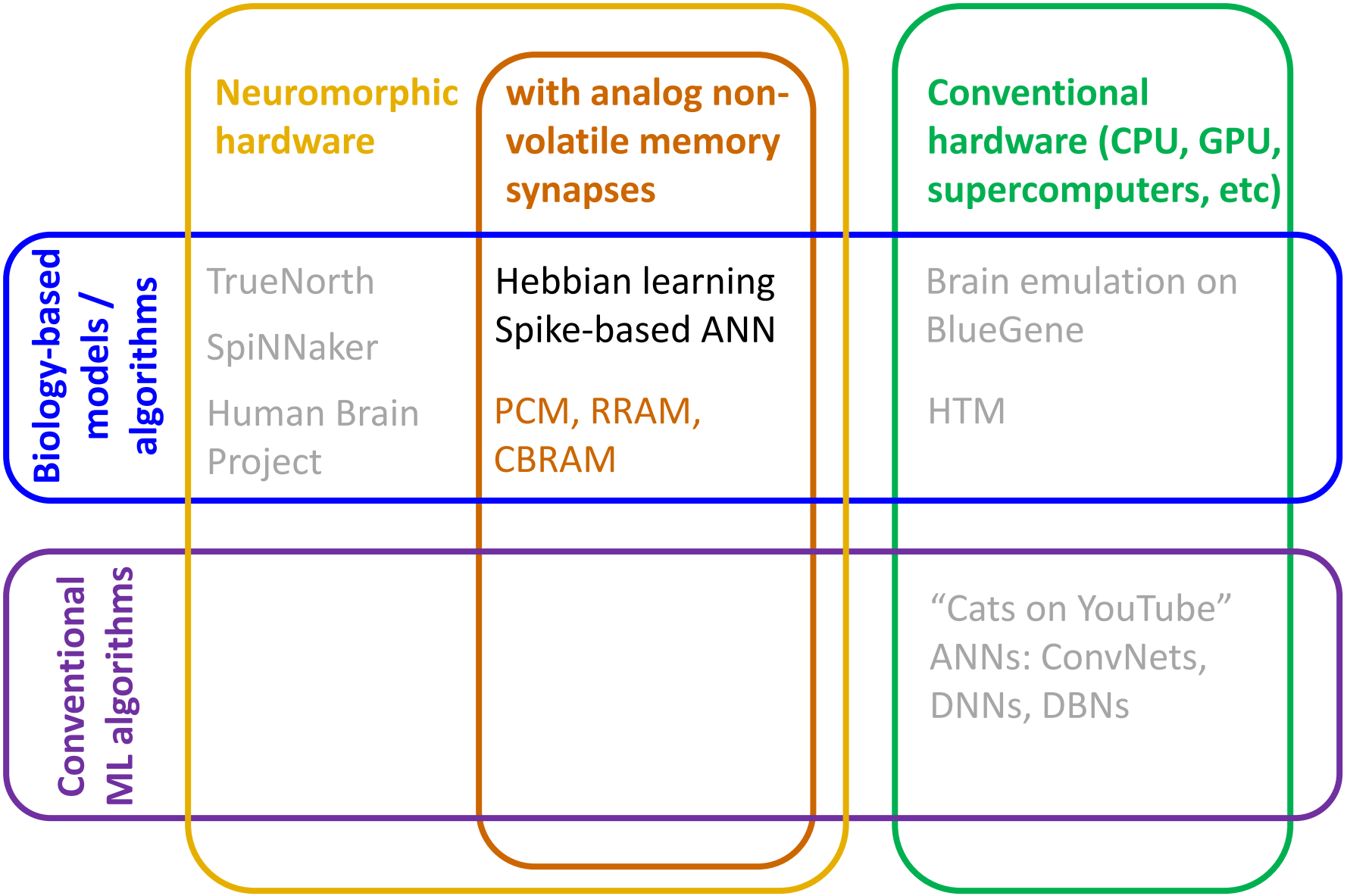
Approaches of Neuromorphic Hardware



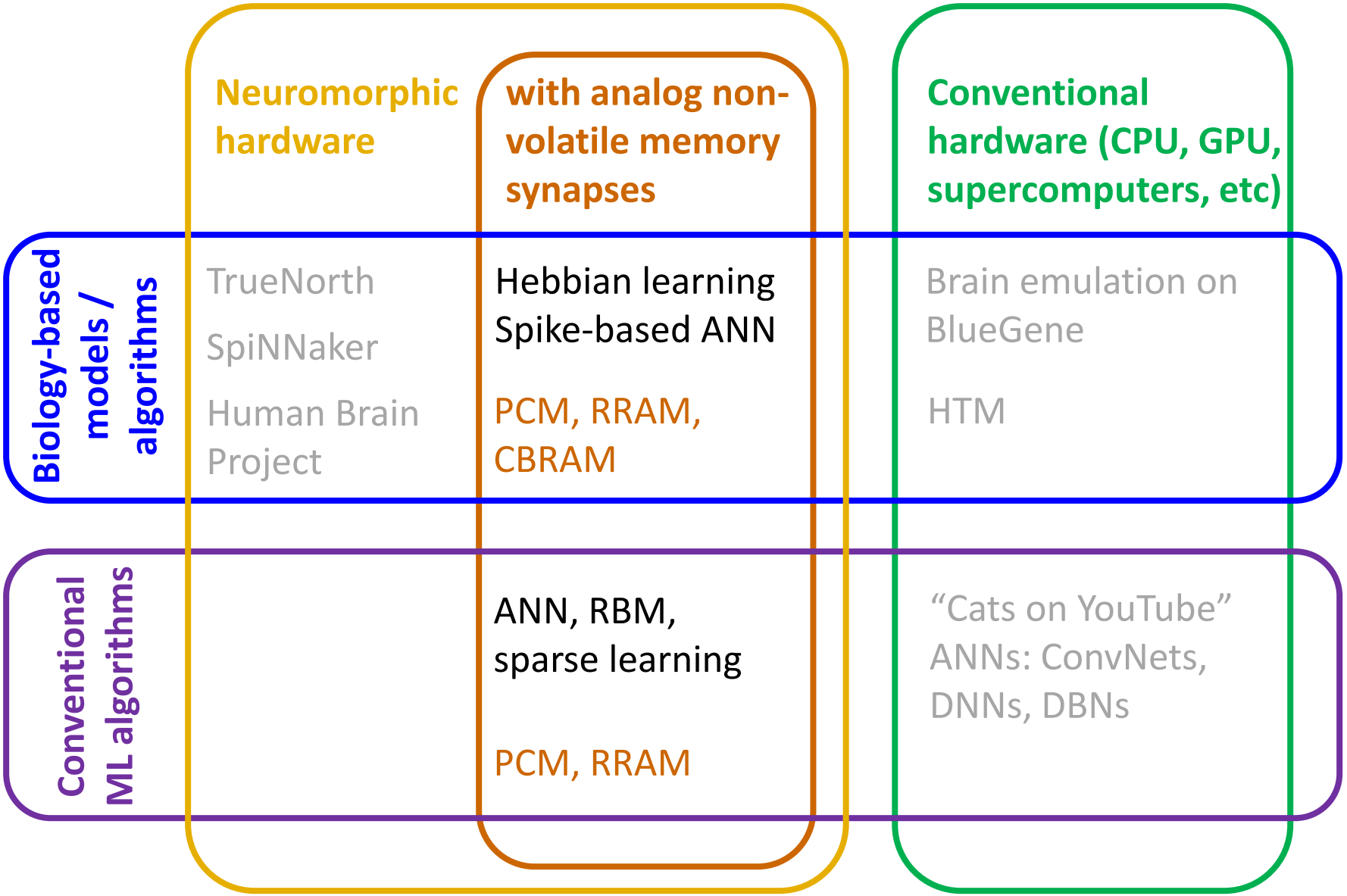
Approaches of Neuromorphic Hardware



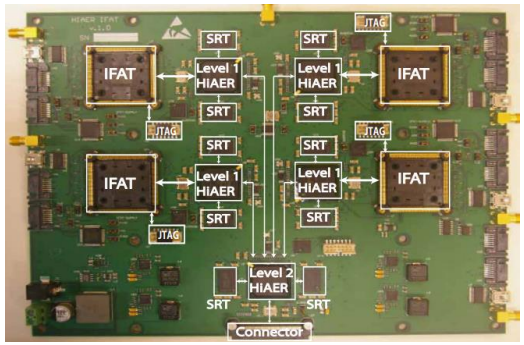
Approaches of Neuromorphic Hardware



Approaches of Neuromorphic Hardware



Today's "Large" Scale Architectures

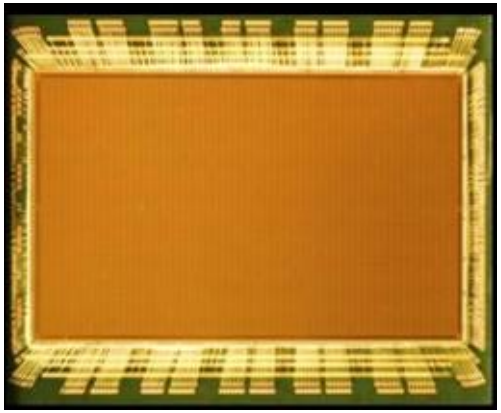


IFAT+HiAER (UCSD)

- Analog neurons
- Tree routing

Neurogrid (Stanford)

- Analog neurons
- Tree routing



TrueNorth (IBM)

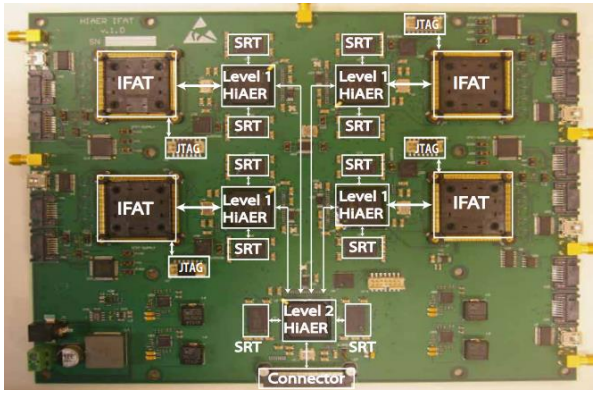
- Digital neurons
- Mesh-based routing

SpiNNaker (U of Manchester)

- Digital neurons
- Mesh routing

Synapses Implemented Today

DRAM (off-chip)



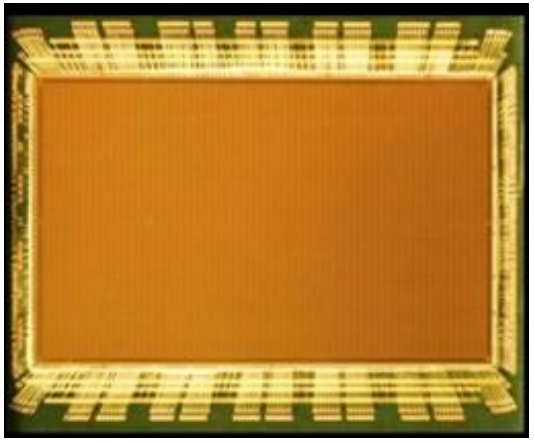
IFAT+HiAER (UCSD)

- Energetically expensive
  - Refresh
  - Off-chip access
- Scalability?

DRAM (off-chip)

Neurogrid (Stanford)

SRAM (on-chip)



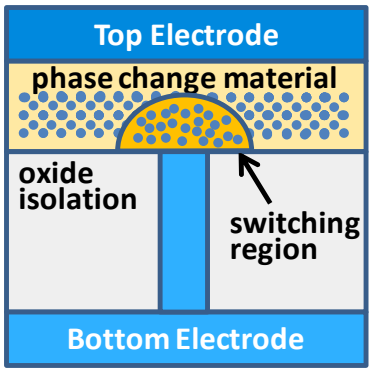
TrueNorth (IBM)

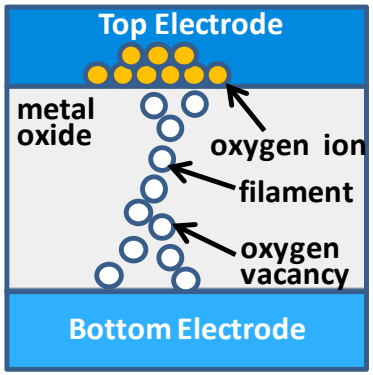
- Area inefficient

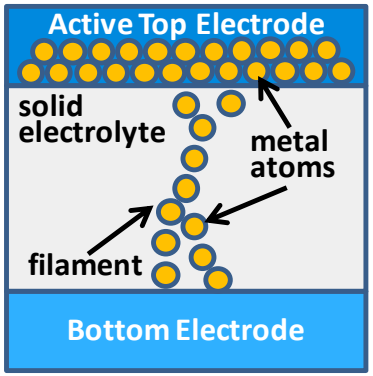
SRAM (on-chip)+
DRAM (off-chip)

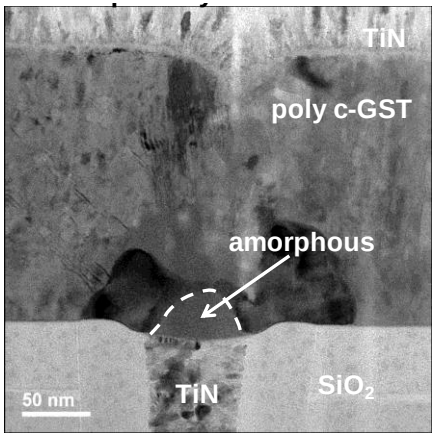
SpiNNaker (U of Manchester)

Non-Volatile Memory (NVM)

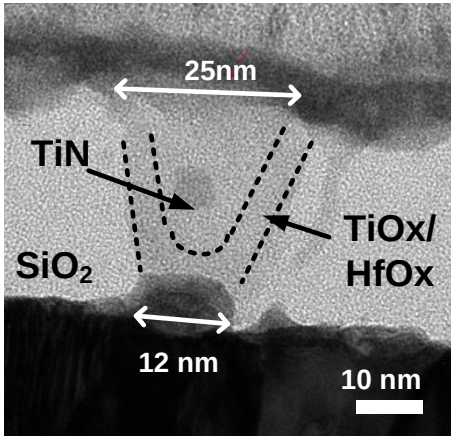




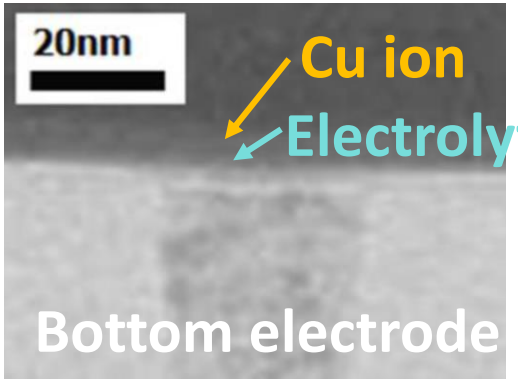




Phase change memory (PCM)



Metal oxide resistive switching memory (RRAM)

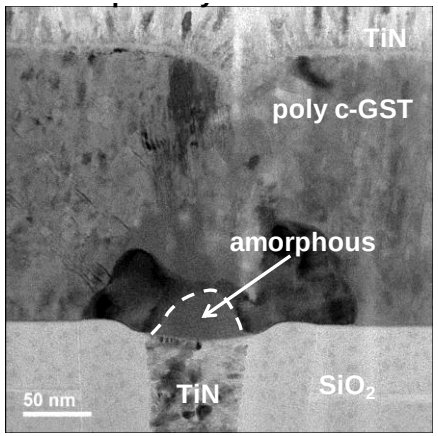


Conductive bridge memory (CBRAM)

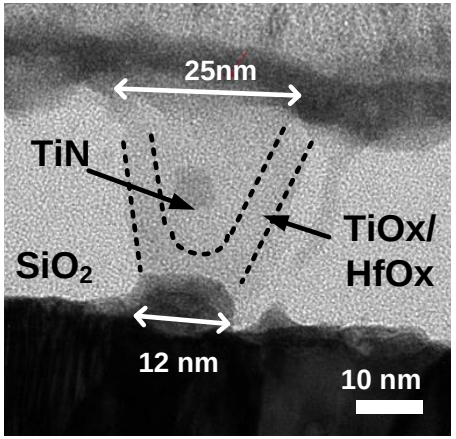
D. Kuzum et al., *Nano Lett*. 2013, Y. Wu et al., *IEDM* 2013; A. Calderoni et al., *IMW* 2014

Non-Volatile Memory (NVM) → Synapse

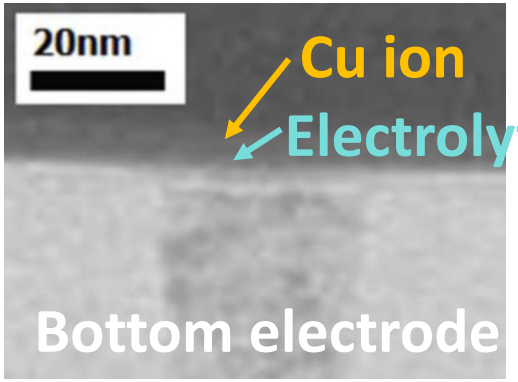
- Analog programmable
- Scalable to a few nm
- Stack in 3D



Phase change memory (PCM)



Metal oxide resistive switching memory (RRAM)



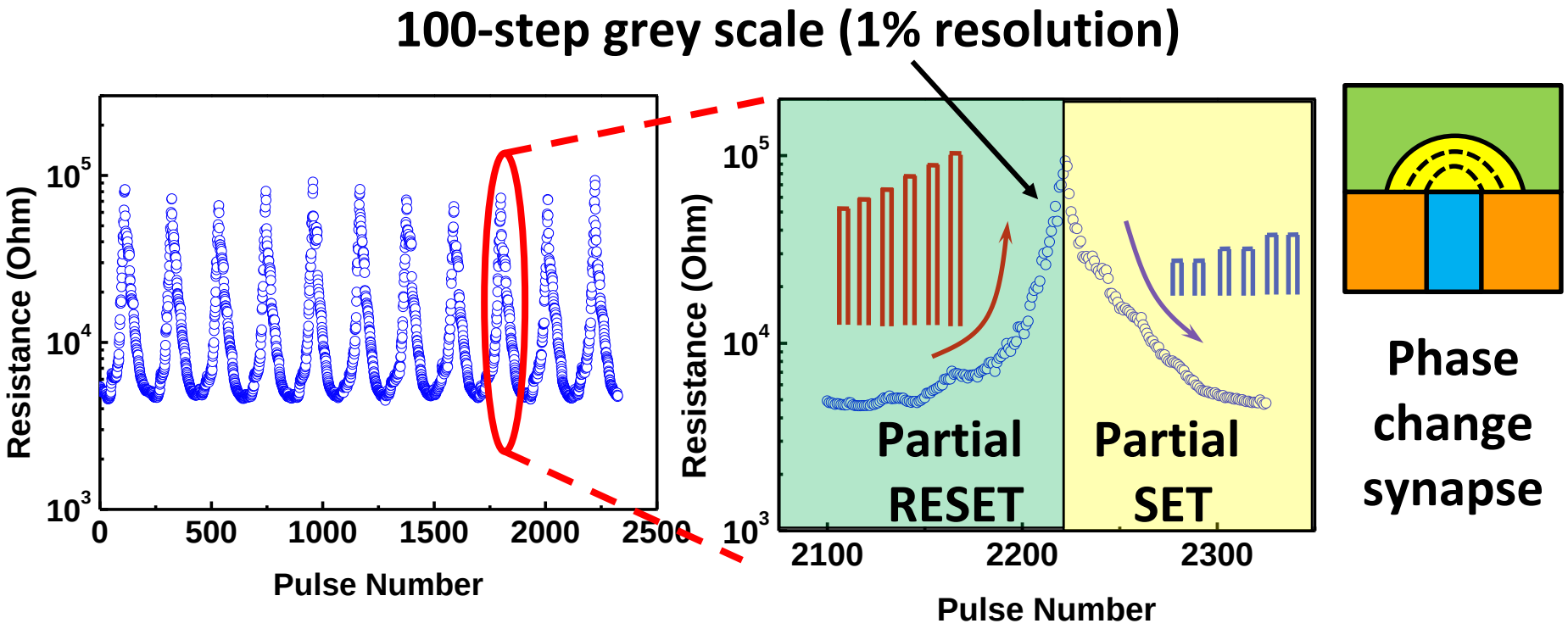
Conductive bridge memory (CBRAM)

D. Kuzum et al., *Nano Lett*. 2013, Y. Wu et al., *IEDM* 2013; A. Calderoni et al., *IMW* 2014

Nanoscale Memory as Synaptic Weights

Synaptic updates in the brain: basis for learning

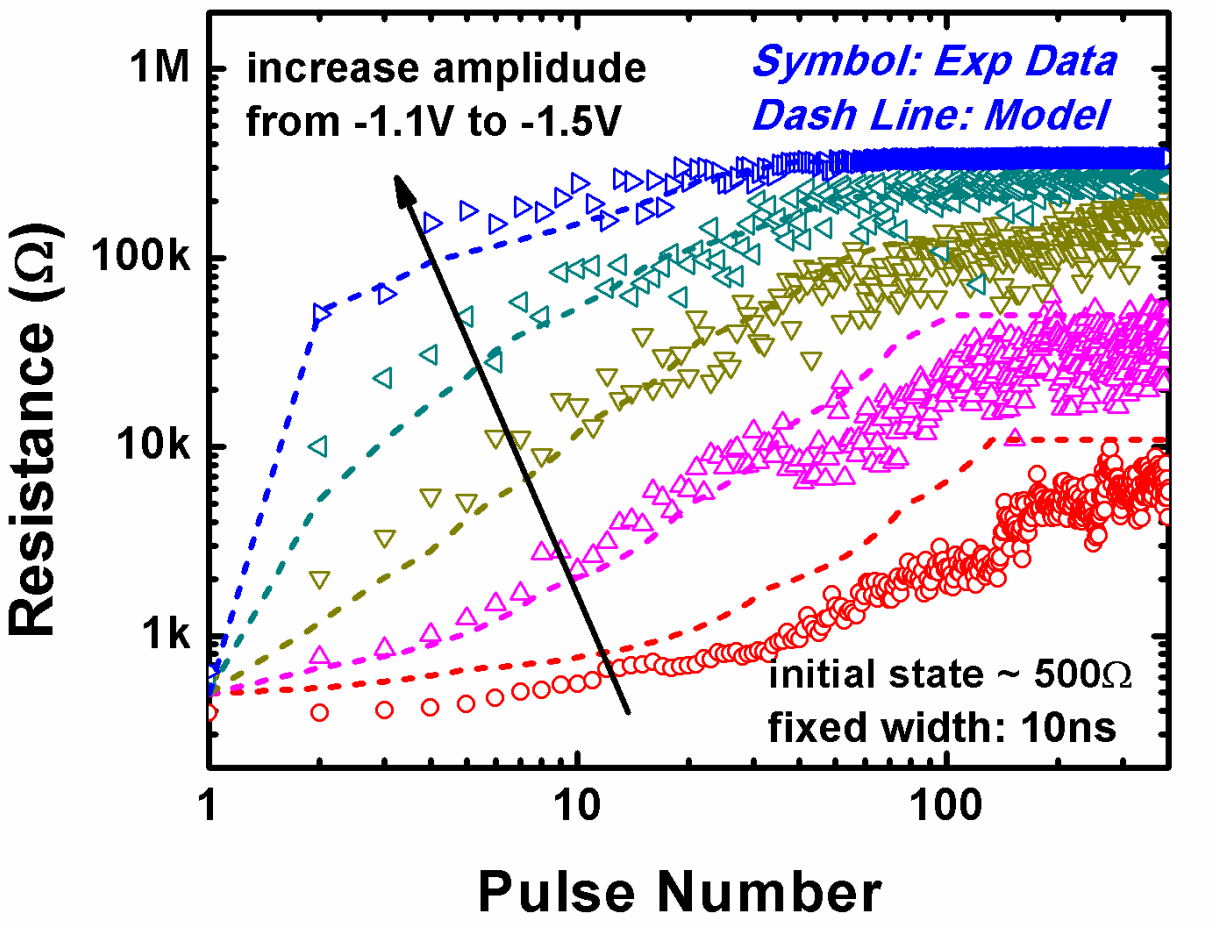
Requirement: analog resistance change



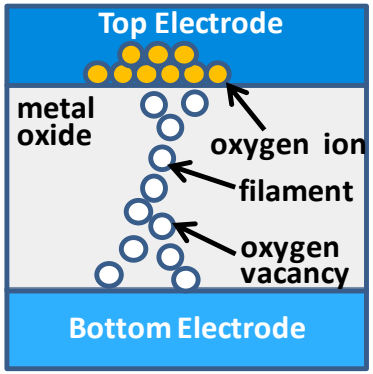
D. Kuzum *et al.*, *Nano Lett.*, p. 2179 (2012)

Gradual Resistance Change

Larger pulse amplitude → Fewer states



RRAM



S. Yu *et al*. *Adv. Mater*. vol. 25, pp. 1774-1779, 2013

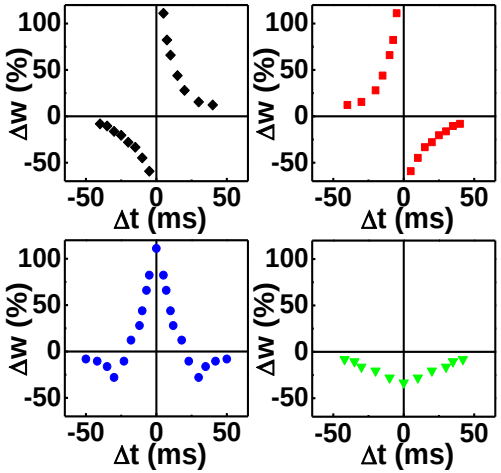
Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)

D. Kuzum *et al.*, *Nano Lett.,* p. 2179 (2012)

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)

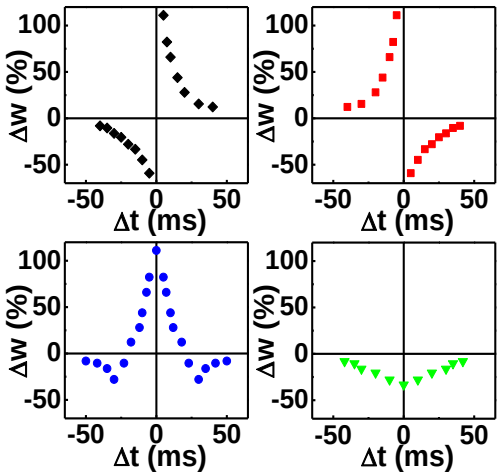


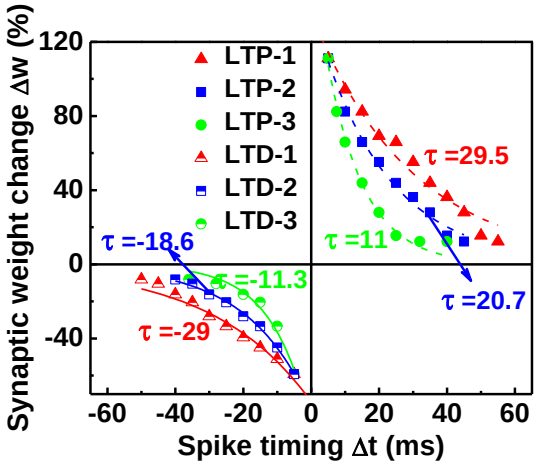
Various STDP kernels

D. Kuzum *et al.*, *Nano Lett.*, p. 2179 (2012)

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)





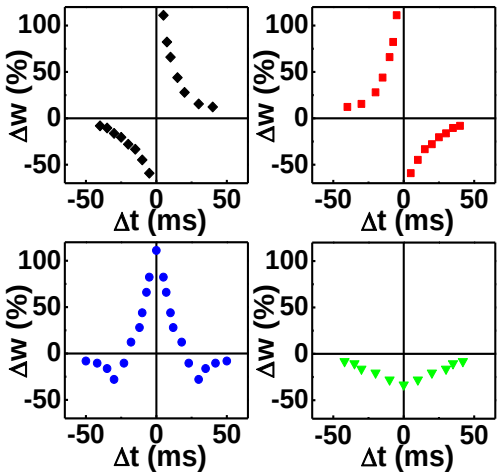
Various STDP kernels

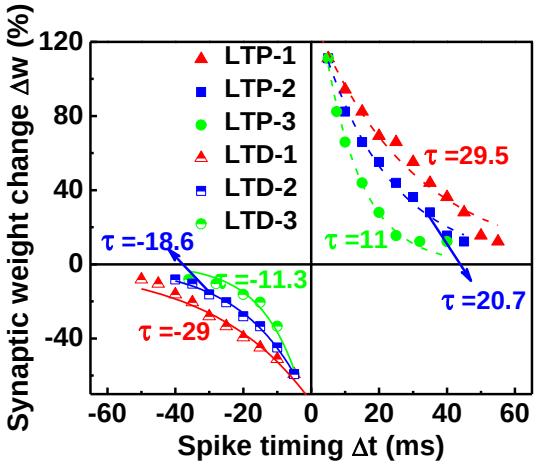
Various time constants

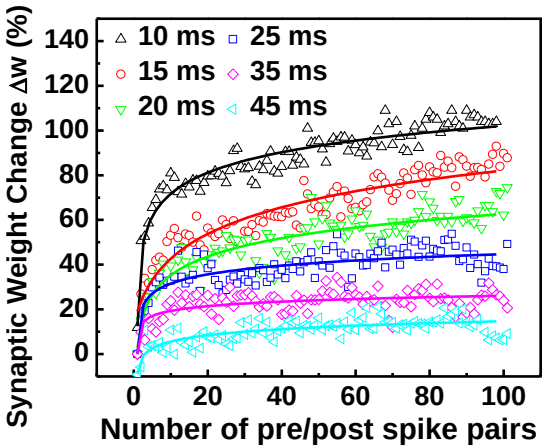
D. Kuzum *et al.*, *Nano Lett.*, p. 2179 (2012)

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)







Various STDP kernels

Various time constants

Weight update saturation

D. Kuzum *et al.*, *Nano Lett.*, p. 2179 (2012)

Stochastic Weight Update

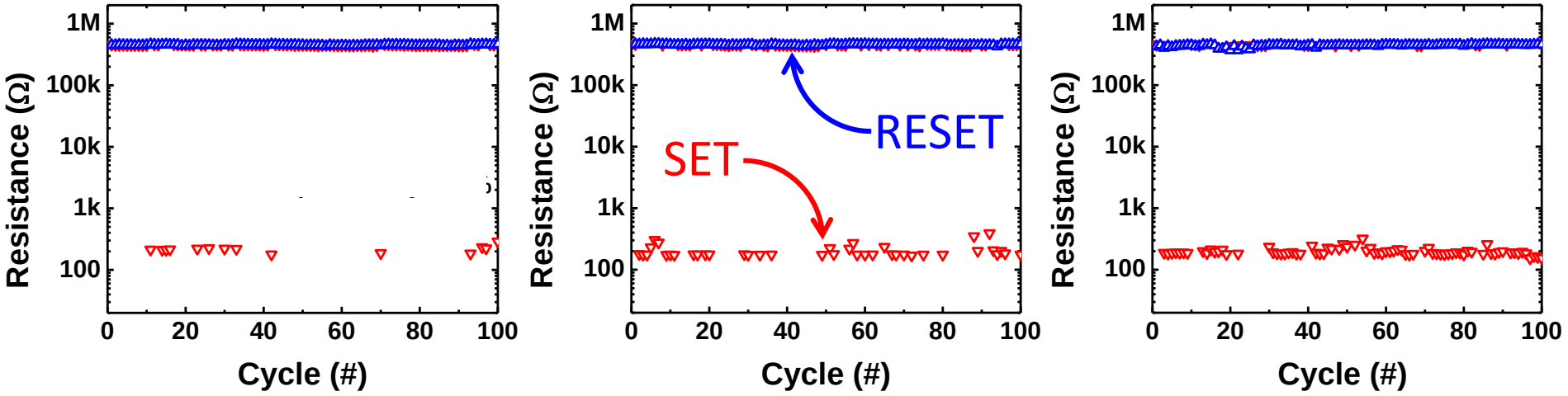
- Program memory close to switching threshold

SET success probability:

14% @ V_{SET} = 1.3 V

40% @ V_{SET} = 1.6 V

76% @ V_{SET} = 1.9 V



S. Yu *et al.*, *Frontiers of Neuroscience*, 2013

Stochastic Weight Update

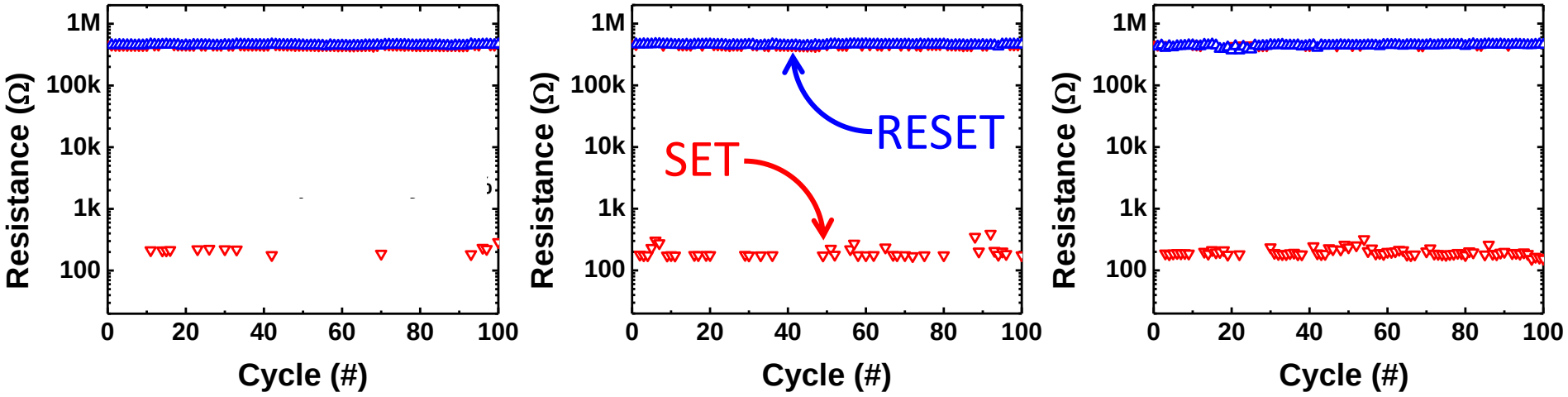
- Program memory close to switching threshold
- Emulate grey scale

SET success probability:

14% @ V_{SET} = 1.3 V

40% @ V_{SET} = 1.6 V

76% @ V_{SET} = 1.9 V



S. Yu *et al.*, *Frontiers of Neuroscience*, 2013

Stochastic Weight Update

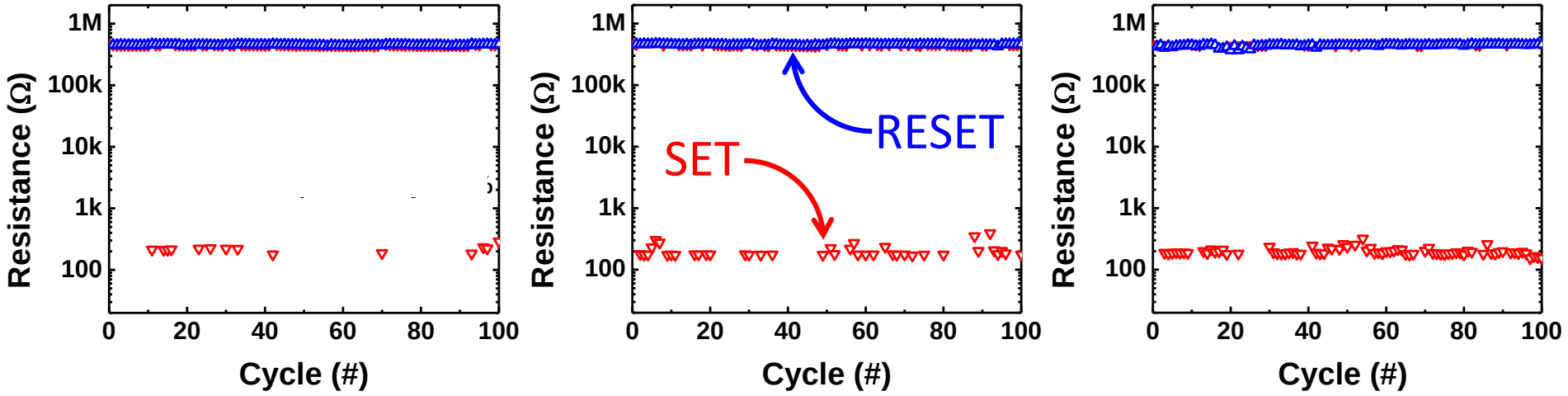
- Program memory close to switching threshold
- Emulate grey scale
- Escape local minima in gradient descent

SET success probability:

14% @ V_{SET} = 1.3 V

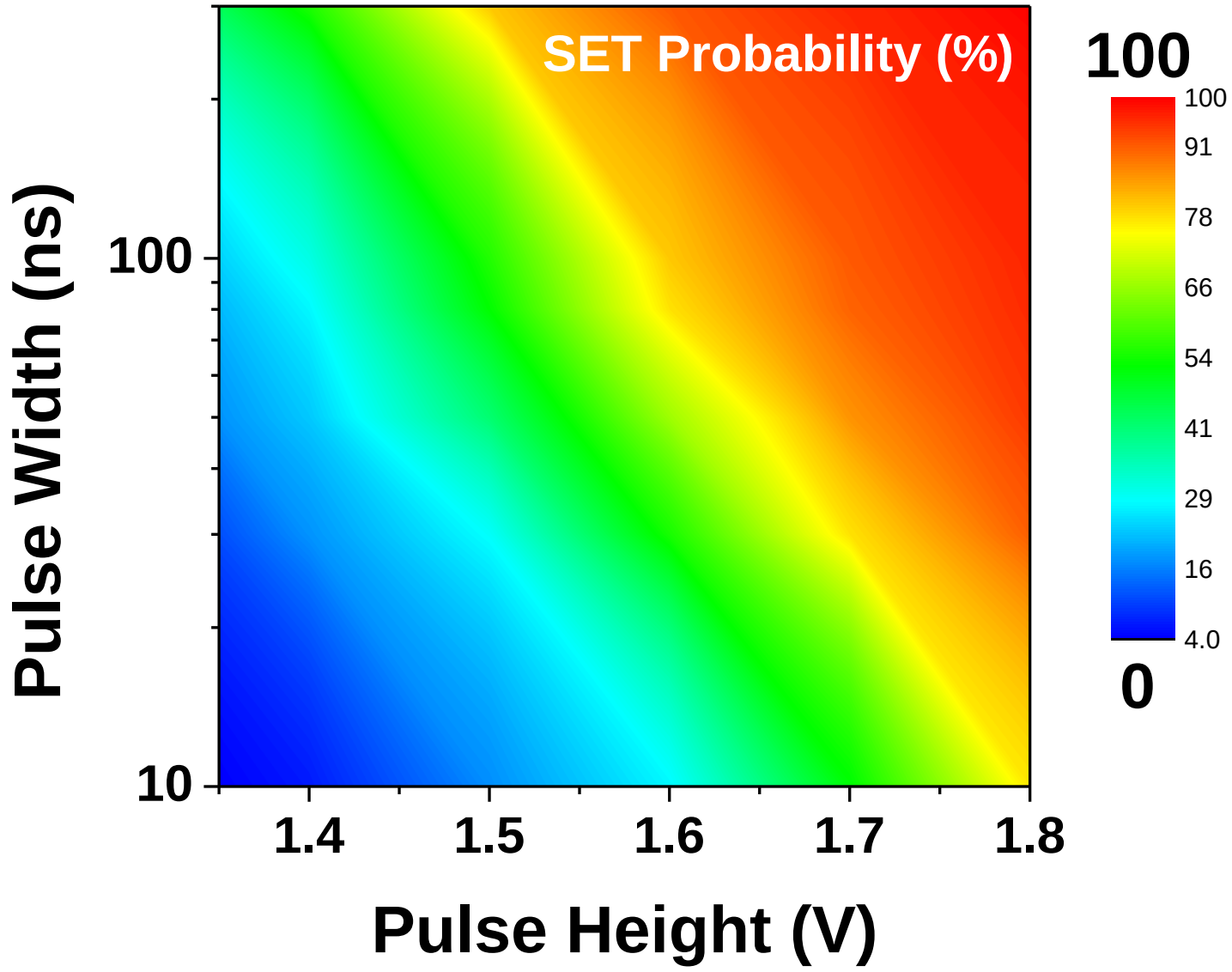
40% @ V_{SET} = 1.6 V

76% @ V_{SET} = 1.9 V



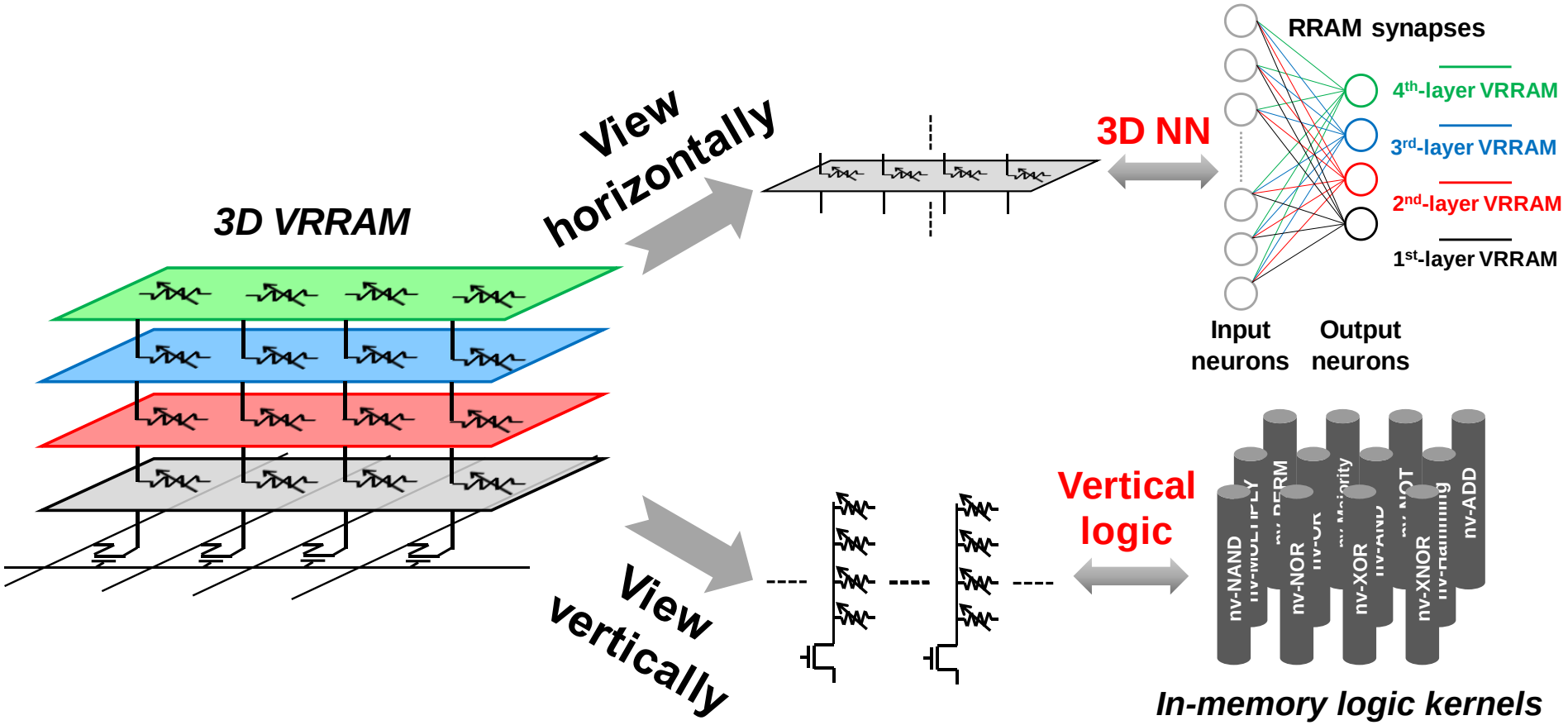
S. Yu *et al.*, *Frontiers of Neuroscience*, 2013

RRAM Stochastic Weight Update



Vertical RRAM In-Memory Computing

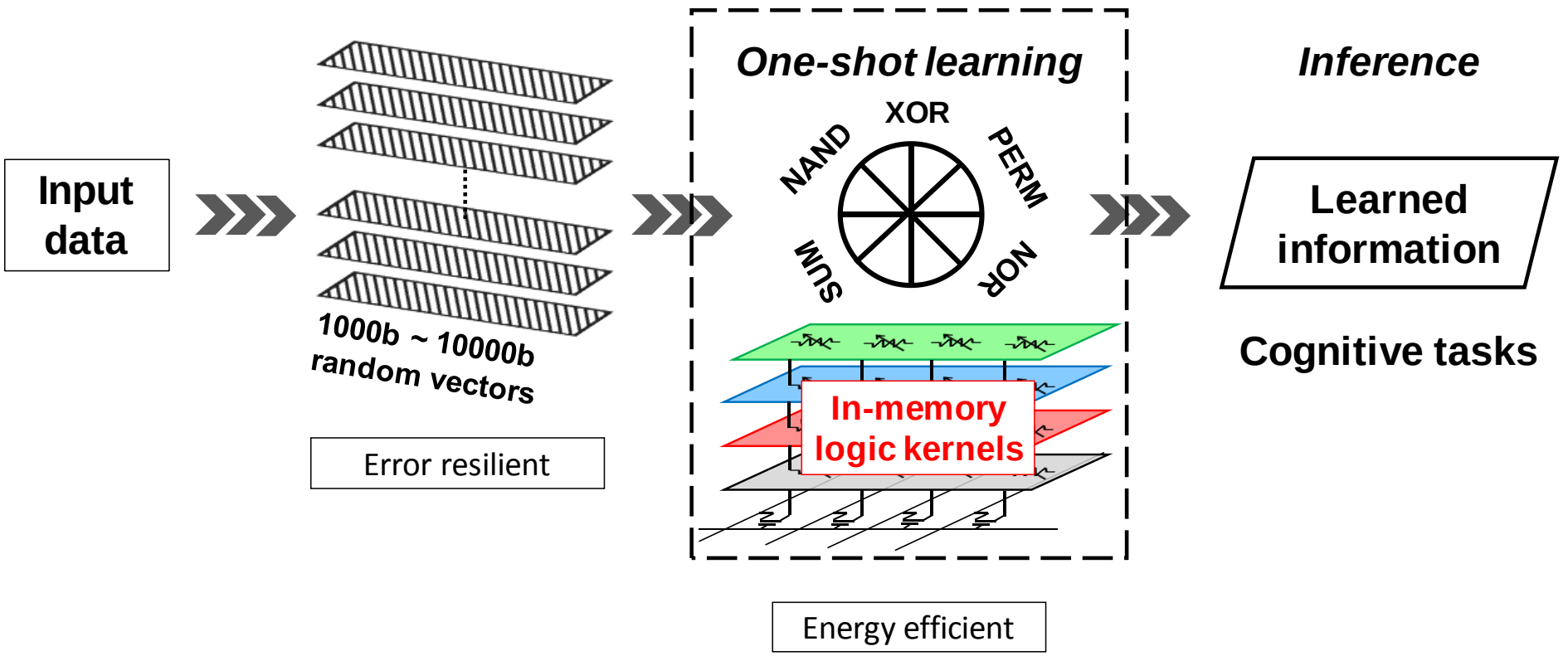
Truly exploit the 3rd dimension



VRRAM: vertical RRAM; NN: neural network

H. Li et al., Symp. VLSI Tech. 2016

In-Memory Logic Kernels for Hyper-Dimensional* Computing

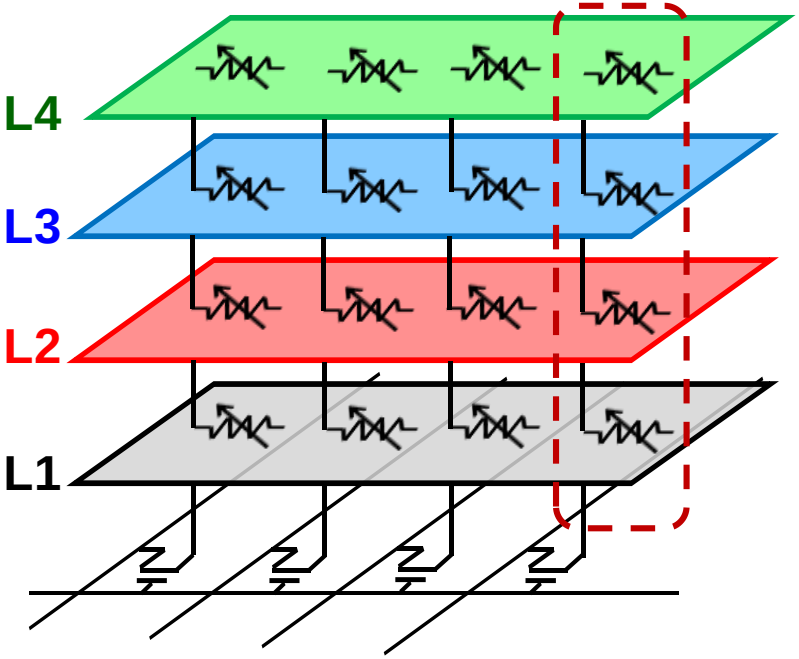


[*P. Kanerva, Cog. Comput.'09]

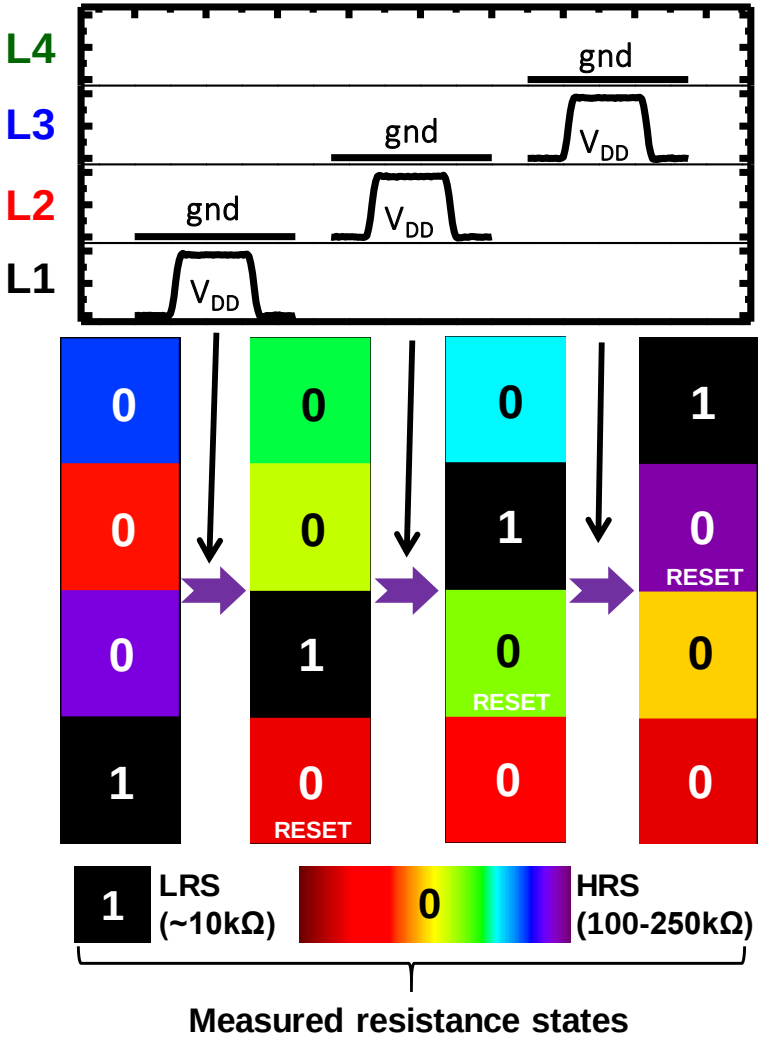
H. Li et al., Symp. VLSI Tech. 2016

In-Memory Logic Kernel: Bit Permutation

Bit permutation is simple and efficient



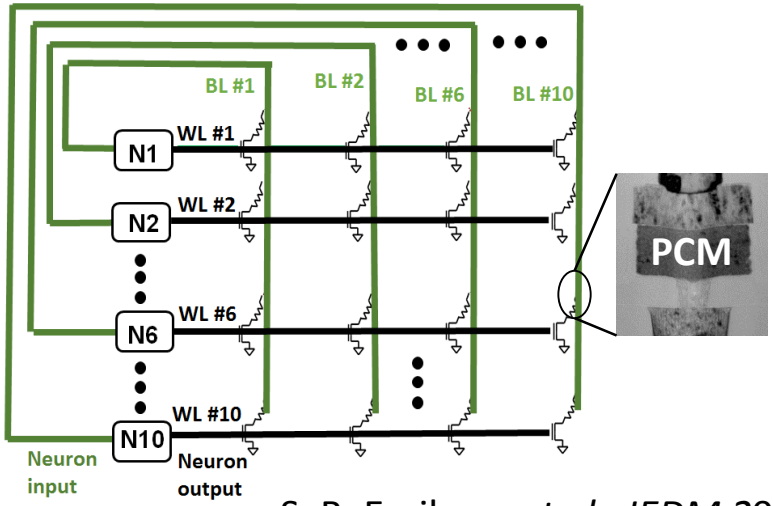
Bit permutation
along vertical pillars



H. Li et al., Symp. VLSI Tech. 2016

Array Level Experimental Demonstrations

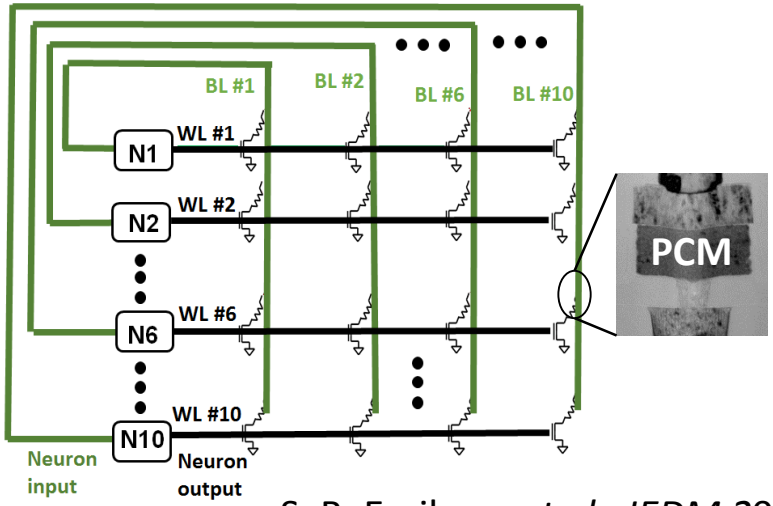
100 PCM synapses:
Biological Hebbian learning



S. B. Eryilmaz *et al.*, *IEDM* 2013

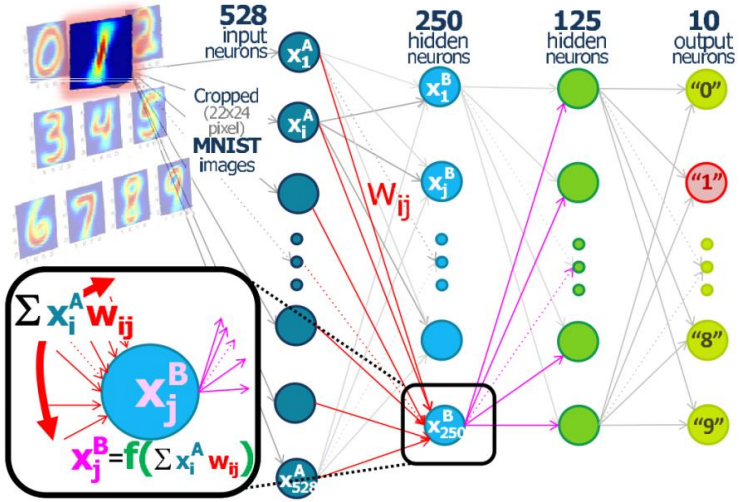
Array Level Experimental Demonstrations

100 PCM synapses:
Biological Hebbian learning



S. B. Eryilmaz *et al.*, *IEDM* 2013

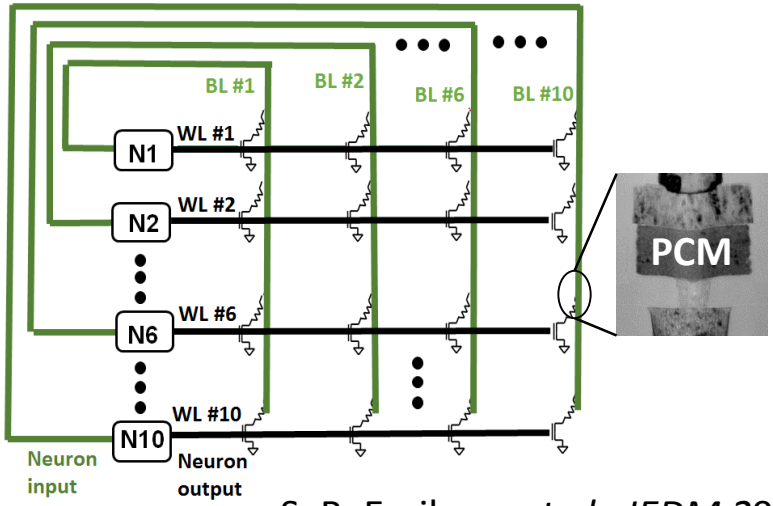
165,000 PCM synapses:
Gradient based backpropagation



G. Burr *et al.*, *IEDM* 2014

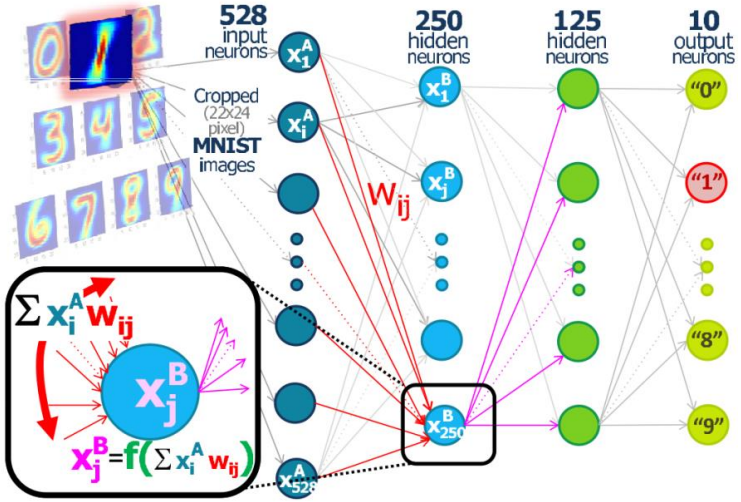
Array Level Experimental Demonstrations

100 PCM synapses:
Biological Hebbian learning



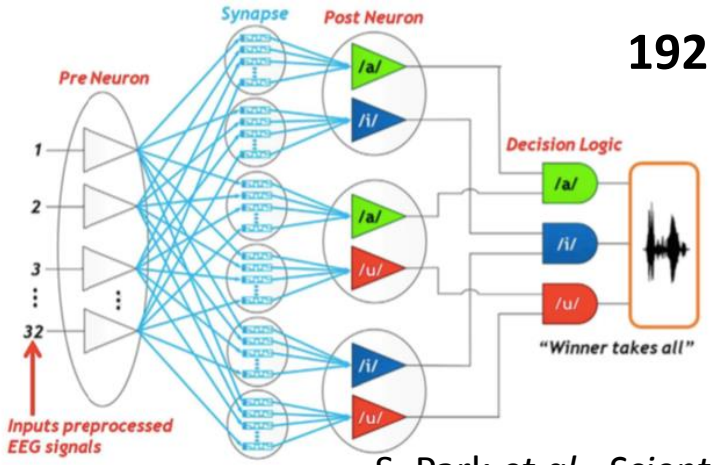
S. B. Eryilmaz *et al.*, *IEDM* 2013

165,000 PCM synapses:
Gradient based backpropagation

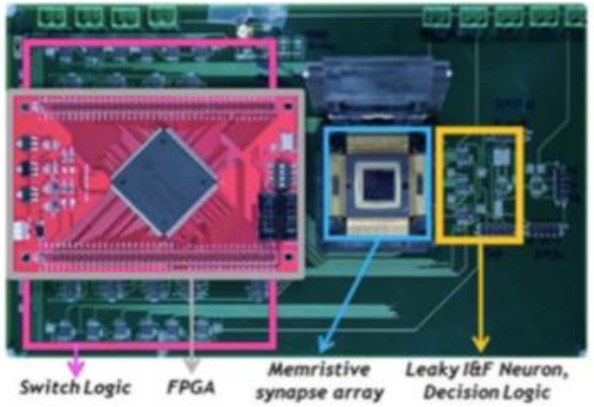


G. Burr *et al.*, *IEDM* 2014

192 synapses with RRAMs: Hebbian learning

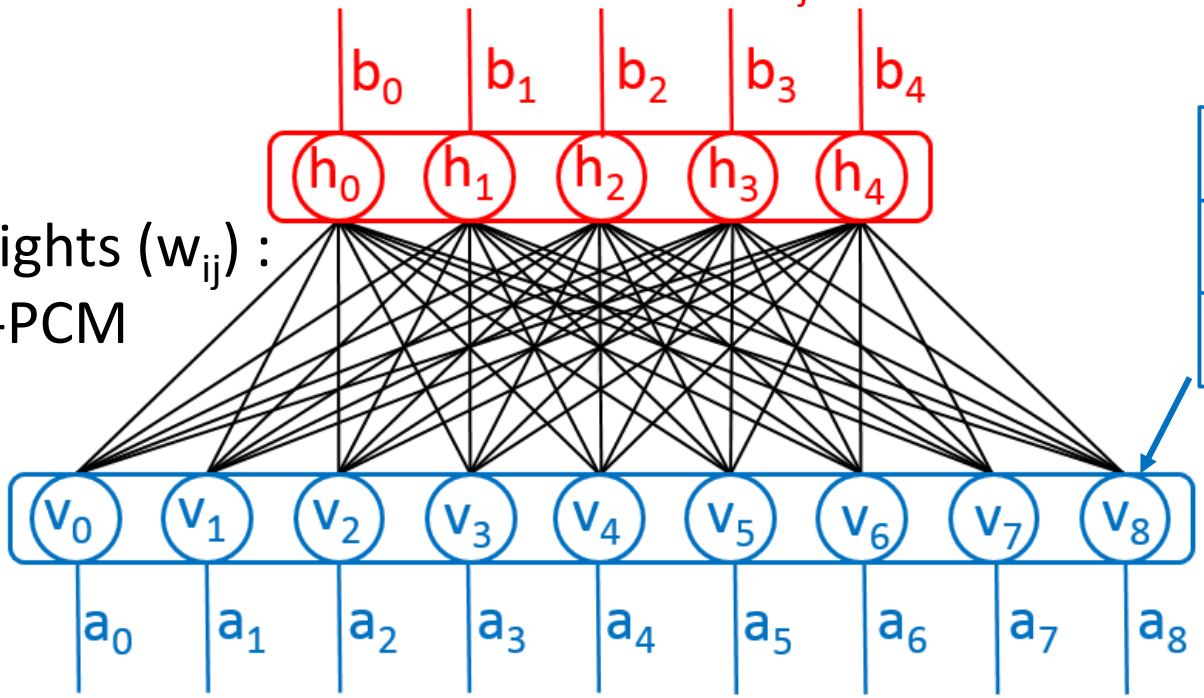


S. Park *et al.*, *Scientific Reports*, 2015



Restricted Boltzmann Machine with Resistive Phase Change Memory Synapses

- Hidden nodes = 0 (b_j, j=0:4)



Input pixels

v_0	v_1	v_2
v_3	v_4	v_5
v_6	v_7	v_8

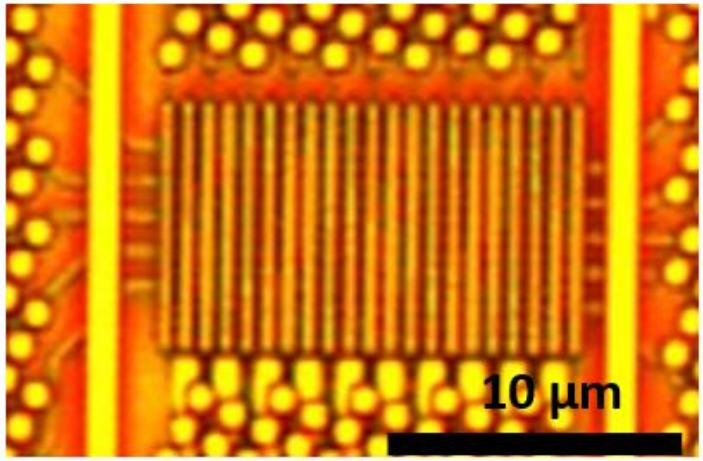
- Pairwise weights (w_{ij}) : train with 2-PCM synapses

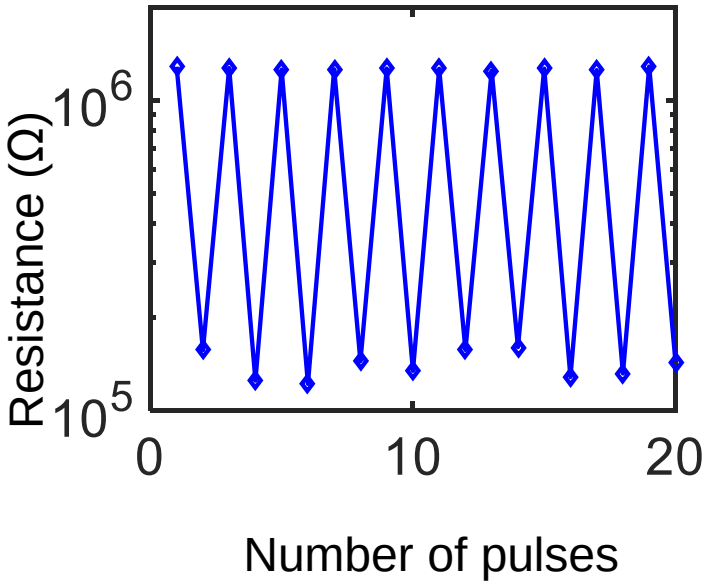
- Visible nodes= 0 (a_i, i=0:8)

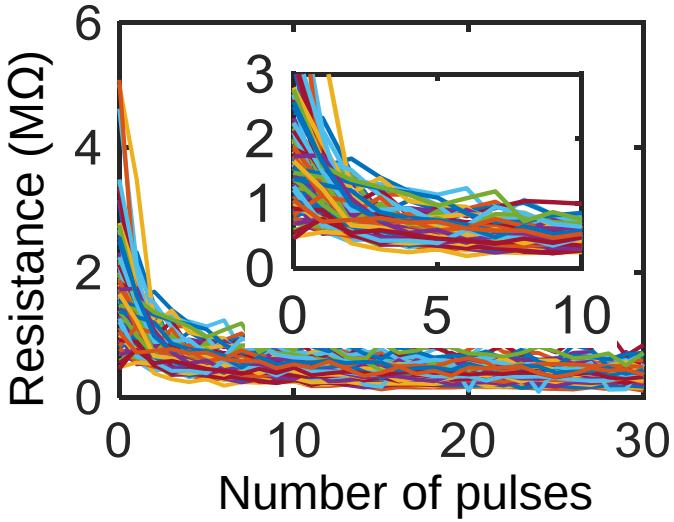
$$\Delta w \propto \langle vh \rangle_{data} - \langle v^k h^k \rangle_{reconstruction}$$

S. B. Eryilmaz et al., submitted 2016

Contrastive Divergence with Resistive Phase Change Memory Synapses

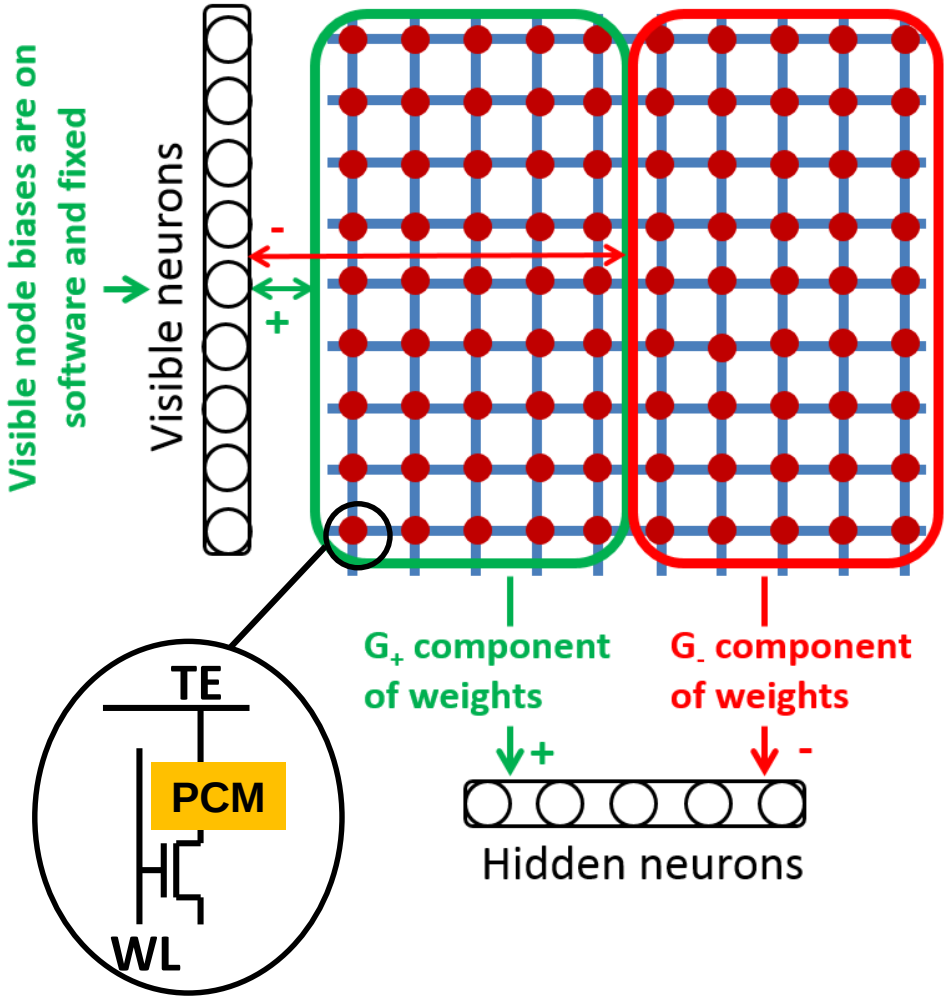


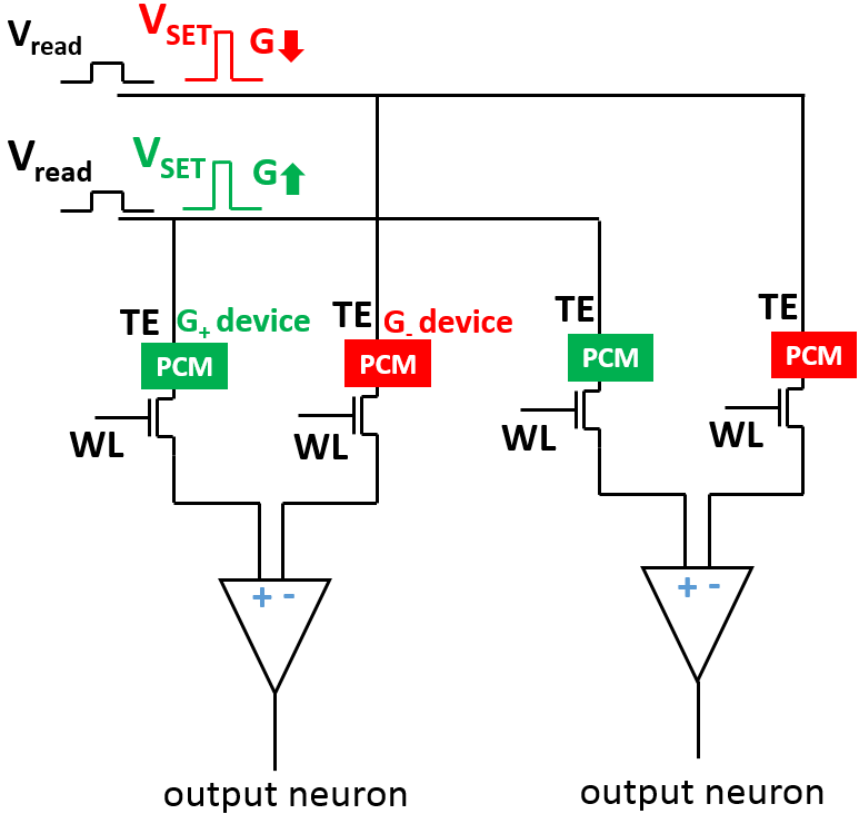




S. B. Eryilmaz et al., submitted 2016

Mapping Contrastive Divergence onto Phase Change Memory Array



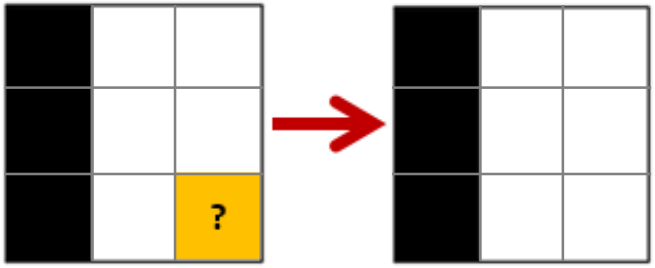


S. B. Eryilmaz et al., submitted 2016

Inference After Training

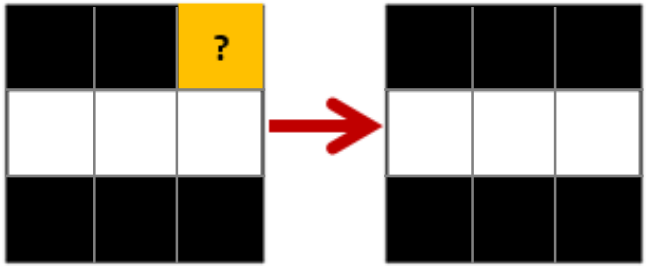
Initial: P(white)=0.07

After training: P(white)=0.87



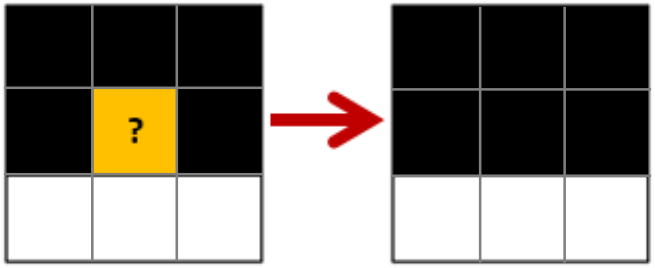
P(black)=0.40

P(black)=0.54

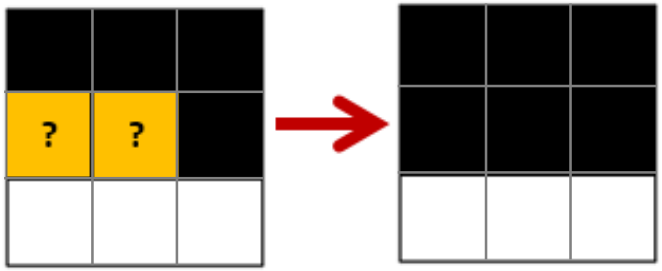


Initial: P(black)=0.42

After training: P(white)=0.67

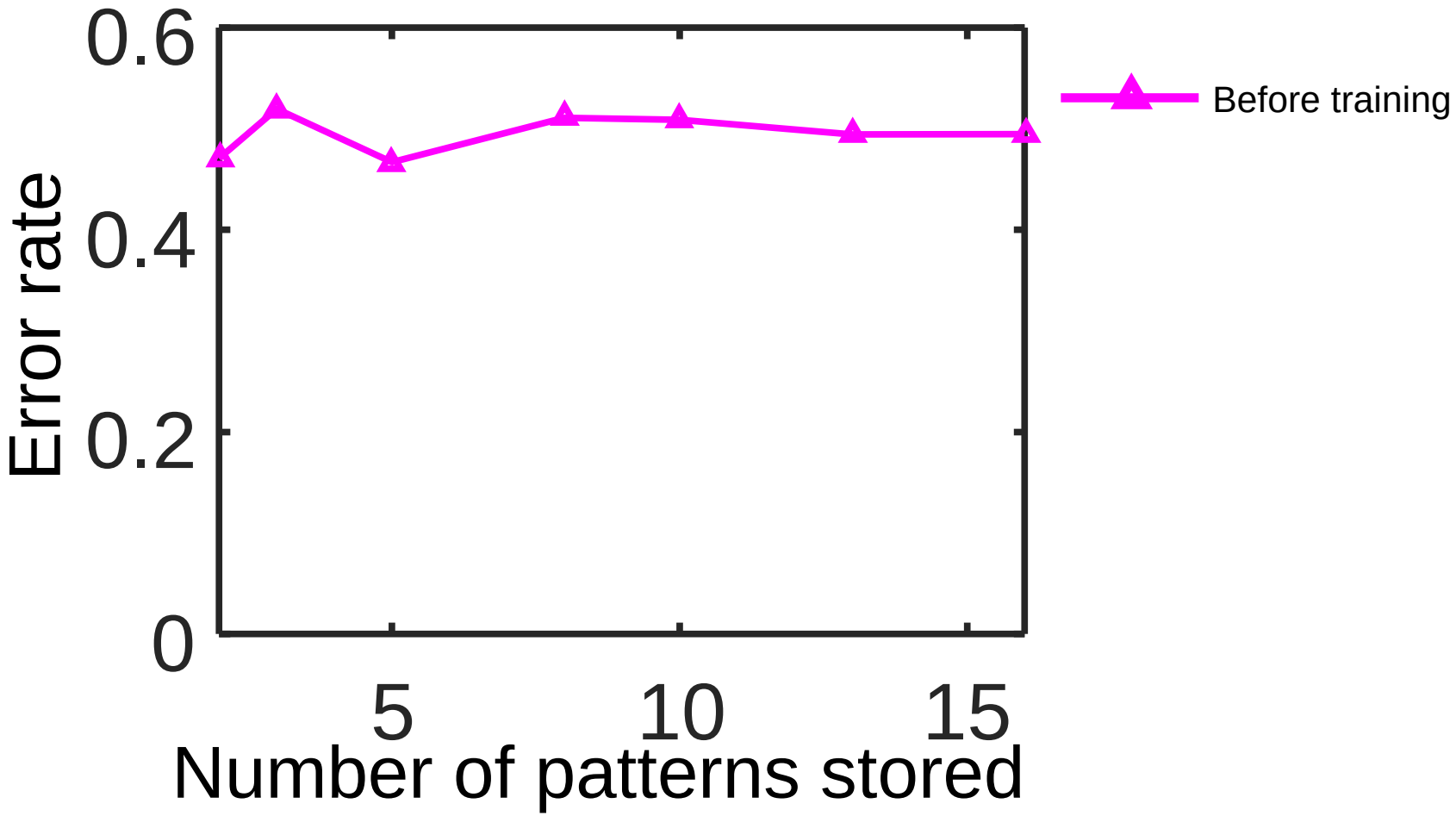


P(white,white) = 0.11 P(black,white) = 0.46
P(white,black) = 0.09 P(black,black) = 0.34
P(white,white) = 0.05 P(black,white) = 0.30
P(white,black) = 0.09 P(black,black) = 0.56



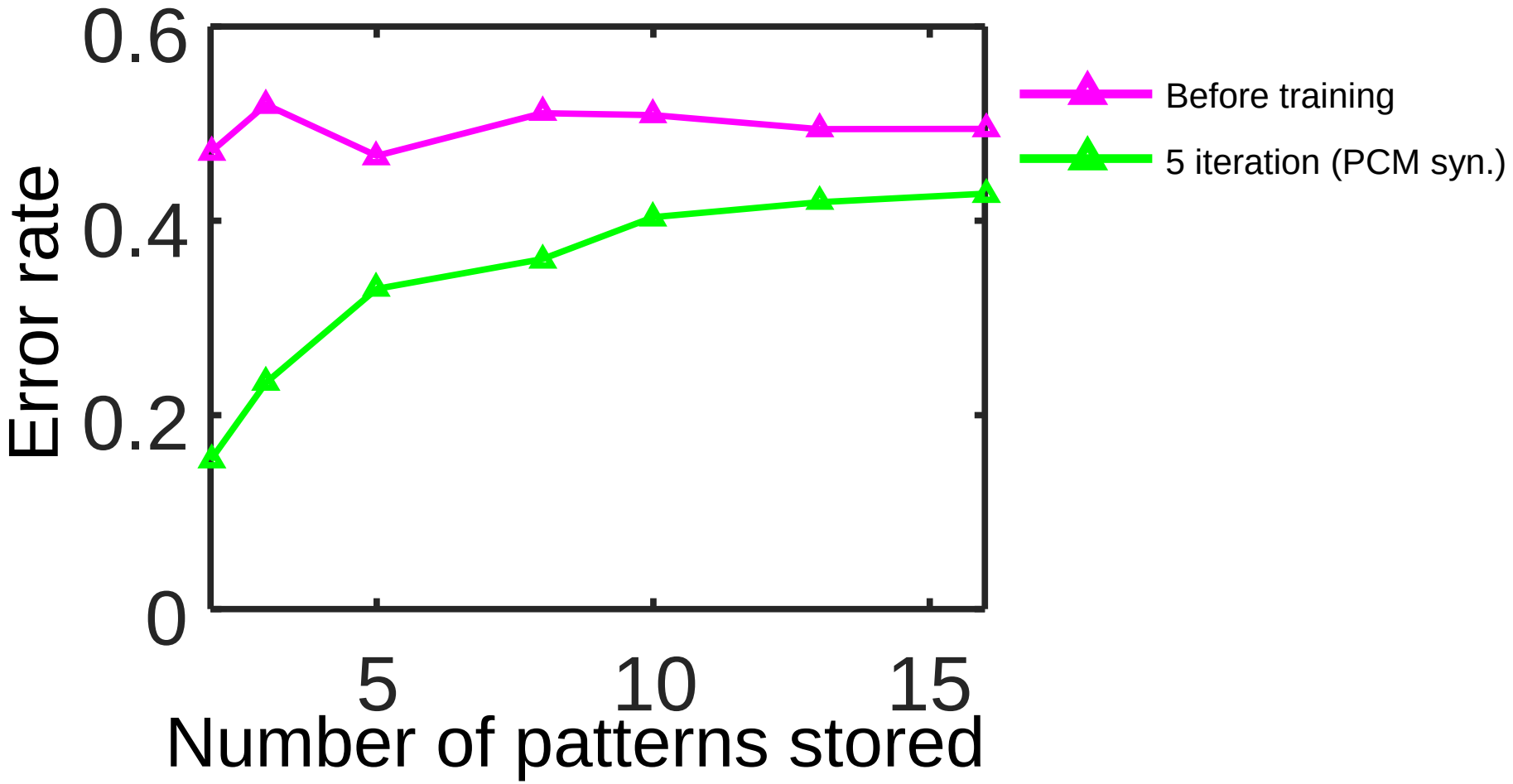
S. B. Eryilmaz et al., submitted 2016

Error Rate vs Number of Patterns Stored



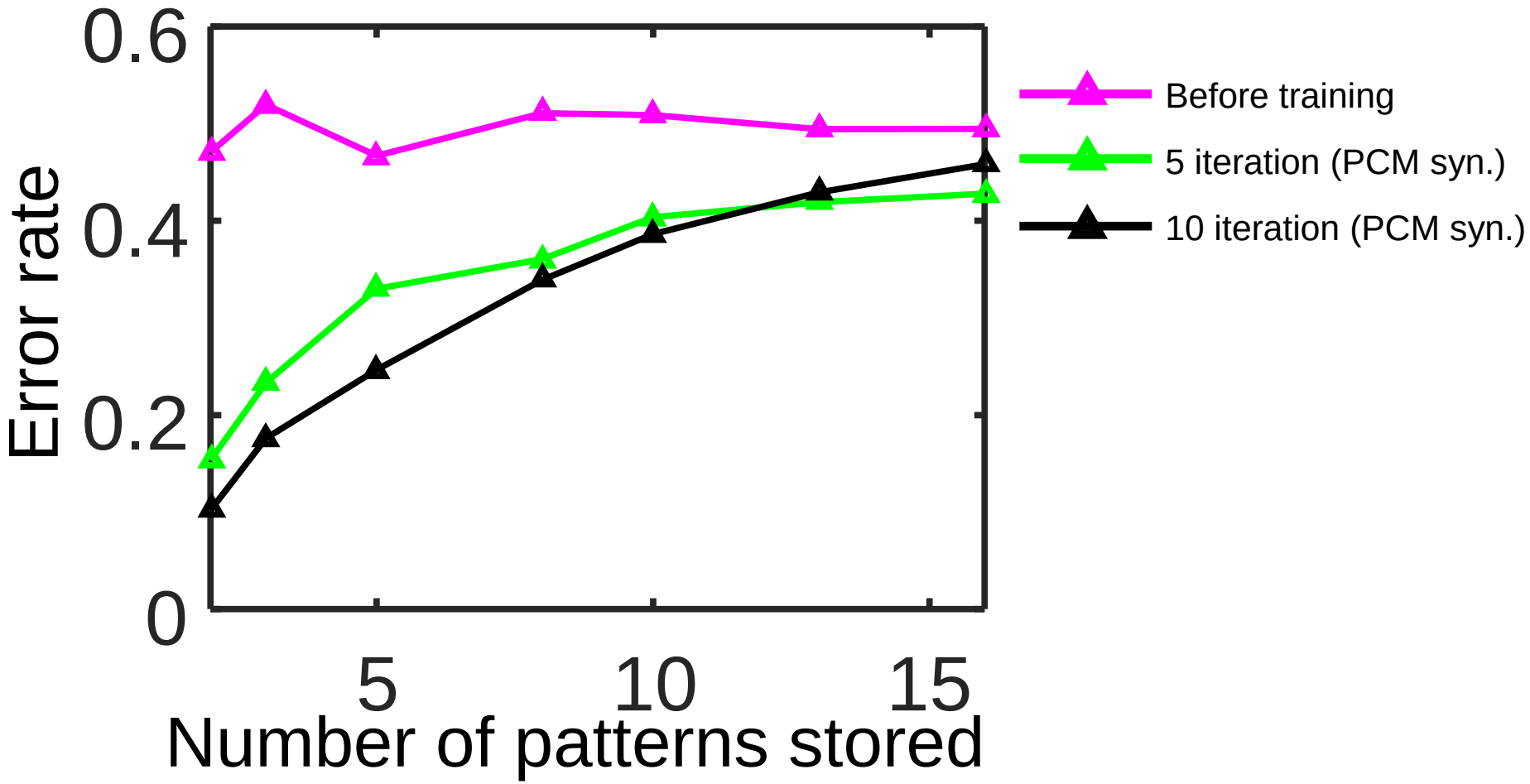
S. B. Eryilmaz et al., submitted 2016

Error Reduces After Training



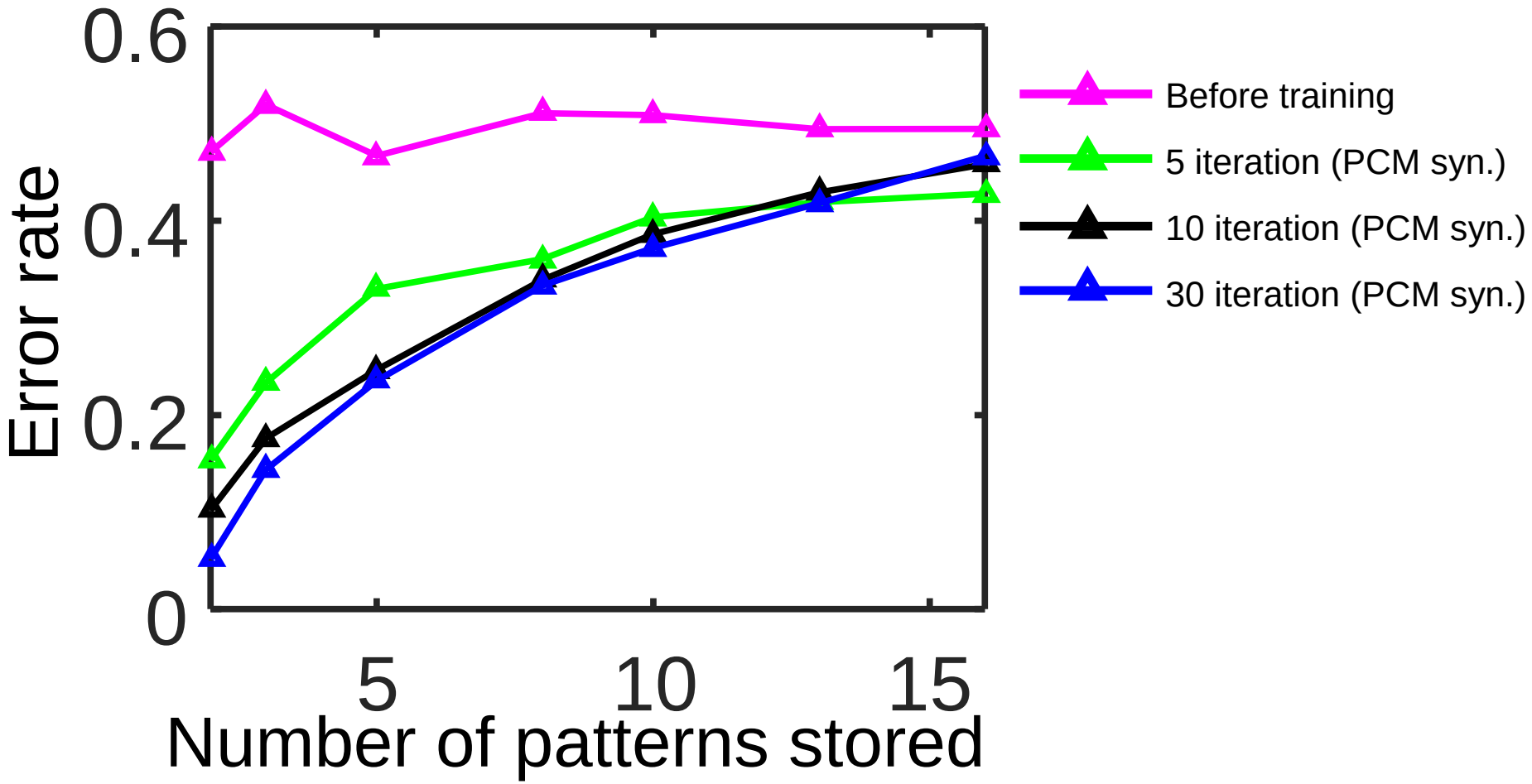
S. B. Eryilmaz et al., submitted 2016

Error Reduces with Further Training



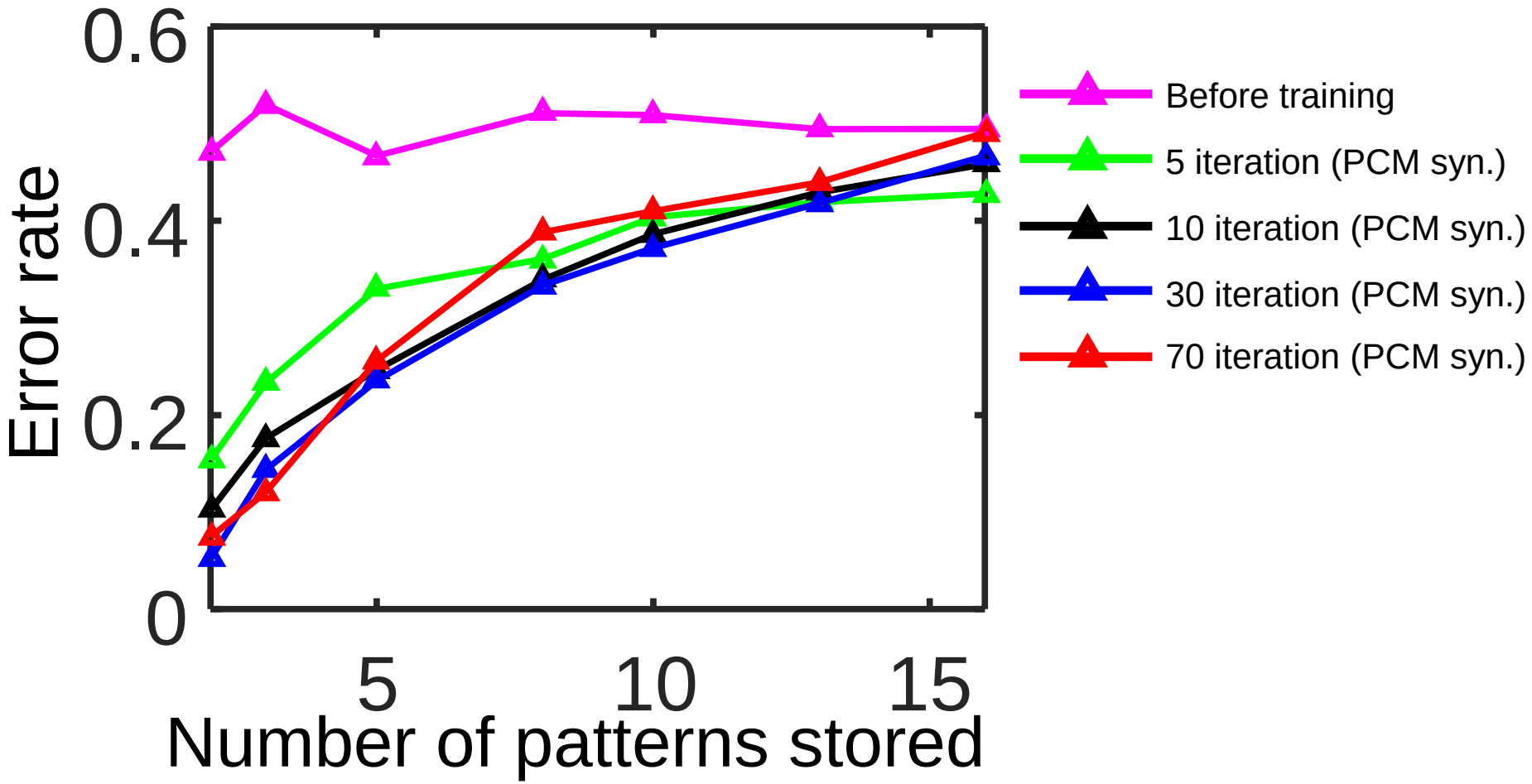
S. B. Eryilmaz et al., submitted 2016

Error Reduction Saturates with Training



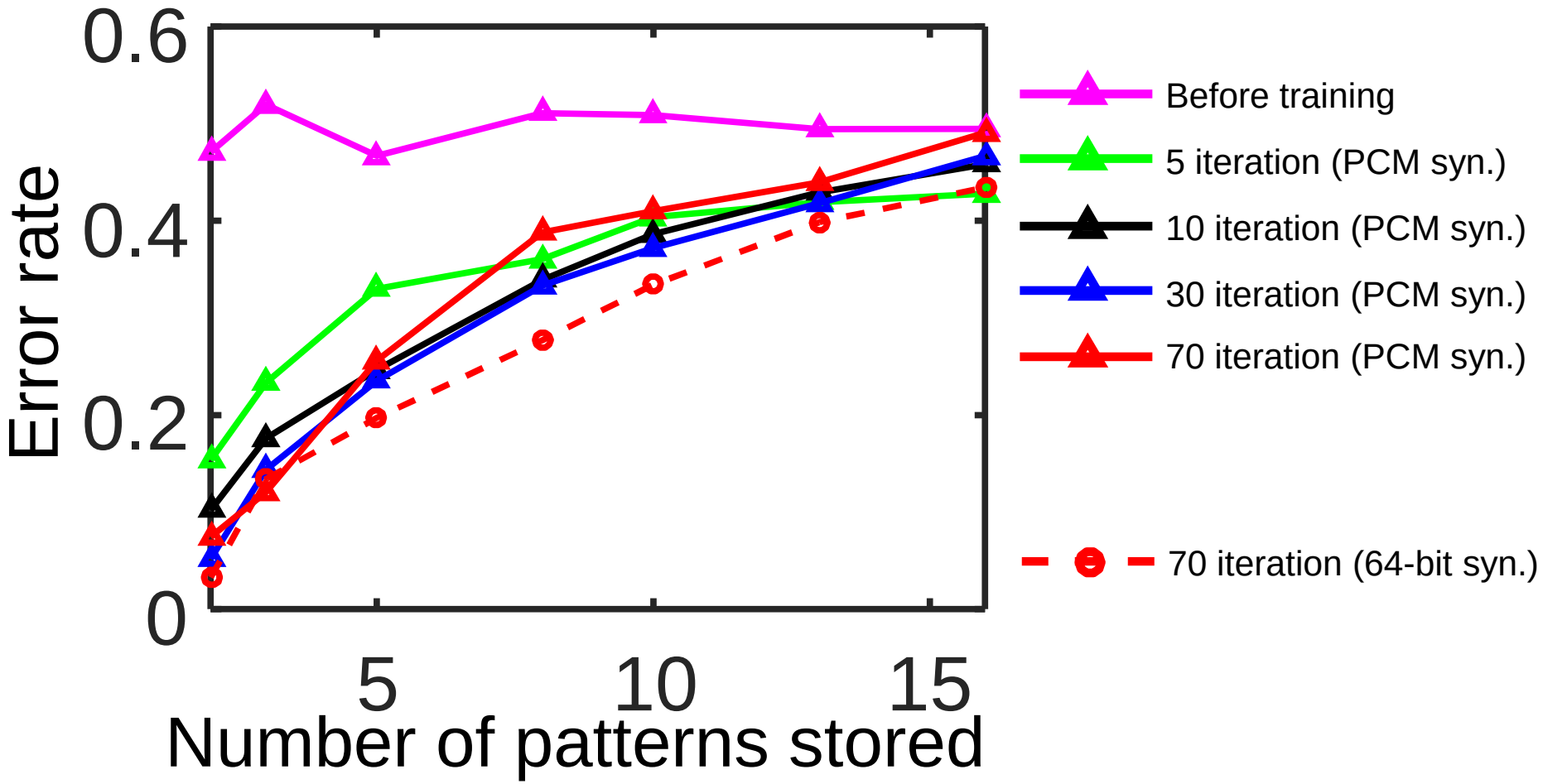
S. B. Eryilmaz et al., submitted 2016

Saturation: PCM Synapse Starts to Unlearn



S. B. Eryilmaz et al., submitted 2016

“Precise” Synapse: Error Continues to Reduce



S. B. Eryilmaz et al., submitted 2016

Energy Consumption per Epoch (Synapses only)

PCM hardware	Conventional hardware*
6.1 nJ	910 nJ (430 nJ logic, 480 nJ memory)

*Energy estimate of Intel Xeon Phi coprocessor (22 nm)

S. B. Eryilmaz et al., submitted 2016

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)
6. Stochastic learning behavior → statistical learning rules

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)
6. Stochastic learning behavior → statistical learning rules
7. Meta-plasticity (internal state variables)

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)
6. Stochastic learning behavior → statistical learning rules
7. Meta-plasticity (internal state variables)
8. Timing as an internal variable

Open Research Questions

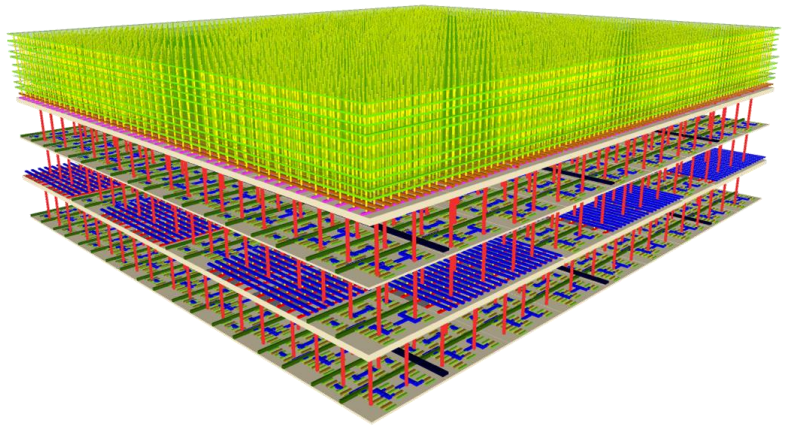
1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)
6. Stochastic learning behavior → statistical learning rules
7. Meta-plasticity (internal state variables)
8. Timing as an internal variable
9. Learning rules: biological? AI?

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)
6. Stochastic learning behavior → statistical learning rules
7. Meta-plasticity (internal state variables)
8. Timing as an internal variable
9. Learning rules: biological? AI?
10. Algorithm-device co-design

Open Research Questions

1. Functionality → performance/Watt, performance/m^2 → variability → reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy ≅ device energy)
6. Stochastic learning behavior → statistical learning rules
7. Meta-plasticity (internal state variables)
8. Timing as an internal variable
9. Learning rules: biological? AI?
10. Algorithm-device co-design
11. Materials/fabrication: monolithic 3D integration **a must**, **MUST** be low temperature



SUPERCELL







Stanford | SystemX Alliance

Focus Area: Computation for Data Analytics

Nano-Engineered Computing Systems Technology

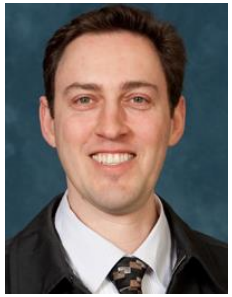
Stanford

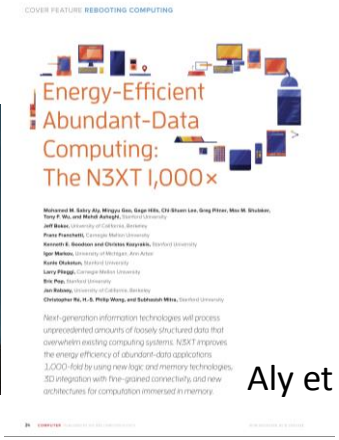
N3XT 0 1 0 1

Berkeley UNIVERSITY OF CALIFORNIA

Carnegie Mellon University

MICHIGAN





Aly et al., *IEEE Computer*, 2015

Collaborators

UCSD

Gert Cauwenberghs
Siddharth Joshi
Emre Neftci
(UC San Diego)

北京大學 PEKING UNIVERSITY

Jinfeng Kang
(Peking U)

IBM

Chung Lam
SangBum Kim
Matt Brightsky

K.S. Lee, J.M. Shieh, W.K. Yen… (NDL, Taiwan)

NDL A Member of NARLabs National Nano Device Laboratories

Students and Post-Docs

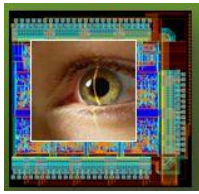
Sponsors

STARnet

SONIC

DARPA

NSF



Visual Cortex on Silicon

http://www.cse.psu.edu/research/visualcortexonsilicon.expedition/

Supported by National Science Foundation Expeditions in Computing Program

Stanford | SystemX Alliance

Stanford | Non-Volatile Memory Technology Research Initiative

Stanford SystemX Alliance

Non-Volatile Memory Technology Research Initiative (NMTRI) @ Stanford University

End of Talk

Questions?