IBM Research
AI Hardware Center

Analog AI @
IBM Research–Almaden
San Jose, CA USA

Accelerating Deep Learning with Analog Memory - A Device, Circuit and Systems Approach

Pritish Narayanan, Geoffrey W. Burr, Stefano Ambrogio, Hsinyu (Sidney) Tsai, Charles Mackin, and An Chen

IBM Almaden – San Jose, CA USA

May 29, 2019

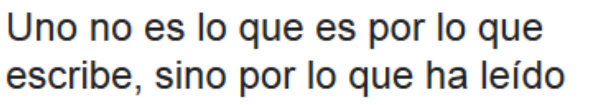
The power of deep neural networks (DNN)

Deep neural networks can solve some problems beyond human level accuracy.

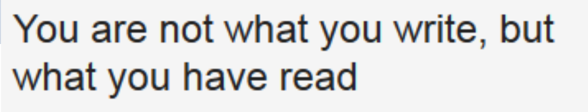
Image recognition:

Speech recognition:

Machine translation:



Uno no es lo que es por lo que escribe, sino por lo que ha leído

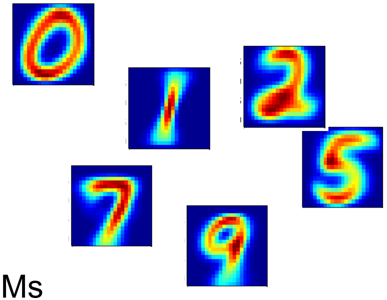


You are not what you write, but what you have read

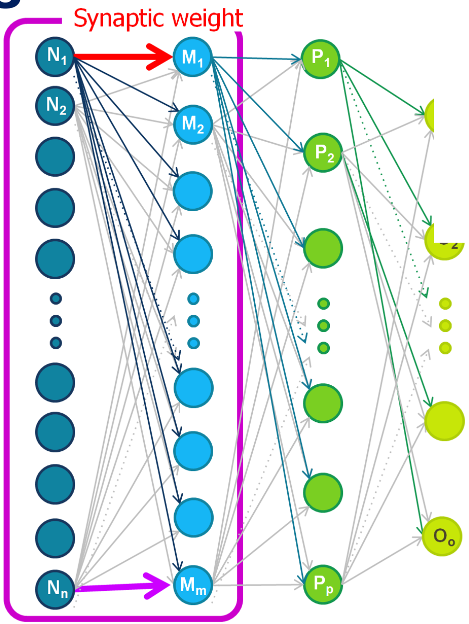
www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Deep Neural Networks

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

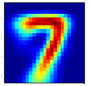


“MNIST” database
~1998
→ check-reading ATMs



A Deep Neural Network contains multiple **layers,** …
each layer containing many **neurons,** …
each neuron driven through many synaptic **weight** connections from other neurons.

Forward inference:

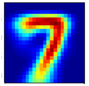


Fully trained network

“This is a seven.”

Hardware opportunity: Efficient, **low-power** deployment → IBM *TrueNorth*

Training:



UN-trained network

“um.. I have no idea?”

“This is a **seven**.”

Hardware opportunity: Train & use big networks FASTER and at LOWER POWER.

AI hardware, present & near-future: high-level view

Forward Inference (in the cloud & at the edge)

Training (mostly in the cloud)

TODAY

CPUs & GPUs

VERY SOON

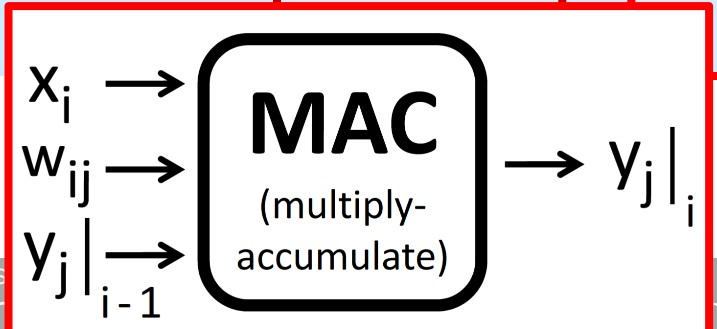
TPU1

Custom digital accelerators

TPU2

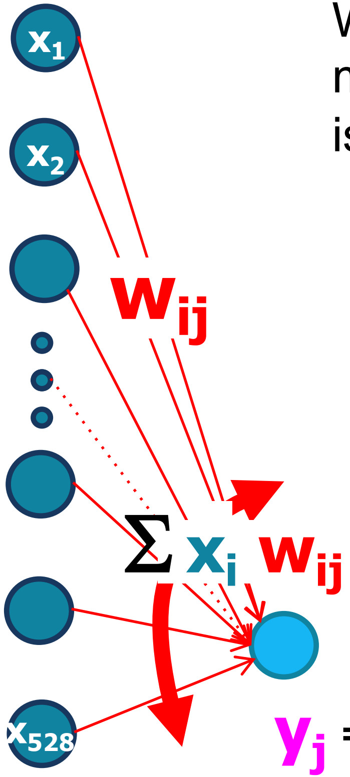
LATER ON…

Analog-memory-based Accelerators ???



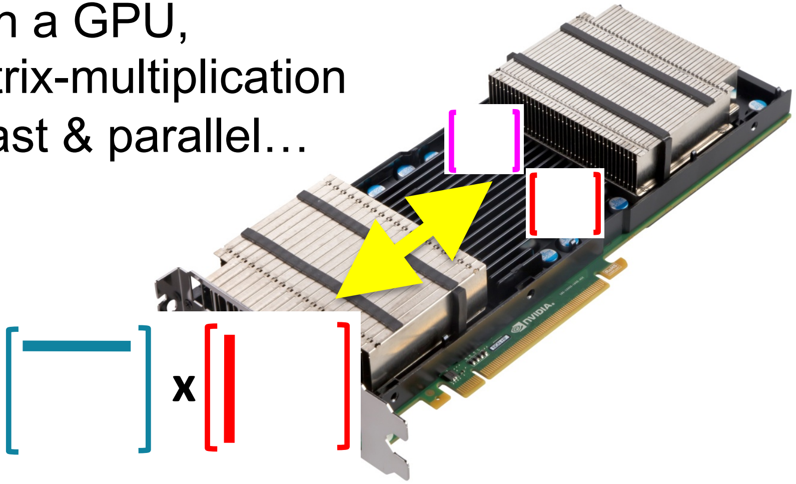
$x_i \rightarrow$, $w_{ij} \rightarrow$, $y_j|_{i-1} \rightarrow$ **MAC** (multiply-accumulate) $\rightarrow y_j|_i$

Computation needed for DNN: “Multiply-accumulate”



$y_j = f(\Sigma\ x_i\ w_{ij})$

With a GPU,
matrix-multiplication
is fast & parallel…

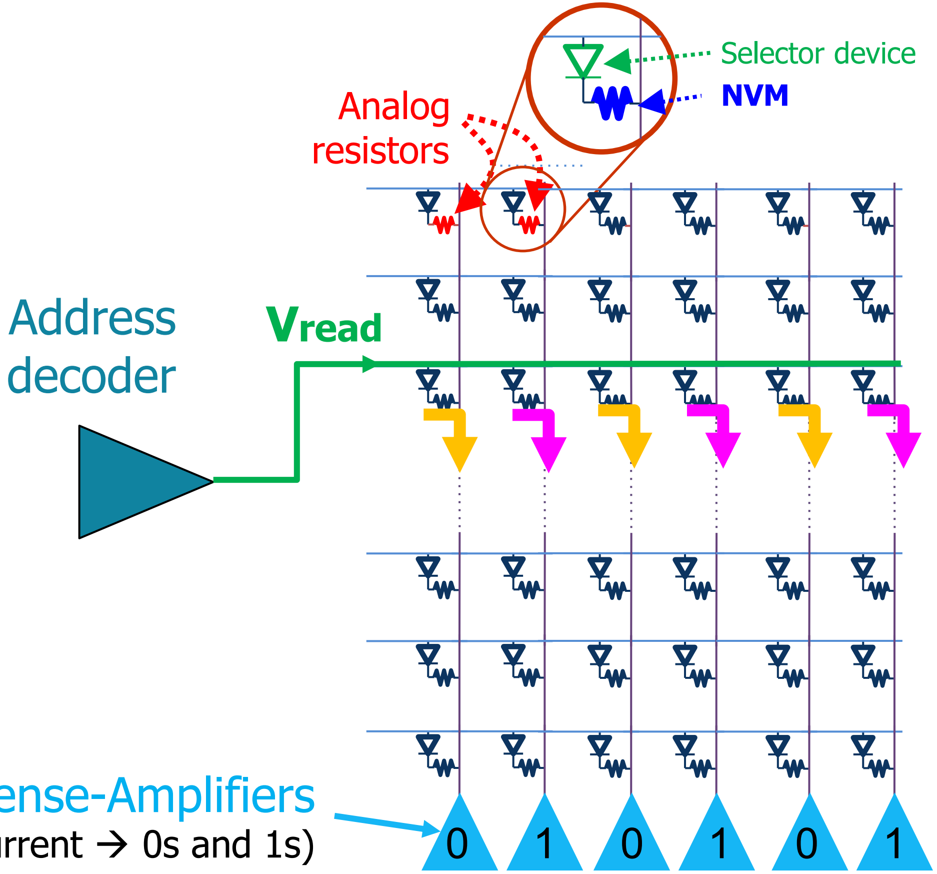


… but **x** and **w** values must arrive from DRAM,
and new **y** values sent back to DRAM

NVM (Non-Volatile Memory): usually for storing digital data (0s and 1s)

NVM technologies include:
MRAM (Magnetic RAM)
PCM (Phase-Change Memory)
RRAM (Resistance RAM)

Like conventional memory (SRAM/DRAM/Flash), an NVM is addressed one row at a time, to retrieve previously-stored digital data.



Multiply-accumulate with Analog Memory

1) Neuron peripheral circuitry

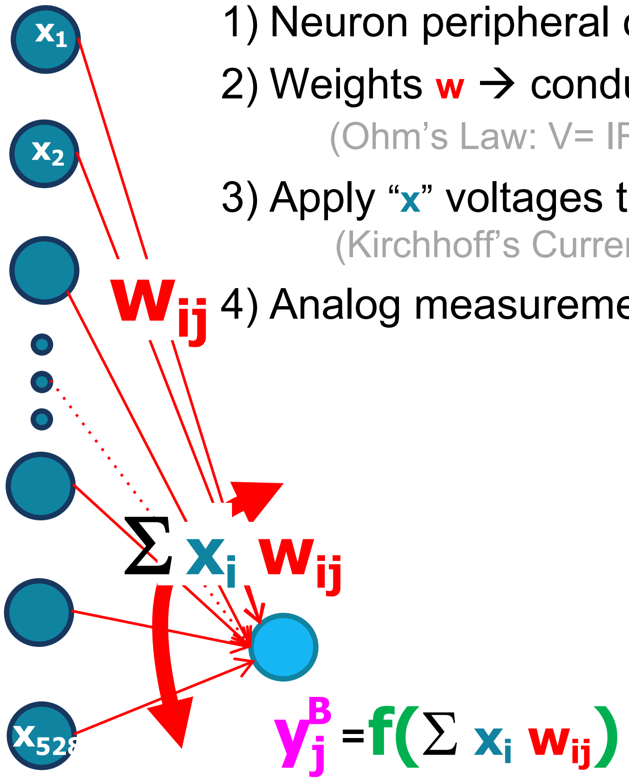
2) Weights **w** → conductances G^+, G^-

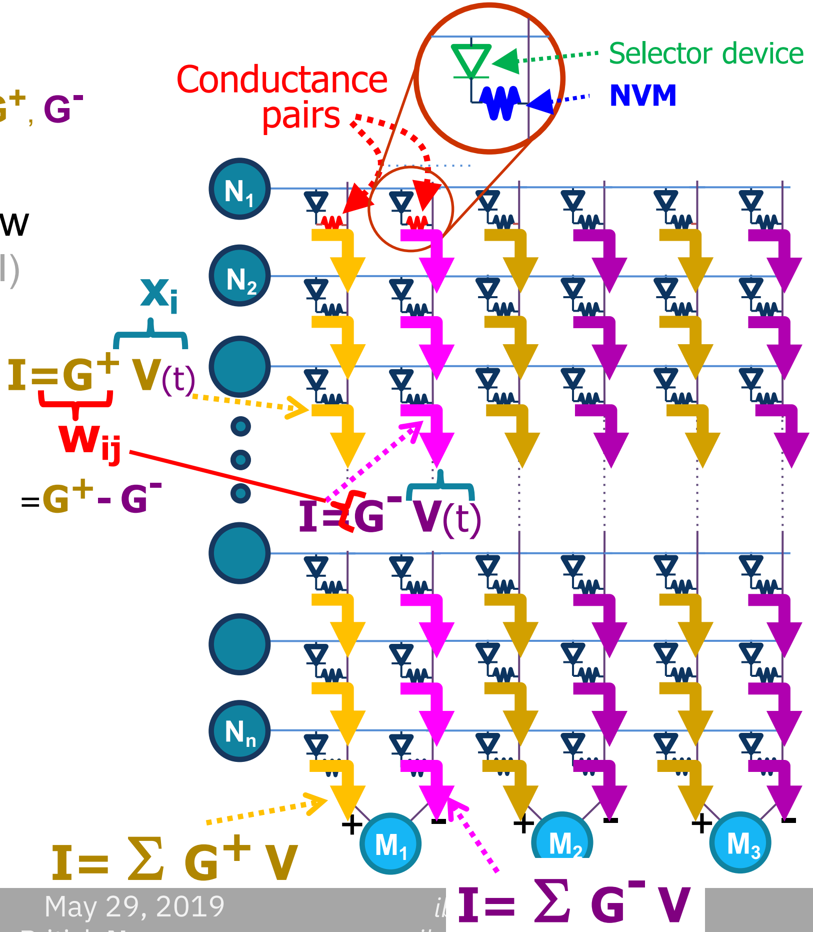
(Ohm's Law: V= IR → I = GV)

3) Apply "x" voltages to **every** row

(Kirchhoff's Current Law → Σ I)

4) Analog measurement

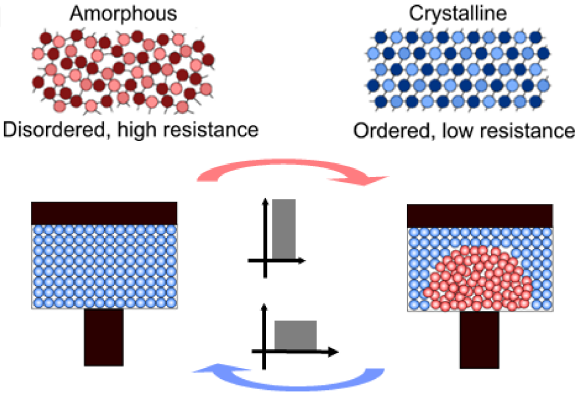


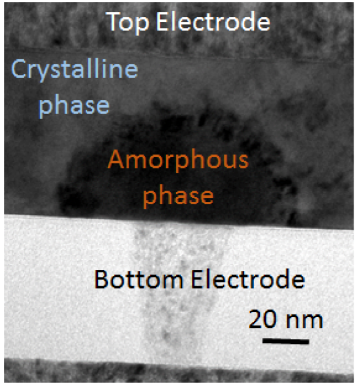


DNN in-situ training using analog memory

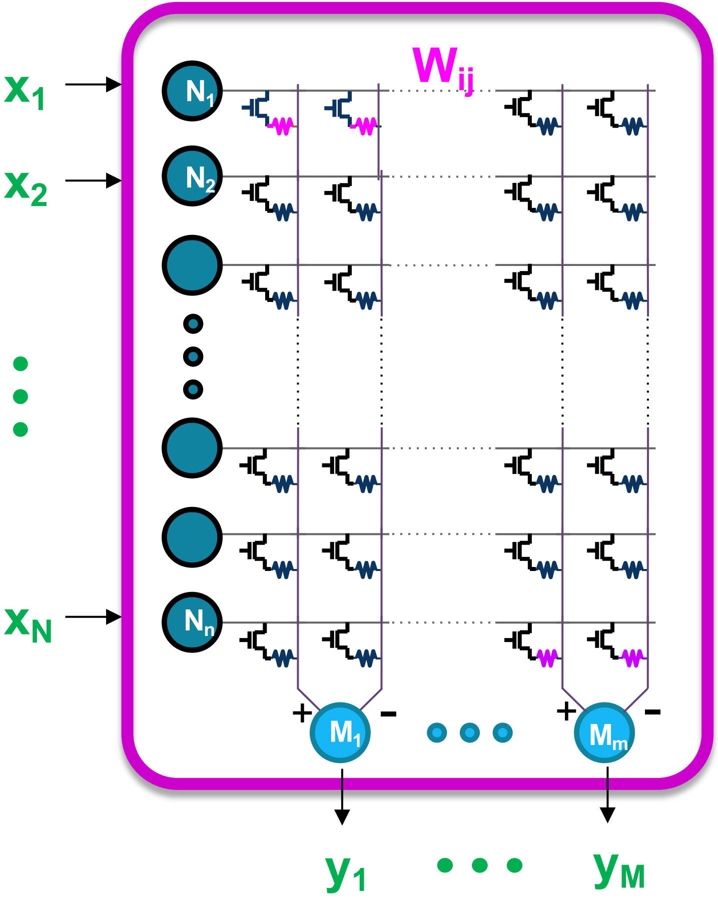
1) Forward Inference

Excitations (**x**) read weights **W**





Phase Change Memory



DNN in-situ training using analog memory

1) Forward Inference

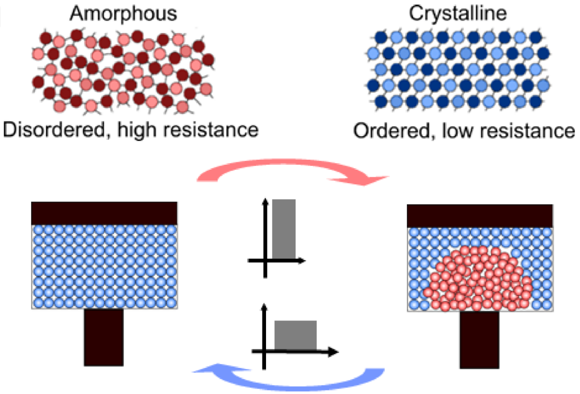
Excitations (**x**) read weights **W**

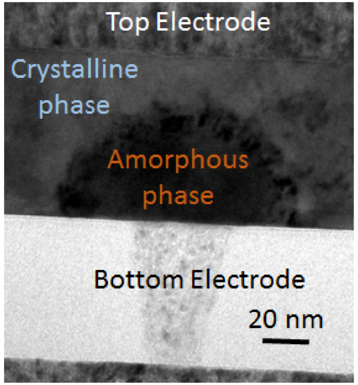
2) Backpropagate errors

Deltas (δ) read weights $\mathbf{W}^T$

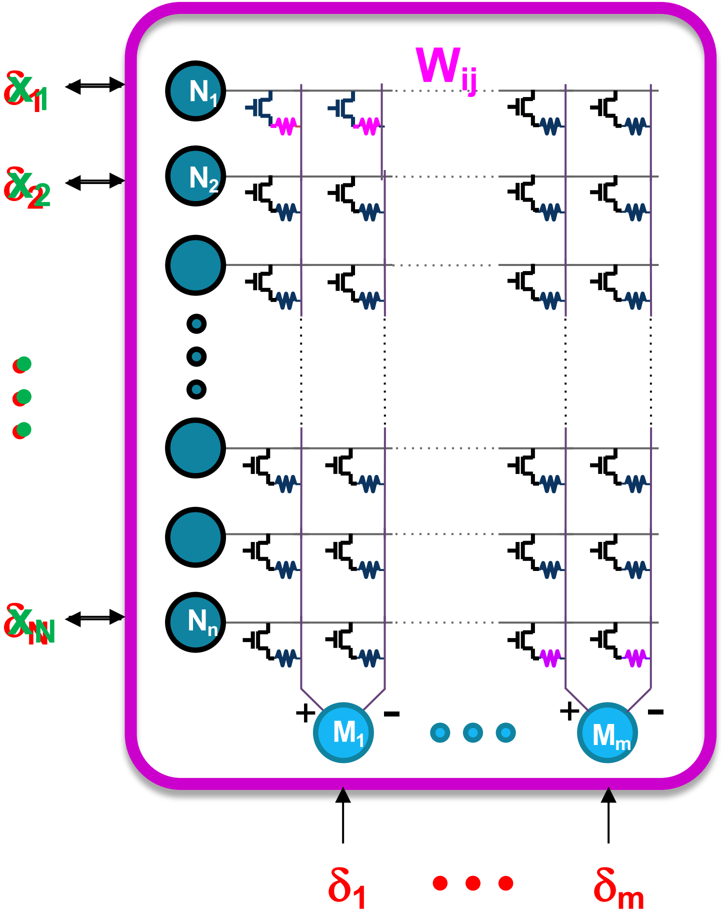
3) Weight update

Combine **x** and δ $\rightarrow \Delta \mathbf{W} \propto x_i * \delta_j$

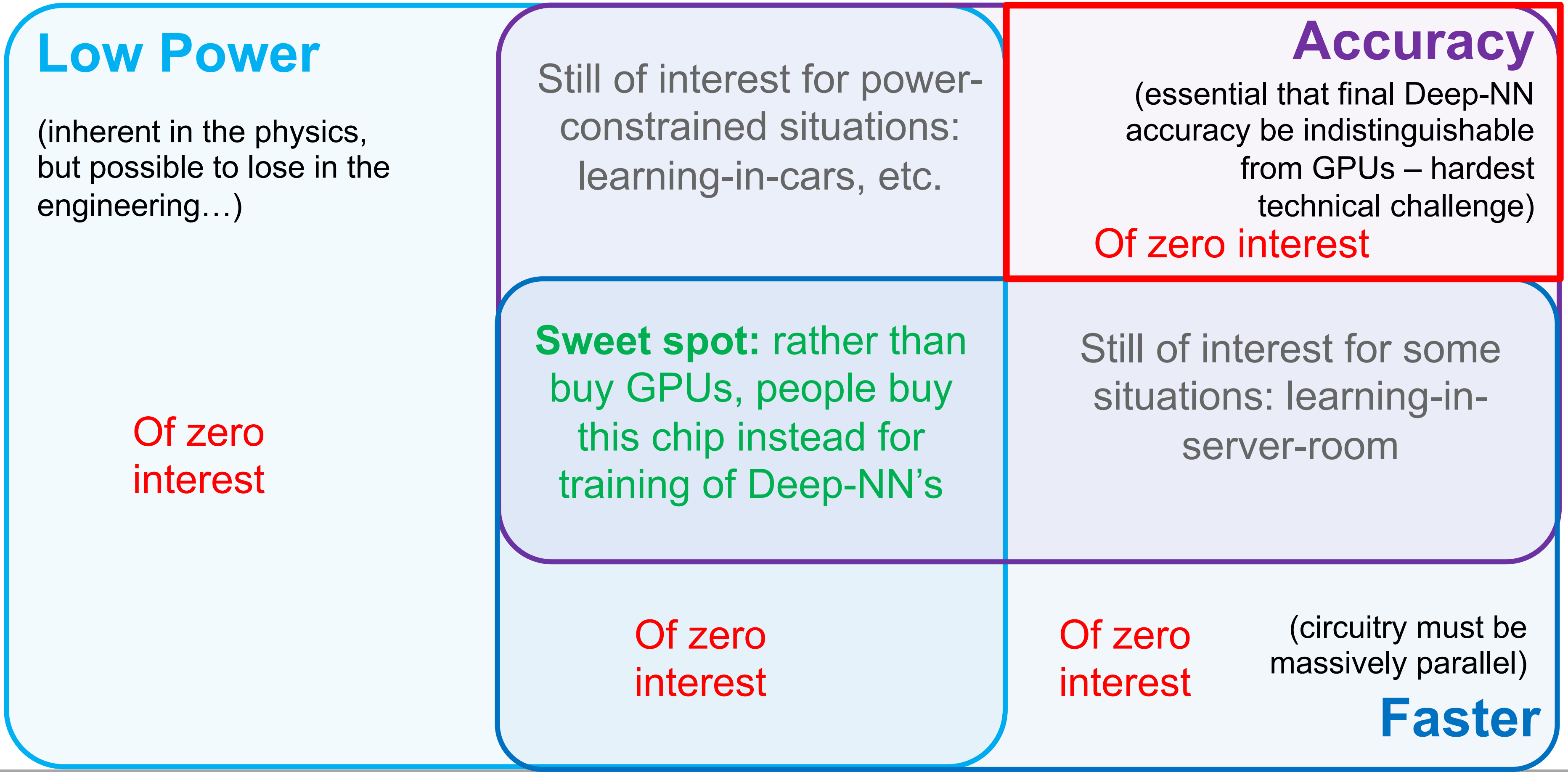




Phase Change Memory



Value Proposition (vs. a GPU)



Low Power

(inherent in the physics, but possible to lose in the engineering…)

Of zero interest

Still of interest for power-constrained situations: learning-in-cars, etc.

Accuracy

(essential that final Deep-NN accuracy be indistinguishable from GPUs – hardest technical challenge)

Of zero interest

Sweet spot: rather than buy GPUs, people buy this chip instead for training of Deep-NN's

Still of interest for some situations: learning-in-server-room

Of zero interest

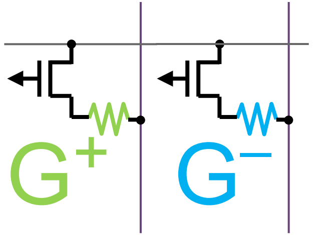
Of zero interest

(circuitry must be massively parallel)

Faster

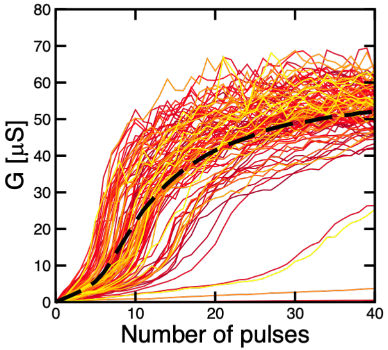
High DNN accuracy despite imperfect PCM devices

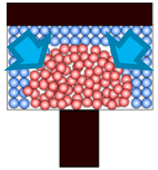
W = G^+ - G^-



Problem: Conductance changes in PCM are …

- uni-directional
- stochastic
- non-linear → asymmetric





What do we really want?

For training…

- Gentle, symmetric conductance changes

Our published results in DNN training w/ PCM

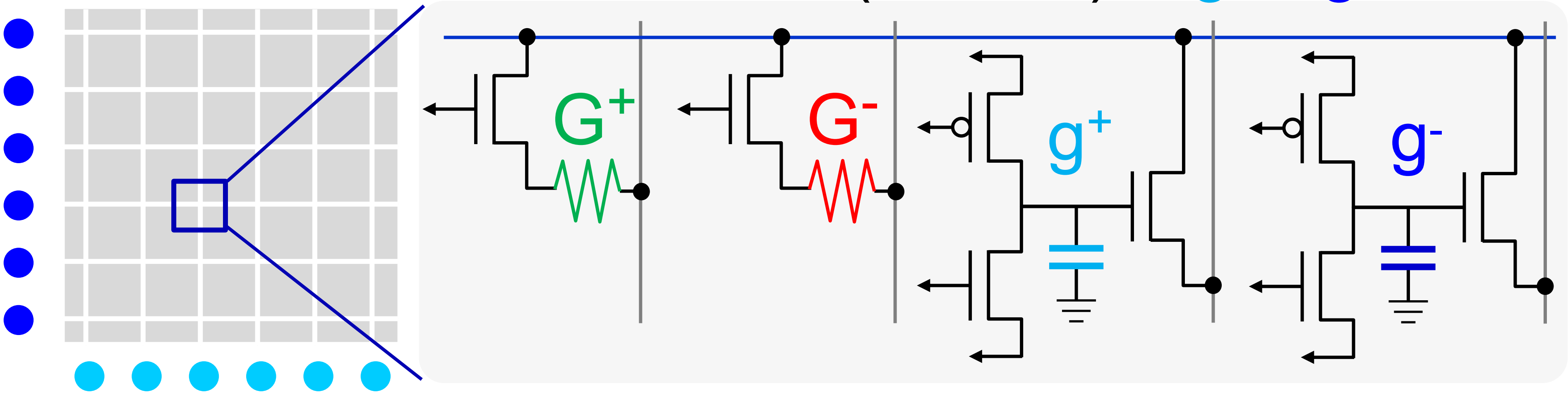
2014 – IEDM → **82%** w/ “mixed-hardware-software” experiment

2018 – ***Nature*** → **98%** (e.g., software-equivalent!) w/ new unit-cell

Novel 2T2R + 3T1C unit cell

More Significant Pair (MSP) Less Significant Pair (LSP)

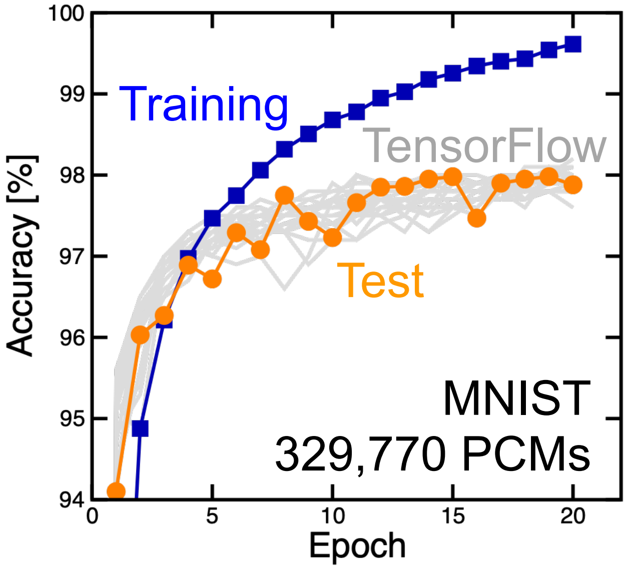
$$W = F * (G^+ - G^-) + g^+ - g^-$$

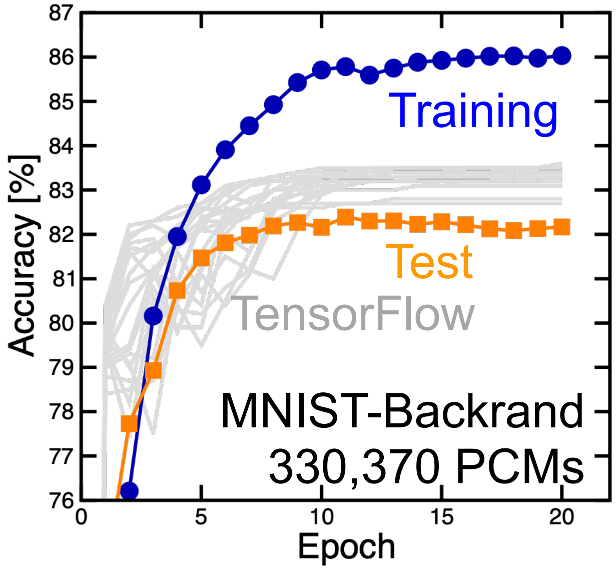


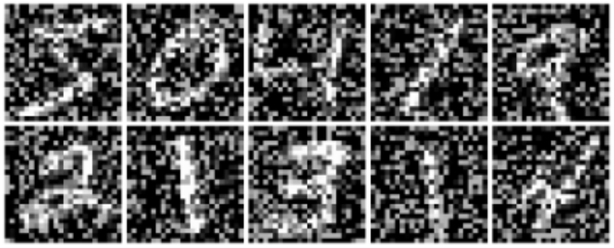
S. Ambrogio et al., *Nature*, 558, 60 (2018)

- Symmetry → Weight update performed on g+ only
  - – g^- shared among many columns (e.g. 128 columns)
- Dynamic Range → Gain factor F (e.g. F = 3)
- Non-Volatility → Weight transferred to PCMs infrequently (every 1000s of images)
- "CMOS variabilities" → Counteracted by "Polarity Inversion" technique

Accuracy on MNIST and MNIST with noise

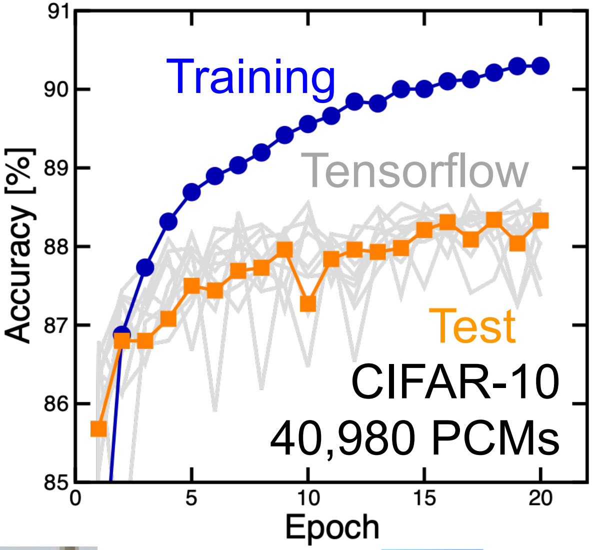


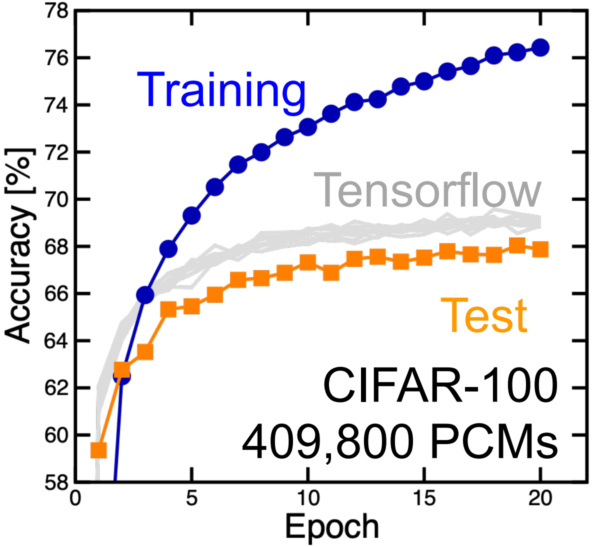


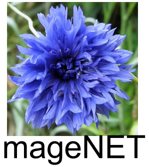


S. Ambrogio et al., *Nature*, 558, 60 (2018)

Transfer learning from ImageNet to CIFAR-10/100

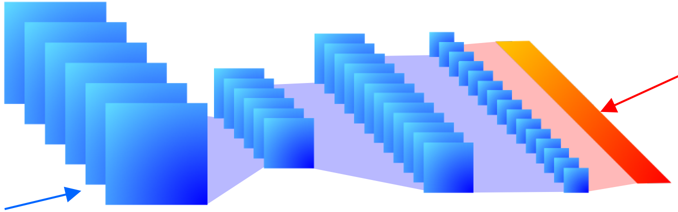





ImageNET

CIFAR-10/100

Transfer Learning: Use pre-trained, scaled weights from ImageNET for convolution layers


Only train last fully-connected layer

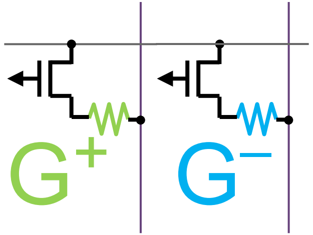
Fully Connected layer

Convolutional and Subsampling layers

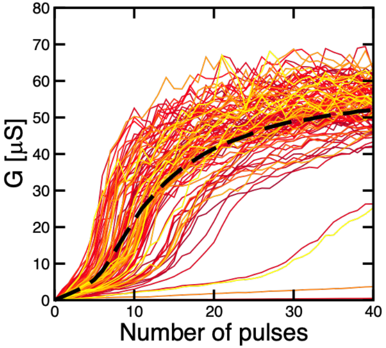
S. Ambrogio et al., *Nature*, 558, 60 (2018)

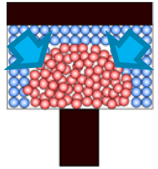
High DNN accuracy despite imperfect PCM devices

$W = G^{+} - G^{-}$



Problem: Conductance changes in PCM are …
- uni-directional
- stochastic
- non-linear → asymmetric





What do we really want?

For training…
- Gentle, symmetric conductance changes

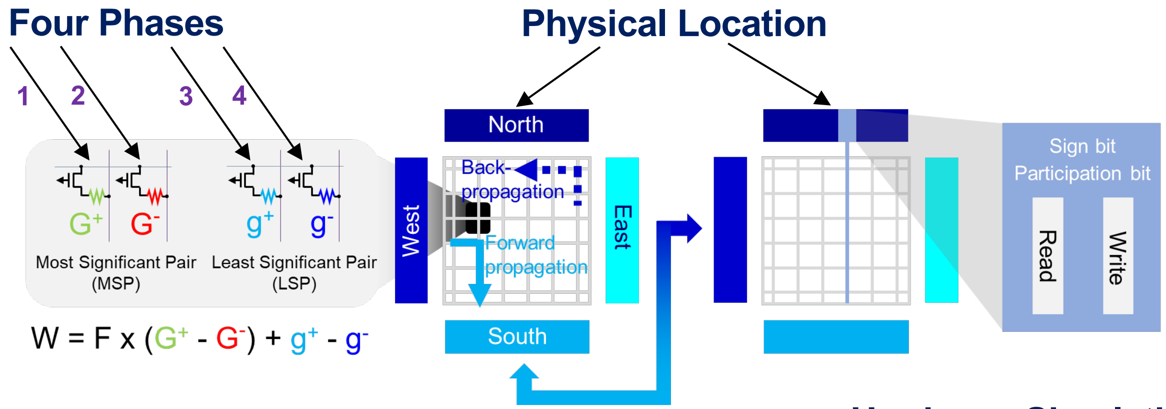
For inference…
- Precise tuning
- High yield
- No change over time

Our recent results in DNN inference w/ PCM

2019 – *Adv. Electr. Mater.* → programming schemes for 4 PCM devices (simulations)

2019 – *VLSI Tech. Symp.* → software-equivalence in “mixed-hardware-software” experiment with Long-Short Term Memory (LSTM) networks
(T8-1: Wed. June 12th, 10:30am)

Programming strategies for multi-PCM weights

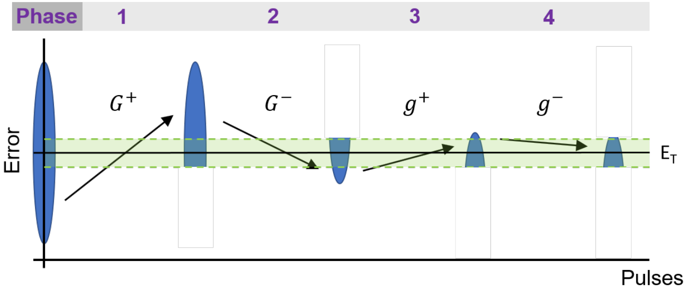


$W = F \times (G^{+} - G^{-}) + g^{+} - g^{-}$

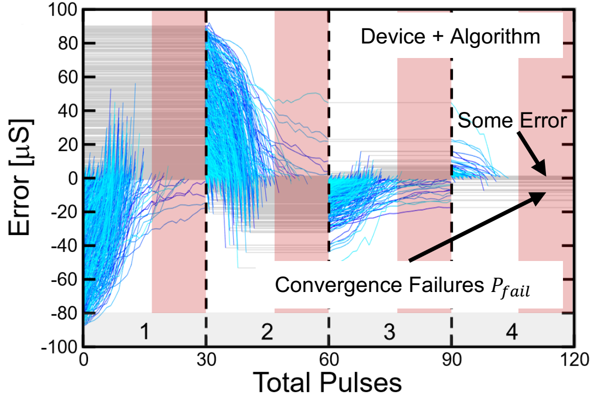
- Minimize computation expense
- Minimize area cost
- 2 bits per weights (p, s)
- Program entire row in parallel

$$Error = W - W_T$$

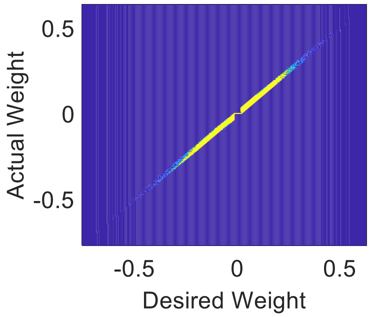
$$\boxed{Error \rightarrow 0}$$



Hardware Simulation



Correlation



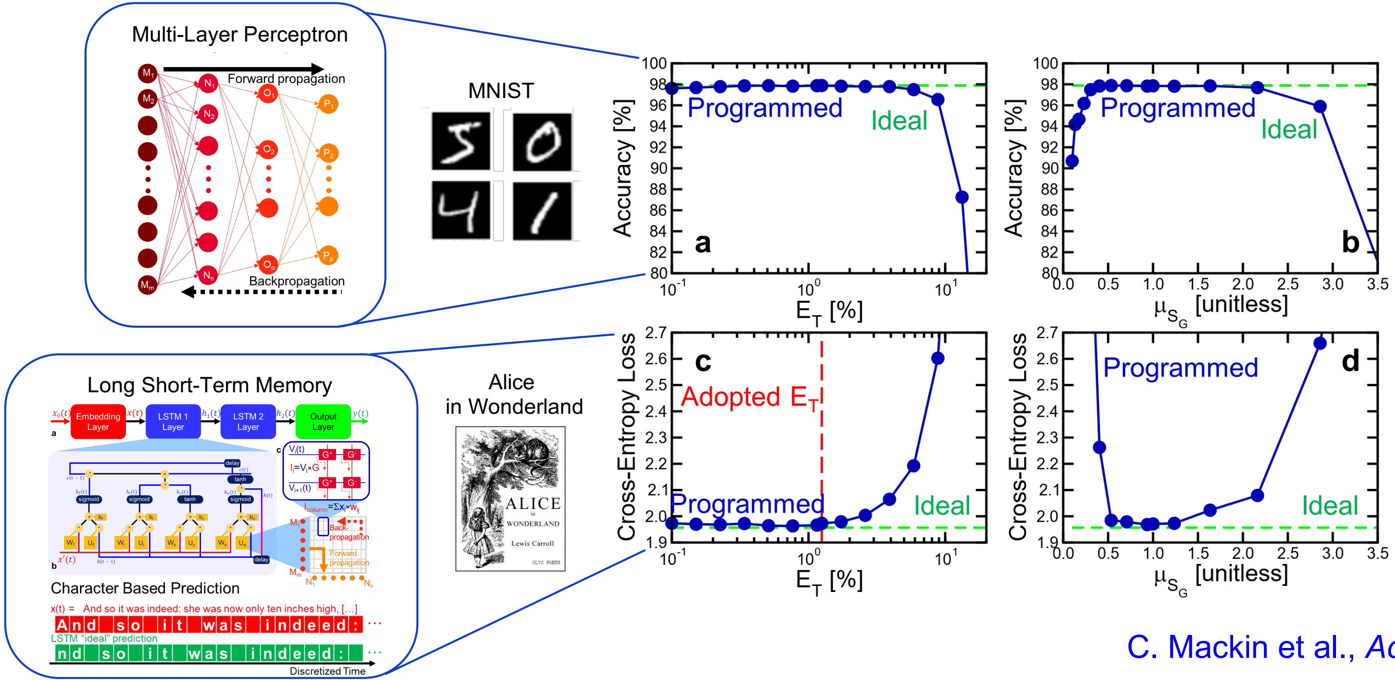
C. Mackin et al., *Adv. Electr. Mater.*, 1900026 (2019)

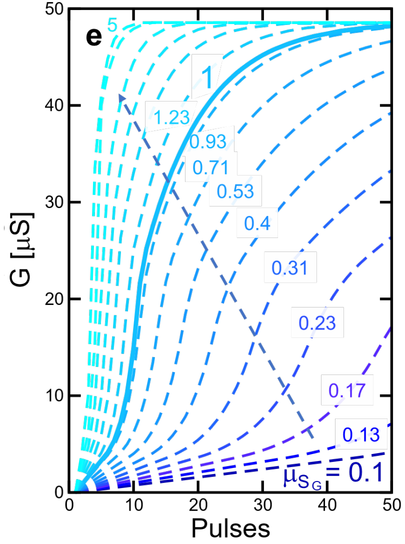
Impact on Network Accuracy

- **Two different types of networks**
- **Multiple parameters**
- **Software-equivalent accuracy despite NVM Variability**

Programming: F, W_{Range}, Total Pulses, E_T

Device: $\mu_{G_{max}}, \sigma_{G_{max}}, \mu_{S_G}, \sigma_{S_G}$

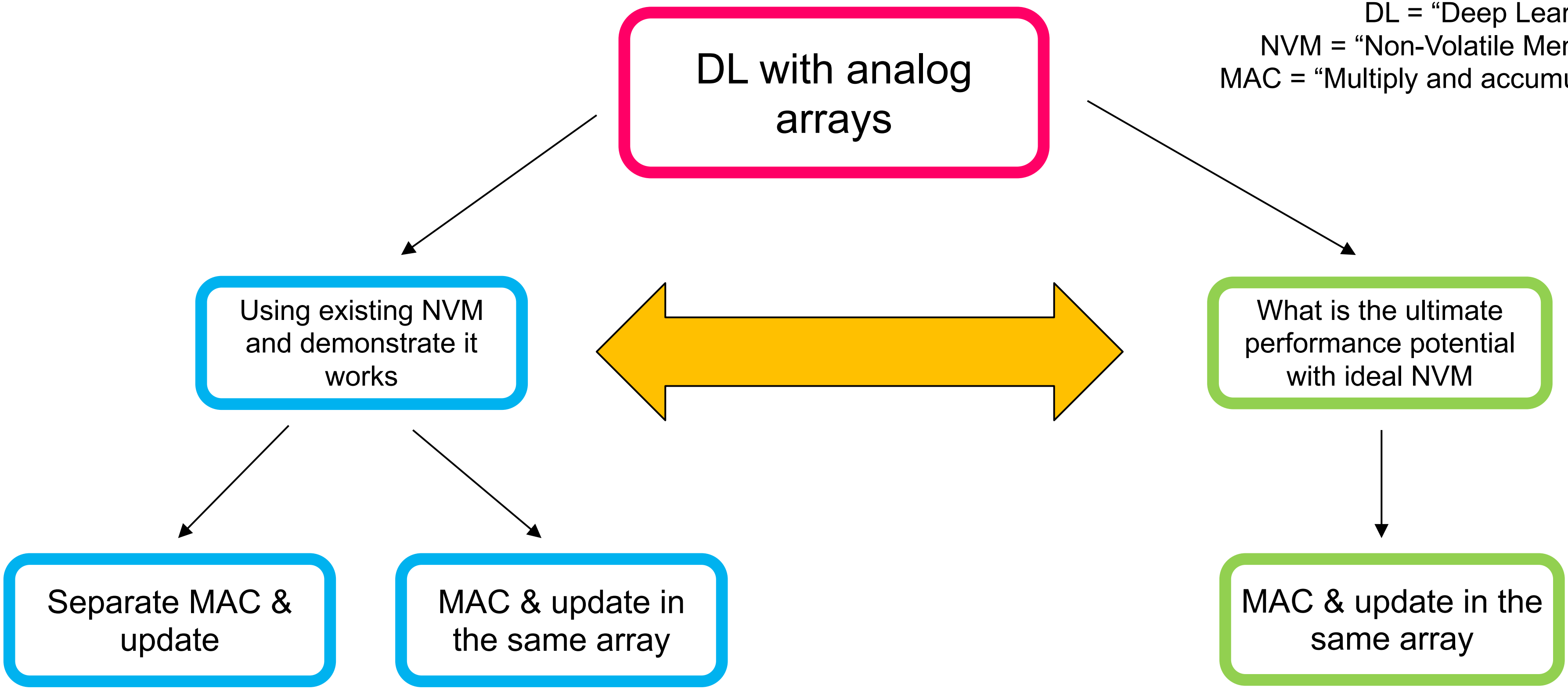




C. Mackin et al., *Adv. Elect. Mater.* 1900026 (2019)

IBM Research worldwide team: A comprehensive approach to Analog AI

DL = “Deep Learning”
NVM = “Non-Volatile Memory”
MAC = “Multiply and accumulate”



DL with analog arrays

Using existing NVM and demonstrate it works

What is the ultimate performance potential with ideal NVM

Separate MAC & update

MAC & update in the same array

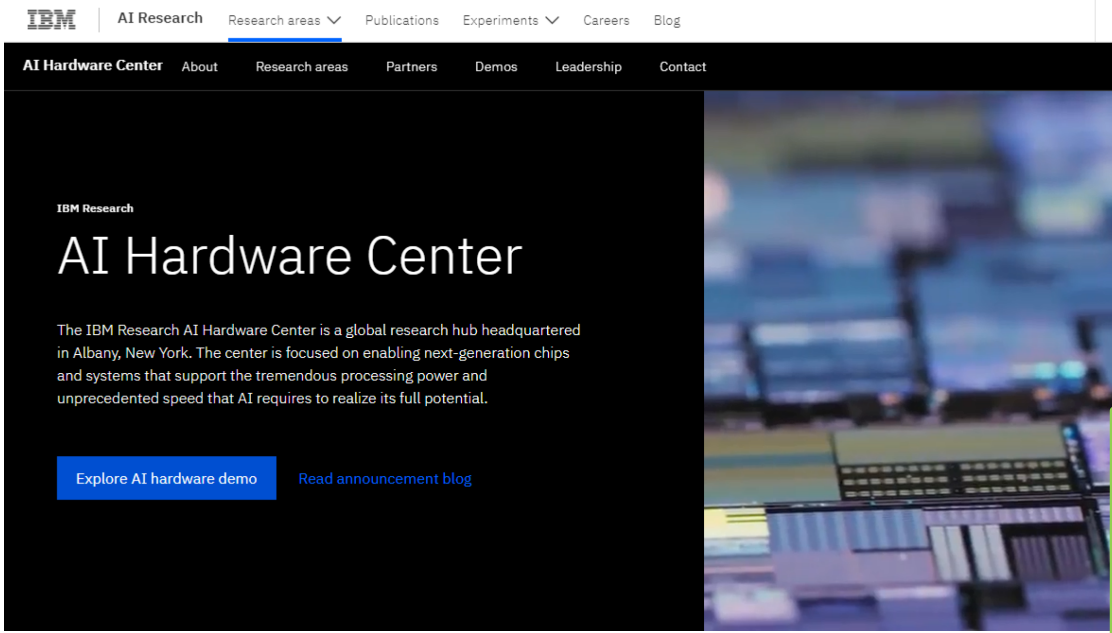
MAC & update in the same array

arXiv:1712.01192v1, 2017

IEDM 2014/2015
Nature 2018

Frontiers of Neuroscience
2016/2017/2018

IBM Research AI Hardware Center

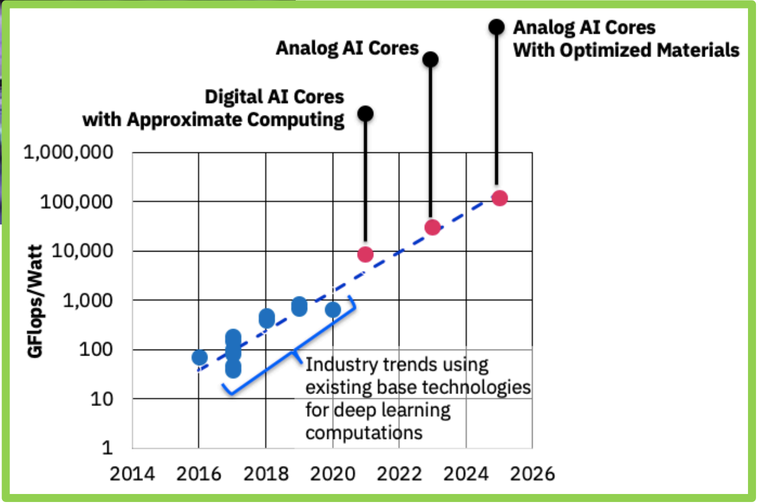


AI Hardware Center

The IBM Research AI Hardware Center is a global research hub headquartered in Albany, New York. The center is focused on enabling next-generation chips and systems that support the tremendous processing power and unprecedented speed that AI requires to realize its full potential.

ibm.biz/AI_hardware

www.ibm.com/blogs/research/2019/02/ai-hardware-center/

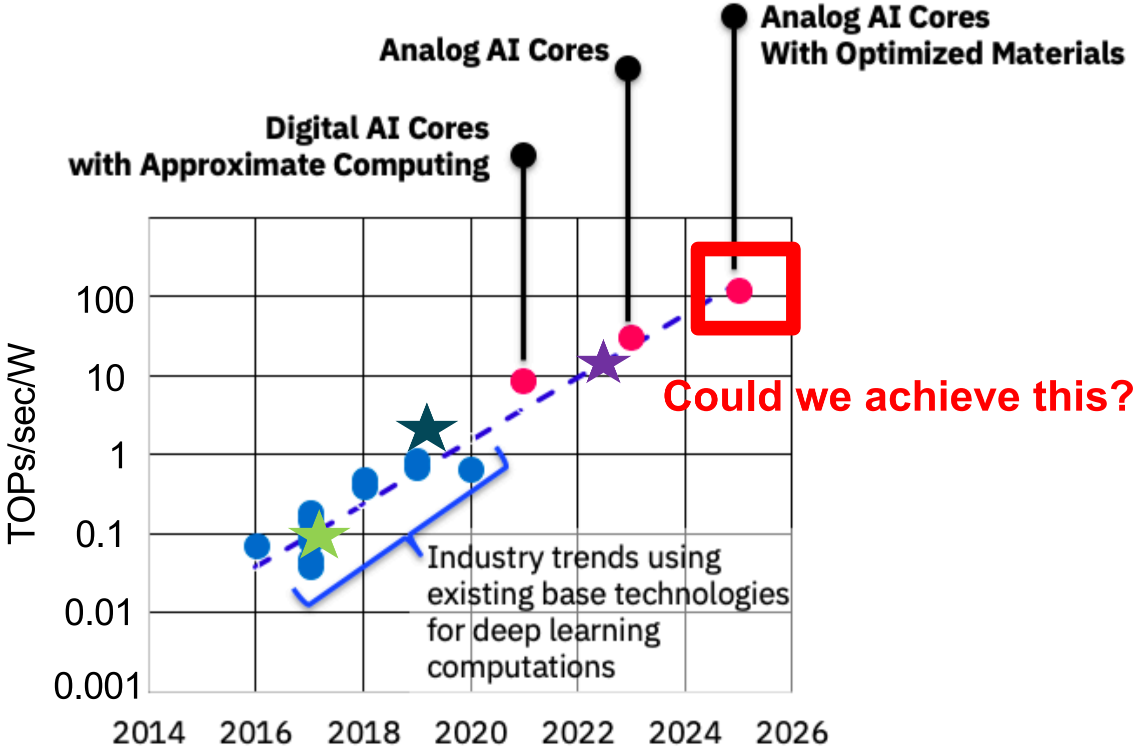


Where are we on the Roadmap?

- NVIDIA V100 : 0.1 TOPs/sec/W

- **Google TPU Gen 1 (Inference):**
  2.3 TOPs/sec/W
  - Inference Only
  - NOT include data movement

- **Circuits from our internal designs**:
  - MNIST : **15.2** TOPs/sec/W
  - PTB LSTM : **14** TOPs/sec/W



AI roadmap from IBM AI Hardware Center announcement
www.ibm.com/blogs/research/2019/02/ai-hardware-center/

H.-Y. Chang et. al, *IBM J. R&D,*
invited paper, accepted May 2019

How can we further improve energy efficiency w/ NVM devices?

1) **Reduce average NVM conductance** → reduces array currents during Multiply-Accumulates

→ Current focus of various material and device design efforts

2) **Reduce technology node**

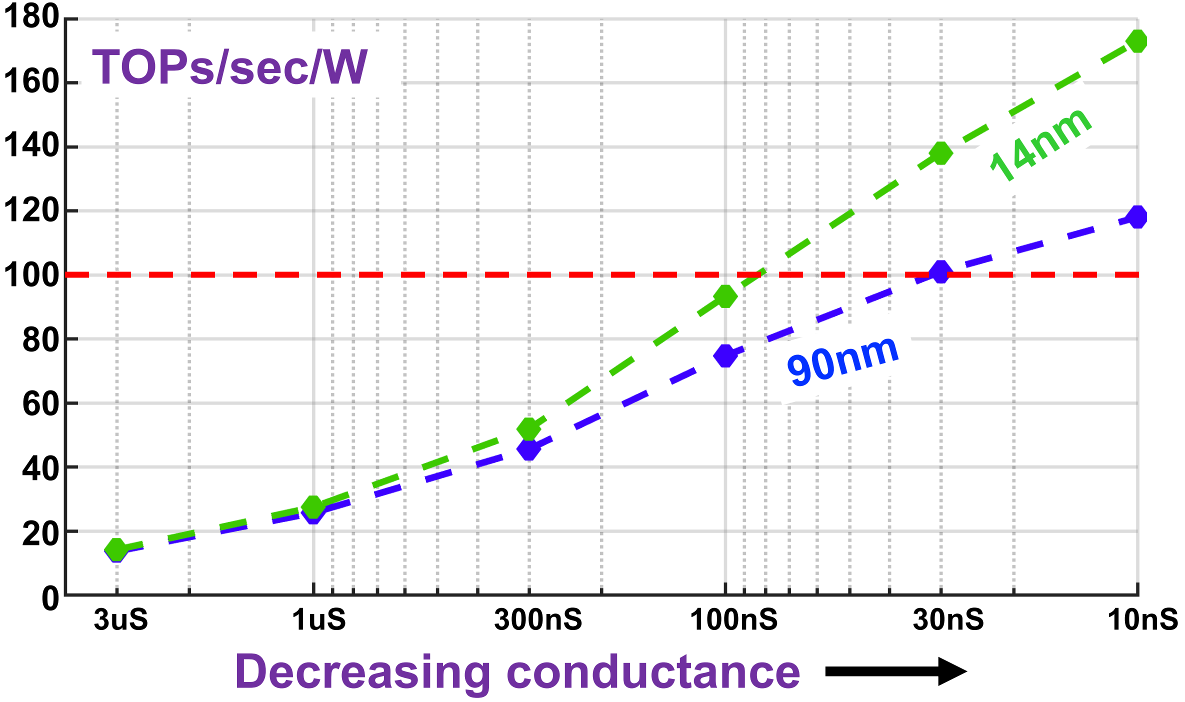
90nm -> 14nm

Benefits even just from scaling of routing energy

Area efficiency for inference: 10—70 TOPs/sec/mm^2

*(vs. **~0.3 TOPs/sec/mm2 for TPU v1**: In-Datacenter Performance Analysis of a Tensor Processing Unit)*

H.-Y. Chang et. al, *IBM J. R&D,* invited paper, accepted May 2019



Conclusion

- NVM-based crossbar arrays can accelerate Deep Machine Learning compared to GPUs
  - Multiply-accumulate performed at the data → saves power and time
  - But conventional NVM devices (like PCM) are imperfect…
- Recent training results
  - Mixed-hardware-software experiments → **software-equivalent training accuracy**
    - 2T2R+3T1C unit cell
    - “polarity inversion” technique
    - MNIST, MNIST-backrand, CIFAR-10 and CIFAR-100 tested (S. Ambrogio et al, ***Nature***, 558, 60 (2018))
- Recent inference results
  - Programming strategies for 4-PCM-based weights (C. Mackin et al., ***Adv. Electr. Mater.***, 1900026 (2019))
  - Mixed-hardware software experiments on LSTM (H. Tsai et al., ***VLSI Tech. Symp.*** (2019))
- Recent power projections based on real circuit designs
  - **100x better energy efficiency** (+ 100x speedup) on fully-connected layers (for LSTM and other networks)

    (H.-Y. Chang et al., ***IBM J. R&D,*** (2019))

pnaraya@us.ibm.com

Acknowledgements

Geoffrey Burr

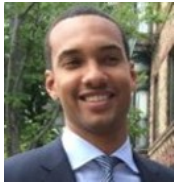
Pritish Narayanan

Bob Shelby

Stefano Ambrogio

Hsinyu Tsai

An Chen


Charles Mackin

Kohji Hosokawa

Scott Lewis

Management Support

Vijay Narayanan

Heike Riel

Wilfried Haensch

Matthew BrightSky

Arvind Kumar

Spike Narayan

Winfried Wilcke

Bulent Kurdi

NVM-for-Machine Learning: Recent & upcoming papers

ibm.biz/analog_AI

1. G. W. Burr, R. M. Shelby et al., “Neuromorphic computing using non-volatile memory," ***Advances in Physics X***, 2(1), 89-124 (**2017**).
   - Review of the NVM-for-neuromorphic field as a whole…
2. P. Narayanan, A. Fumarola, et al., “Towards on-chip acceleration of the backpropagation algorithm using non-volatile memory,” ***IBM Journal of Research and Development***, **61**(4/5), 11:1-11 (**2017**)
   - Summarizes the circuit design challenges
3. H. Tsai, S. Ambrogio, et al., “Recent progress in analog memory-based accelerators for Deep Learning,” ***Journal of Physics D,* 51**(28), 283001 (**2018**)
   - Review & overview paper
4. S. Ambrogio, P. Narayanan, et al., “Equivalent-accuracy Neuromorphic Hardware Acceleration of Neural Network Training using Analog Memory,” ***Nature,* 558**(7708), 60 (**2018**)
   - Demonstrate software-equivalent accuracy on training of Fully-Connected networks w/ PCM-based mixed hardware-software experiment
5. G. Cristiano, M. Giordano, et al., “Perspective on training fully connected networks with resistive memories: Device requirements for multiple conductances of varying significance,” ***Journal of Applied Physics,* 124**(15), 151901 (**2018**)
   - How does our multiple-conductance idea change the specifications for NVM devices needed for training?
6. C. Mackin, H. Tsai,, et al., “Weight Programming in DNN Analog Hardware Accelerators in the Presence of NVM Variability,” ***Advanced Electronic Materials,*** 1900026 (2019)
   - How to accurately program multiple-conductance weights using NVM devices with device-to-device variability?
7. H. Tsai, S. Ambrogio, et al., “Inference of Long-Short Term Memory networks at software-equivalent accuracy using 2.5M analog Phase Change Memory devices,” ***VLSI Technology Symposium,*** to be given (2019)
   - Demonstrate software-equivalent accuracy on inference of LSTM networks w/ PCM-based mixed hardware-software experiment
8. H.-Y. Chang, P. Narayanan, et al., “AI hardware acceleration with analog memory: micro-architectures for low energy at high speed,” ***IBM Journal of Research and Development,*** to appear (2019)
   - Micro-architectural approaches that lead to both high energy efficiency AND large DNN acceleration

NVM-for-Machine Learning: IBM Collaborators

9. S. Kim et al., “Analog CMOS-based Resistive Processing Unit for Deep Neural Network Training”, **arXiv**, preprint 1706.06620
10. T. Gokmen et al., “Acceleration of deep neural network training with resistive cross-point devices: design considerations”, **Frontiers in Neuroscience**, vol. 10, page 333, Jul 2016
11. Y. Li et al., “Capacitor-based Cross-point Array for Analog Neural Network with Record Symmetry and Linearity”, **VLSI Technology Symposium** 2018
12. M.L. Gallo et al., “Mixed-precision training of deep neural networks using computational memory”, **arXiv preprint** 1712.01192
13. I. Boybat et al., “Neuromorphic computing with multi-memristive synapses”, **Nature communications**, vol. 9(1), page 2514, June 2018
14. A. Sebastian et al., “Temporal correlation detection using Computational Phase Change Memory, **Nature Communications**, vol. 8, page 1115, Oct 2017
15. S. R. Nandakumar et al., “Supervised learning in spiking neural networks with MLC PCM synapses”, **Device Research Conference**, 2017
16. Gong et al., “Signal and Noise Extraction from Analog Memory Elements for Neuromorphic Computing”, **Nature communications**, vol. 9(1), page 2102, May 2018
17. M. Salinga et al., “Monatomic phase change memory”, **Nature Materials**, vol. 17, page 681-695, June 2018
18. I. Giannopoulos et al., “8-bit Precision In-Memory Multiplication with Projected Phase-Change Memory”, **IEDM** 2018
19. J. Tang et al., “ECRAM as Scalable Synaptic Cell for High-Speed, Low-Power Neuromorphic Computing”, **IEDM** 2018

NVM-for-Machine Learning: (Some) Non-IBM Work

- S.B. Erylimaz et al., “Neuromorphic architectures with electronic synapses”, International Symposium on Quality Electronic Design (ISQED), Mar 2016
- S. B. Eryilmaz, et al., "Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures," IEEE International Electron Devices Meeting (IEDM) 2015, pp. 4.1.1-4.1.4.
- S. Yu, "Neuro-Inspired Computing With Emerging Nonvolatile Memorys," in Proceedings of the IEEE, vol. 106, no. 2, pp. 260-285, Feb. 2018.
- P. Y. Chen, et al., "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," IEEE International Electron Devices Meeting (IEDM) 2017, pp. 6.1.1-6.1.4.
- E. J. Fuller et al., “Li-Ion Synaptic Transistor for Low Power Analog Computing”, Advanced Materials, 29(4), 2017
- S. Agarwal *et al*., "Achieving ideal accuracies in analog neuromorphic computing using periodic carry," Symposium on VLSI Technology, 2017, pp. T174-T175.
- https://cross-sim.sandia.gov/
- X. Guo *et al*., "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology,” IEEE International Electron Devices Meeting (IEDM) 2017, pp. 6.5.1-6.5.4.
- M. Prezioso et al., “Training and operation of an integrated neuromorphic network based on metal-oxide memristors” *Nature,* vol. 521, pp. 61–64, 2015
- K. Moon *et al*., "High density neuromorphic system with Mo/Pr0.7Ca0.3MnO3 synapse and NbO2 IMT oscillator neuron," IEEE International Electron Devices Meeting (IEDM) 2015, pp. 17.6.1-17.6.4.
- C. LI, Analogue signal and image processing with large memristor crossbars, Nature Electronics*,* vol. 1, pp. 52–59, 2018.
- S. Ambrogio, “Spike-timing dependent plasticity in a transistor-selected resistive switching memory”, Nanotechnology 24 384012, 2013.
- E. Vianello *et al*., "Resistive Memories for Spike-Based Neuromorphic Circuits," 2017 IEEE International Memory Workshop (IMW), 2017