The N3XT Technology for Brain-Inspired Computing

H.-S. Philip Wong
Stanford University

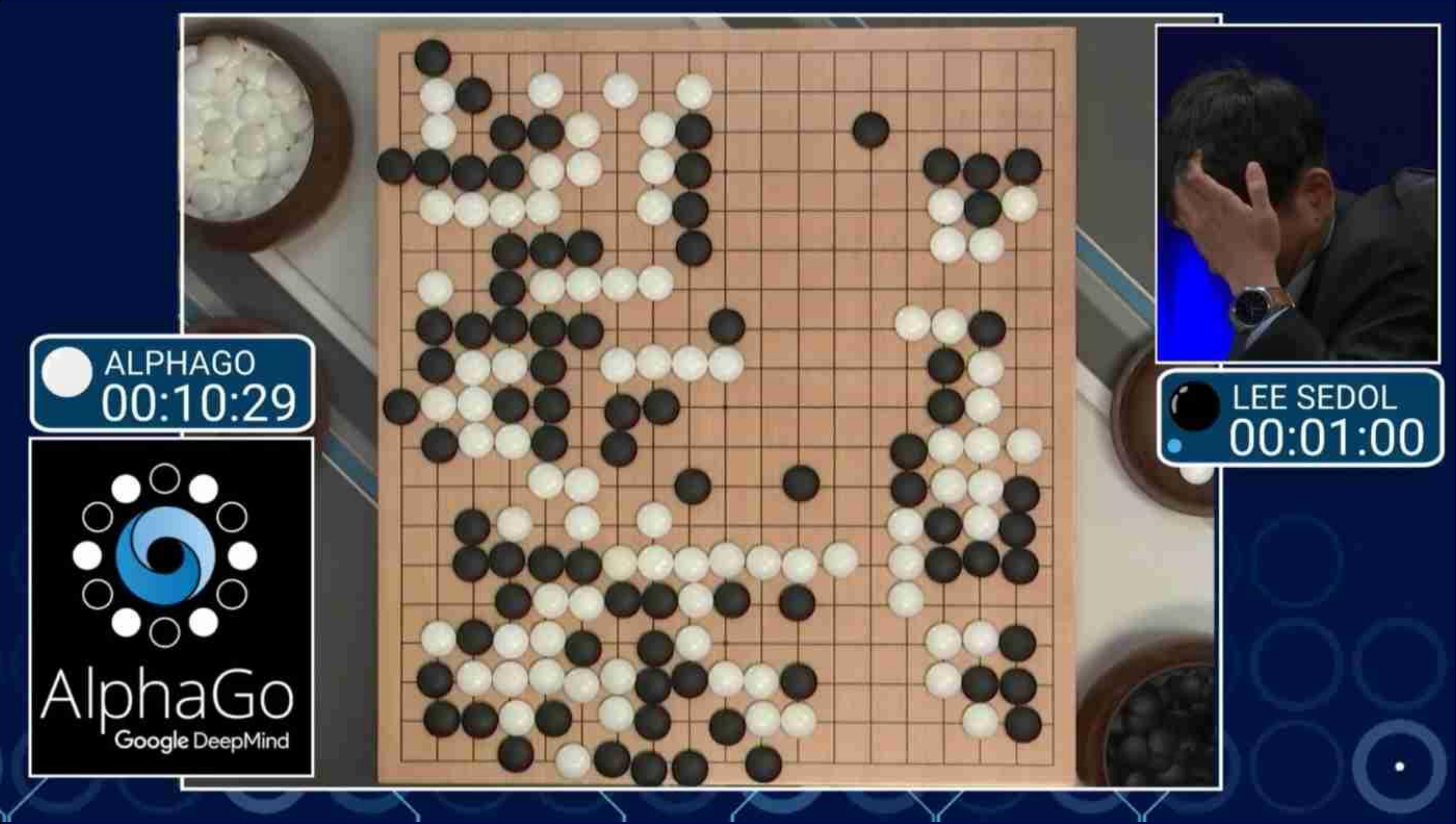
Source: Google

Source: vrworld.com

Source: BDC Magazine

1988 Winter Olympic Games in Calgary, Canada

Inbox
4 Unread Messages
VIP
3 Unread Messages
Search
New Message
4



Source: Gogamguro.com

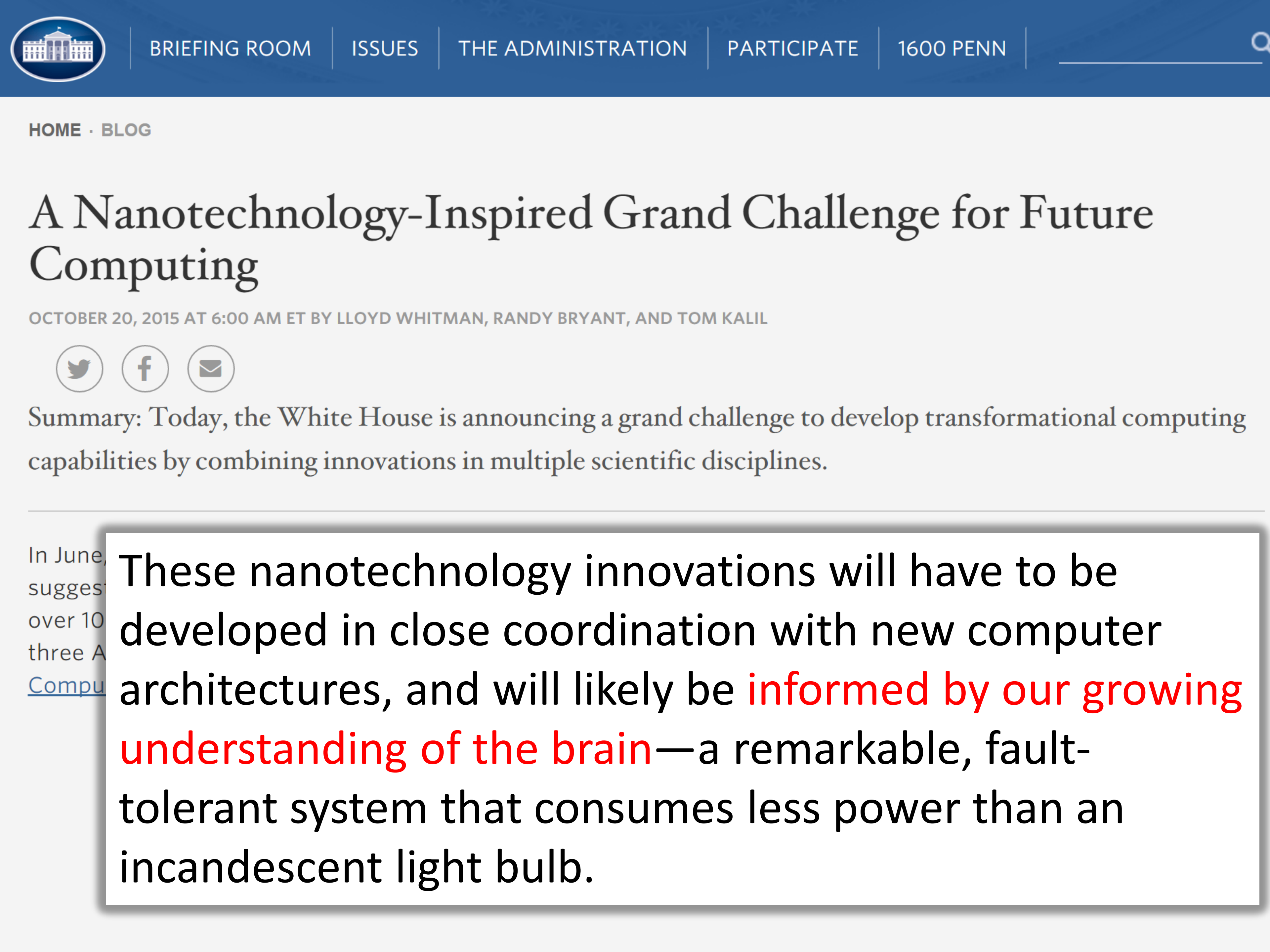
100’s of kW

Source: Google

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



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,
sugges
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.

Approaches of Neuromorphic Hardware

Approaches of Neuromorphic Hardware

Biology-based models / algorithms

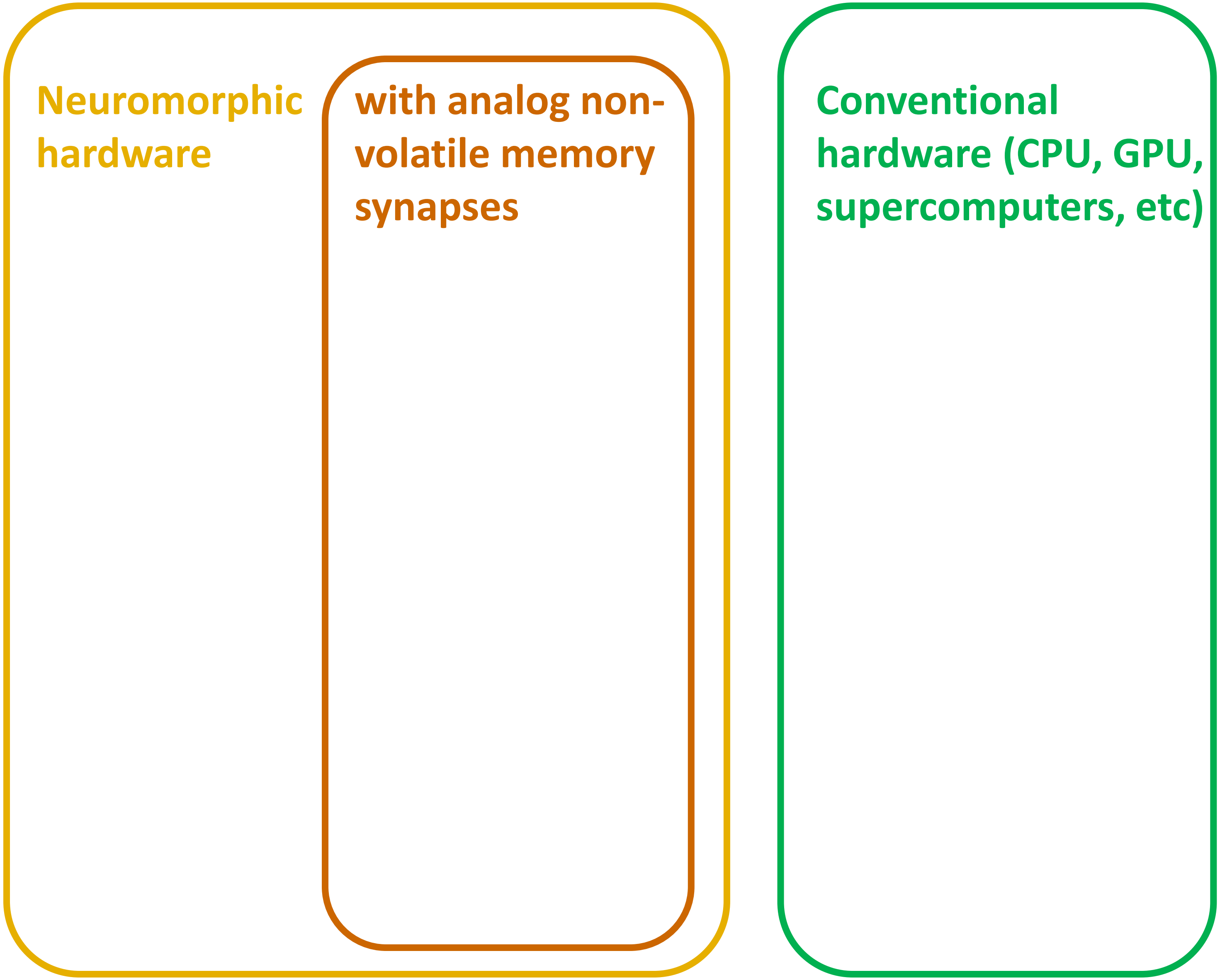
Conventional ML algorithms

Approaches of Neuromorphic Hardware

Neuromorphic hardware

Conventional hardware (CPU, GPU, supercomputers, etc)

Approaches of Neuromorphic Hardware

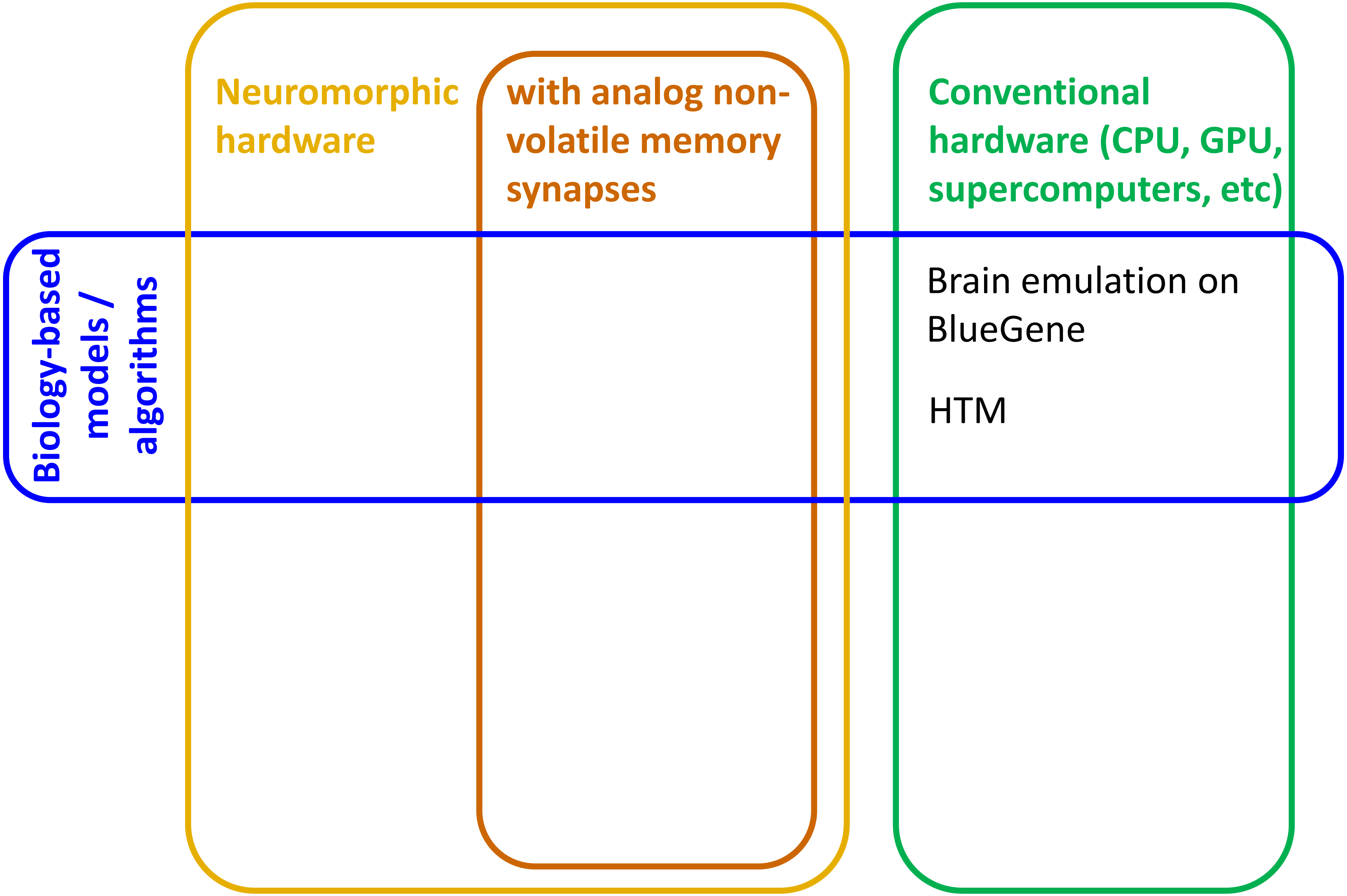


Neuromorphic hardware

with analog non-volatile memory synapses

Conventional hardware (CPU, GPU, supercomputers, etc)

Approaches of Neuromorphic Hardware



Neuromorphic hardware

with analog non-volatile memory synapses

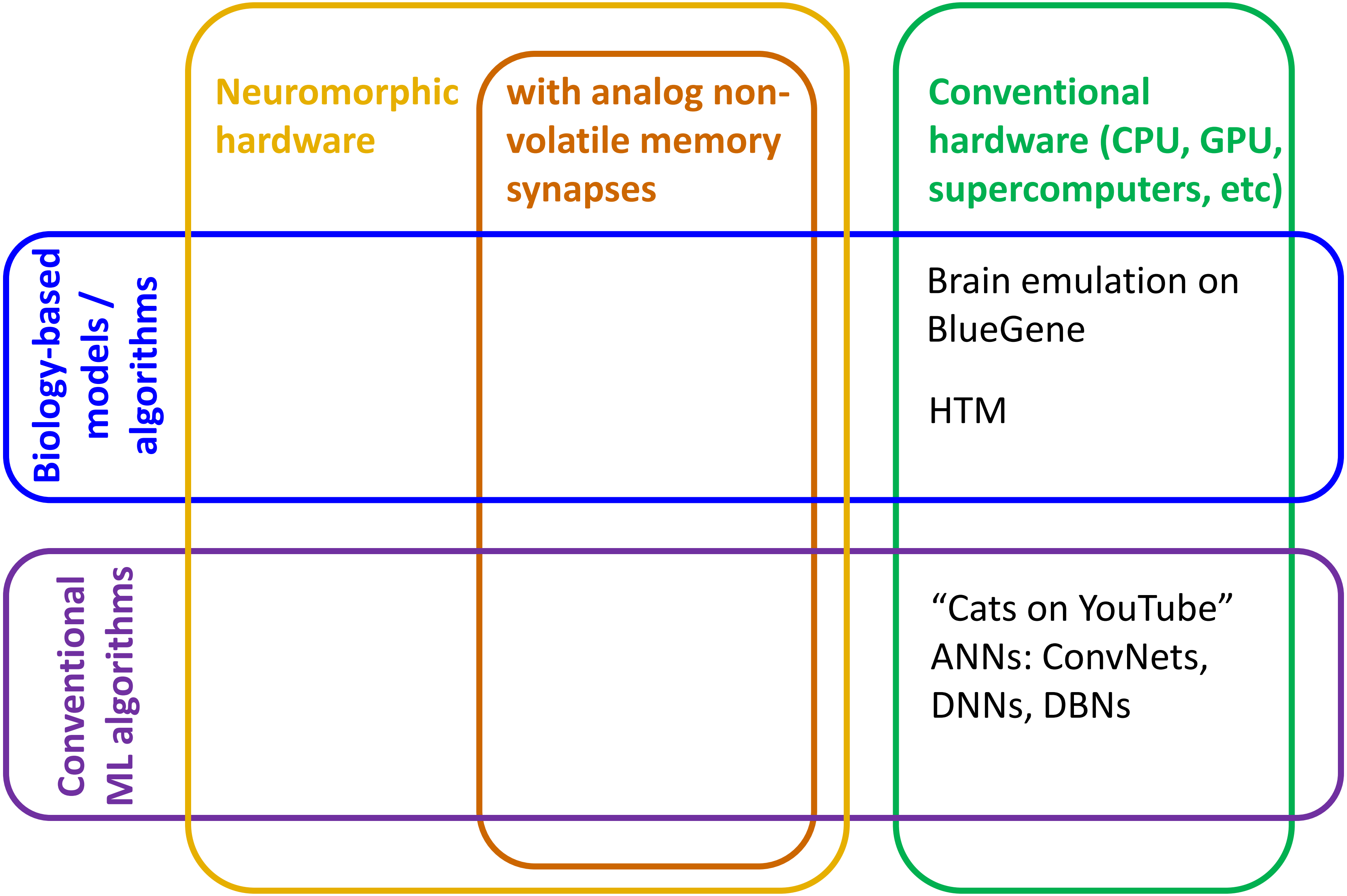
Conventional hardware (CPU, GPU, supercomputers, etc)

Biology-based models / algorithms

Brain emulation on BlueGene

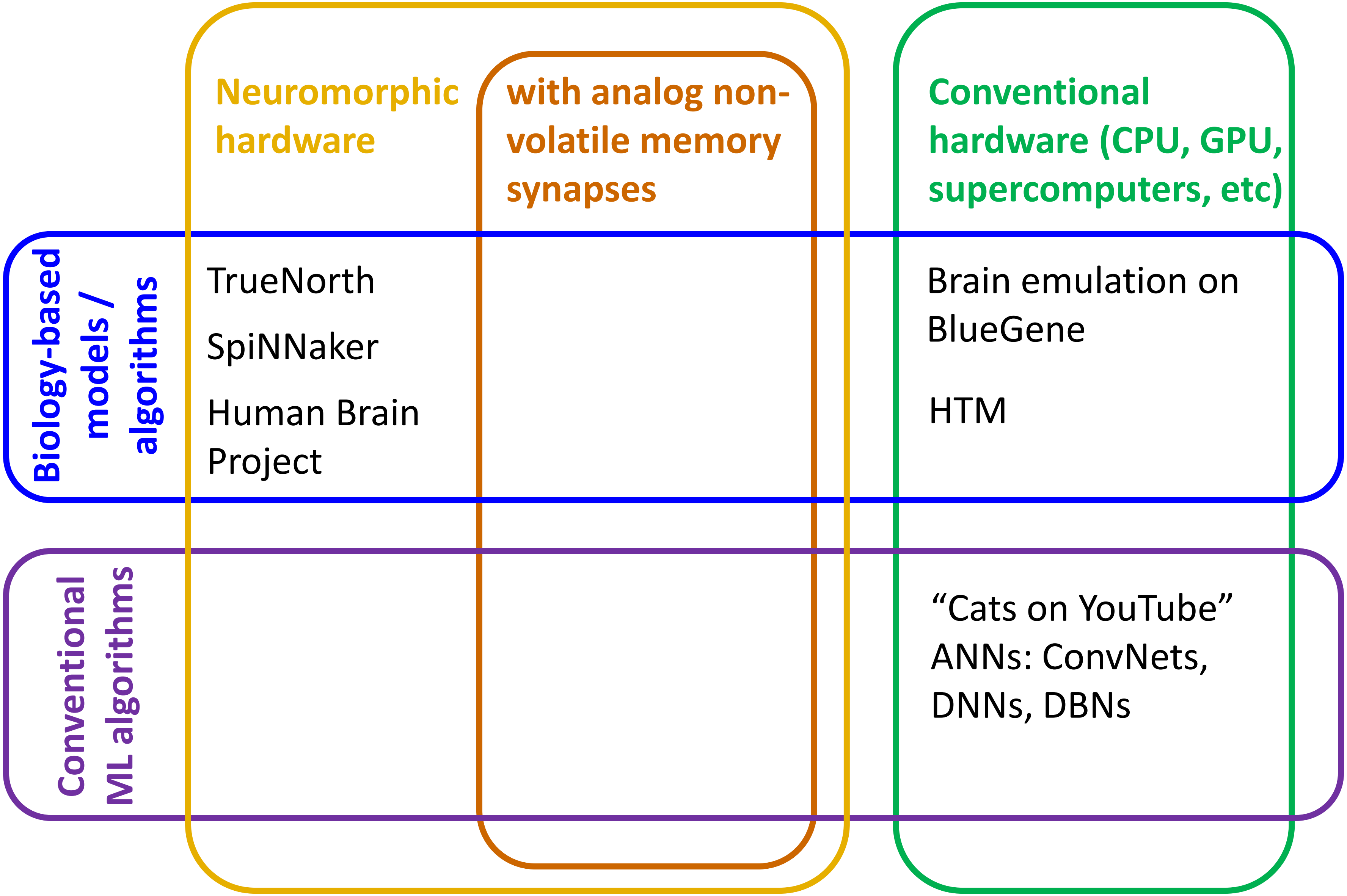
HTM

Approaches of Neuromorphic Hardware



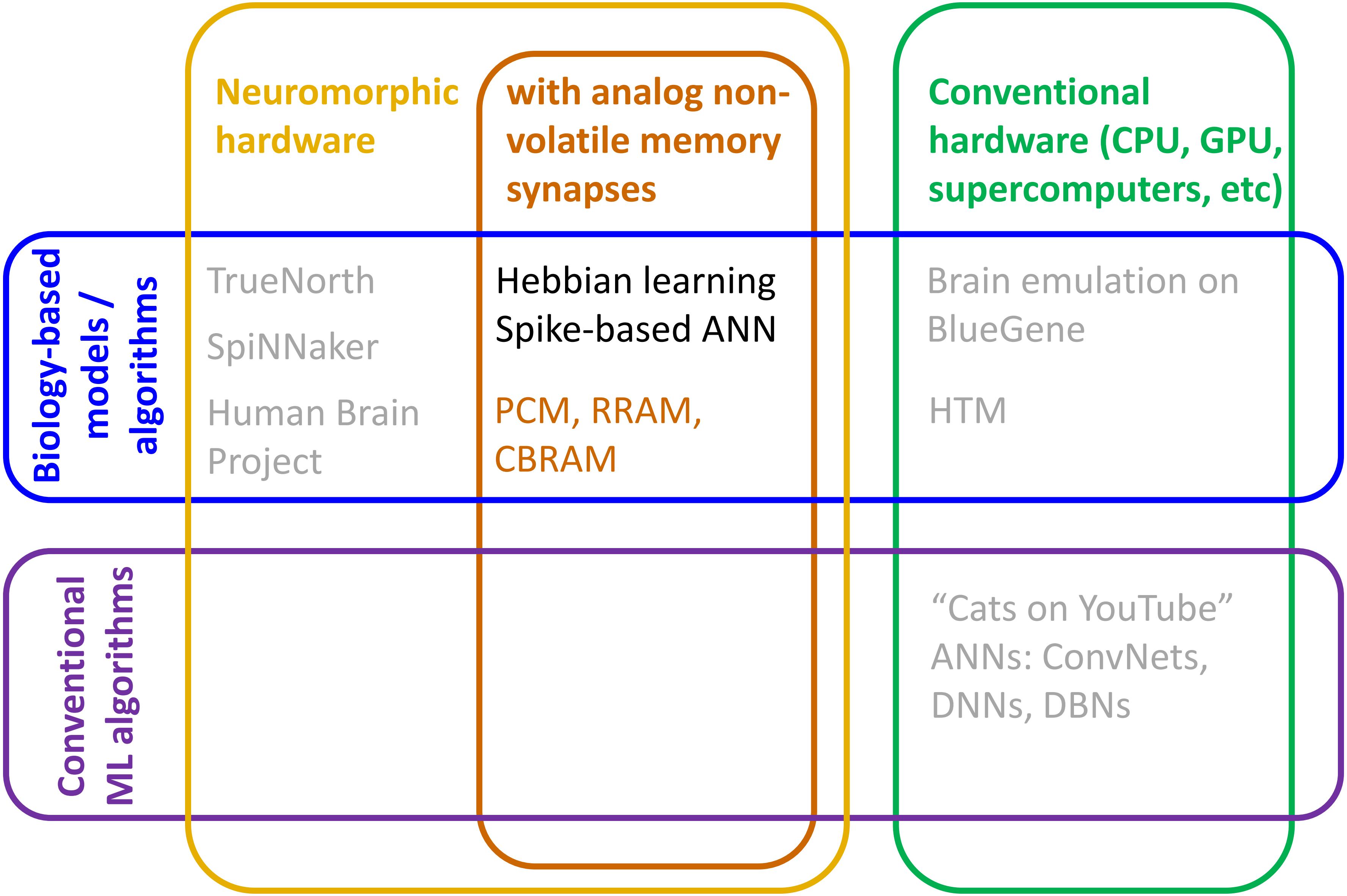
	Neuromorphic hardware	with analog non-volatile memory synapses	Conventional hardware (CPU, GPU, supercomputers, etc)
Biology-based models / algorithms			Brain emulation on BlueGene HTM
Conventional ML algorithms			“Cats on YouTube” ANNs: ConvNets, DNNs, DBNs

Approaches of Neuromorphic Hardware



	Neuromorphic hardware	with analog non-volatile memory synapses	Conventional hardware (CPU, GPU, supercomputers, etc)
Biology-based models / algorithms	TrueNorth SpiNNaker Human Brain Project		Brain emulation on BlueGene HTM
Conventional ML algorithms			"Cats on YouTube" ANNs: ConvNets, DNNs, DBNs

Approaches of Neuromorphic Hardware



Neuromorphic hardware

with analog non-volatile memory synapses

Conventional hardware (CPU, GPU, supercomputers, etc)

Biology-based models / algorithms

TrueNorth

SpiNNaker

Human Brain Project

Hebbian learning
Spike-based ANN

PCM, RRAM, CBRAM

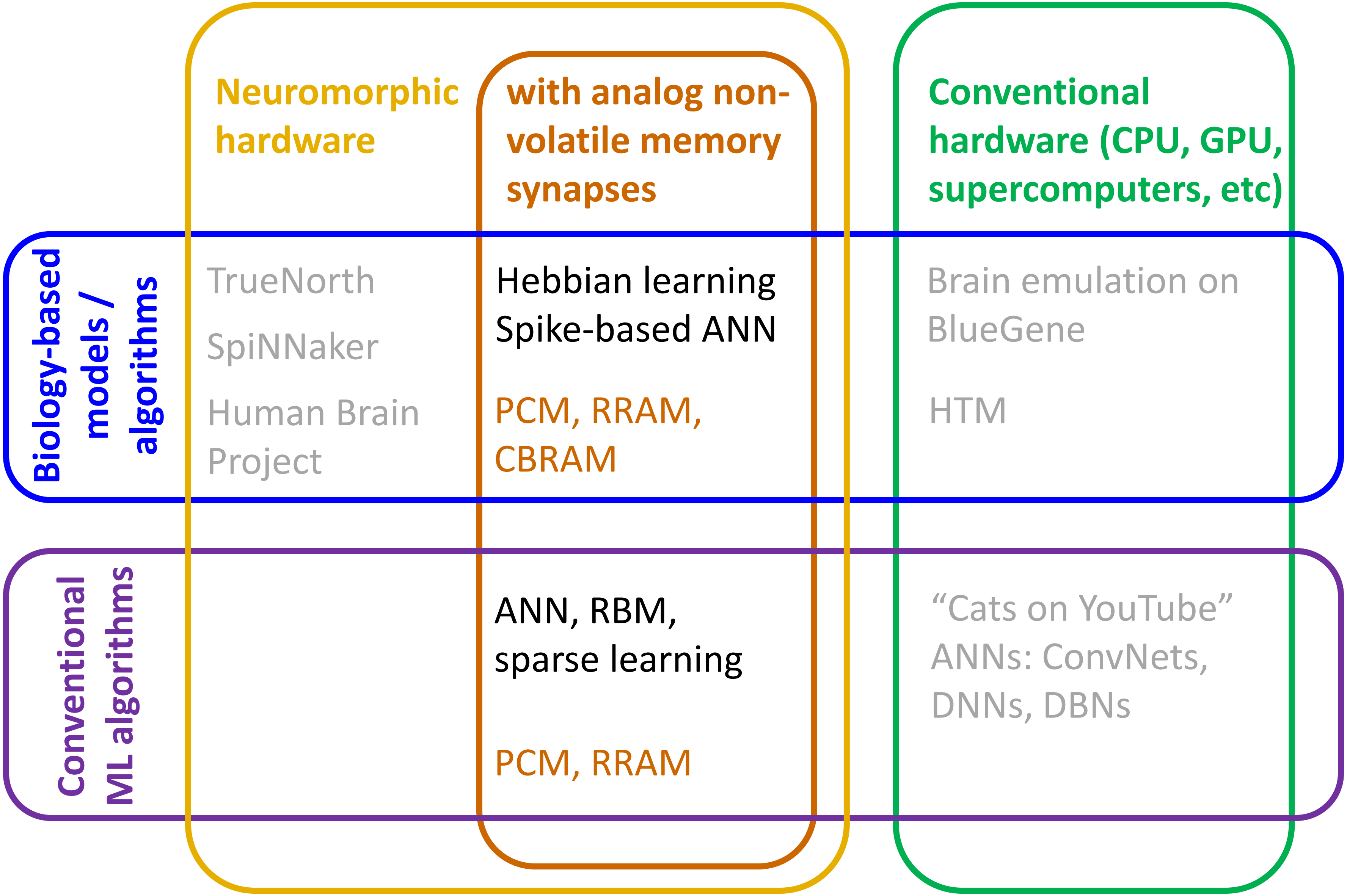
Brain emulation on BlueGene

HTM

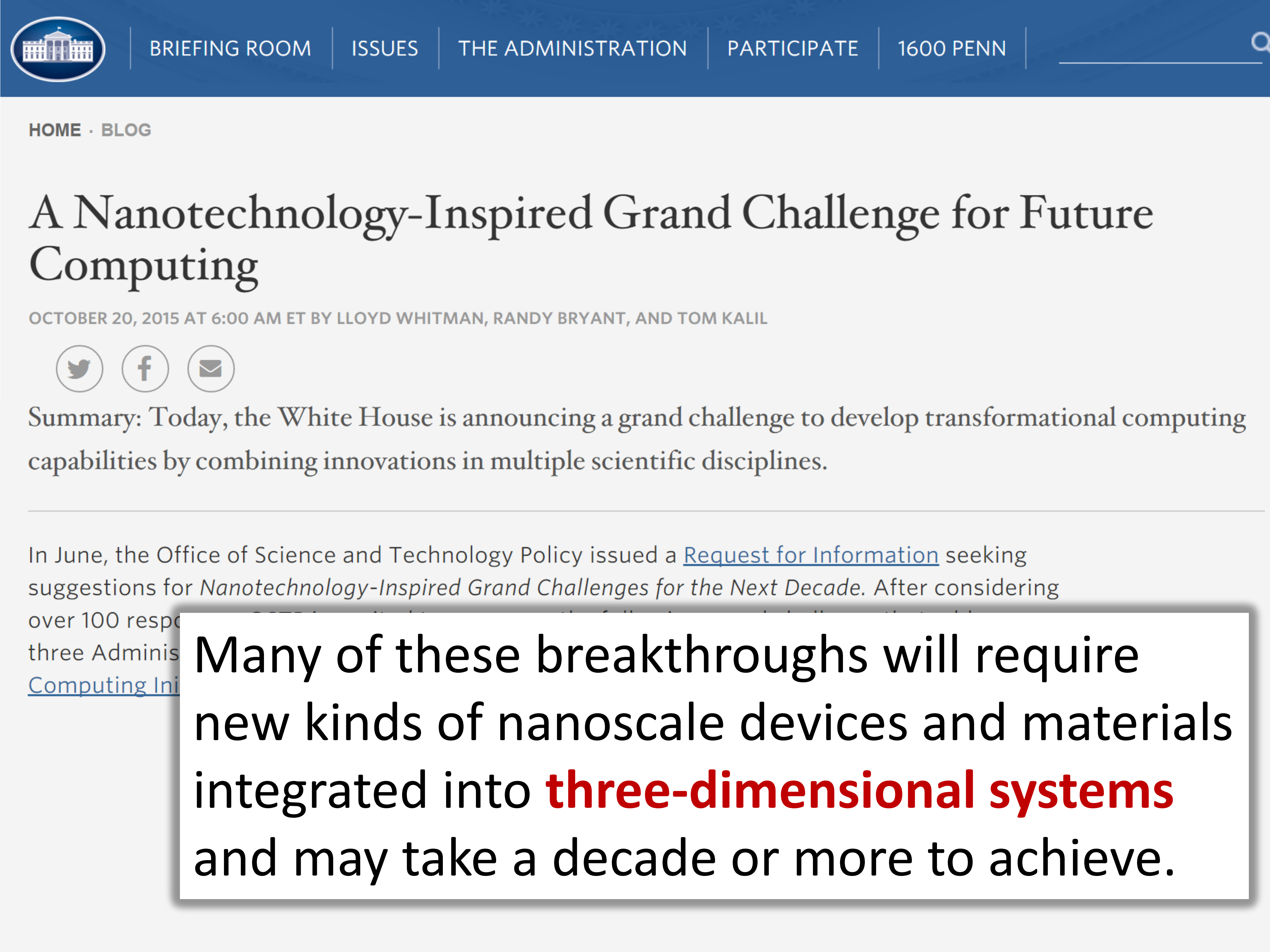
Conventional ML algorithms

"Cats on YouTube"
ANNs: ConvNets, DNNs, DBNs

Approaches of Neuromorphic Hardware



	Neuromorphic hardware	with analog non-volatile memory synapses	Conventional hardware (CPU, GPU, supercomputers, etc)
Biology-based models / algorithms	TrueNorth SpiNNaker Human Brain Project	Hebbian learning Spike-based ANN PCM, RRAM, CBRAM	Brain emulation on BlueGene HTM
Conventional ML algorithms		ANN, RBM, sparse learning PCM, RRAM	"Cats on YouTube" ANNs: ConvNets, DNNs, DBNs



BRIEFING ROOM | ISSUES | THE ADMINISTRATION | PARTICIPATE | 1600 PENN

HOME · BLOG

A Nanotechnology-Inspired Grand Challenge for Future Computing

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

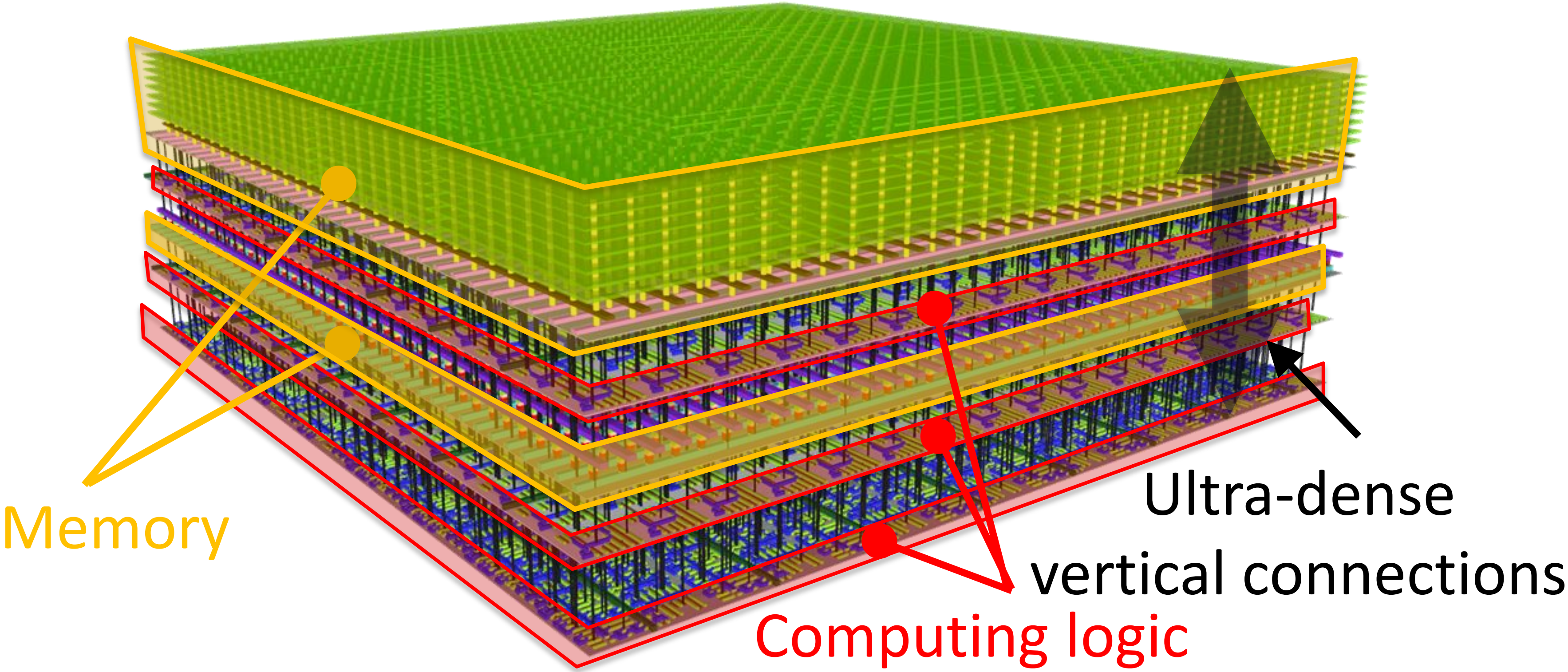
Summary: Today, the White House is announcing a grand challenge to develop transformational computing capabilities by combining innovations in multiple scientific disciplines.

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 resp
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.

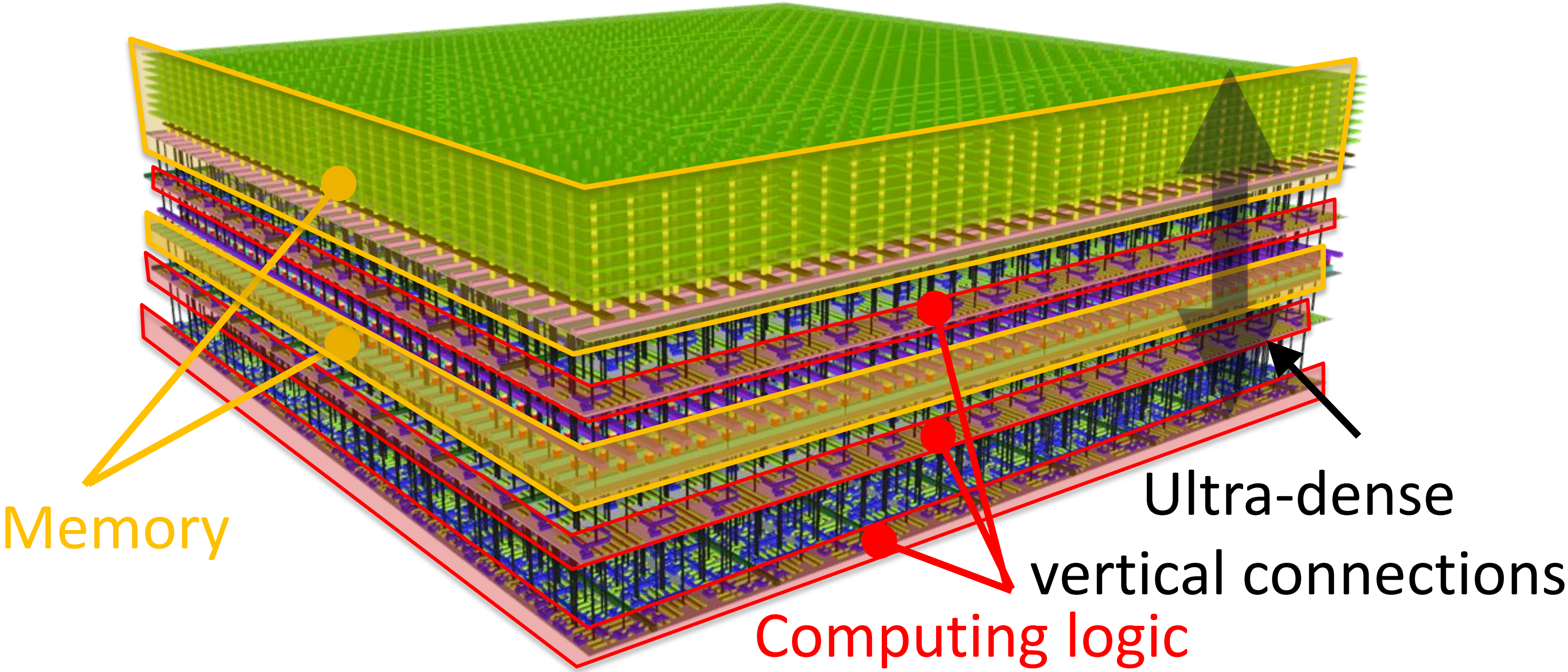
N3XT Nanosystems

Computation immersed in memory



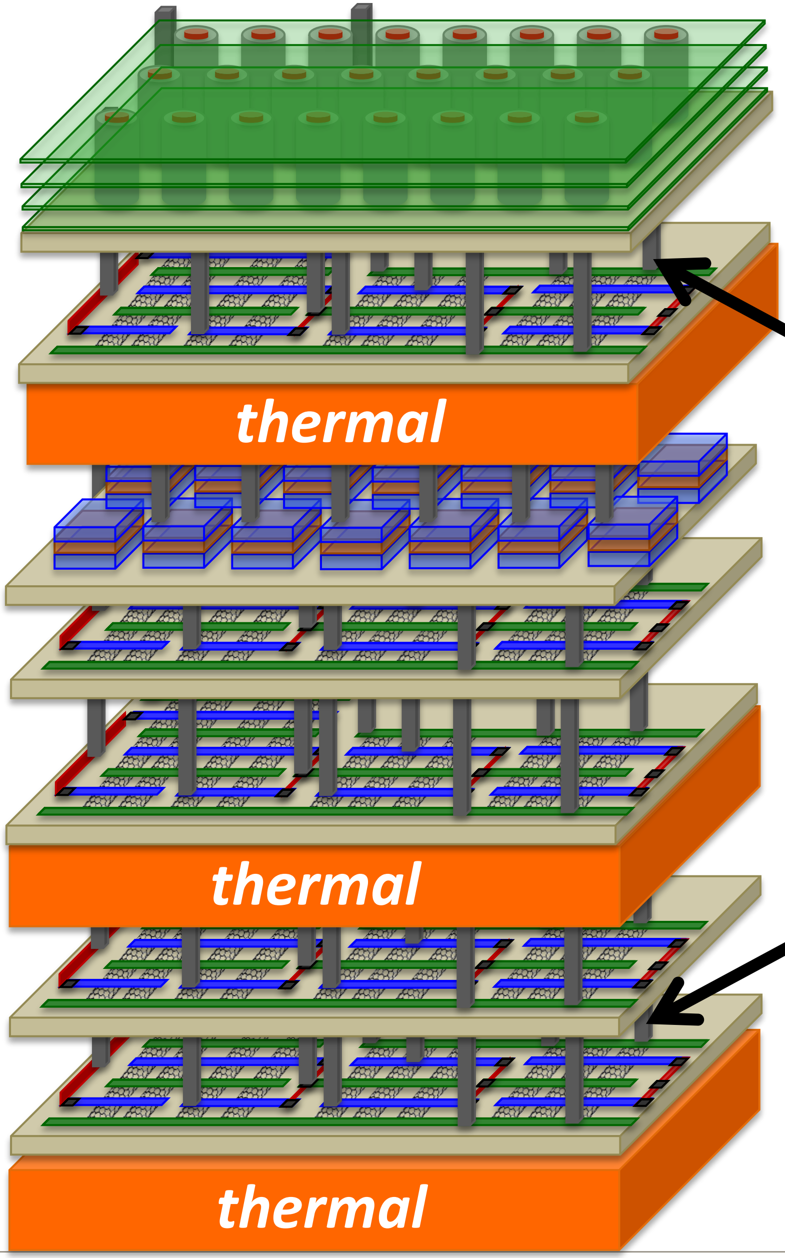
N3XT Nanosystems

Computation immersed in memory



Impossible with today's technologies

N3XT: Computation Immersed in Memory



3D Resistive RAM
Massive storage

1D CNFET, 2D FET
Compute, RAM access

MRAM
Quick access

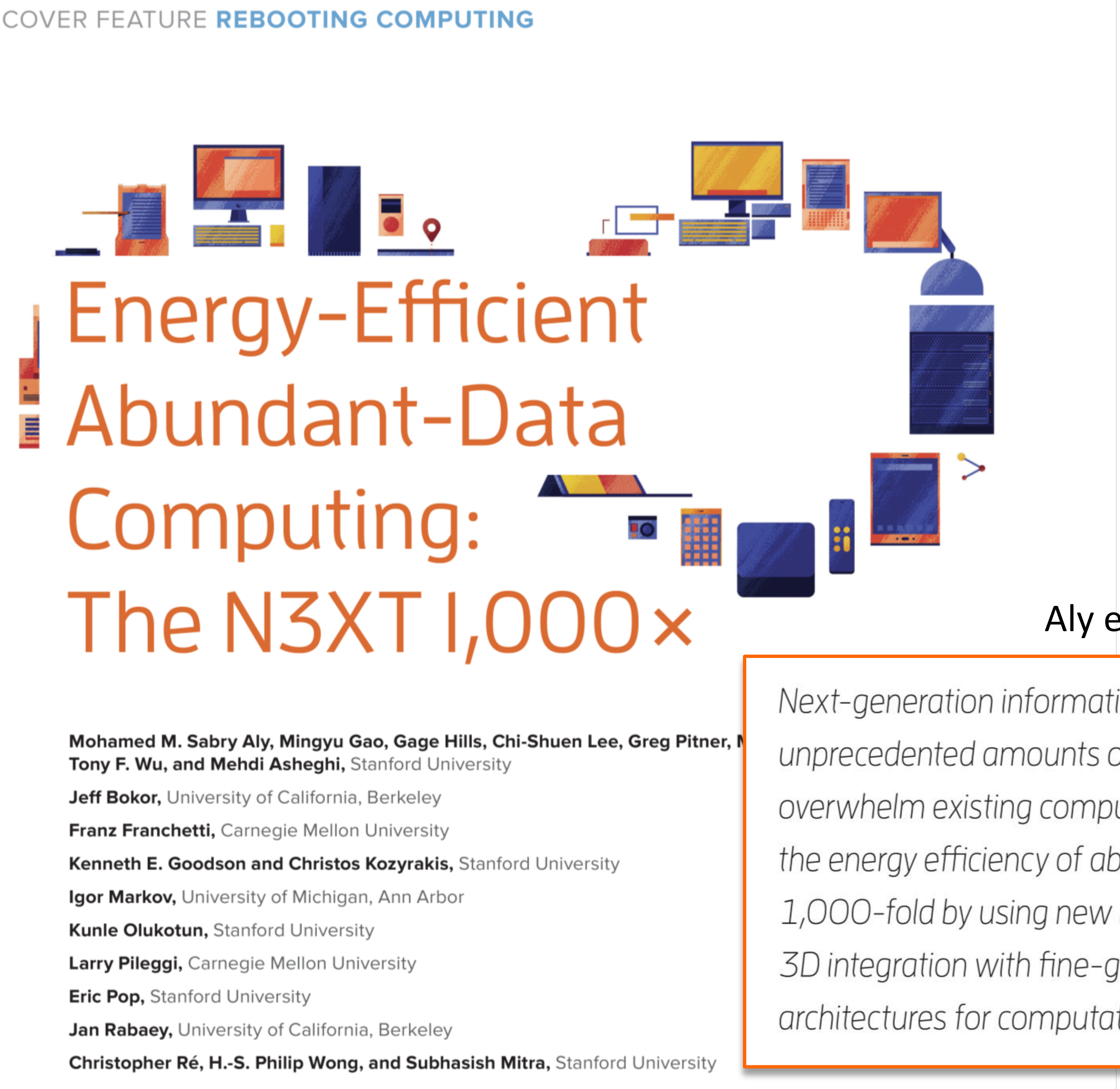
1D CNFET, 2D FET
Compute, RAM access

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

Not TSV

Ultra-dense, fine-grained vias

Silicon compatible



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, 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

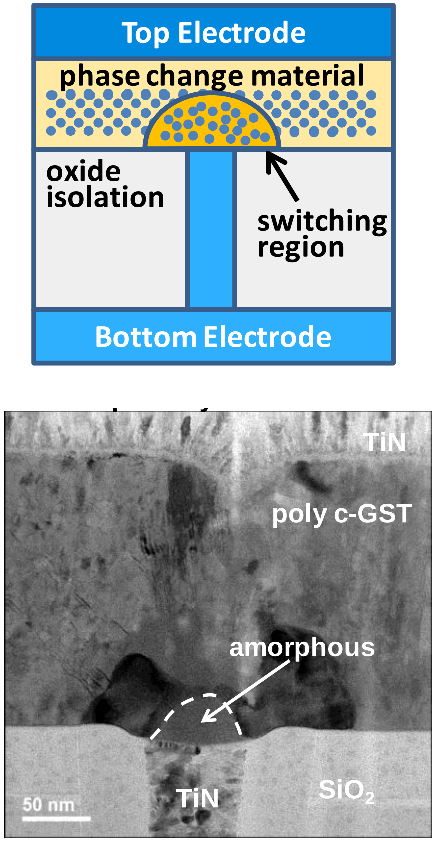
Jan Rabaey, University of California, Berkeley

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

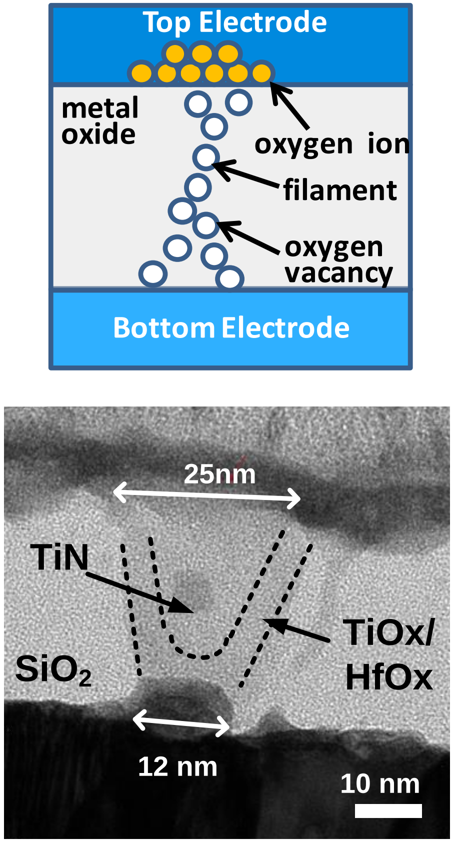
Aly et al., *IEEE Computer*, 2015

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.

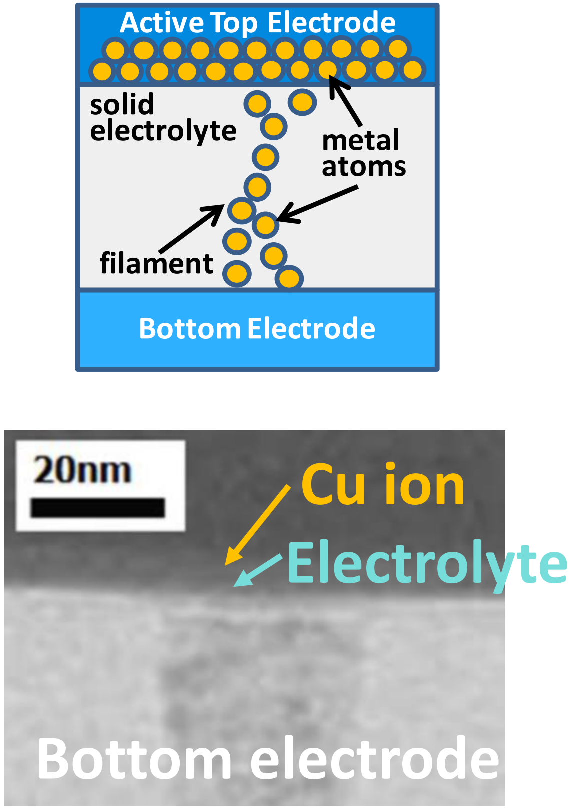
Non-Volatile Memory (NVM)



Phase change memory (PCM)



Metal oxide resistive switching memory (RRAM)

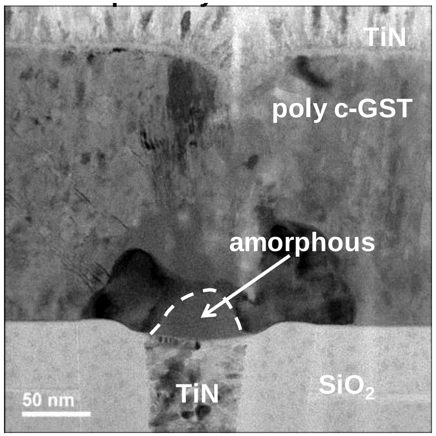


Conductive bridge memory (CBRAM)

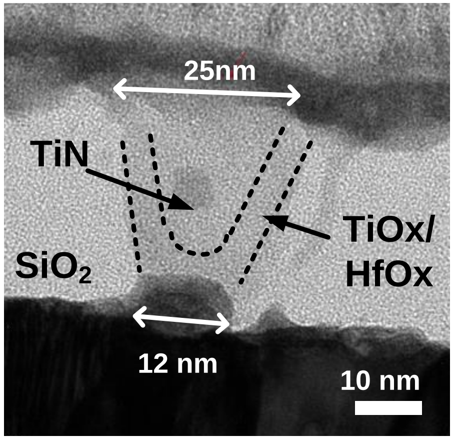
D. Kuzum et al., *Nano Lett*. 2013, Y. Wu et al., *IEDM* 2013; A. Calderoni et al., *IMW* 2014

Non-Volatile Memory (NVM) → Synapse

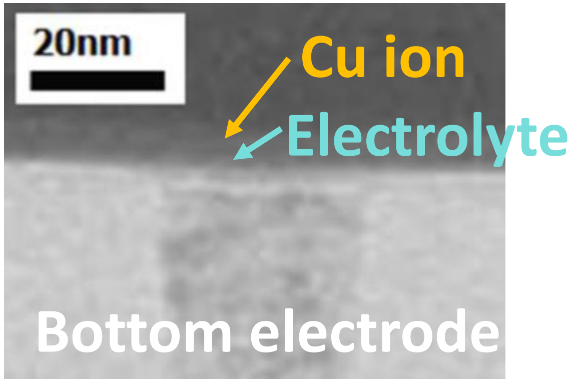
- Analog programmable
- Scalable to a few nm
- Stack in 3D



Phase change memory (PCM)



Metal oxide resistive switching memory (RRAM)



Conductive bridge memory (CBRAM)

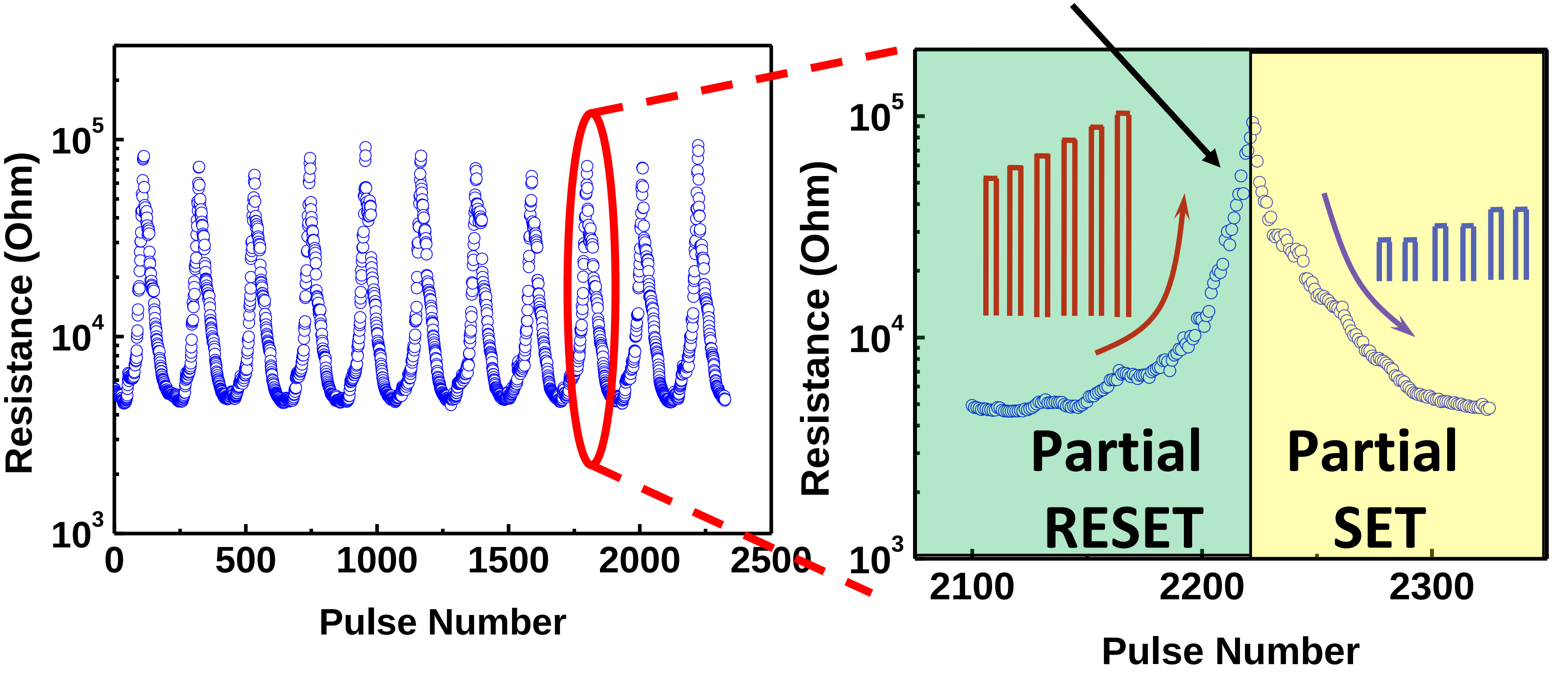
D. Kuzum et al., *Nano Lett*. 2013, Y. Wu et al., *IEDM* 2013; A. Calderoni et al., *IMW* 2014

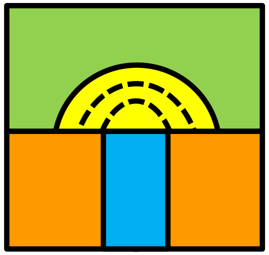
Nanoscale Memory as Synaptic Weights

Synaptic updates in the brain: basis for learning

Requirement: analog resistance change

100-step grey scale (1% resolution)





Phase change synapse

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

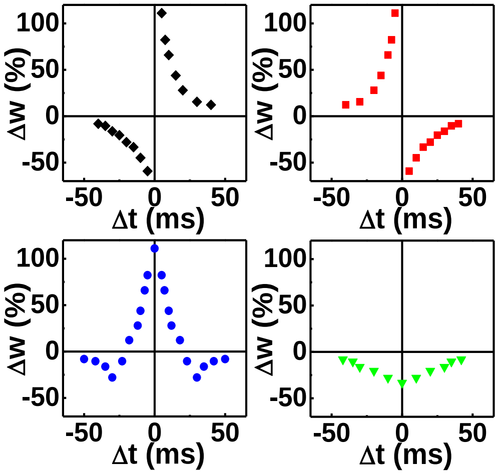
Nanoscale Memory Can Emulate Biological Synaptic Behavior

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)

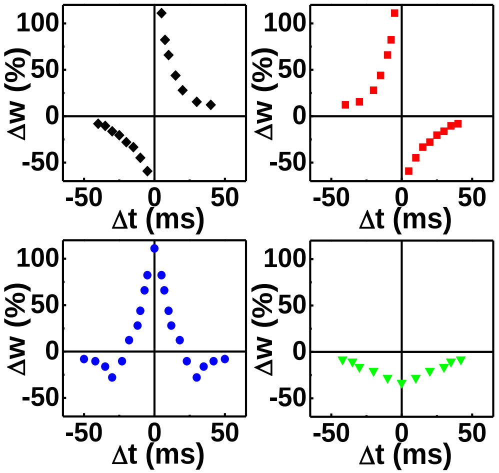


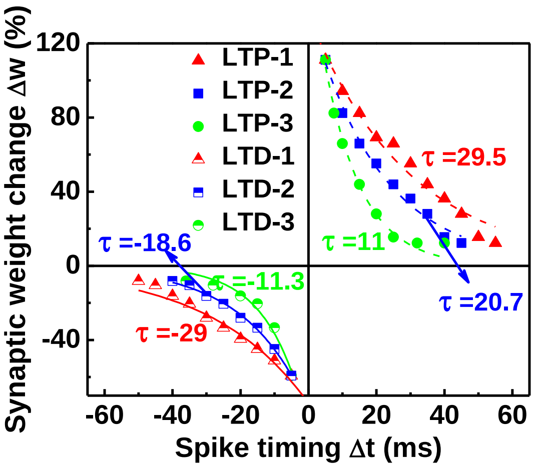
Various STDP kernels

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

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)





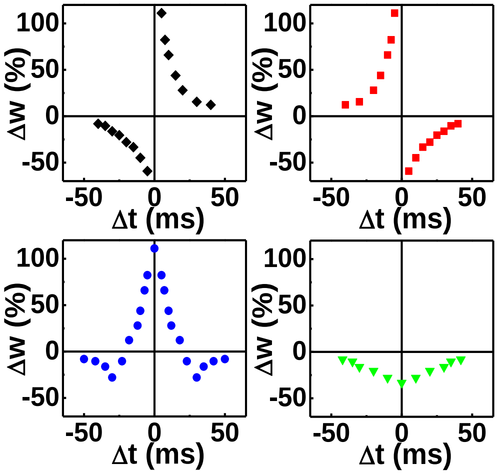
Various STDP kernels

Various time constants

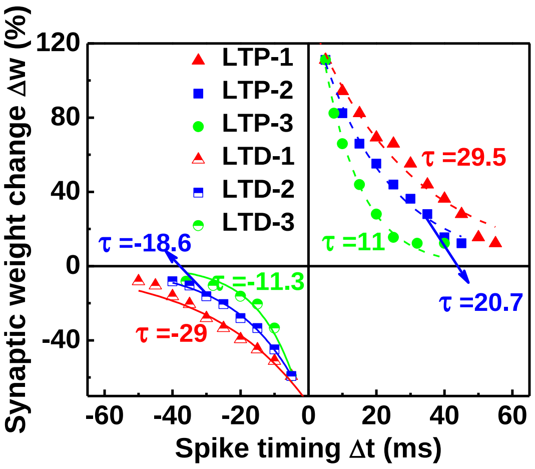
D. Kuzum *et al.*, *Nano Lett.*, p. 2179 (2012)

Nanoscale Memory Can Emulate Biological Synaptic Behavior

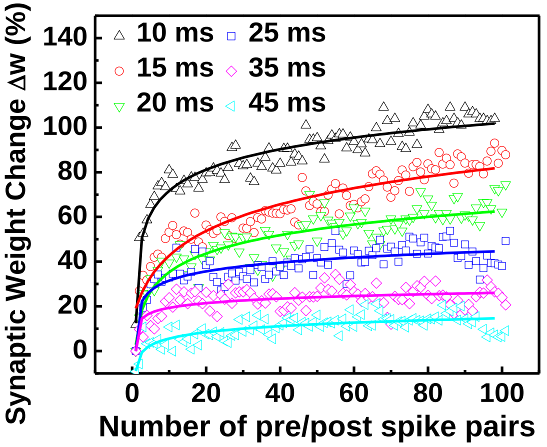
STDP (spike-timing-dependent plasticity)



Various STDP kernels



Various time constants



Weight update saturation

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

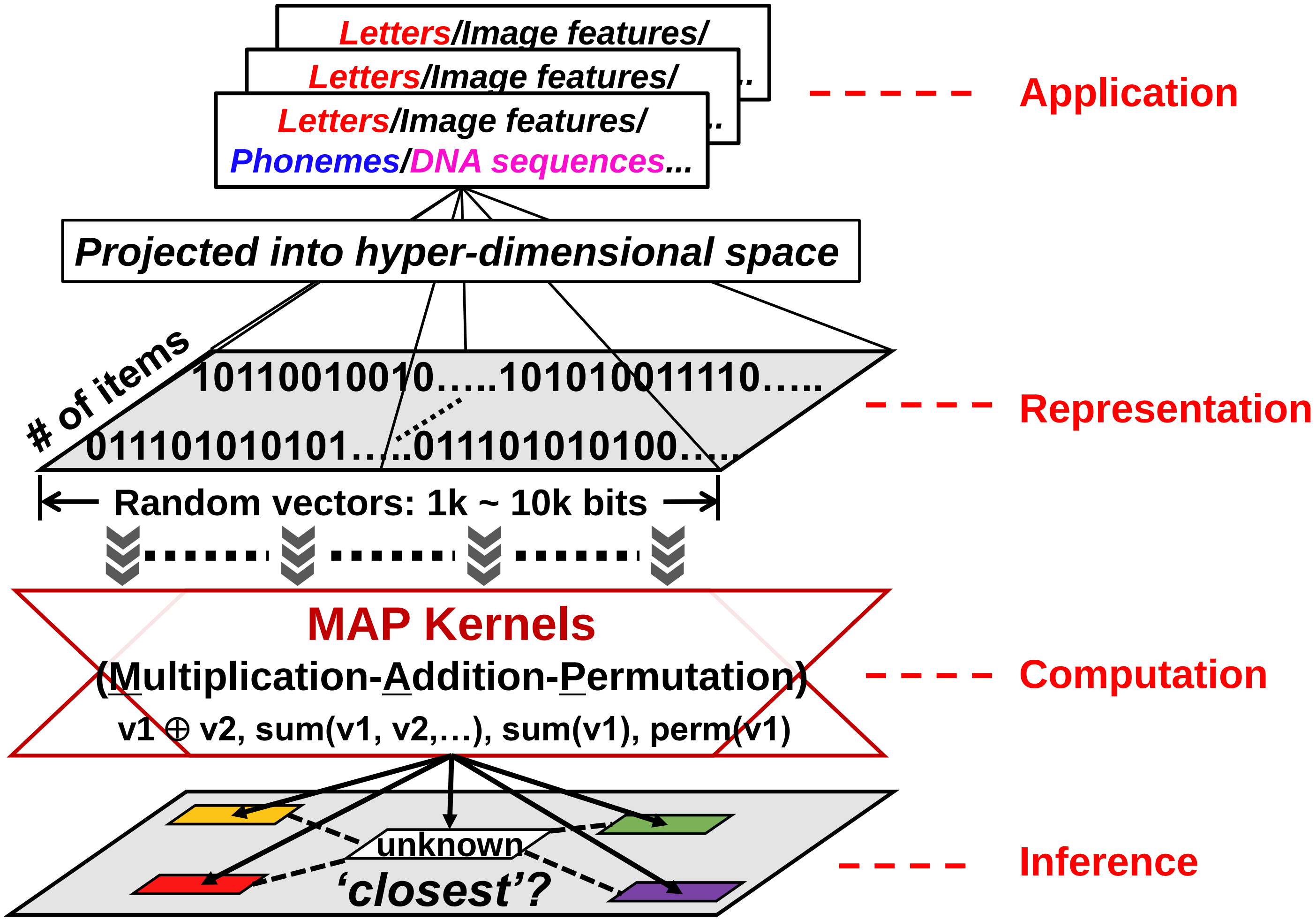
Hyper Dimensional (HD) Computing

Elements of Hyper dimensional Computing:

Also known as vector symbolic architectures or holographic reduced representation (Kanerva, *Cognitive Computation, 1(2):139-159,2009*)

- Information is represented by High-dimensional representation (e.g., D = 10,000)
- Variables and values are combined into a “holistic” record using vector algebra:
  - Multiplication for Binding
  - Addition for Bundling
- Composed vector can in turn become a component in further composition
- Holistic record is decoded with (inverse) multiplication
- Approximate results of vector operations are identified with exact ones using content-addressable memory

HD Computing layers



Letters/Image features/

Letters/Image features/

Letters/Image features/ Phonemes/DNA sequences...

Projected into hyper-dimensional space

# of items

10110010010…..101010011110…..

011101010101…..011101010100…..

Random vectors: 1k ~ 10k bits

MAP Kernels

(Multiplication-Addition-Permutation)

v1 ⊕ v2, sum(v1, v2,…), sum(v1), perm(v1)

unknown

'closest'?

Application

Representation

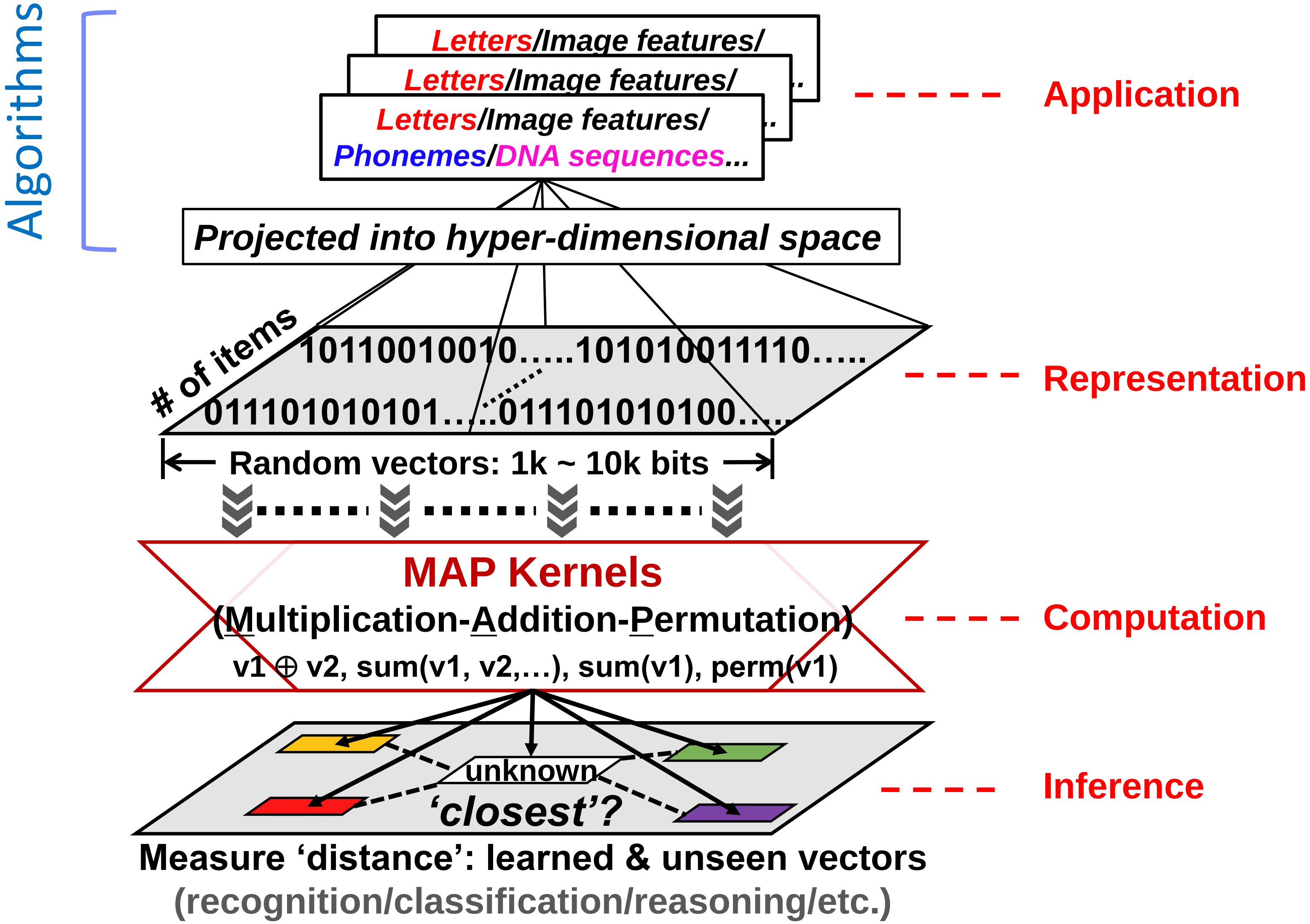
Computation

Inference

Measure 'distance': learned & unseen vectors

(recognition/classification/reasoning/etc.)

HD Computing layers



Algorithms

Letters/Image features/
Letters/Image features/
Letters/Image features/
Phonemes/DNA sequences...

Application

Projected into hyper-dimensional space

# of items

10110010010.....101010011110.....

011101010101.....011101010100.....

Representation

Random vectors: 1k ~ 10k bits

MAP Kernels

(Multiplication-Addition-Permutation)

v1 ⊕ v2, sum(v1, v2,…), sum(v1), perm(v1)

Computation

unknown

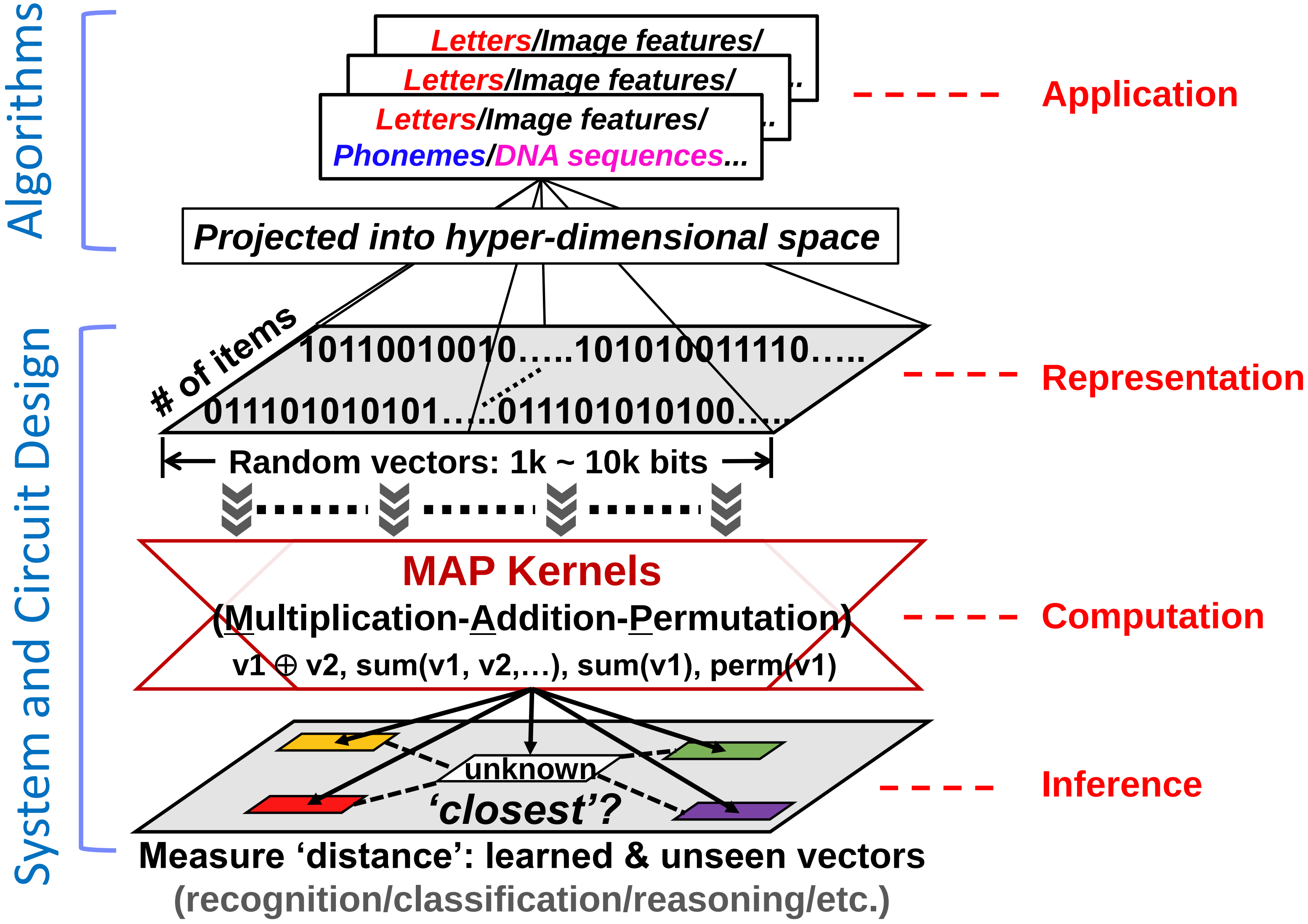
'closest'?

Inference

Measure 'distance': learned & unseen vectors

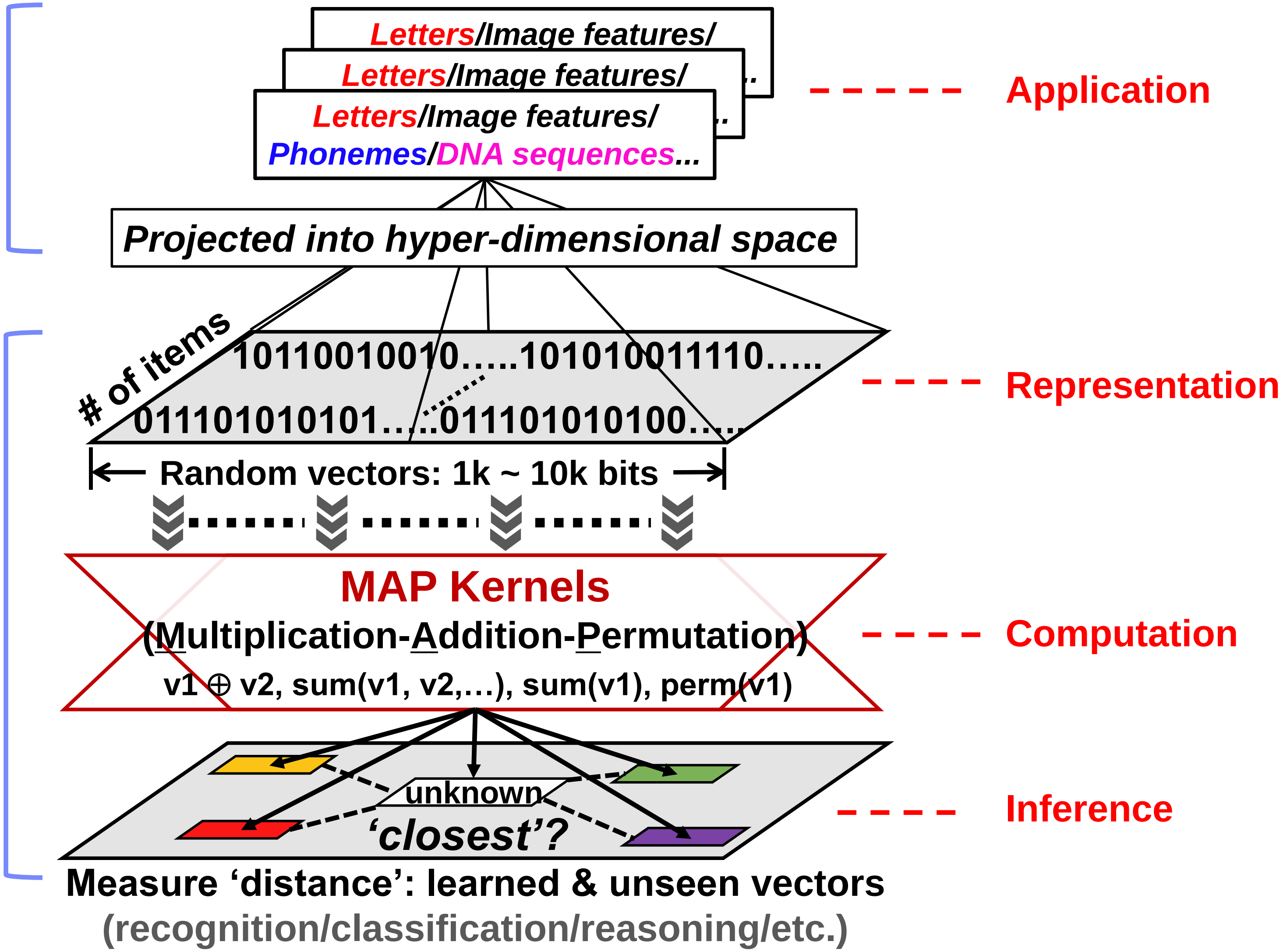
(recognition/classification/reasoning/etc.)

HD Computing layers


Algorithms
System and Circuit Design
Letters/Image features/
Phonemes/DNA sequences...
Projected into hyper-dimensional space
# of items
10110010010…..101010011110…..
011101010101…..011101010100…..
Random vectors: 1k ~ 10k bits
MAP Kernels
(Multiplication-Addition-Permutation)
v1 ⊕ v2, sum(v1, v2,…), sum(v1), perm(v1)
unknown
'closest'?
Measure 'distance': learned & unseen vectors
(recognition/classification/reasoning/etc.)
Application
Representation
Computation
Inference

HD Computing layers

Associative memory enabled by novel device technologies



Algorithms

System and Circuit Design

Letters/Image features/
Phonemes/DNA sequences...

Projected into hyper-dimensional space

Application

# of items

10110010010…..101010011110…..
011101010101…..011101010100…..

Random vectors: 1k ~ 10k bits

Representation

MAP Kernels

(Multiplication-Addition-Permutation)

v1 ⊕ v2, sum(v1, v2,…), sum(v1), perm(v1)

Computation

unknown

'closest'?

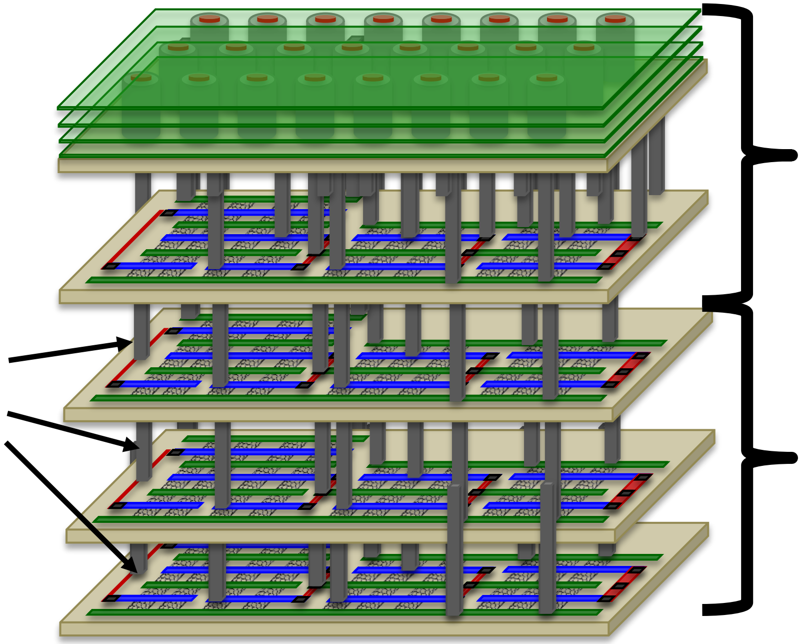
Inference

Measure 'distance': learned & unseen vectors

(recognition/classification/reasoning/etc.)

Hyperdimensional (HD) Computing

- Monolithic 3D enables
  - – Energy-efficient classification
  - – Area efficient HD projection
    - → use RRAM variability & stochastic switching



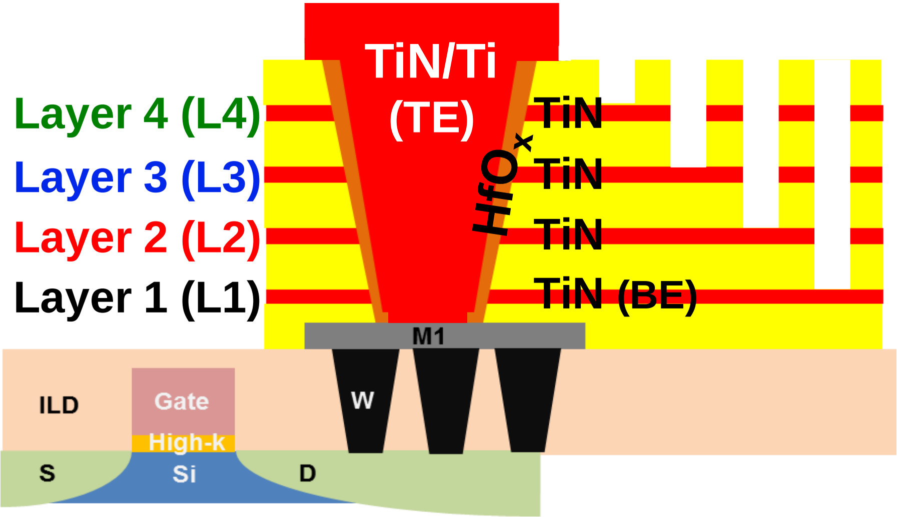
3D RRAM
+ low power access transistors
+ address decoders

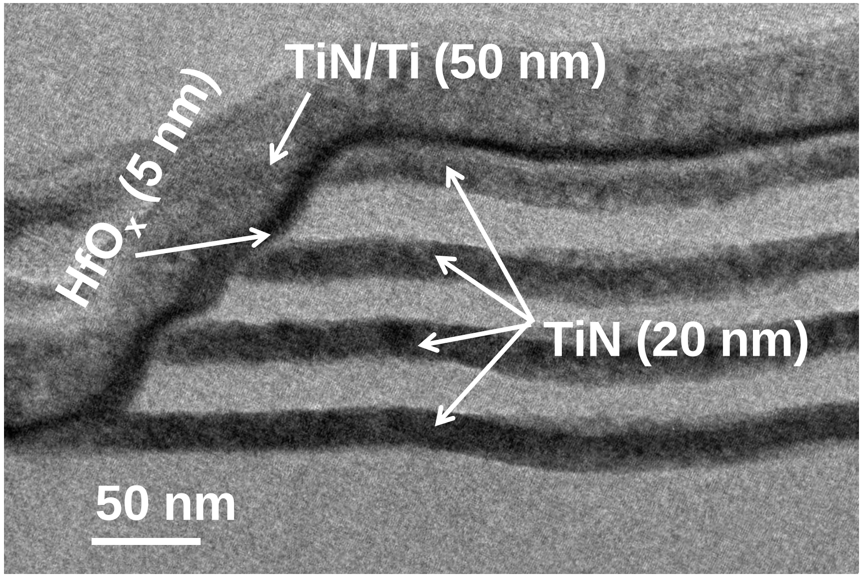
Low power computation

High-density
Inter-layer vias

3D Enables In-Memory Computing

3D RRAM with FinFET BL select

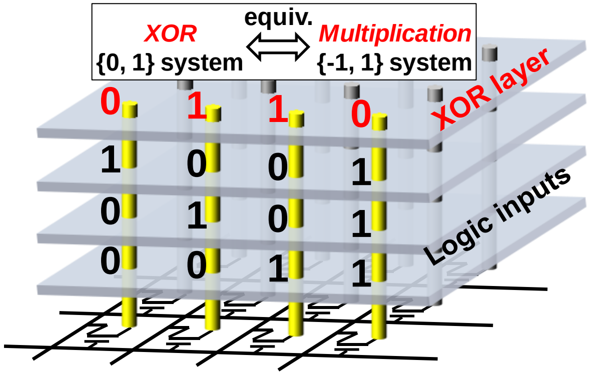




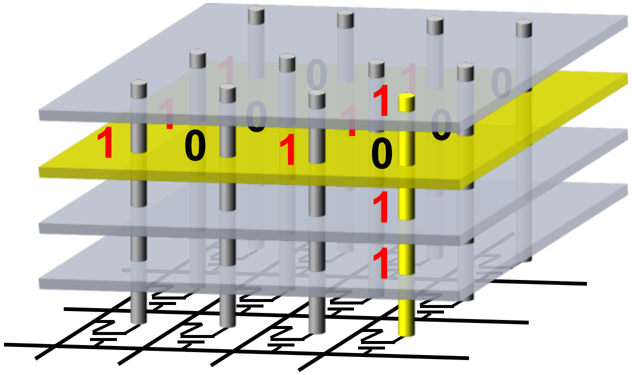
H. Li *et al.*, *Symp. VLSI Tech.*, 2016

MAP Kernels: 3D RRAM Approach

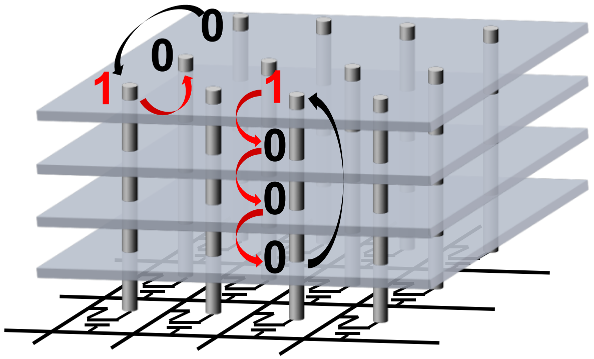
Key HD operations: **m**ultiplication, **a**ddition, **p**ermutation



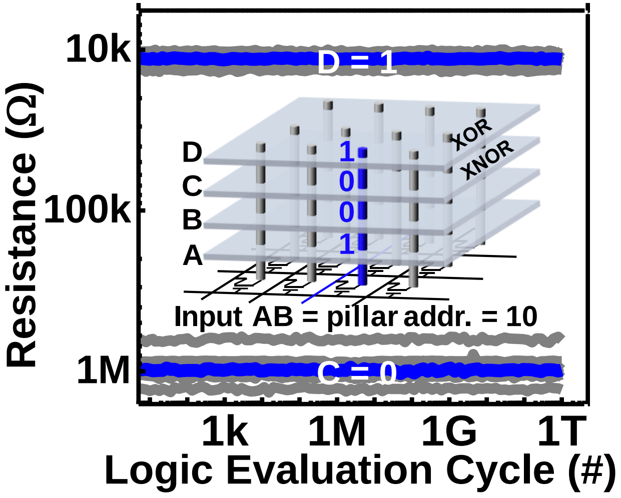
Multiplication

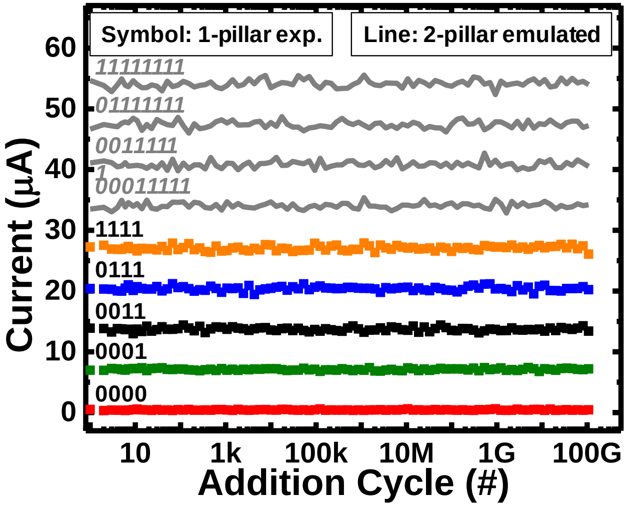


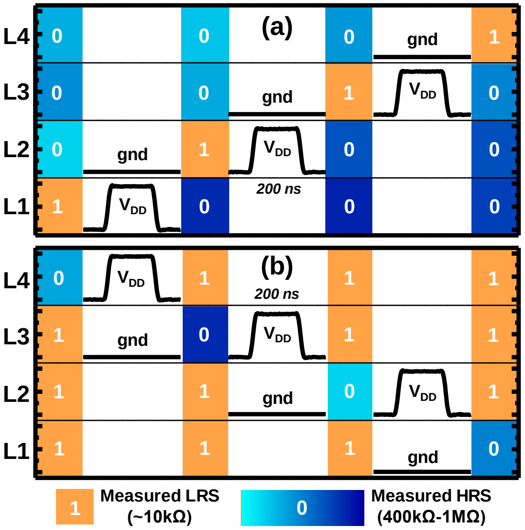
Addition



Permutation



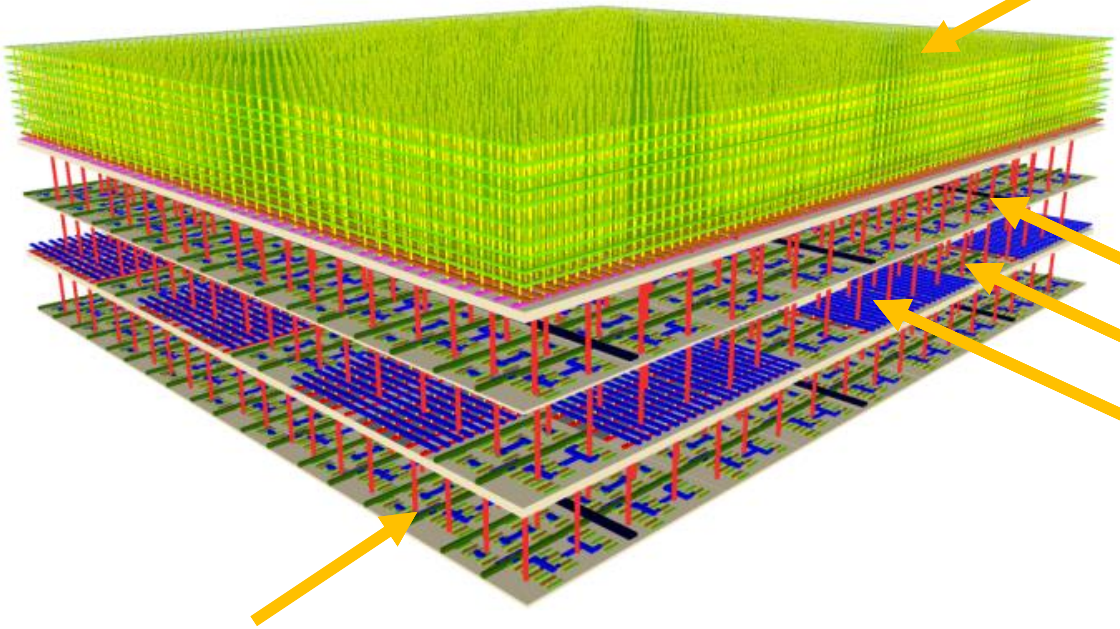




Measured data on 4-layer 3D vertical RRAM

3D Integration of Memory and Logic Circuits within the Same Layer & Across Layers

N3XT



Synapses/Weights (3D RRAM)

Neuron circuits

Communication

Synapses/Weights (CNFET/2D FET) + RRAM

Logic (Si CMOS)

Nano-Engineered Computing Systems Technology

N3XT

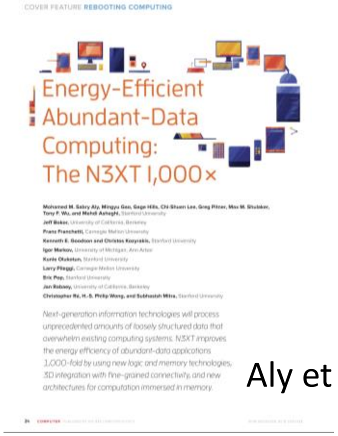
0 1 0 1

Stanford

Berkeley UNIVERSITY OF CALIFORNIA

Carnegie Mellon University

MICHIGAN



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

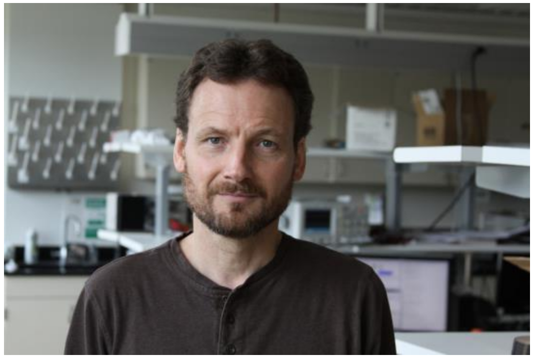
Next-generation information technologies will process unprecedented amounts of loosely structured data that overwhelm existing computing systems. N3XT improves the energy efficiency of abundant-data applications 1,000-fold by using new logic and memory technologies, 3D integration with fine-grained connectivity,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

24 COMPUTER

Aly et al., *IEEE Computer*, 2015

Students and Post-Docs

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

Sponsors

STARnet

SONIC

NSF

E2CDA “Engima”

Visual Cortex on Silicon

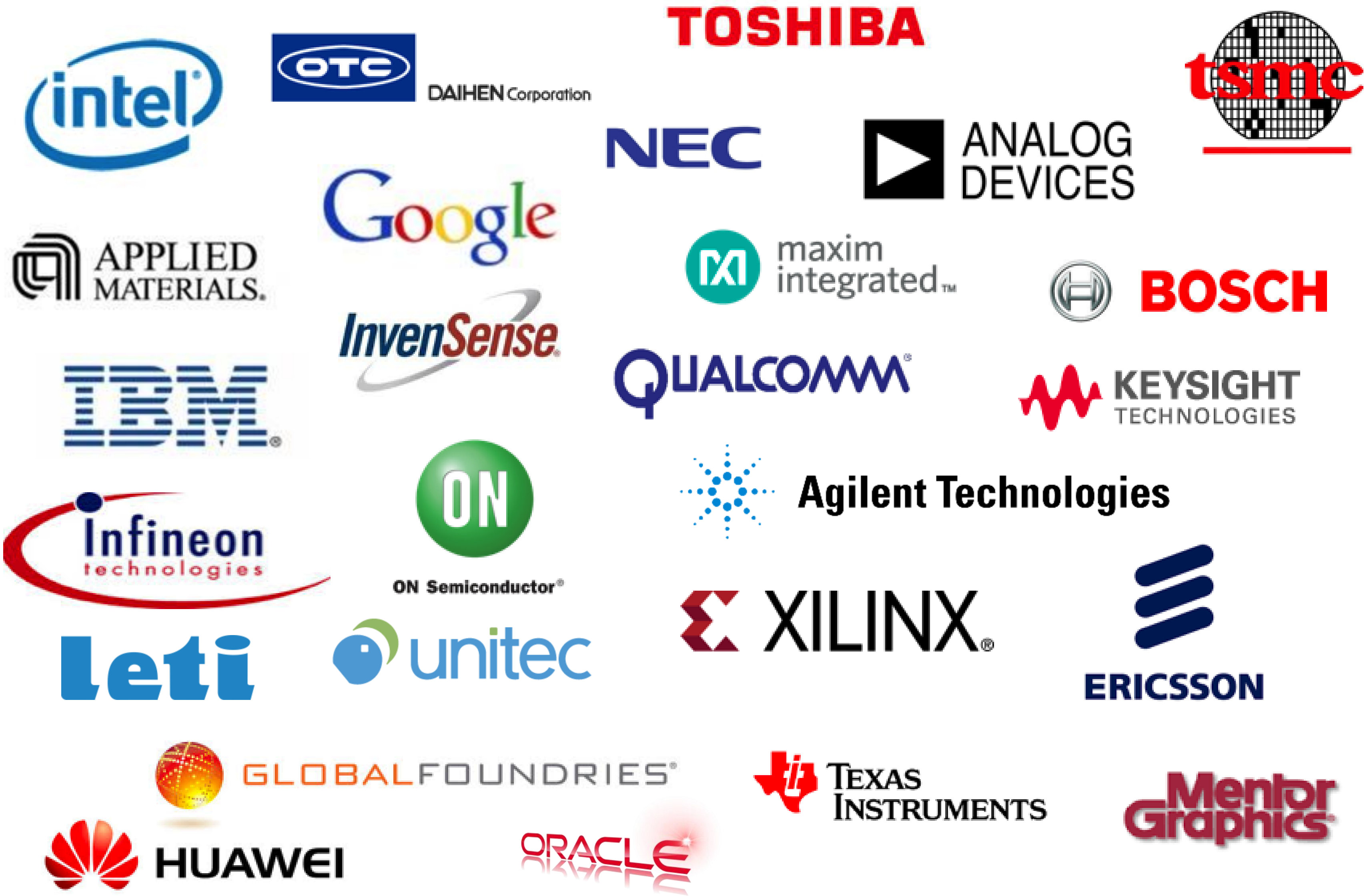
http://www.cse.psu.edu/research/visualcortexonsilicon.expedition/

Supported by National Science Foundation Expeditions in Computing Program

Stanford | SystemX Alliance

Stanford | Non-Volatile Memory Technology Research Initiative

Stanford SystemX Alliance

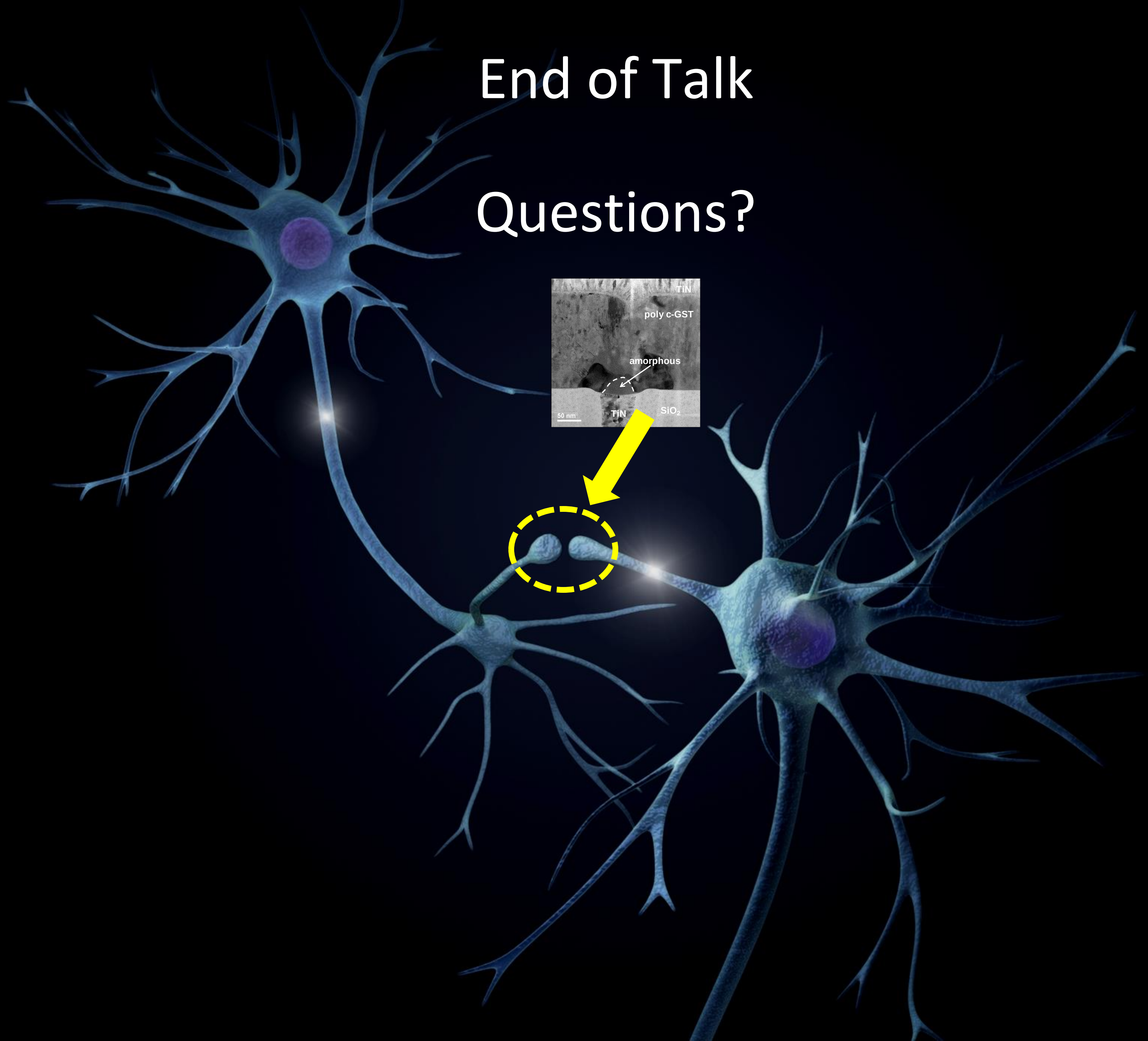


Non-Volatile Memory Technology Research Initiative (NMTRI) @ Stanford University

SanDisk
Micron
TOSHIBA
Lam RESEARCH
intel
SAMSUNG
tsmc

End of Talk

Questions?



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