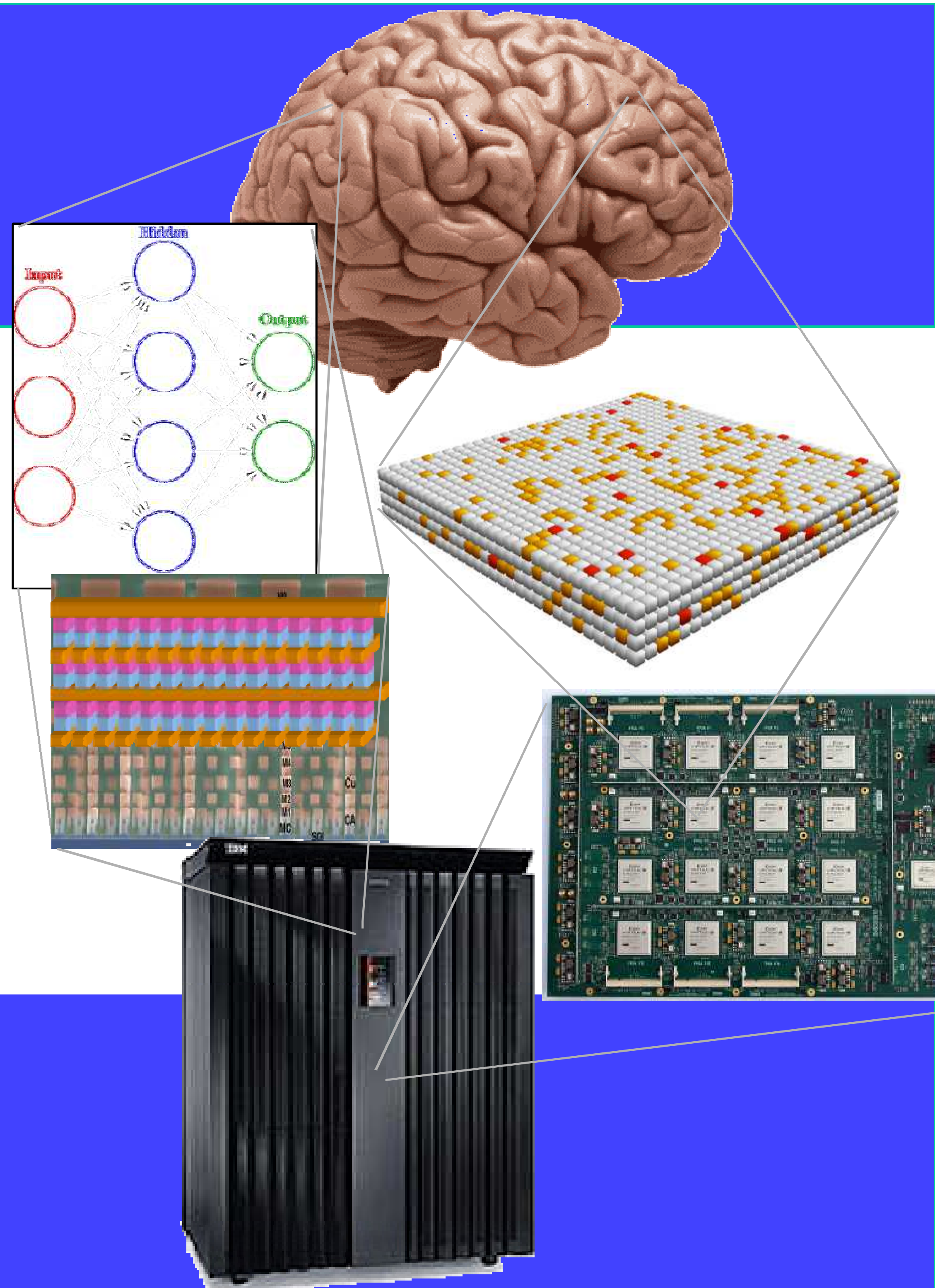
Crossbar array research @ IBM



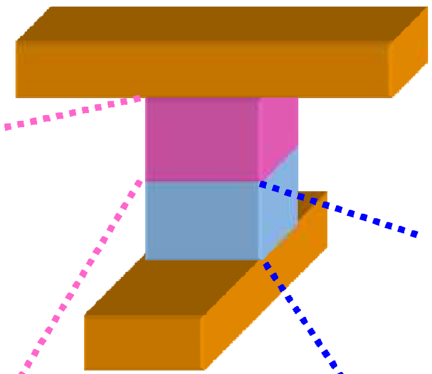
IBM Research – Almaden

Geoffrey W. Burr

July 15, 2016

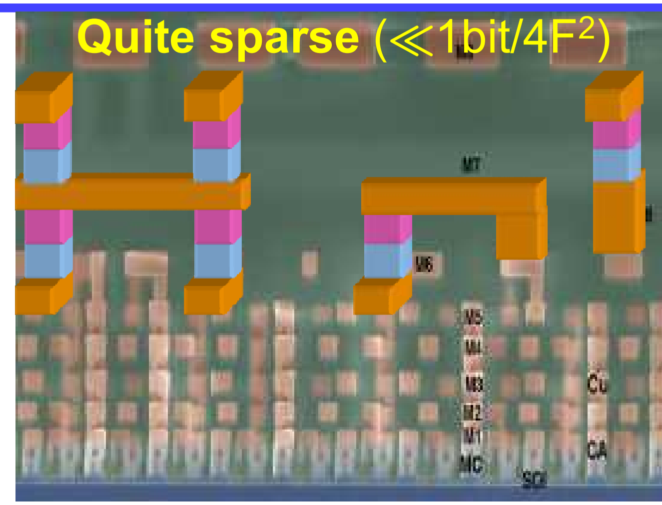
Back-End-Of-the-Line-compatible Non-Volatile Memory:

a fundamental "building block" enabling a range of applications



Access Device
(highly nonlinear or transient switch)

NVM
(long-term digital storage)



Programmable e-fuses
(FPGAs, reconfigurable computing)

Embedded storage
(Automotive)

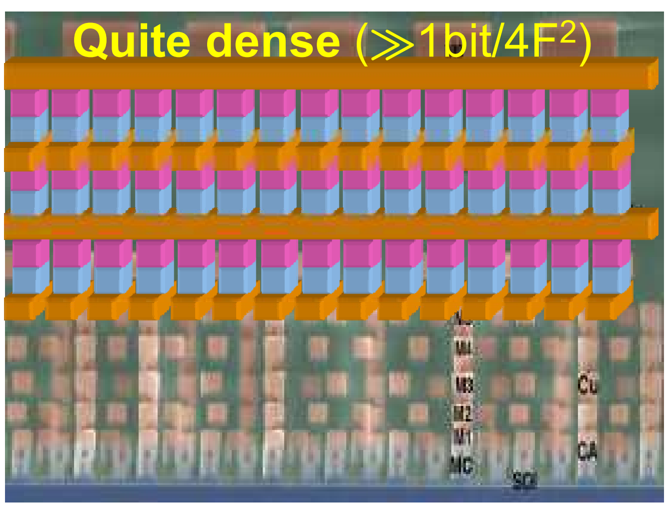
Embedded memory
(Low-power, mobile computing)

Standalone M-class SCM
(Hybrid memory)

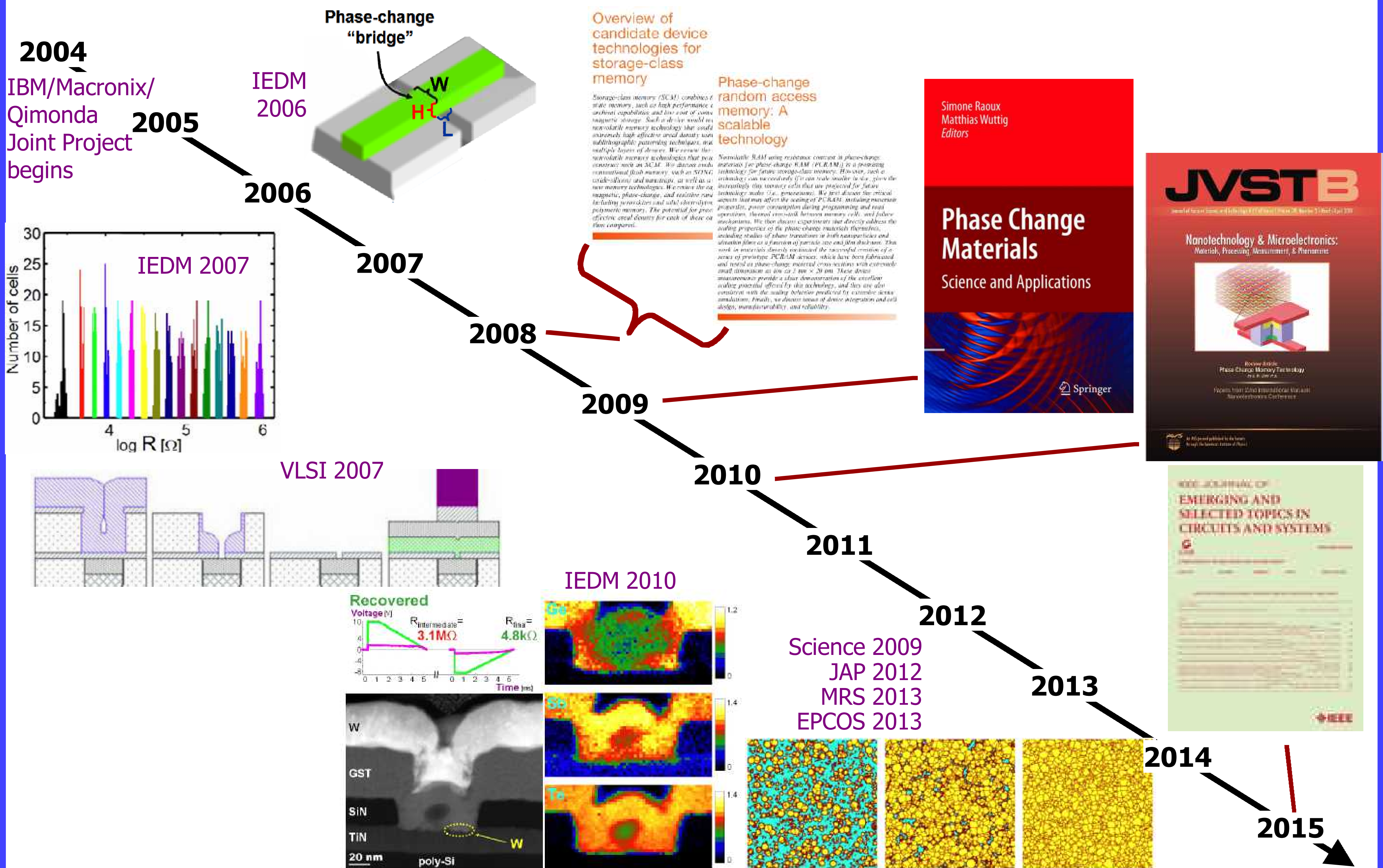
Computation-in-Memory
(Distributed computing)

Standalone S-class Storage Class Memory
(Enhanced Flash)

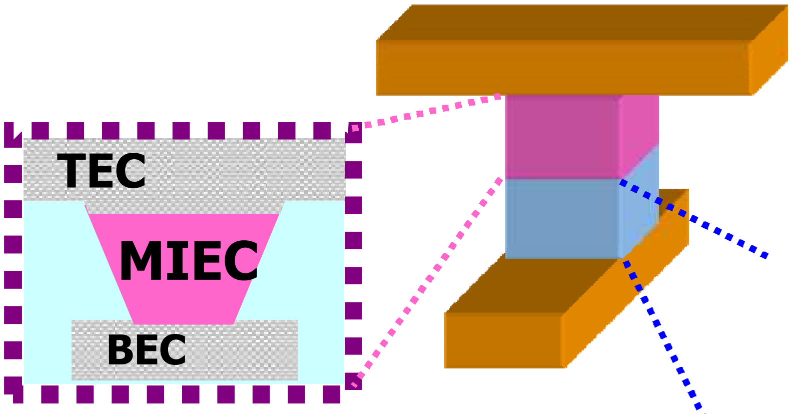
Artificial synapses
(Non-VN Computing)



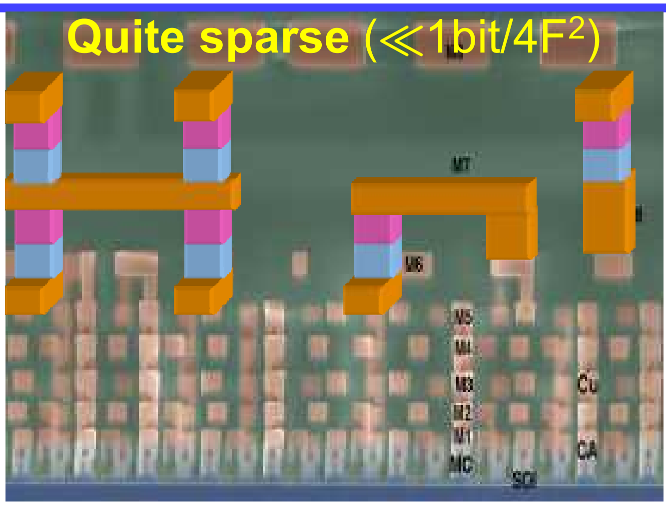
History of **Phase Change Memory** at IBM Almaden



MIEC-based "access device" +NVM: a fundamental, BEOL-compatible "building block"



PCM
RRAM
CBRAM
MRAM



Programmable e-fuses
(FPGAs, reconfigurable computing)

Embedded storage
(Automotive)

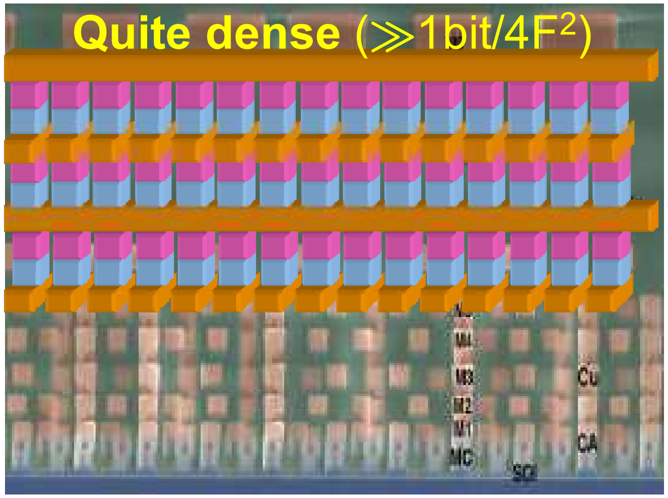
Embedded memory
(Low-power, mobile computing)

Standalone M-class SCM
(Hybrid memory)

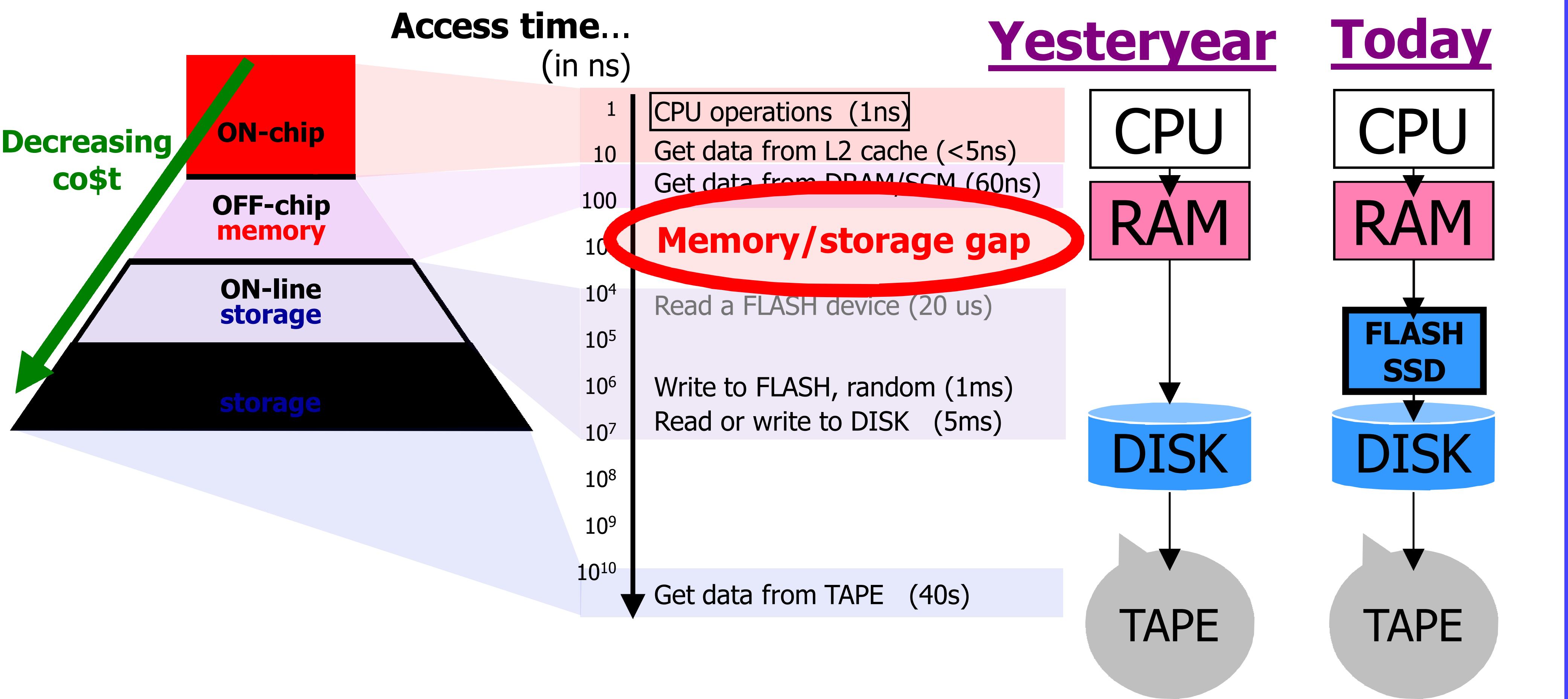
Computation-in-Memory
(Distributed computing)

Standalone S-class SCM
(Enhanced Flash)

Artificial synapses
(Non-VN Computing)



Problem (& opportunity): The access-time gap between memory & storage



- Today, **Solid-State Disks** based on NAND Flash can offer fast ON-line storage, and storage capacities are increasing as devices scale down to smaller dimensions…

…but while prices are dropping, the **performance gap** between memory and storage remains significant, and the already-**poor device endurance** of Flash is getting worse.

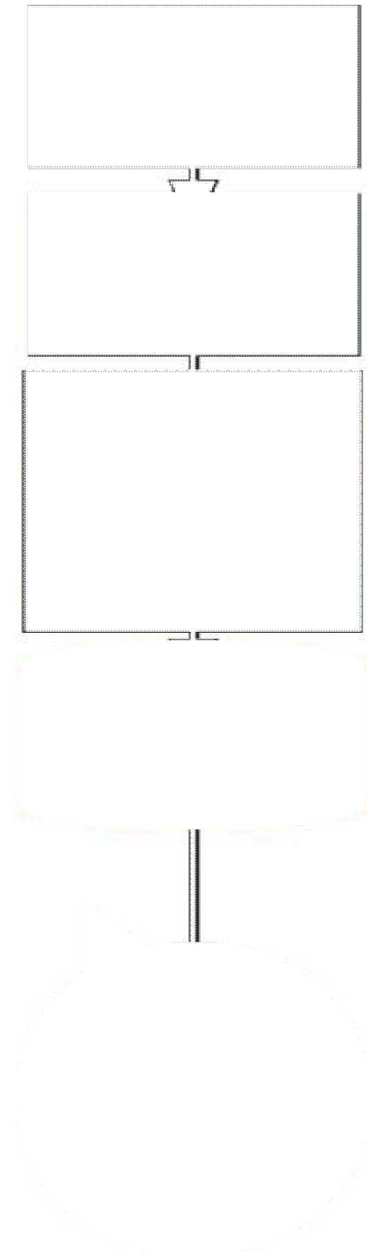
Storage Class Memory (SCM)

DESIRED FEATURES

- **Solid-state** → no moving parts
- **Nonvolatile** → retains data on power-off
- **Fast access speed** → approaching DRAM
- **High endurance** → many program/erase cycles
- **Low cost per bit** → approaching hard disk

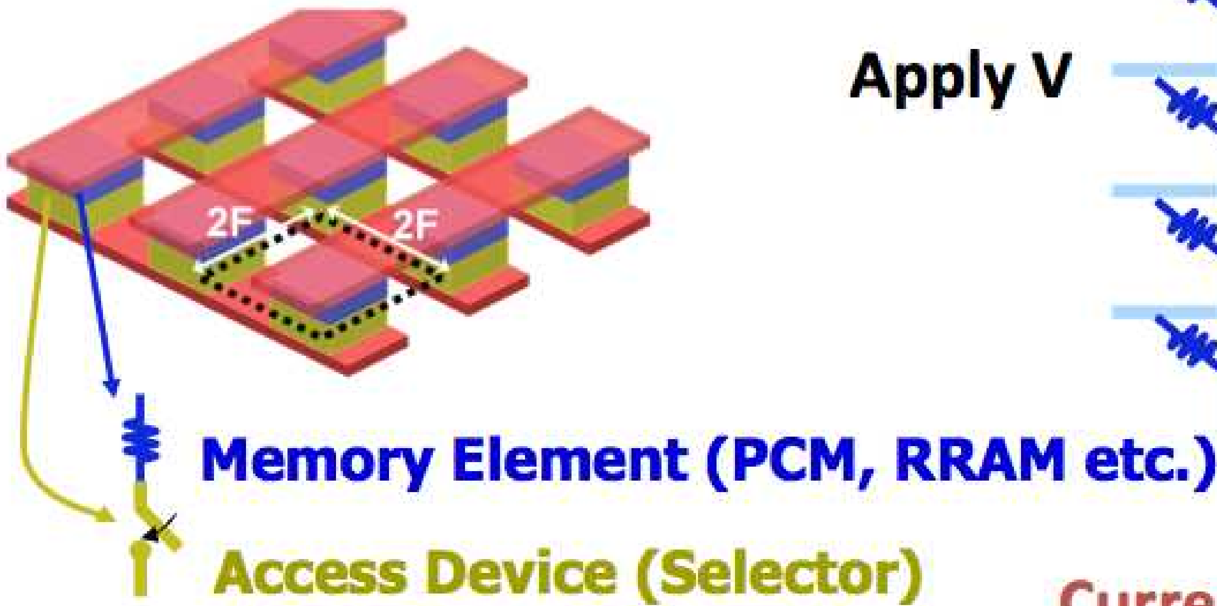
A new class of storage/memory devices that blurs the distinctions between …

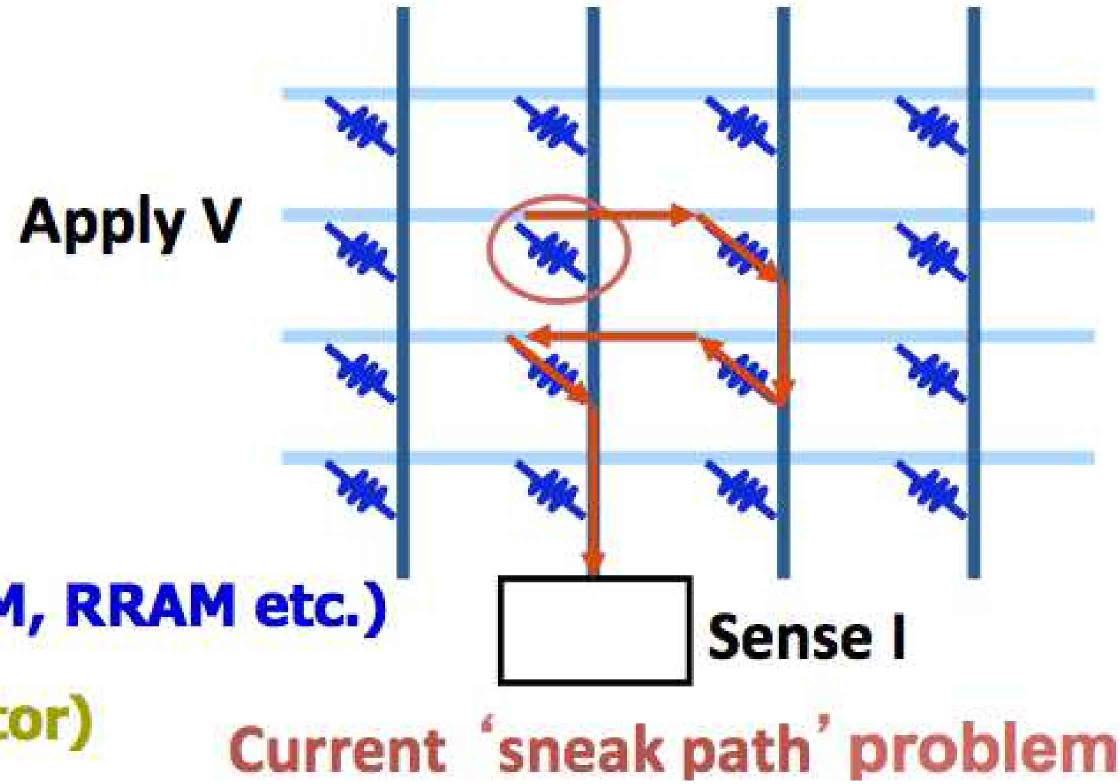
Memory (fast, expensive, volatile)
and
Storage (slow, cheap, nonvolatile)



(Wilcke, USENIX FAST tutorial, 2009)

Need for an Access Device





Access device needed in series with memory element

- Cut off current 'sneak paths'
  that lead to incorrect sensing and wasted power
- Typically diodes used as access devices
- Could also use devices with highly non-linear I-V curves

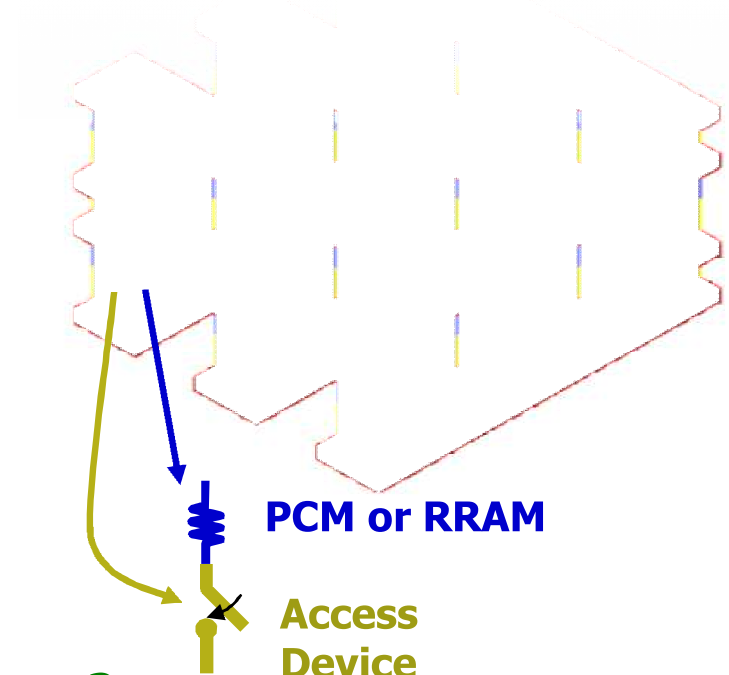
Requirements for an Access Device for 3D Crosspoint Memory

✓ High ON-state current density
>10 MA/cm² for PCM / RRAM RESET

✓ Low OFF-state leakage current
>10⁷ ON/OFF ratio, and
wide low-leakage (**< 100pA**) voltage zone to accommodate half-selected cells in large arrays

✓ Back-End process compatible
<400C processing to allow 3D stacking

✓ Bipolar operation
needed for optimum RRAM operation

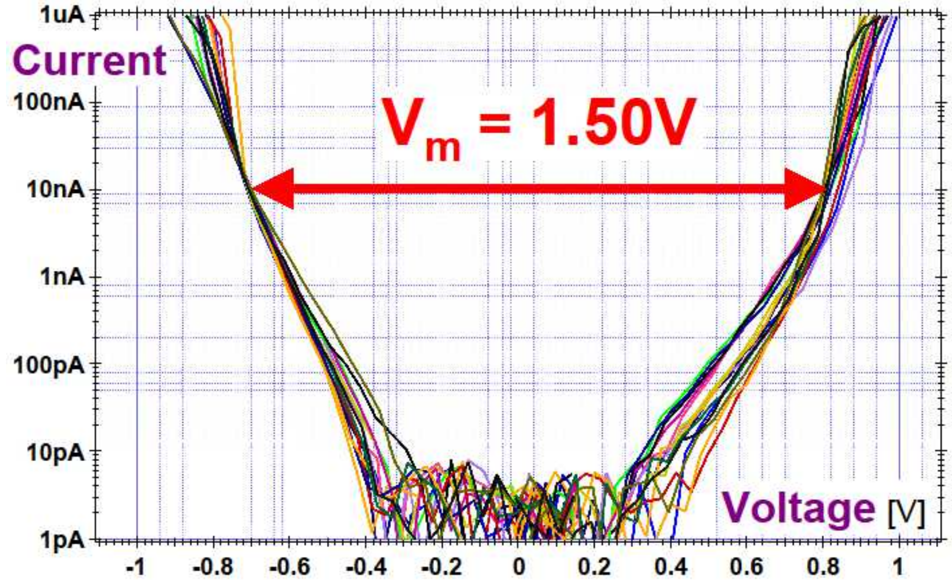


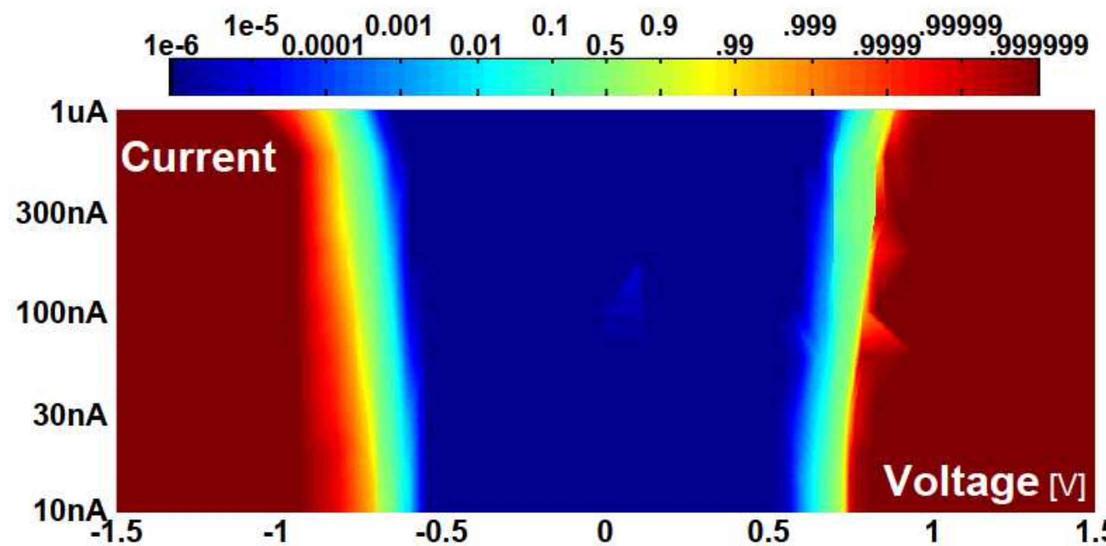
- ✓ variability?
- ✓ yield?
- ✓ co-integration with NVM?
- ✓ turn-ON speed for write?
- ✓ endurance?
- ✓ manufacturability?
- ✓ scalability?
- ✓ long-term leakage?
- ✓ turn-OFF speed?
- ✓ turn-ON speed for read?
- ✓ quantitative modeling?
- ✓ array design (interplay between NVM & selector characteristics)

Novel **M**ixed-**I**onic-**E**lectronic-**C**onduction (MIEC) Access Device

Strengths

- **High** enough **ON currents** for PCM – cycling of PCM has been demonstrated
- **Low** enough **OFF current** for large arrays
- Very large (>>1e10) endurance for typical 5uA read currents
- Voltage margins > 1.5V with tight distributions → sufficient for large arrays
- CMP process demonstrated
- 512kBit arrays demonstrated w/ 100% yield
- Scalable to <30nm CD, <12nm thickness
- Capable of 15ns write, 50ns read
- Highly stable in un-/half-select conditions





Weaknesses

- Maximum voltage across companion NVM during switching must be low (1-2V) → influences half-select condition and thus achievable array size
- **Endurance during NVM programming** is strongly dependent on programming current

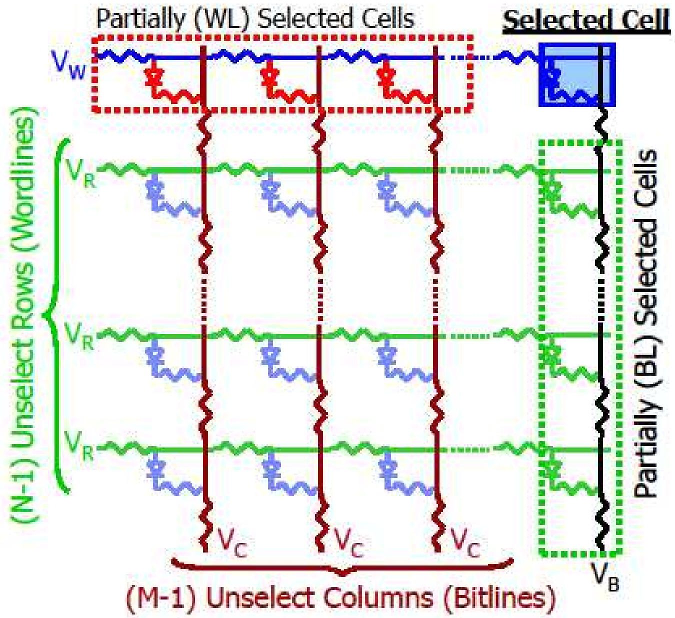
Gopalakrishnan, VLSI 2010
Shenoy, VLSI 2011
Burr, VLSI 2012
Virwani, IEDM 2012
Burr, VLSI 2013

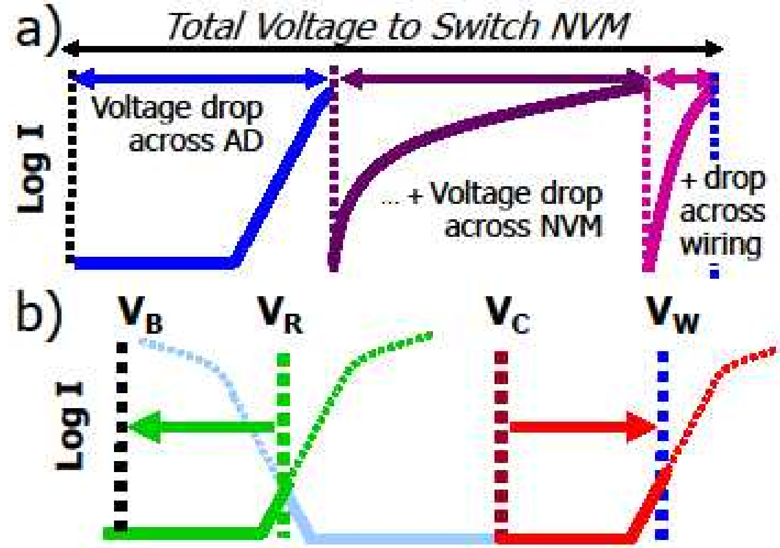
Shenoy, *Semi. Sci. Tech.* **29**/104005 (2014)
Burr, *JVST-B* **32**/040802 (2014)

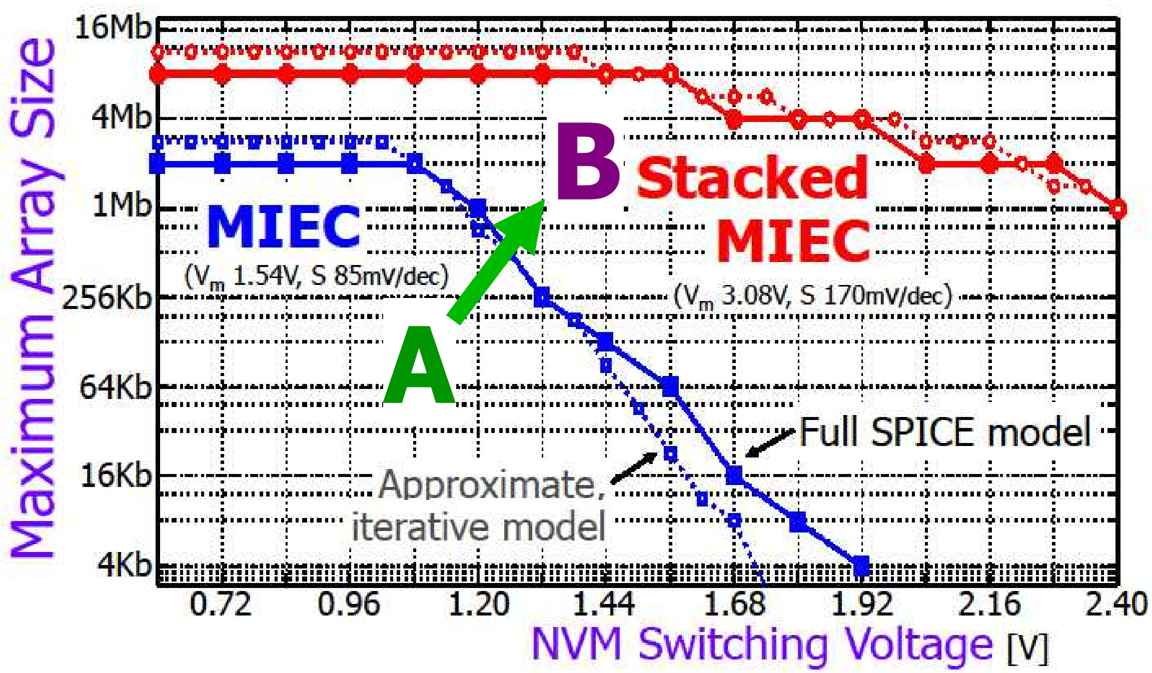
Narayanan, DRC & IEDM 2014, *J-EDS* **3**/423 (2015), *IEEE J.ESTC&S* (2016)

Padilla, IEEE-TED 62/963 (2015)

DRC 2014 – Crossbar array design using SPICE modeling







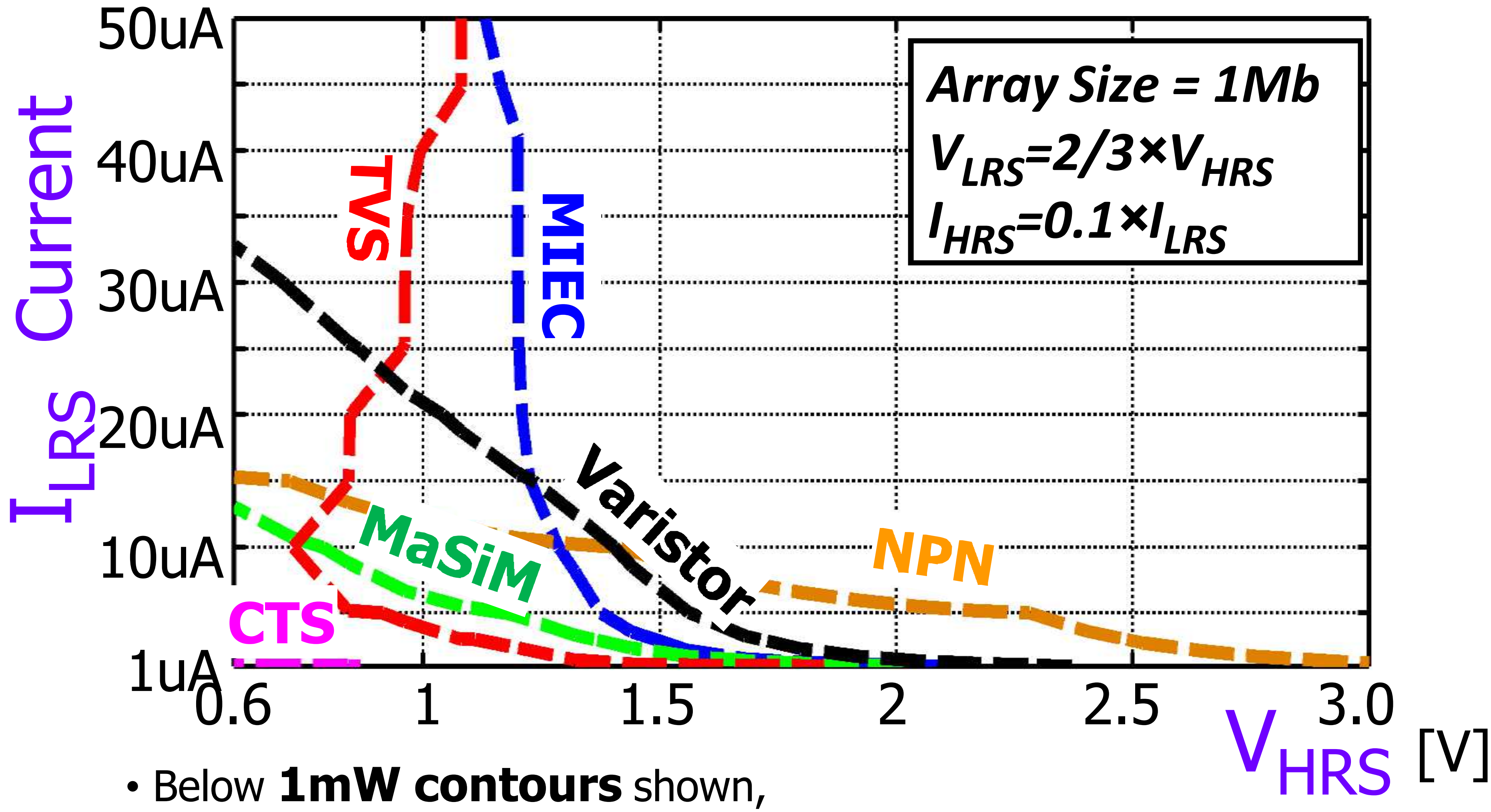
A) Efficient design point:
nearly all injected power delivered to "selected" device(s)

B) Inefficient design point:
much more injected power, which is mostly **dissipated** in "unselected" devices!!

IEDM 2014 paper: compare access devices using SPICE

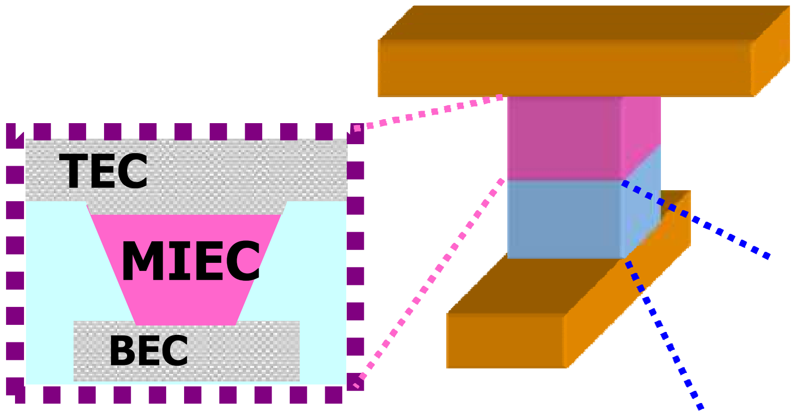
Circuit-Level Benchmarking of Access Devices for Resistive Nonvolatile Memory Arrays

P. Narayanan, G. W. Burr, R. S. Shenoy, K. Virwani, and B. Kurdi

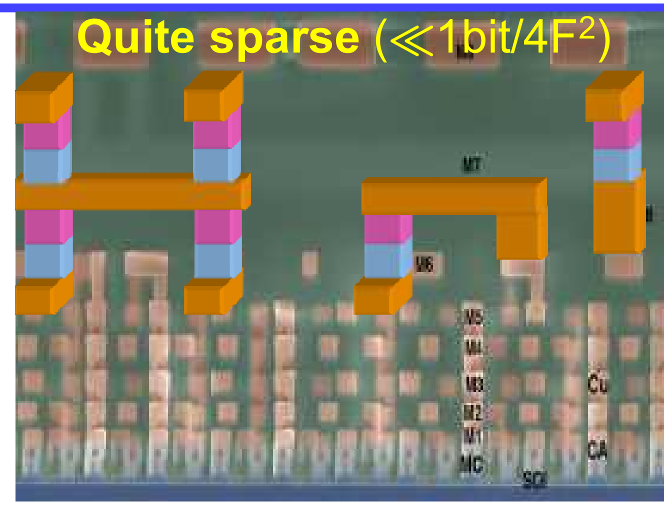


- Below **1mW contours** shown, parallel writes are still a viable option…

MIEC+NVM: a fundamental, BEOL-compatible "building block"



PCM
RRAM
CBRAM
MRAM



Programmable e-fuses
(FPGAs, reconfigurable computing)

Embedded storage
(Automotive)

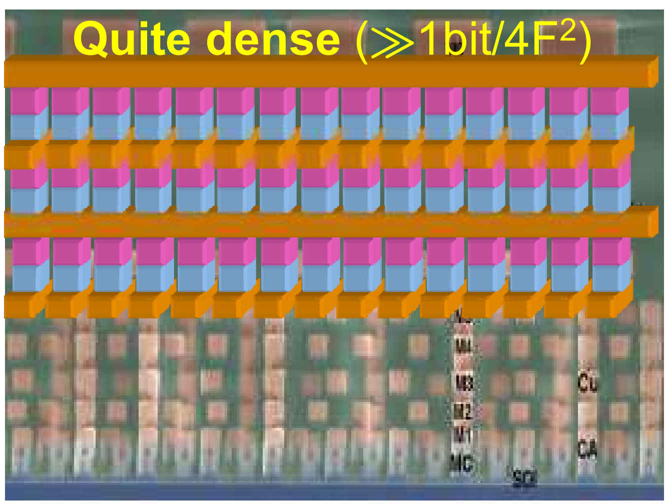
Embedded memory
(Low-power, mobile computing)

Standalone M-class SCM
(Hybrid memory)

Computation-in-Memory
(Distributed computing)

Standalone
S-class SCM
(Enhanced Flash)

Artificial synapses
(Non-VN Computing)



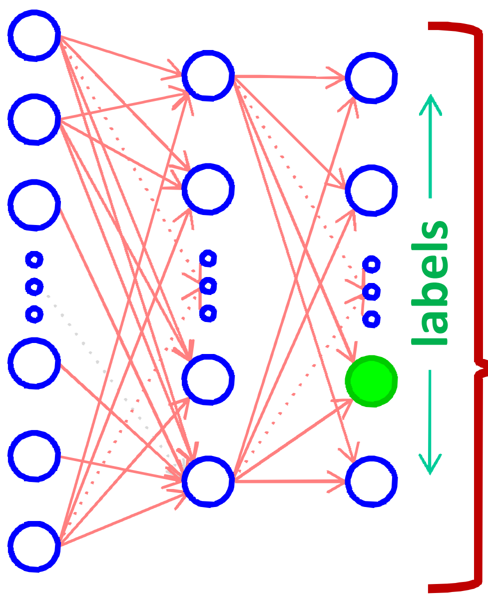
Cognitive computing

systems that learn at scale,
reason with purpose &
interact with humans naturally.

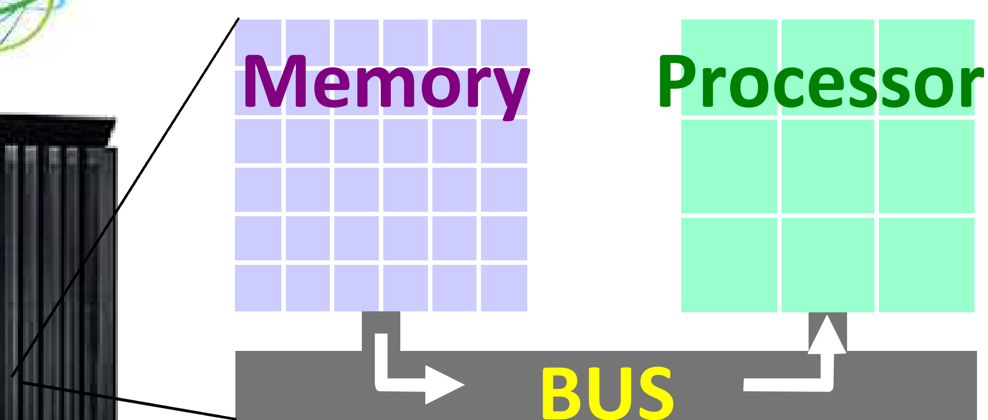
Neuromorphic Devices and Architectures
- **accelerate** today's machine learning

Machine Intelligence
- create flexible systems that **learn continuously**

“Deep” Neural Networks



Von Neumann Architecture



from Oct 2, 2015 IBM whitepaper, “Co… cognition, and the future of knowing" https://ibm.biz/BdHErb

Cognitive computing

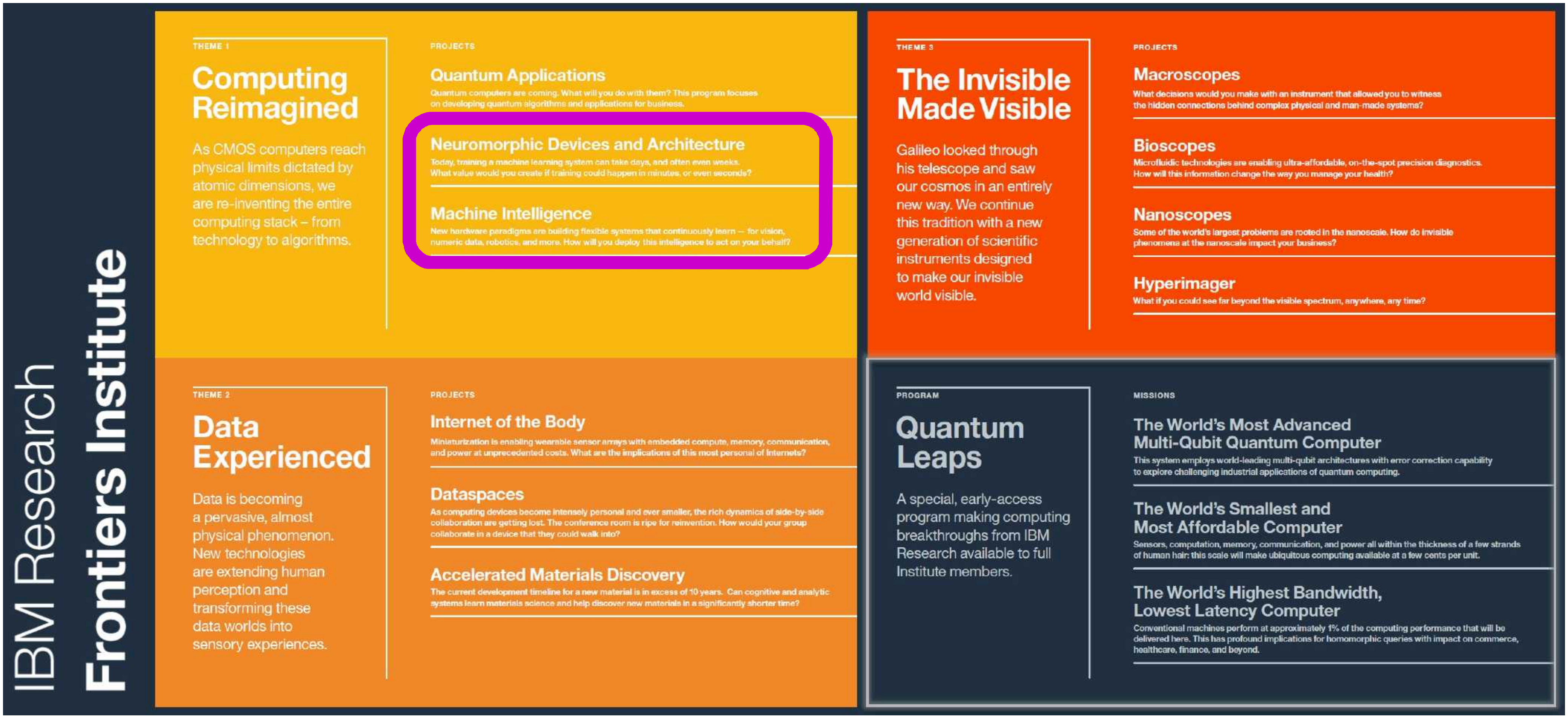
systems that learn at scale,
reason with purpose &
interact with humans naturally.

Neuromorphic Devices and Architectures

- **accelerate** today's machine learning

Machine Intelligence

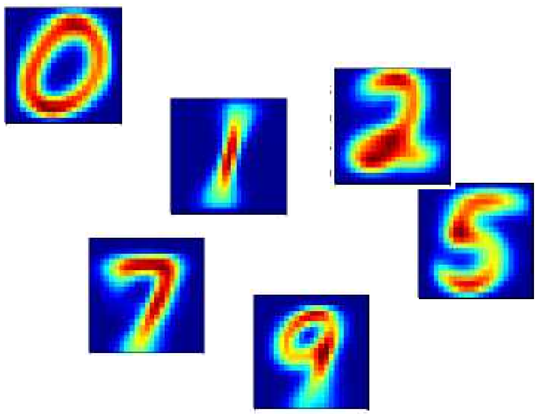
- create flexible systems that **learn continuously**

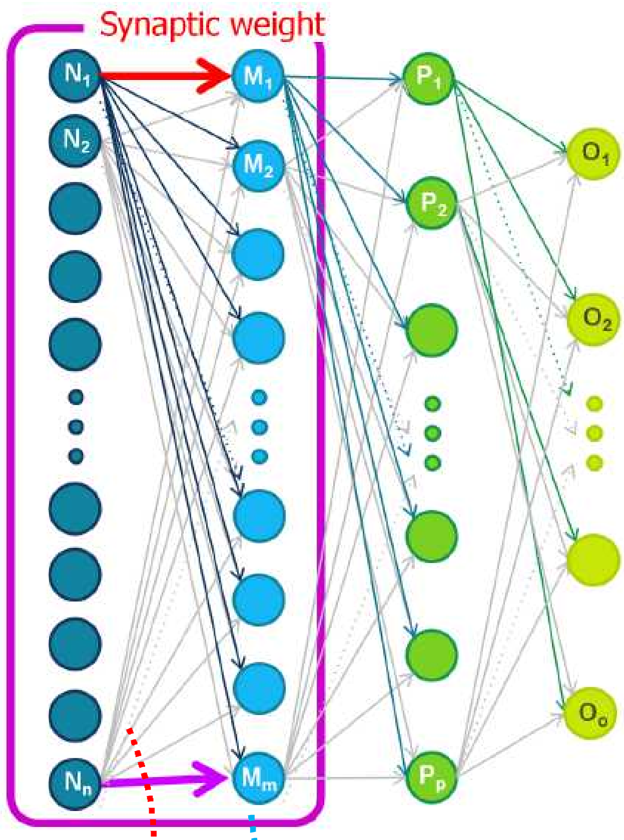


IBM + partner companies

"Deep Learning" on GPUs

1) Input data (images, raw speech data, etc.) input to neural network





2) classification results compared to labels

3) corrections "backpropagated" & all weights updated

Combine 100-1000 input **vectors** into an input **matrix** ***("mini-batch")***

… multiply by current **weight matrix,**

→ excitation into next **hidden neurons**

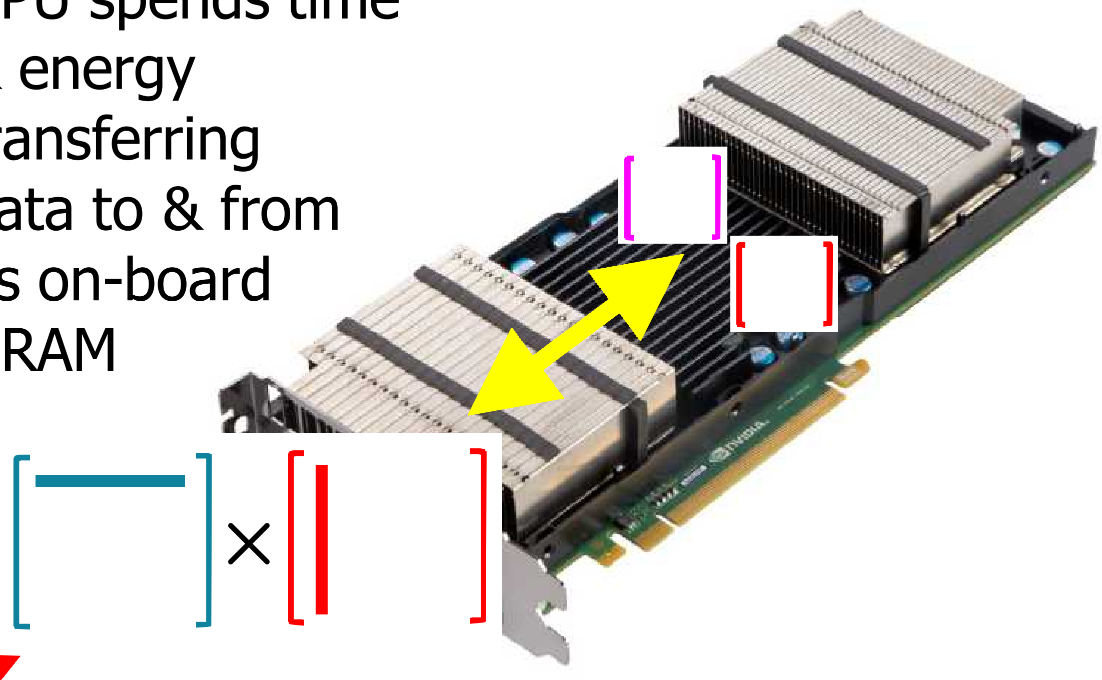
$$[\quad] \times [\quad] = [\quad]$$

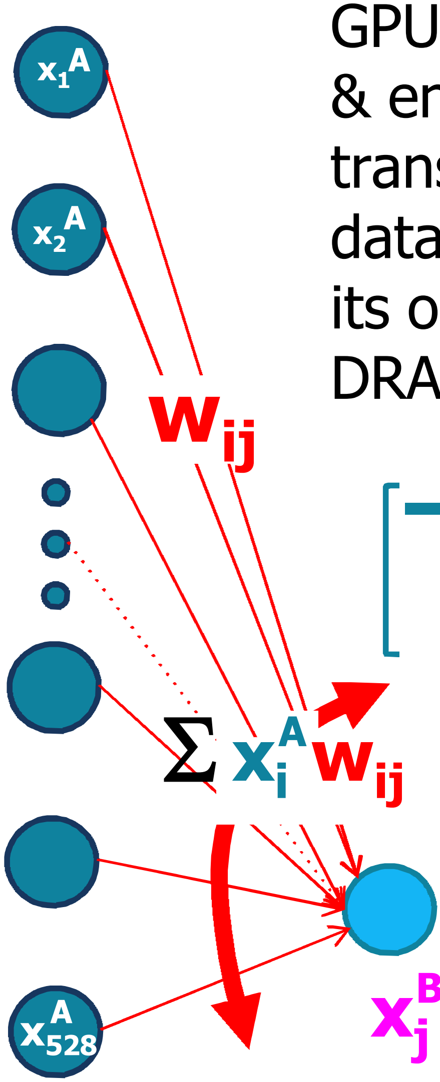
All steps can be mapped to matrix multiplications

→ can run very fast on GPUs

Multiply-accumulate: in GPU matrix-mult, but then **move** data

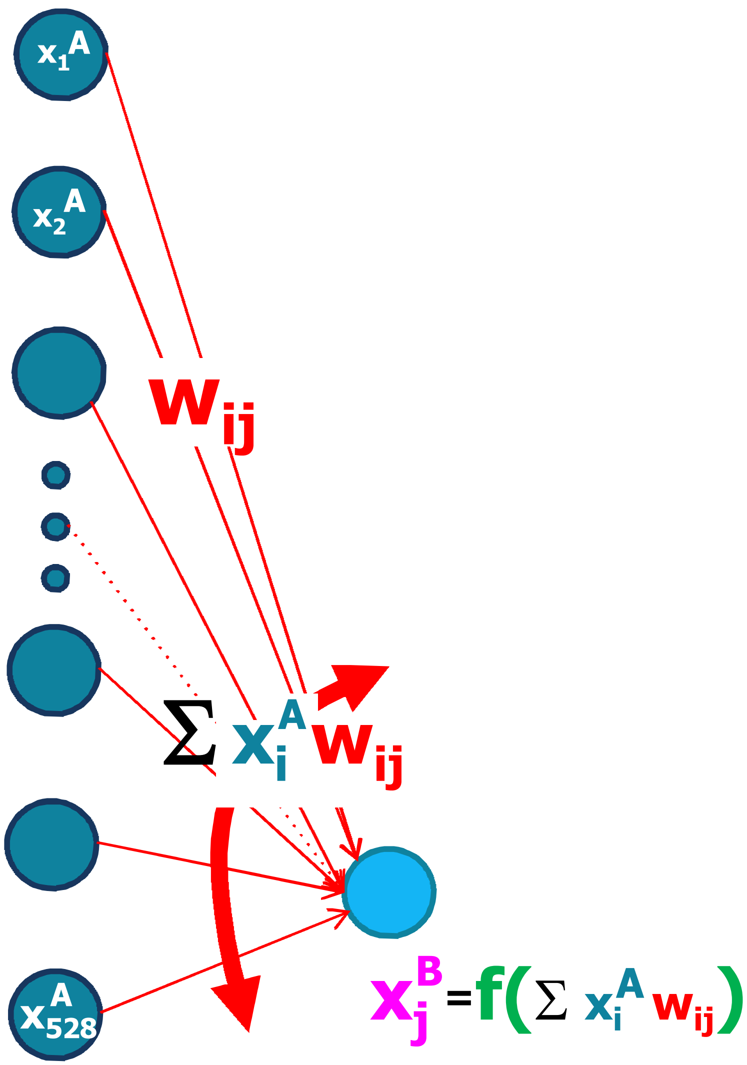
GPU spends time & energy transferring data to & from its on-board DRAM



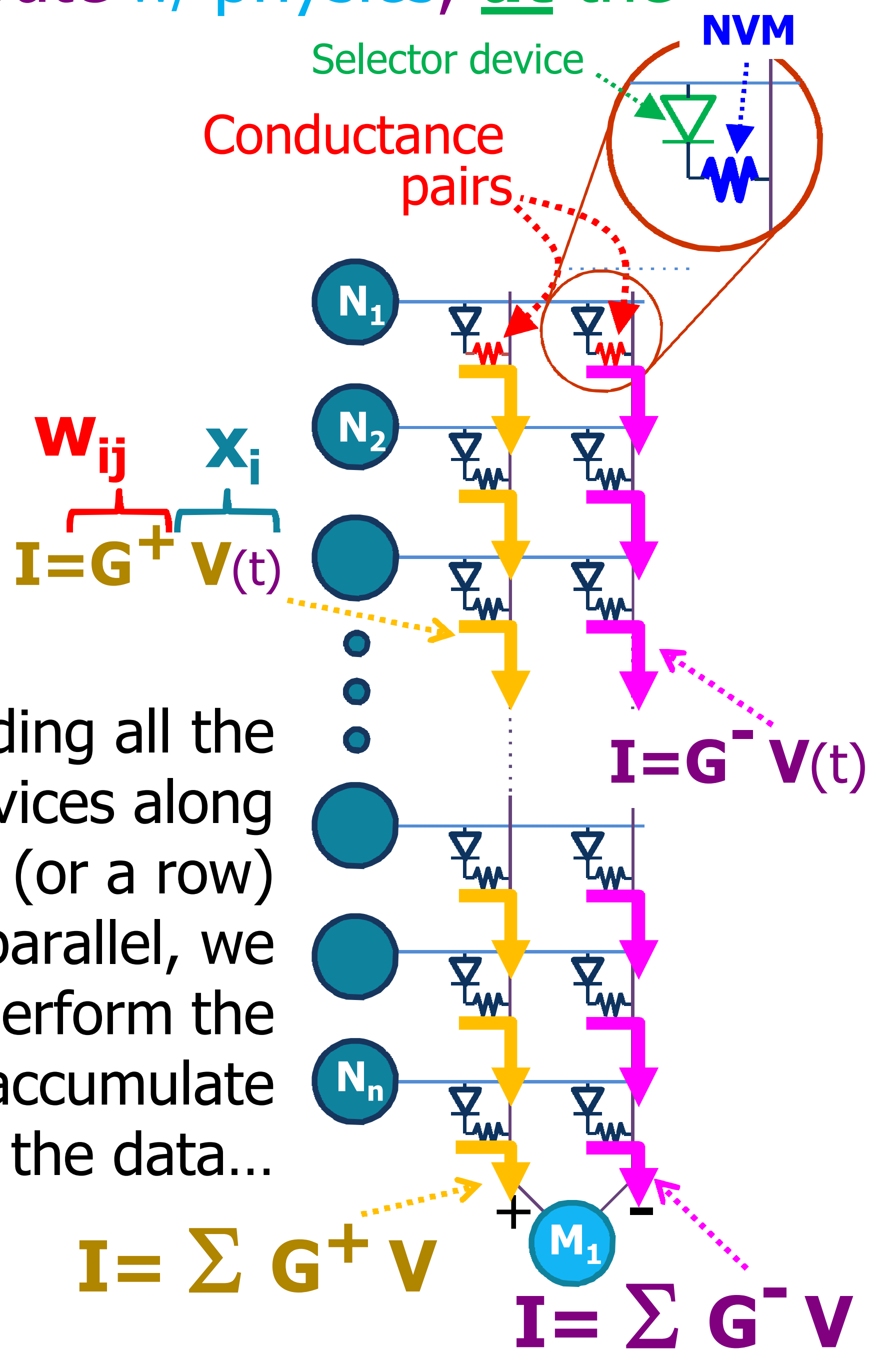


$$x_j^B = f(\sum x_i^A w_{ij})$$

Multiply-accumulate: NVM → compute w/ physics, at the data



By reading all the NVM devices along a column (or a row) in parallel, we perform the multiply-accumulate AT the data…



NVM-for-Machine-Learning

Like TrueNorth: compute **AT** the weight data

Unlike TrueNorth: learning performed **on-chip**

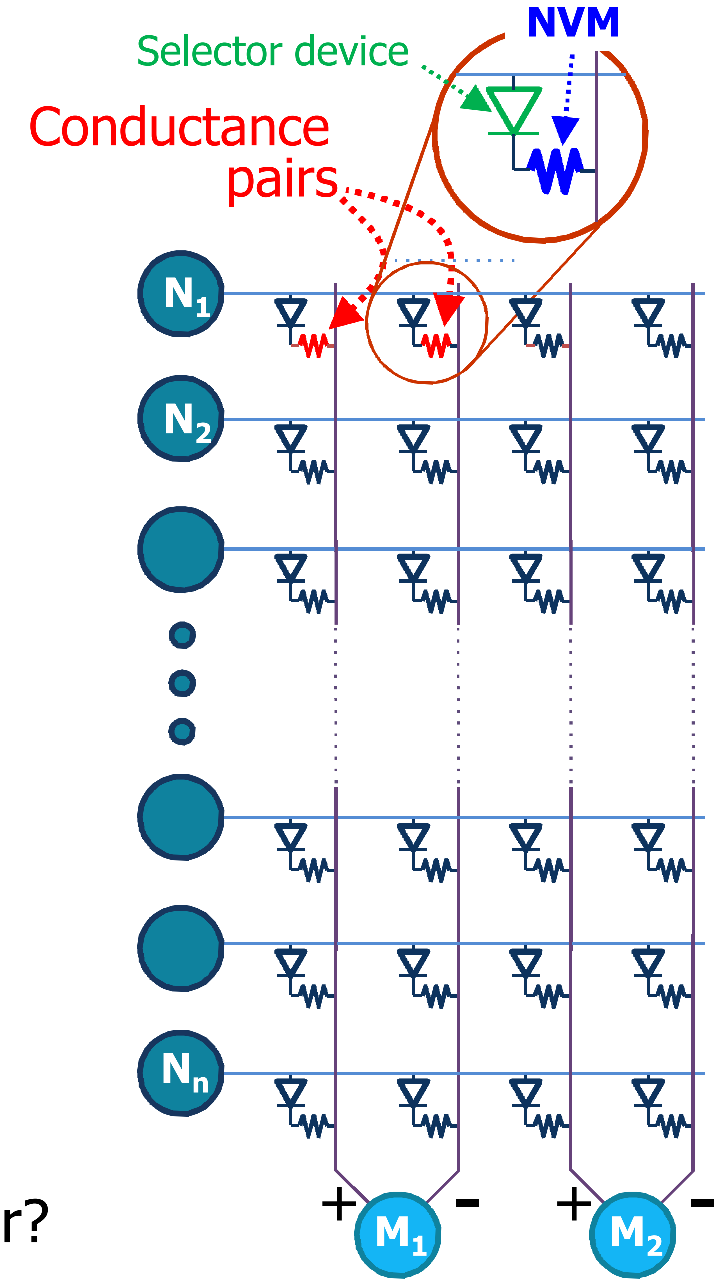
For TrueNorth, **power** is everything

For NVM-for-ML, need **speed-up** over GPUs

Research challenges

1) What do we really need from the NVM devices?
   - Recap of our *IEDM2014*, *IEEE-TED2015* work
     → Need competitive ML performance

2) What are the potential benefits, in speed & power?
   - Speed → Parallelism → Area-efficient circuits



Published work on “what do we need from the NVM?”

[1] **IEDM 2014**

Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element

G. W. Burr, R. M. Shelby, C. di Nolfo, J. W. Jang‡, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, and H. Hwang‡

→ First large-scale mixed hardware-software demonstration + tolerancing

→ ~82% accuracy on MNIST with 5000 examples

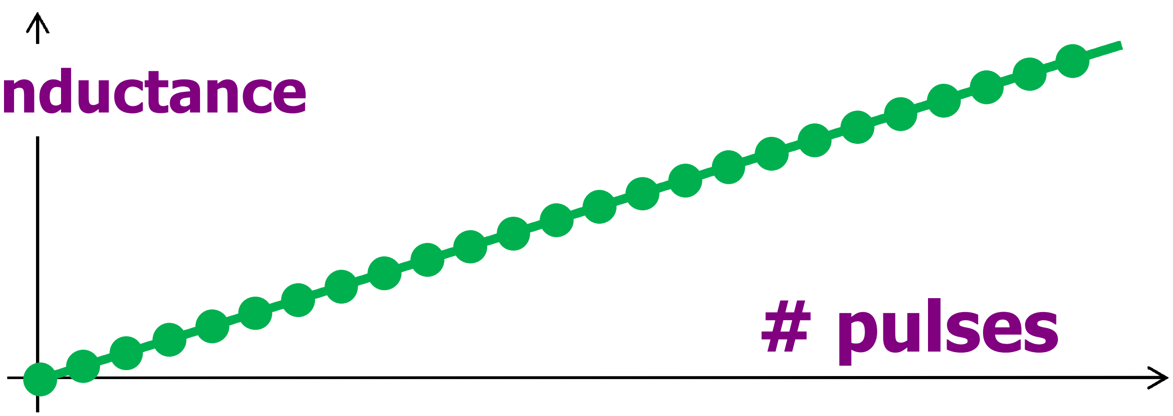
[2] Invited paper in **IEEE-TED** (v**62**(11), 3498 (2015).)

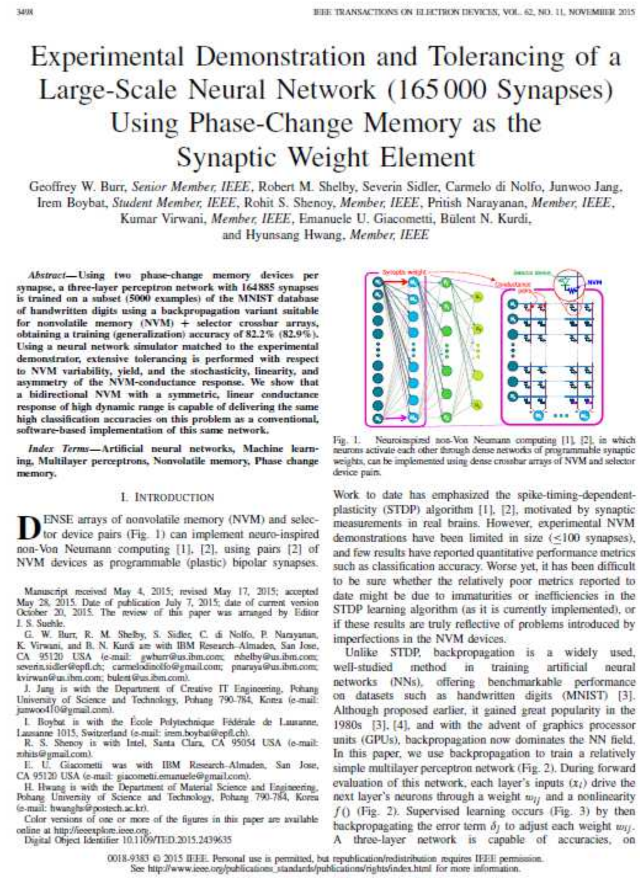
→ Showed that high accuracy

(~94% w/ 5,000 examples,

97-98% w/ 60,000 examples)

is possible – NVM just needs a **linear** conductance response w/ **small** steps





Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element

Geoffrey W. Burr, *Senior Member, IEEE*, Robert M. Shelby, Severin Sidler, Carmelo di Nolfo, Junwoo Jang, Irem Boybat, *Student Member, IEEE*, Rohit S. Shenoy, *Member, IEEE*, Pritish Narayanan, *Member, IEEE*, Kumar Virwani, *Member, IEEE*, Emanuele U. Giacometti, Bülent N. Kurdi, and Hyunsang Hwang, *Member, IEEE*

***Abstract*—Using two phase-change memory devices per synapse, a three-layer perceptron network with 164885 synapses is trained on a subset (5000 examples) of the MNIST database of handwritten digits using a backpropagation variant suitable for nonvolatile memory (NVM) + selector crossbar arrays, obtaining a training (generalization) accuracy of 82.2% (82.9%). Using a neural network simulator matched to the experimental demonstrator, extensive tolerancing is performed with respect to NVM variability, yield, and the stochasticity, linearity, and asymmetry of the NVM-conductance response. We show that a bidirectional NVM with a symmetric, linear conductance response of high dynamic range is capable of delivering the same high classification accuracies on this problem as a conventional, software-based implementation of this same network.**

***Index Terms*—Artificial neural networks, Machine learning, Multilayer perceptrons, Nonvolatile memory, Phase change memory.**

I. Introduction

DENSE arrays of nonvolatile memory (NVM) and selector device pairs (Fig. 1) can implement neuro-inspired non-Von Neumann computing [1], [2], using pairs [2] of NVM devices as programmable (plastic) bipolar synapses.

Manuscript received May 4, 2015; revised May 17, 2015; accepted May 28, 2015. Date of publication July 7, 2015; date of current version October 20, 2015. The review of this paper was arranged by Editor J. S. Suehle.

G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, P. Narayanan, K. Virwani, and B. N. Kurdi are with IBM Research–Almaden, San Jose, CA 95120 USA (e-mail: gwburr@us.ibm.com; rshelby@us.ibm.com; severin.sidler@epfl.ch; carmelodinolfo@gmail.com; pnaraya@us.ibm.com; kvirwan@us.ibm.com; bulent@us.ibm.com).

J. Jang is with the Department of Creative IT Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea (e-mail: junwoo410@gmail.com).

I. Boybat is with the École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: irem.boybat@epfl.ch).

R. S. Shenoy is with Intel, Santa Clara, CA 95054 USA (e-mail: rohits@gmail.com).

E. U. Giacometti was with IBM Research–Almaden, San Jose, CA 95120 USA (e-mail: giacometti.emanuele@gmail.com).

H. Hwang is with the Department of Material Science and Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea (e-mail: hwanghs@postech.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2015.2439635

Fig. 1. Neuroinspired non-Von Neumann computing [1], [2], in which neurons activate each other through dense networks of programmable synaptic weights, can be implemented using dense crossbar arrays of NVM and selector device pairs.

Work to date has emphasized the spike-timing-dependent-plasticity (STDP) algorithm [1], [2], motivated by synaptic measurements in real brains. However, experimental NVM demonstrations have been limited in size (≤100 synapses), and few results have reported quantitative performance metrics such as classification accuracy. Worse yet, it has been difficult to be sure whether the relatively poor metrics reported to date might be due to immaturities or inefficiencies in the STDP learning algorithm (as it is currently implemented), or if these results are truly reflective of problems introduced by imperfections in the NVM devices.

Unlike STDP, backpropagation is a widely used, well-studied method in training artificial neural networks (NNs), offering benchmarkable performance on datasets such as handwritten digits (MNIST) [3]. Although proposed earlier, it gained great popularity in the 1980s [3], [4], and with the advent of graphics processor units (GPUs), backpropagation now dominates the NN field. In this paper, we use backpropagation to train a relatively simple multilayer perceptron network (Fig. 2). During forward evaluation of this network, each layer's inputs (x_i) drive the next layer's neurons through a weight w_{ij} and a nonlinearity $f()$ (Fig. 2). Supervised learning occurs (Fig. 3) by then backpropagating the error term δ_j to adjust each weight w_{ij}. A three-layer network is capable of accuracies, on

Published work on "what do we need from the NVM?"

[1] **IEDM 2014**

Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element

G. W. Burr, R. M. Shelby, C. di Nolfo, J. W. Jang‡, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, and H. Hwang‡

→ First large-scale mixed hardware-software demonstration + tolerancing
→ ~82% accuracy on MNIST with 5000 examples

[2] Invited paper in **IEEE-TED** (v**62**(11), 3498 (2015).)

→ Showed that high accuracy
(~94% w/ 5,000 examples,
97-98% w/ 60,000 examples)
is possible – NVM just needs a **linear** conductance response w/ **small** steps

[3] Invited talk **@IEDM 2015** (Neuromorphic Focus Session)

Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: comparative performance analysis (accuracy, speed, and power)

G. W. Burr, P. Narayanan, R. M. Shelby, S. Sidler, I. Boybat, C. di Nolfo, and Y. Leblebici†

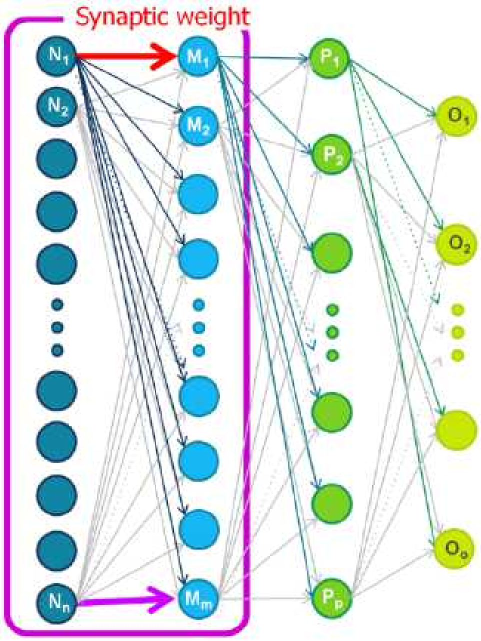
→ showed prospects for speedup (up to 25x) and lower power (100x to 3000x)

Summary of NVM-for-Machine-Learning

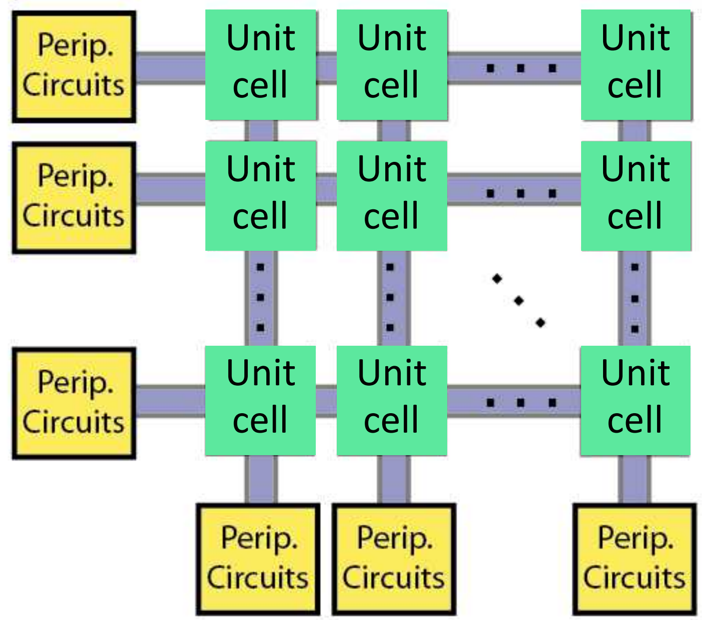
- **NVM-based crossbar arrays CAN accelerate Machine Learning** compared to GPU-based training
  - → Multiply-accumulate performed AT the data
  - → Prospect for **25x speedup** & **120-2850x lower power**
- **Need: competitive ML accuracy**
  - ✓ **experimental results**: ~82% on "minor-league" MNIST using PCM
  - ✓ "ideal" NVM w/ linear G-response of high dynamic range → sufficient!
  - → Our plan: better NVM + innovations to protect network from real NVM
- **Need: area-efficient peripheral circuitry**
  - ✓ power benefits are quite significant
  - ✓ but **design must preserve** speedup benefits
    - → **Aggressive** timing & **minimal circuit sharing**
- **More rigorous power/speed analysis** → based on real circuit designs
- **Flexible, reconfigurable interconnectivity** between arrays
- Need to also support **convolutional neural networks**

IBM Research – multiple paths to faster ML training

<u>Accelerate</u> backpropagation training (e.g., Deep-NN, Conv-NN, and LSTM)…



…by performing **multiply-accumulates** *<u>on-chip</u>* using ***analog*** resistive memory elements.



Unit cell =

Existing NVM
(e.g., PCM, "PCMO")

- Available now
- Truly non-volatile
- Compact cell
- Nonlinear + asymmetric

Capacitors
(CMOS-RPU)

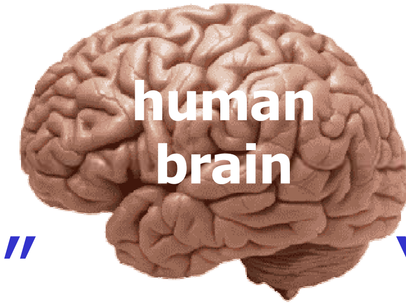
- Available now
- Leaky → need refresh?
- Larger cell
- Suitably linear

Improved NVM
(Device-RPU)

- Yet to be developed
- Non-volatile
- Compact cell
- Linearity is key (asymmetry can be dealt with)

Tayfun Gokmen (IBM Yorktown)
Seyoung Kim (IBM Yorktown)

"Machine Learning" vs. "Machine Intelligence"



"Brain-inspired" computing
(1940's understanding of the brain)

"Machine Learning"

solving a specific task on **labeled** data by defining & optimizing an objective function

PRO:

- can follow gradient descent thru backpropagation
  → convergence to "good" solutions
- mapping to matrix manipulation → GPUs!!
- great progress in ML thanks to competitions
  - Many datasets created
  - Focus on **quantifying** performance

CON:

- we're sure the brain doesn't do backpropagation
- can only handle **static, labelled** data
- insistence on quantifying performance
  may now be stifling innovation

"Brain-inspired" computing
(modern understanding of the brain)

"Machine Intelligence"

flexible systems that continuously learn from **unlabeled** data, and that perform (motor) actions, predict consequences of those actions, and then plan ahead to reach goals

PRO:

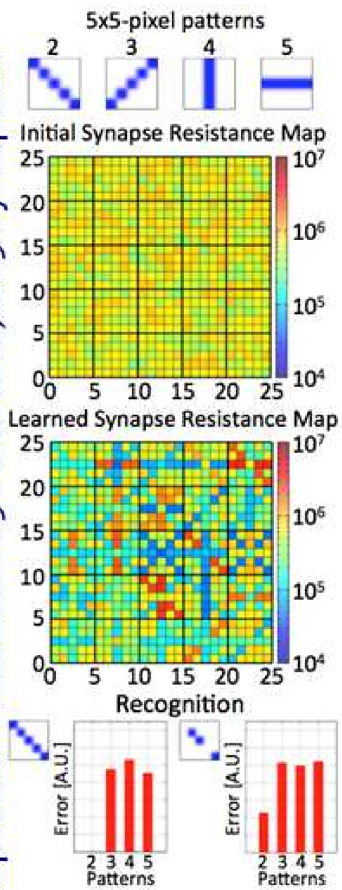
- we're sure this is what the brain does
- MI should be able to handle
  unlabelled & **temporal** data
- MI should enable continuous learning

CON:

- we don't know (yet) how the brain guarantees
  robust, stable convergence in learning
- we have to figure out how to appropriately
  quantify "performance"

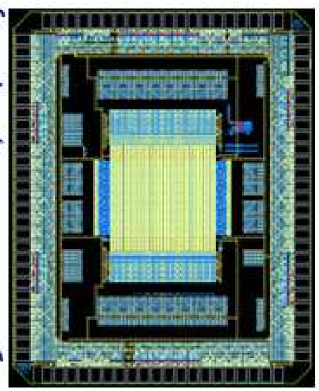
Smart Memory Roadmap

PCM-based Smart Memory Concept
Spice Simulation: 25 neurons, 625 synapses



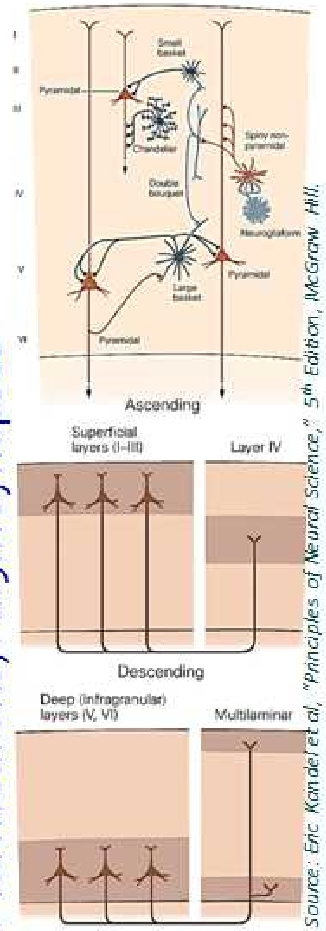
2013

90nm CMOS Bat'leth
256 neurons, 64K synapses



2014

90nm CMOS Bárðarbunga
Brain-like microcircuit wiring
>10K neurons, >2.5M synapses



???

Introducing your Personal Assistance in Smart Memory

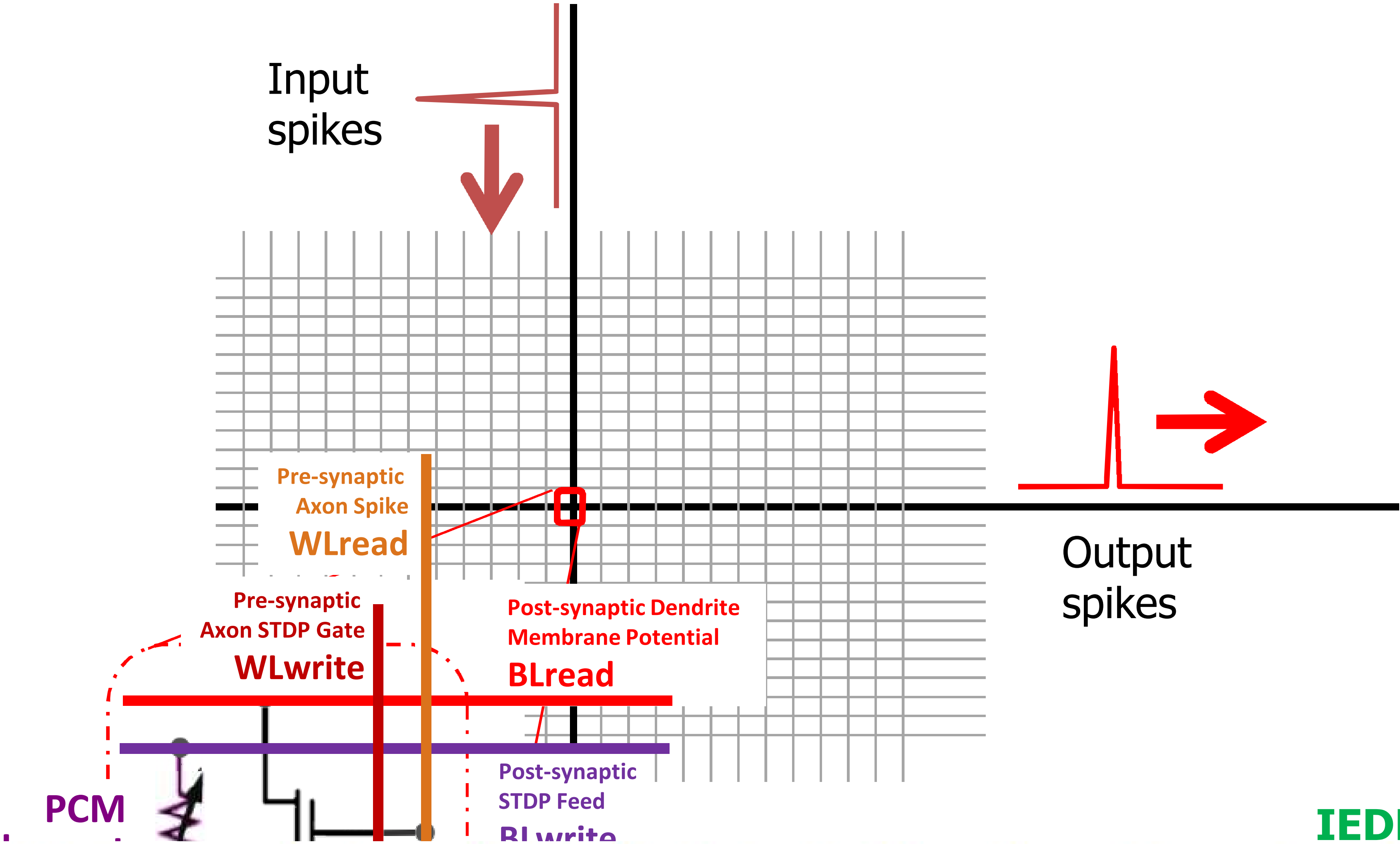
64-bit architecture
Motion coprocessor
4GB Mobile DRAM
256GB Flash Memory
……
Smart Memory with
~1M neurons
~256M synapses
……

???

• Spike-Timing-Dependent-Plasticity (STDP) using Phase Change Memory

Chung Lam (*clam@us.ibm.com*)
Sangbum Kim *(sangbum.kim@us.ibm.com)*

2T1R PCM design for **S**pike-**T**iming-**D**ependent-**P**lasticity



IEDM2015

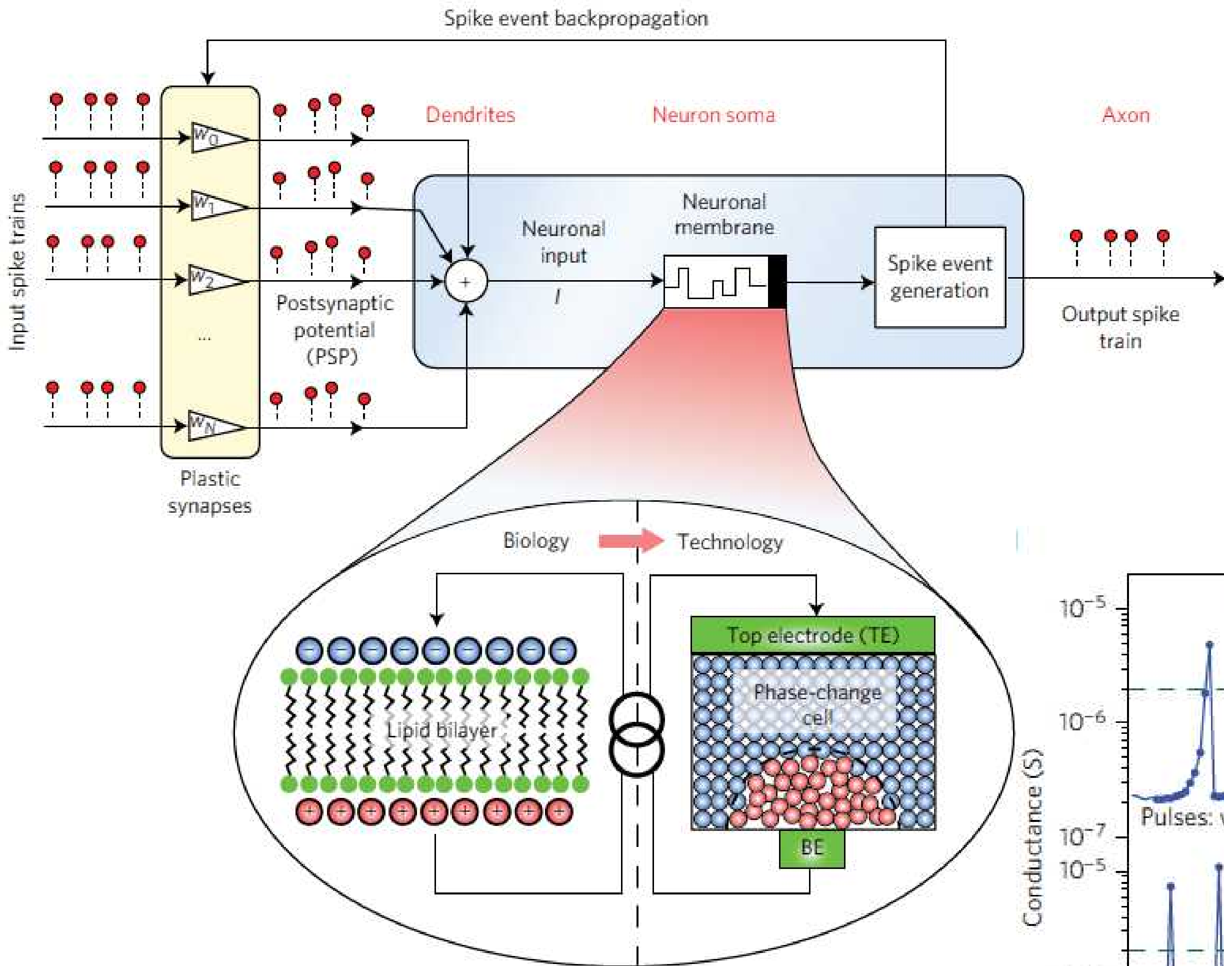
NVM Neuromorphic Core with 64k-cell (256-by-256) Phase Change Memory Synaptic Array with On-Chip Neuron Circuits for Continuous In-Situ Learning

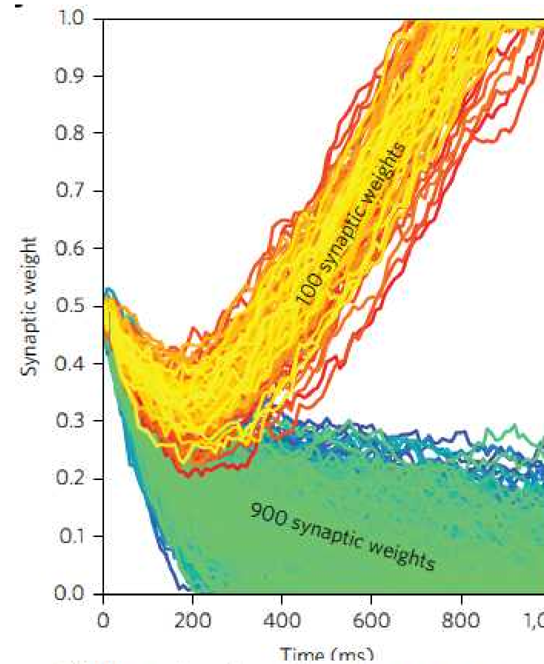
S. Kim, M. Ishii[†], S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. W. Burr[#], N. Sosa, A. Ray, J.-P. Han, C. Miller, K. Hosokawa[†], and C. Lam
IBM T. J. Watson Research Center, 1101 Kitchawan Rd., Yorktown Heights, NY, 10598, USA
[†]IBM Tokyo Research Lab, Tokyo, Japan, [#]IBM Research–Almaden, San Jose, CA, USA.
Tel: +1(914)945-2530, Fax: +1(914)945-4256, email: SangBum.Kim@us.ibm.com

nature nanotechnology
ARTICLES
PUBLISHED ONLINE: 16 MAY 2016 | DOI: 10.1038/NNANO.2016.70

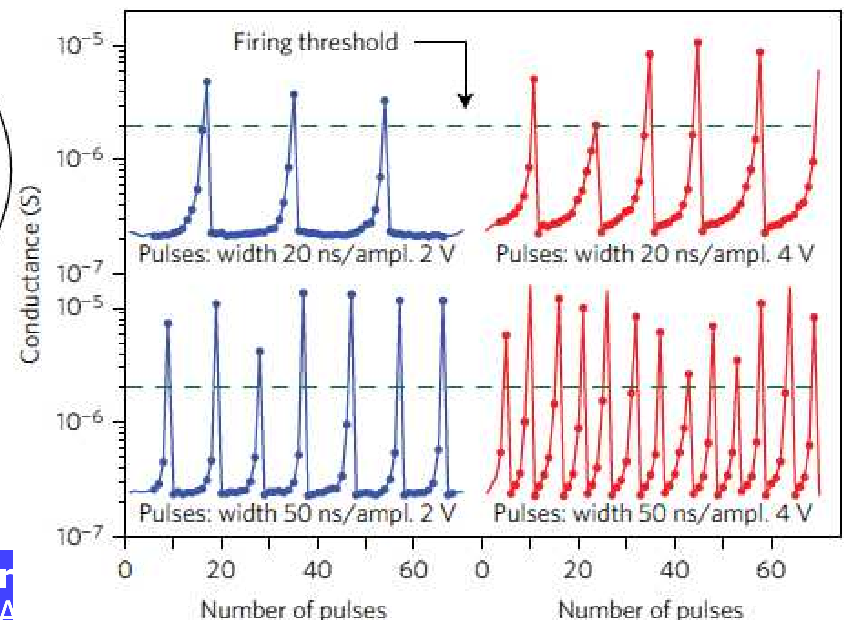
Stochastic phase-change neurons

Tomas Tuma[1]*, Angeliki Pantazi[1], Manuel Le Gallo[1,2], Abu Sebastian[1] and Evangelos Eleftheriou[1]*

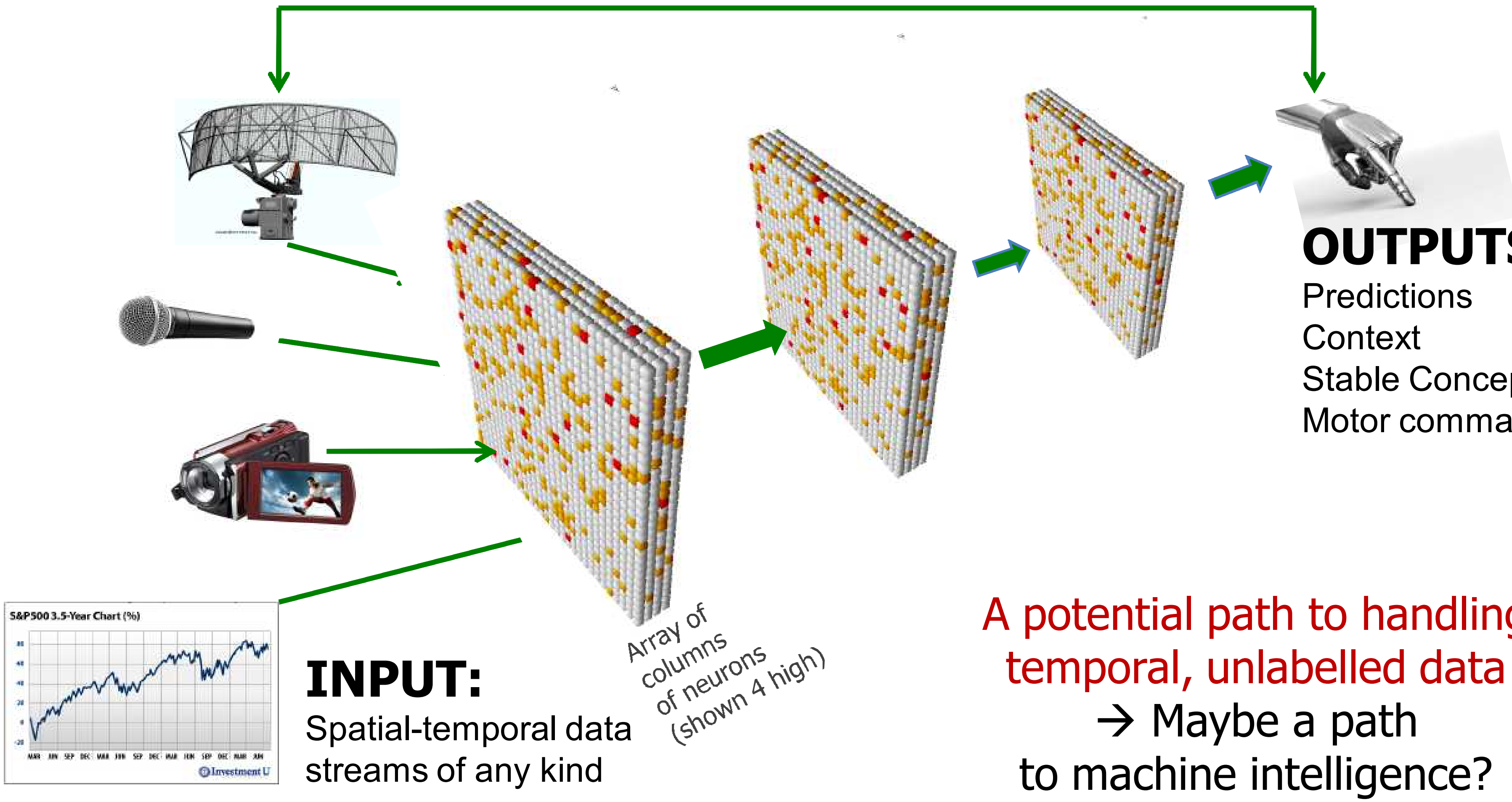




At the steady state, the synapses corresponding to the correlated input streams are potentiated, whereas the synapses corresponding to the uncorrelated input streams are depressed.



Machine Intelligence based on sequences of **S**parse **D**istributed **R**epresentations



OUTPUTS:
Predictions
Context
Stable Concepts (SDR)
Motor commands

INPUT:
Spatial-temporal data streams of any kind

A potential path to handling temporal, unlabelled data
→ Maybe a path to machine intelligence?

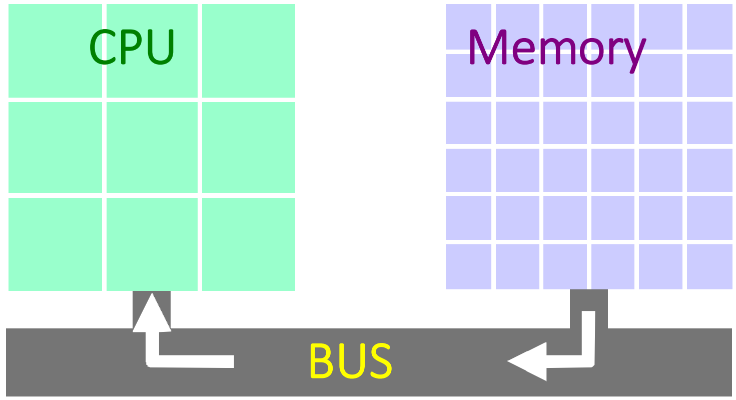
Requires HUGE fanout:
many POTENTIAL synapses
(internally analog, externally binary)

"**C**ontext-**A**ware **L**earning"

winfriedwilcke@us.ibm.com

Von Neumann architecture: aspects we're likely to miss (a LOT)

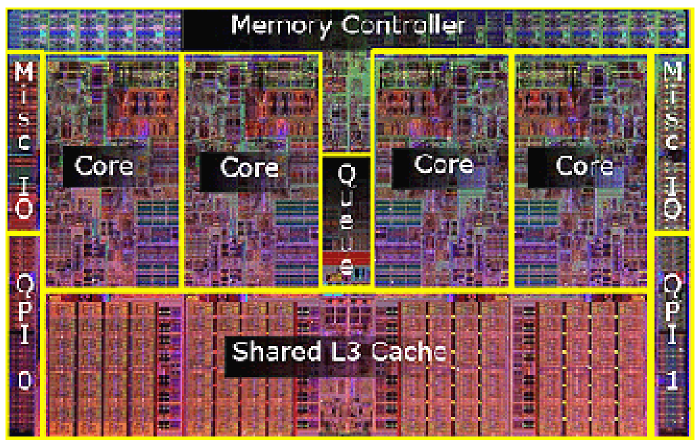
1) Programmable
→ adaptable



OK, fine: let's research devices to enable energy-efficient **non-**Von Neumann architectures

2) Great co$t model

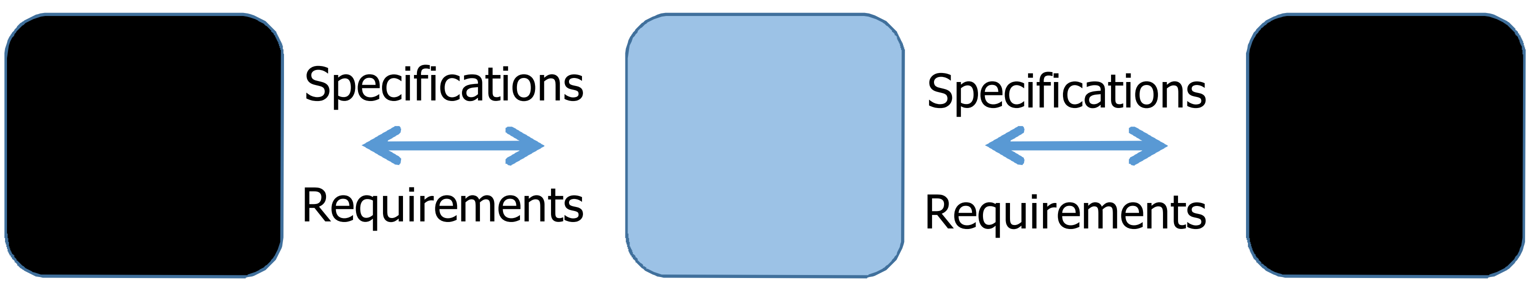
Design 1 piece of hardware…



Sell it to LOTS of people for vastly different purposes…

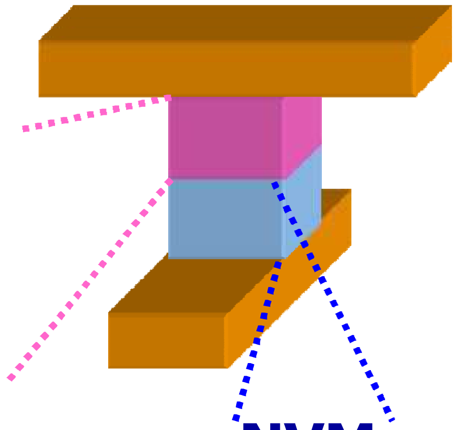
3) Modularity of design

…



…

Research needs for “NanoCrossbars” (1/2)



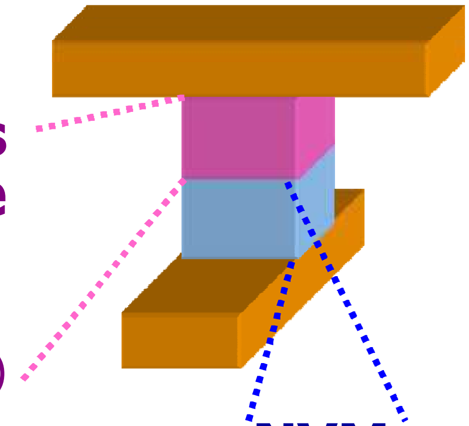
1) **NVM devices**

- any flaws must be addressable by engineering
- good SNR (resistance-range / variability)
- I_{prog} < 50uA, V_{prog} < 2.0V,
  t_{prog} < 10usec, t_{read} << 1usec
- (neuro)
  - When they fail → fail to OPEN (not SHORT)
  - Yield >90-95%

2) **Access devices**

- < 10nA half-select at ($V_{total\text{-}applied}$/2) → can ONLY be evaluated for NVM+AD pair!!
- extremely tight variability
  (variability in nonlinear IV or holding voltage
  → uncertainty in $V_{access\ device}$ at I_{read}, I_{write}
  → loss of read SNR + requires device-overwrite→ endurance-loss)
- When they fail → fail to OPEN (not SHORT)
  –OR– ~100% yield & high endurance

Research needs for "NanoCrossbars" (2/2)



3) **Neuromorphic applications**

- **Accelerating backprop:**
  - NVM devices with LINEAR conductance change, from G_{min} to G_{max}
  - Area-efficient circuit design
  - Methods to protect ANN from nonlinear NVM devices

- **STDP-based NN:** (e.g., spikes for learning not just communication)
  - Killer app that requires learning-from-timing
  - Architecture/global-algorithm that harnesses STDP-like local learning rule for **robust learning** to support/enable above killer-app

- **Machine Intelligence:**
  - Significant algorithm development needed → **too early for crossbars!**

Message: Device researchers who want to have an impact here **MUST** also **learn/know/advance** the **circuits/systems/algorithms** module(s)

NVM-for-Machine Learning: Acknowledgements


Geoffrey Burr


Pritish Narayanan


Bob Shelby


Scott Lewis


Kohji Hosokawa


Masatoshi Ishii


Atsuya Okazaki


Moriyoshi Ohara


Hiroshi Inoue

Collaborators:

Prof. Hyunsang Hwang, POSTECH

Prof. Yusuf Leblebici, EPFL

Students:

Junwoo Jang (now at Samsung)

Kibong Moon

Carmelo di Nolfo (now at IBM Almaden)
Irem Boybat (student @ IBM Zurich)
Severin Sidler (student @ IBM Zurich)

Alessandro Fumarola
Lucas Sanches (from USP, Brazil)