RRAM Fabric for Neuromorphic and Reconfigurable Compute-In-Memory Systems

Wei D. Lu

University of Michigan
Electrical Engineering and Computer Science
Ann Arbor, MI, USA

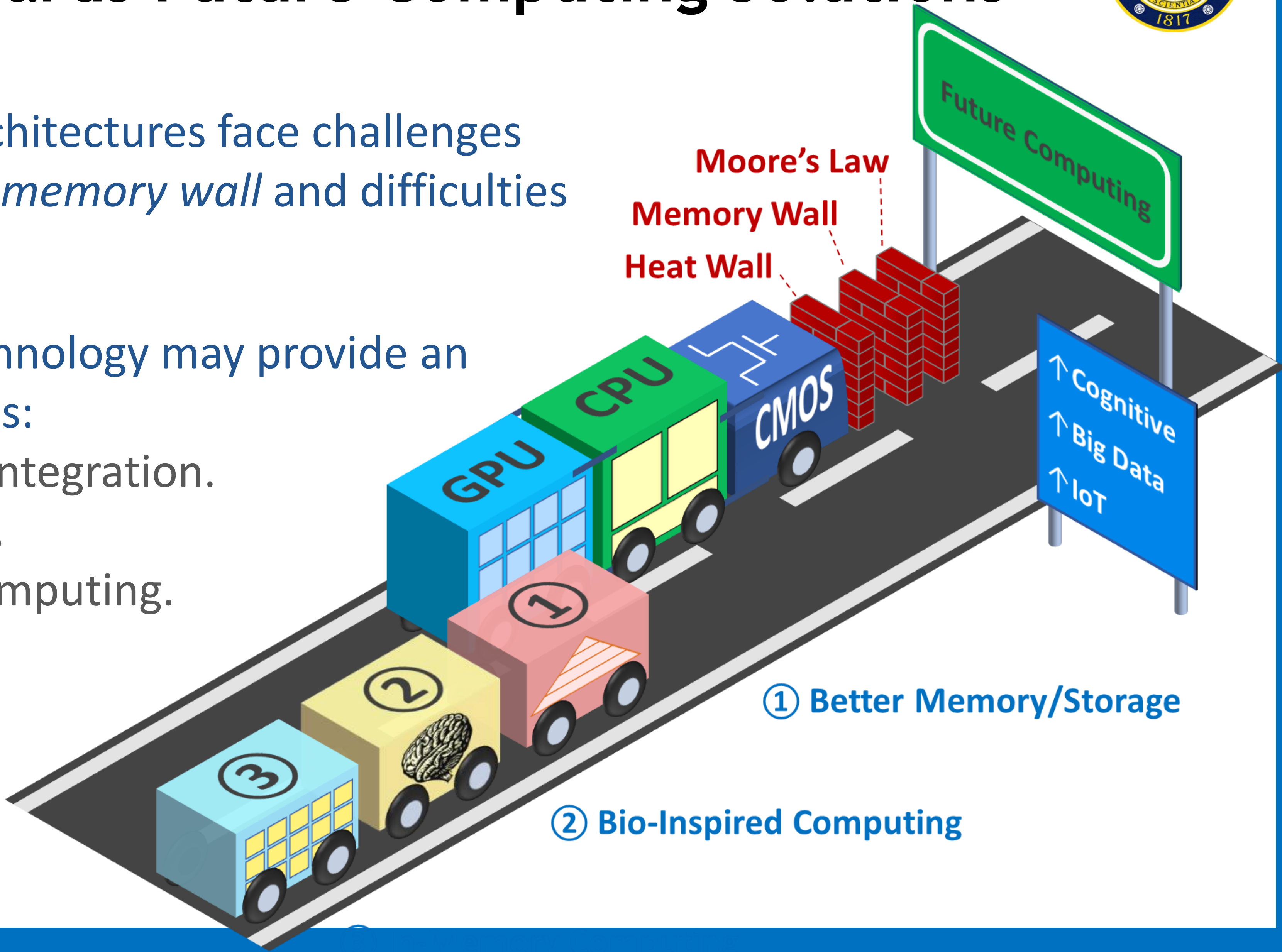
Outline

- Introduction - RRAM Devices
- Improving Computing Efficiency using RRAM Arrays
  - Bring memory as close to logic as possible
  - Neuromorphic computing in artificial neural networks
  - More bio-inspired networks, taking advantage of the internal ionic dynamics
  - In-memory computing for logic and arithmetic operations
- Future - reconfigurable systems based on a common physical fabric

The Race Towards Future Computing Solutions

- Conventional computing architectures face challenges including the *heat wall*, the *memory wall* and difficulties in continued device scaling.

- Developments in RRAM technology may provide an alternative path that enables:
  - Hybrid memory–logic integration.
  - Bioinspired computing.
  - Efficient in-memory computing.



① Better Memory/Storage

② Bio-Inspired Computing

M. A. Zidan, J. P. Strachan, and W. D. Lu, Nature Electronics 1: 22–29 (2018)

Need to Rethink Computing

Observations

- Compute is cheap (< 1pJ)
- Programming an instruction is very expensive (70pJ - fetching an instruction alone is 25pJ)
- DRAM access is another 10-100x more expensive

Solutions

- Keep data local
- Non-instruction based? – how?
- Get rid of DRAM!

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1.1pJ
32 bit	3.7pJ

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ

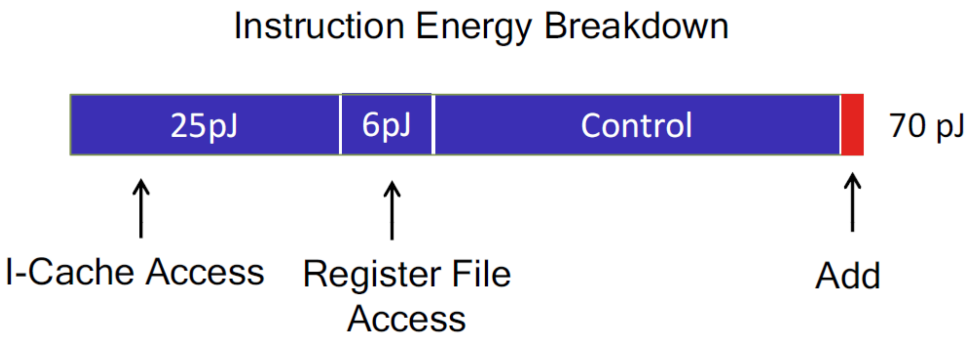


Figure 1.1.9: Rough energy costs for various operations in 45nm 0.9V.

Mark Horowitz, ISSCC 2017

Rethinking Computing

IT'S TIME

TO RETHINK

COMPUTING

Before

Optimize architecture and circuit design to minimize compute cost

Now

Compute is *cheap*
Data and data routing are *expensive*

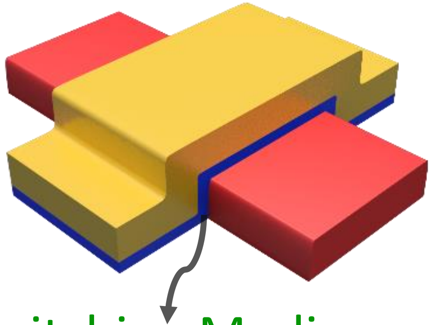
→ **Fundamentally redesigning the architecture from data-routing point of view, not from compute point of view**

Two-Terminal Memory Devices and Crossbar Arrays

» Resistive memory (RRAM), *memory* + *resistor* (memristor)

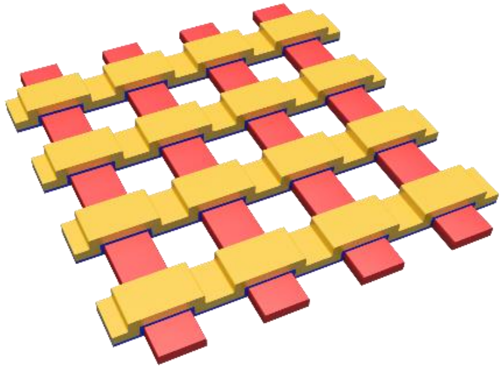
- Simple structure
  - Formed by two-terminal devices
  - Not limited by transistor scaling
- Ultra-high density
  - NAND-like layout, cell size 4F^2
  - Terabit potential
- Large connectivity
- Application:
  - Memory
  - Neuromorphic
  - General Purpose Computing

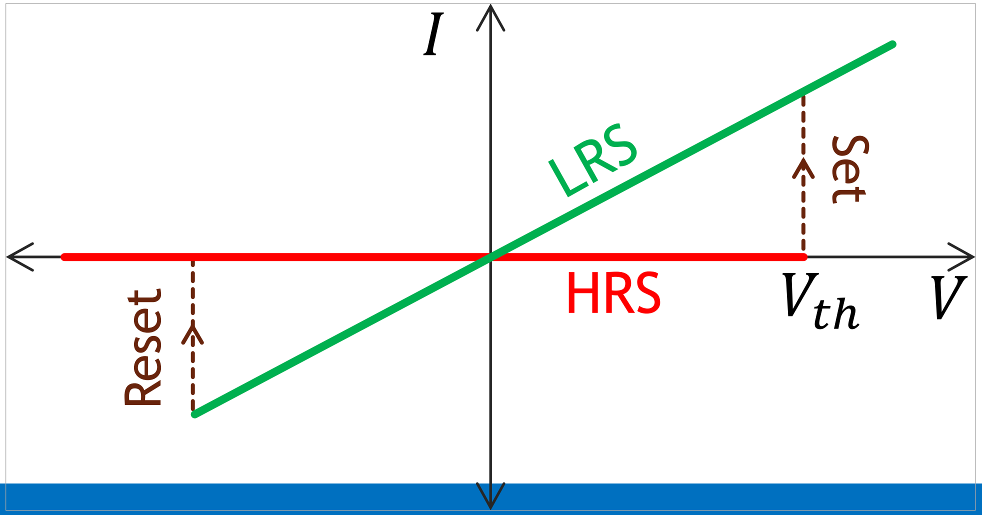
Single Cell Structure



Switching Medium

Crossbar Structure



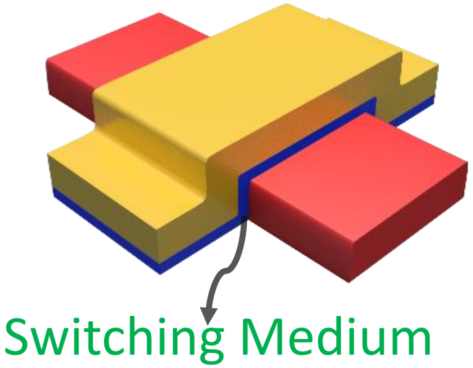


Two-Terminal Memory Devices and Crossbar Arrays

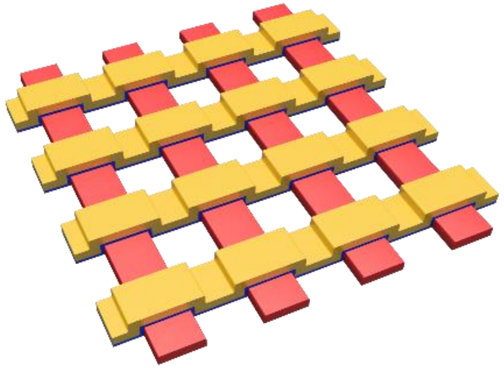
» Resistive memory (RRAM), *memory* + *resistor* (memristor)

- Simple structure
  - Formed by two-terminal devices
  - Not limited by transistor scaling
- Ultra-high density
  - NAND-like layout, cell size $4F^2$
  - Terabit potential
- Large connectivity
- Application:
  - Memory
  - Neuromorphic
  - General Purpose Computing

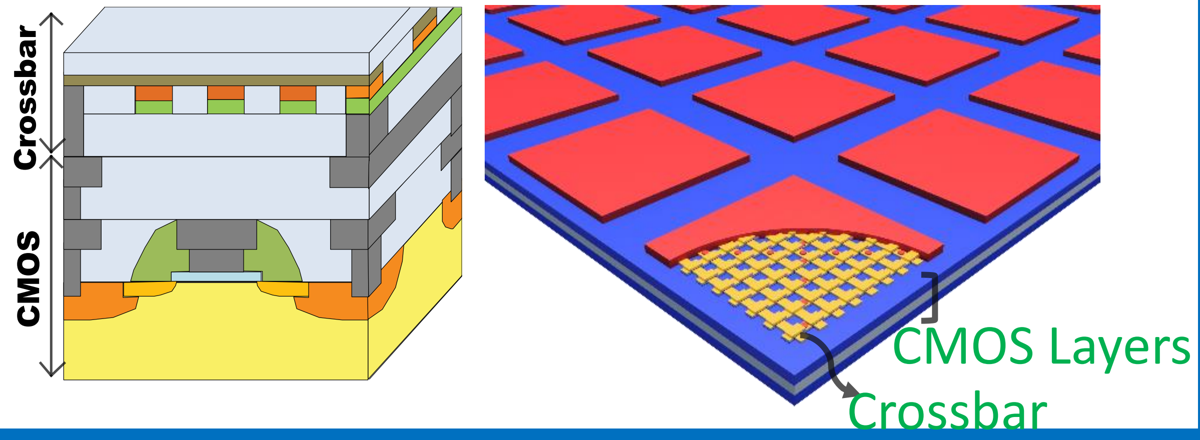
Single Cell Structure



Crossbar Structure

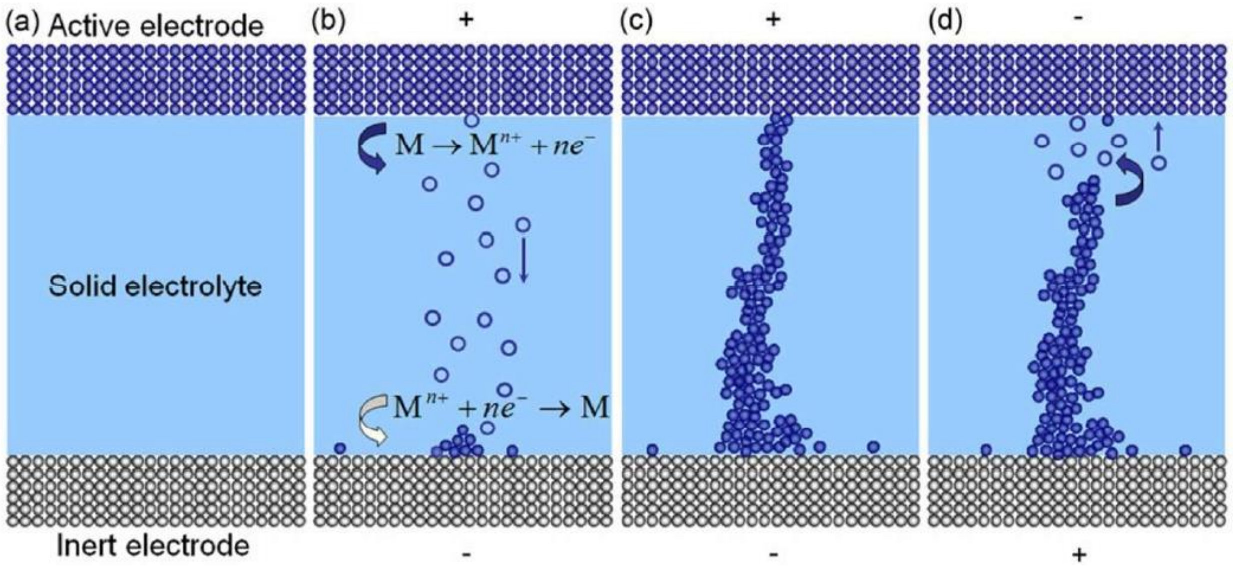


CMOS Integration



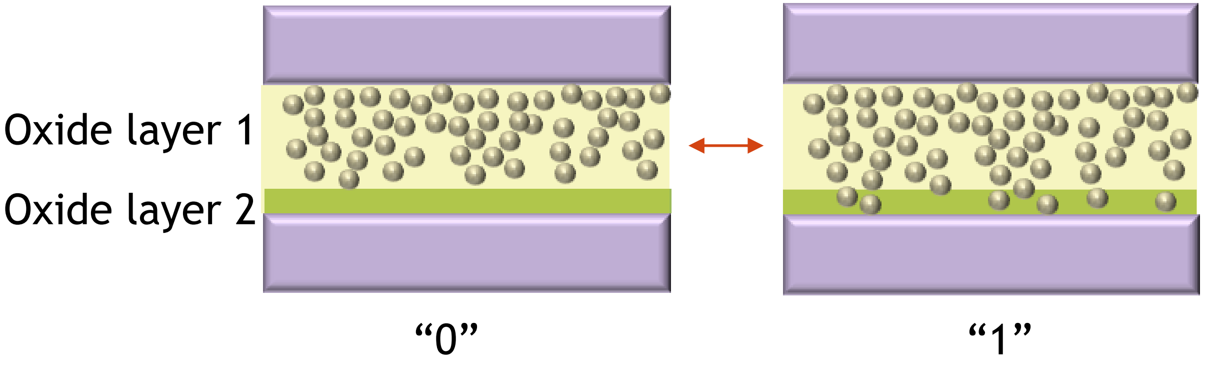
Coupled electronic/ionic effects

ElectroChemical Metallization Cell (ECM, CBRAM)



- **Creating “new” materials on the fly**
- Active electrode material + inert dielectric
- “Filament” based on electrode material injection and redox at electrodes
- Switching layer facilitates ionic movement

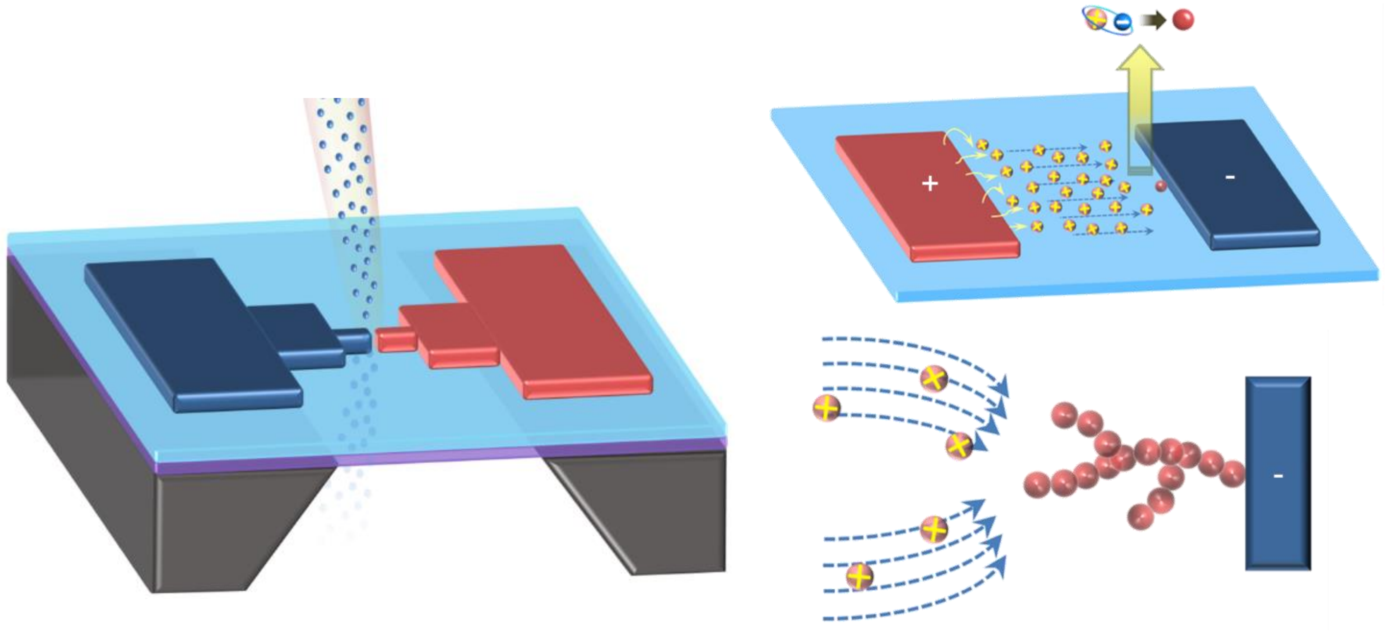
Valency Change Cell (VCM)

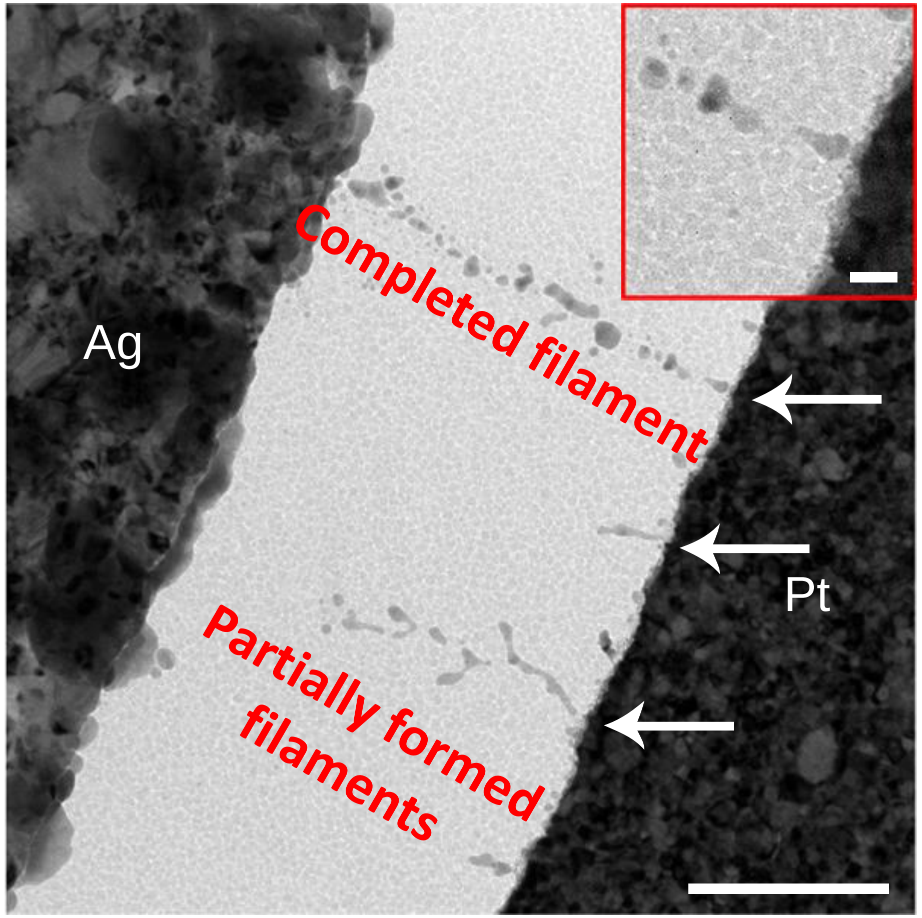


- **Modulating exiting material properties**
- Filament based on oxygen exchange between two oxide layers
- Electrode plays minor role

Y. Yang and W. Lu, Nanoscale, 5, 10076 (2013)

Visualization of Filament

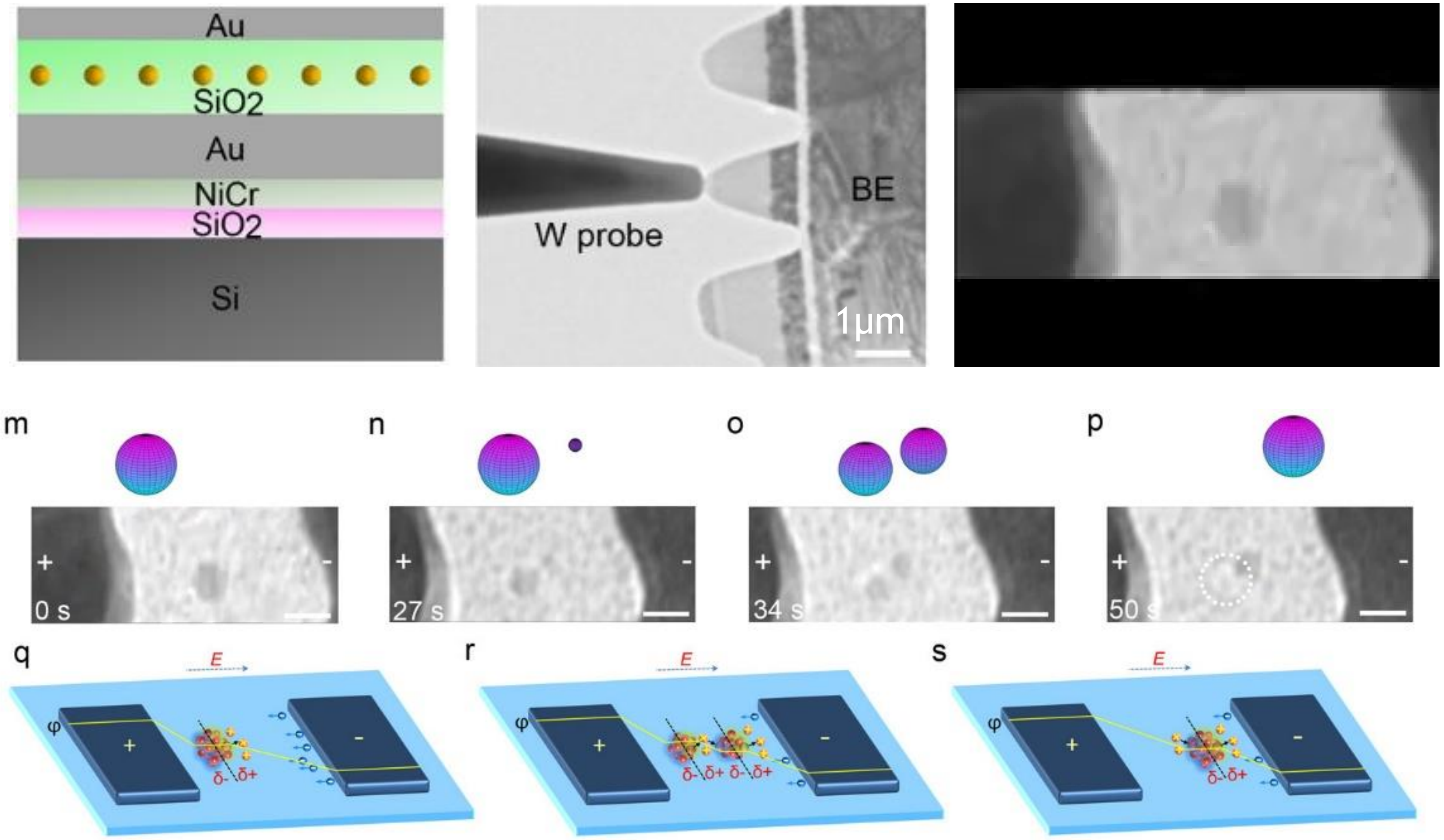




- Ag/SiO_2/Pt structure, sputtered SiO2 film
- The filament grows from the IE backwards toward the AE
- Branched structures were observed with wider branches pointing to the AE

Y. Yang, Gao, Chang, Gaba, Pan, and W. Lu, Nature Communications, 3, 732, 2012.

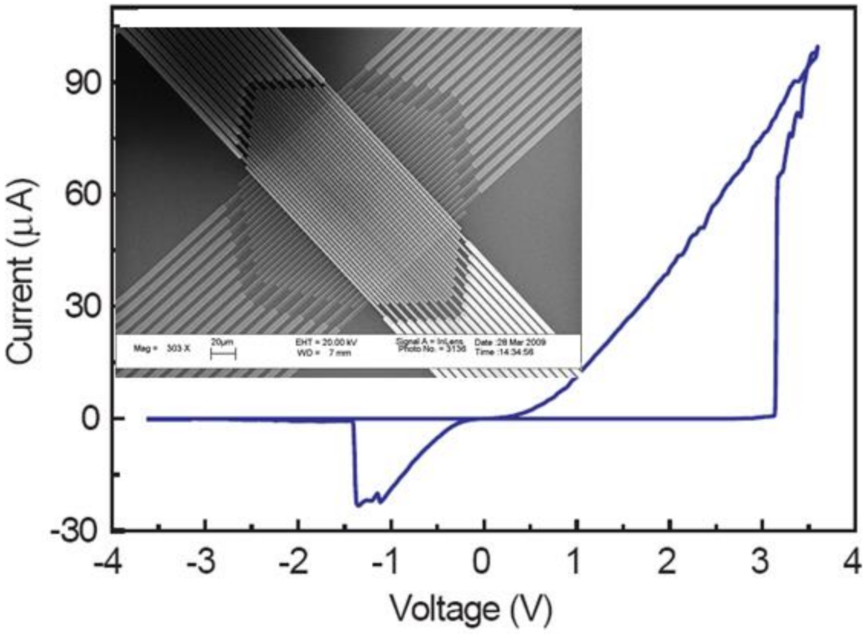
Microscopic Origin of Dynamic Filament Formation

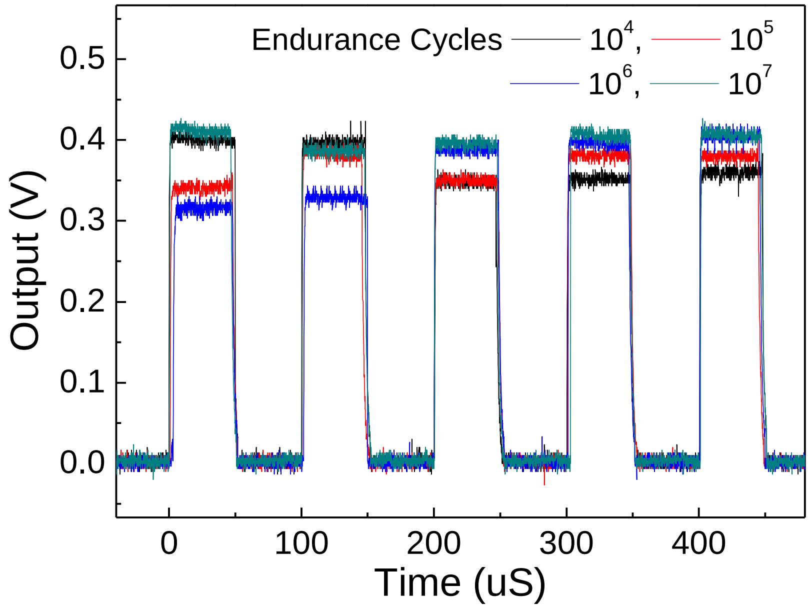


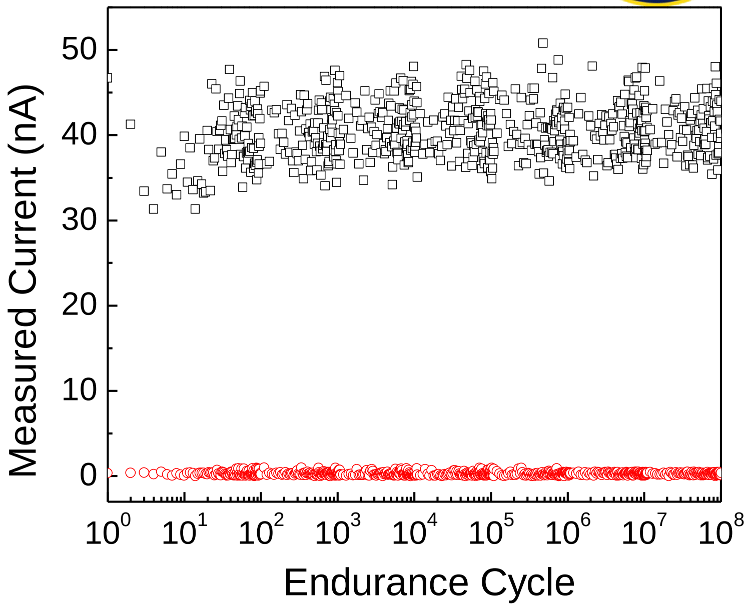
- Metal inclusions form bipolar electrodes, with redox processes happening at opposite sides
- Dissolution of the original Ag particle leads to new particle nucleation and growth at downstream position
- Resulting in effective Ag particle migration in the electric field direction

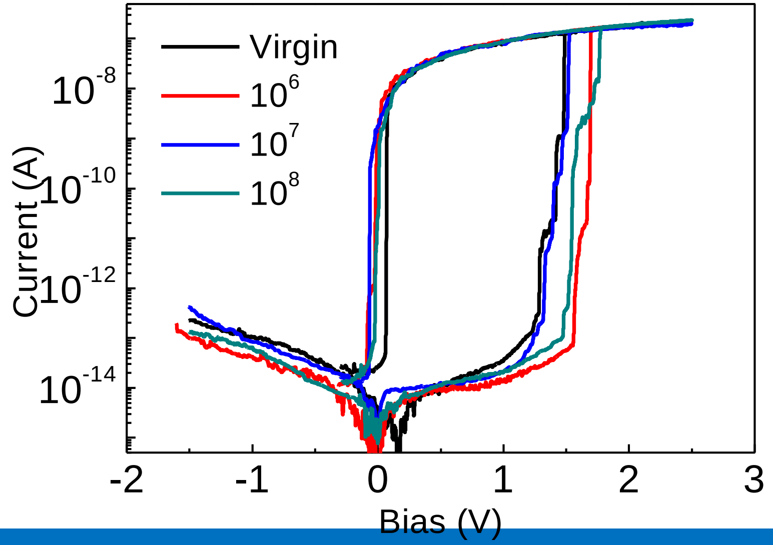
Y. Yang, Gao, Li, Pan, Tappertzhofen, Choi, Waser, Valov, W. Lu, Nature Communications 5, 4232 , 2014

RRAM Resistance Switching Characteristics









» 1e6 on/off

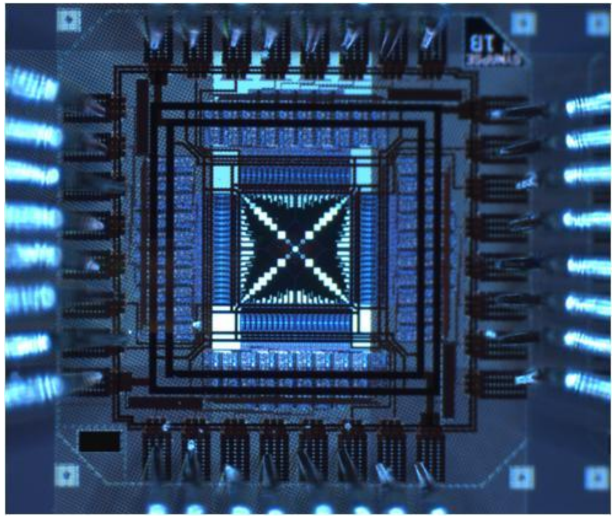
» 1e8 W/E endurance

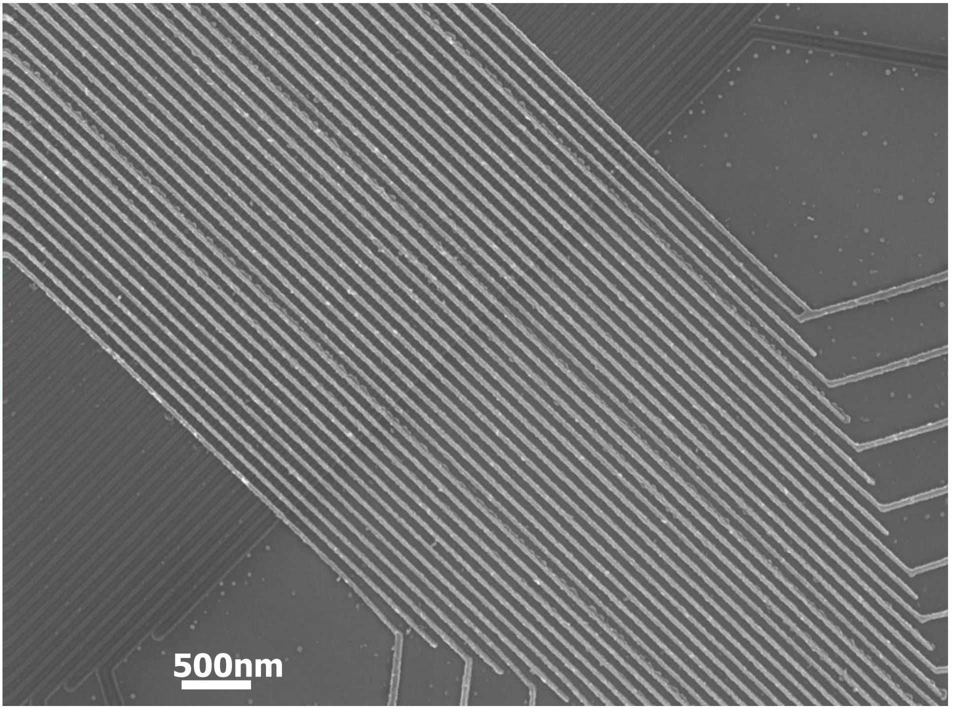
» Switching speed ~10ns

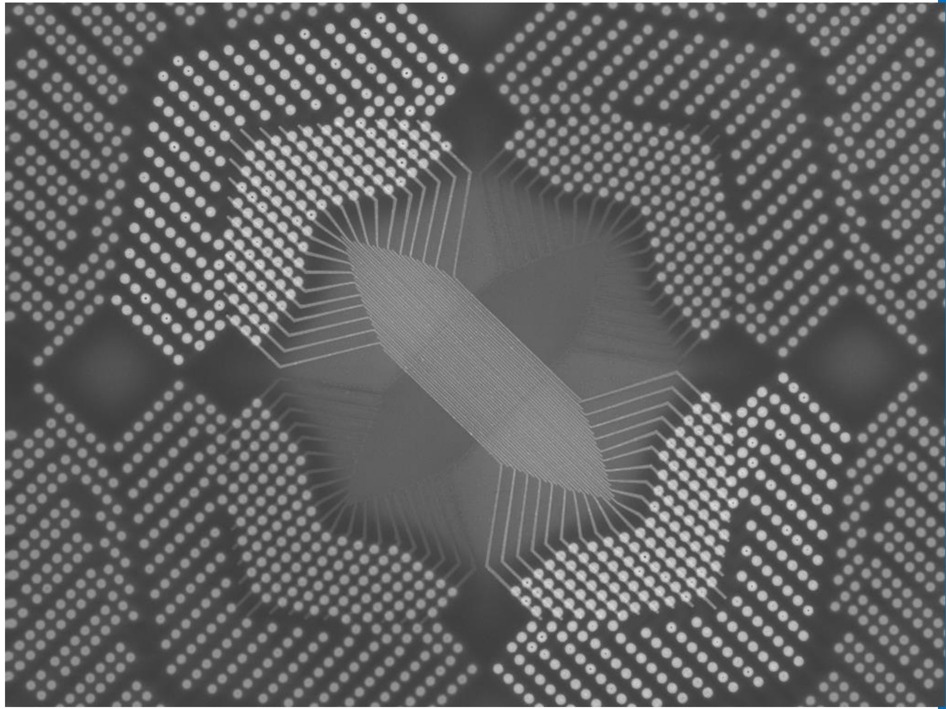
Jo, Kim, W. Lu, Nano Lett., 8, 392 (2008)

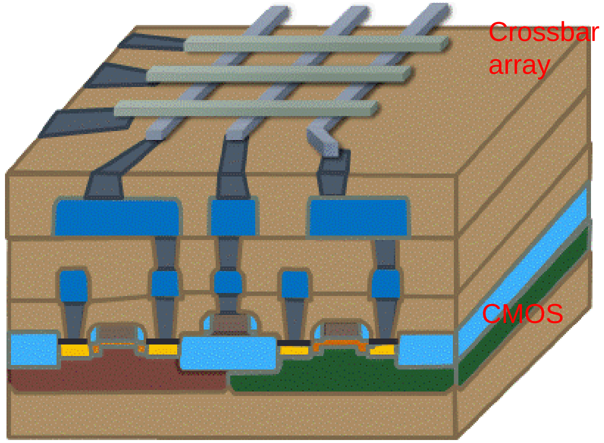
Kim, Jo, W. Lu, Appl. Phys. Lett. 96, 053106 (2010)

Integrated RRAM Crossbar/CMOS System









- Low-temperature process, RRAM array fabricated on top of CMOS
- CMOS provides address mux/demux
- RRAM array: 100nm pitch, 50nm linewidth with density of 10Gbits/cm^2
- CMOS units – larger but fewer units needed. 2n CMOS cells control n2 memory cells

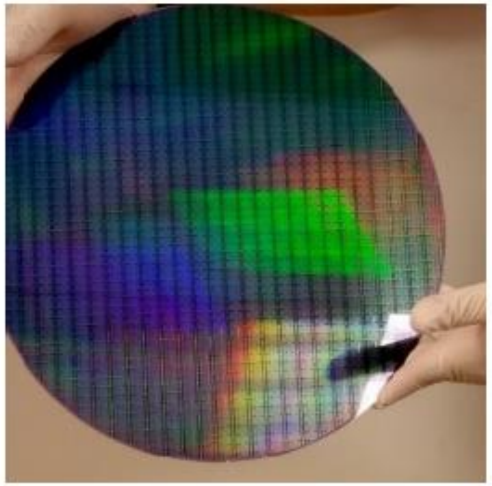
Kim, Gaba, Wheeler, Cruz-Albrecht, Srivinara, W. Lu Nano Lett., 12, 389–395 (2012).

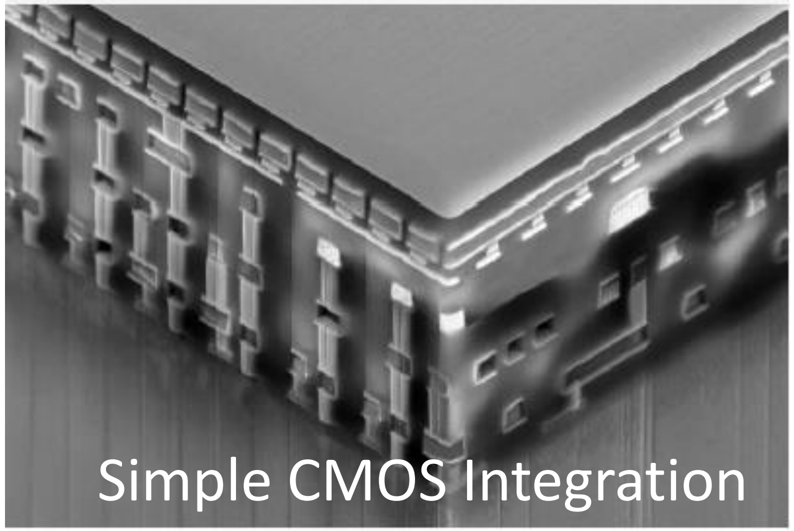
Outline

- Introduction - RRAM Devices
- Improving Computing Efficiency using RRAM Arrays
  - Bring memory as close to logic as possible
  - Neuromorphic computing in artificial neural networks
  - More bio-inspired networks, taking advantage of the internal ionic dynamics
  - In-memory computing for logic and arithmetic operations
- Future - reconfigurable systems based on a common physical fabric

RRAM as Embedded NVM

- **CMOS** Compatible
- **3D** Stackable, Scalable Architecture – Low thermal budget process
- **Architectures** proven include multiple Via schemes and Subtractive etching
- **Commercial Products** offered by several fabs



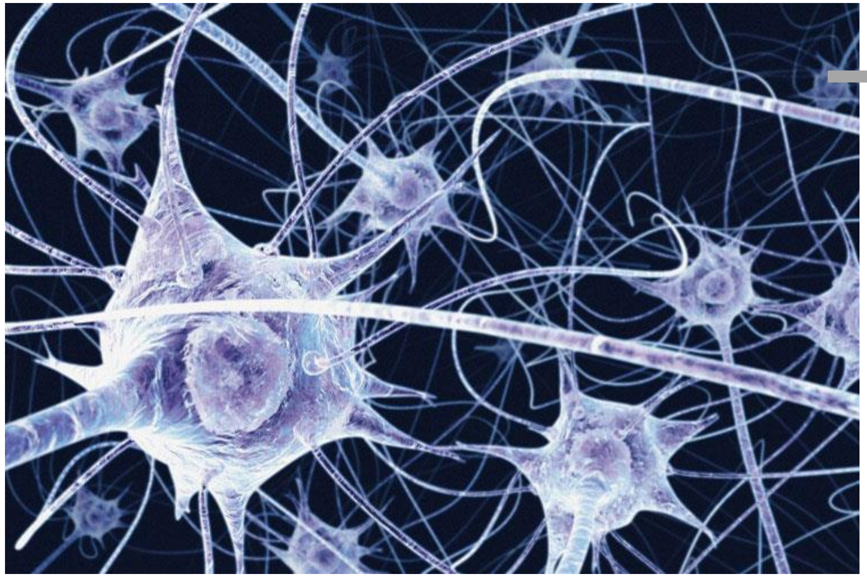

Simple CMOS Integration

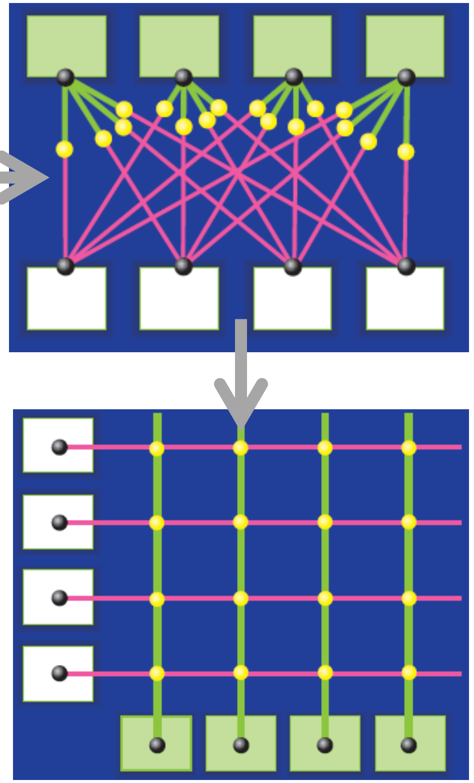
Outline

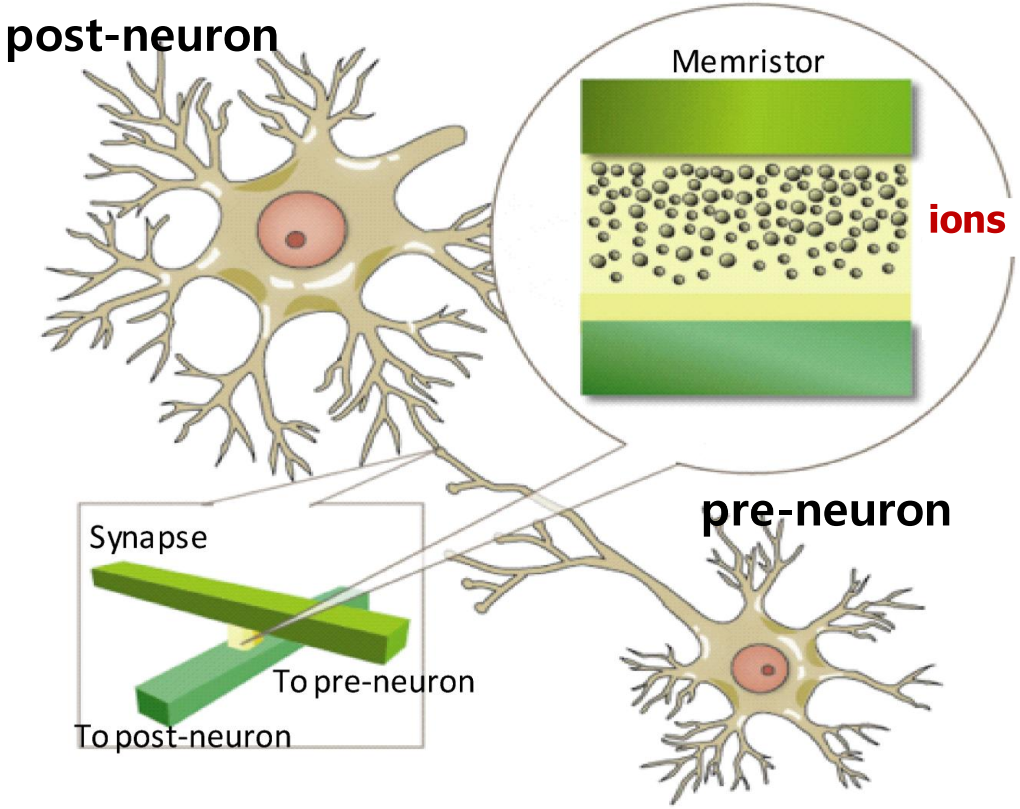
- Introduction - RRAM Devices
- Improving Computing Efficiency using RRAM Arrays
  - Bring memory as close to logic as possible
  - Neuromorphic computing in artificial neural networks
  - More bio-inspired networks, taking advantage of the internal ionic dynamics
  - In-memory computing for logic and arithmetic operations
- Future - reconfigurable systems based on a common physical fabric

RRAM Based Neural Network Hardware

- Synapse – reconfigurable two-terminal resistive switches
  - **Goal:** building bio-inspired, efficient artificial neural networks







Neuromorphic Computing with RRAM Arrays

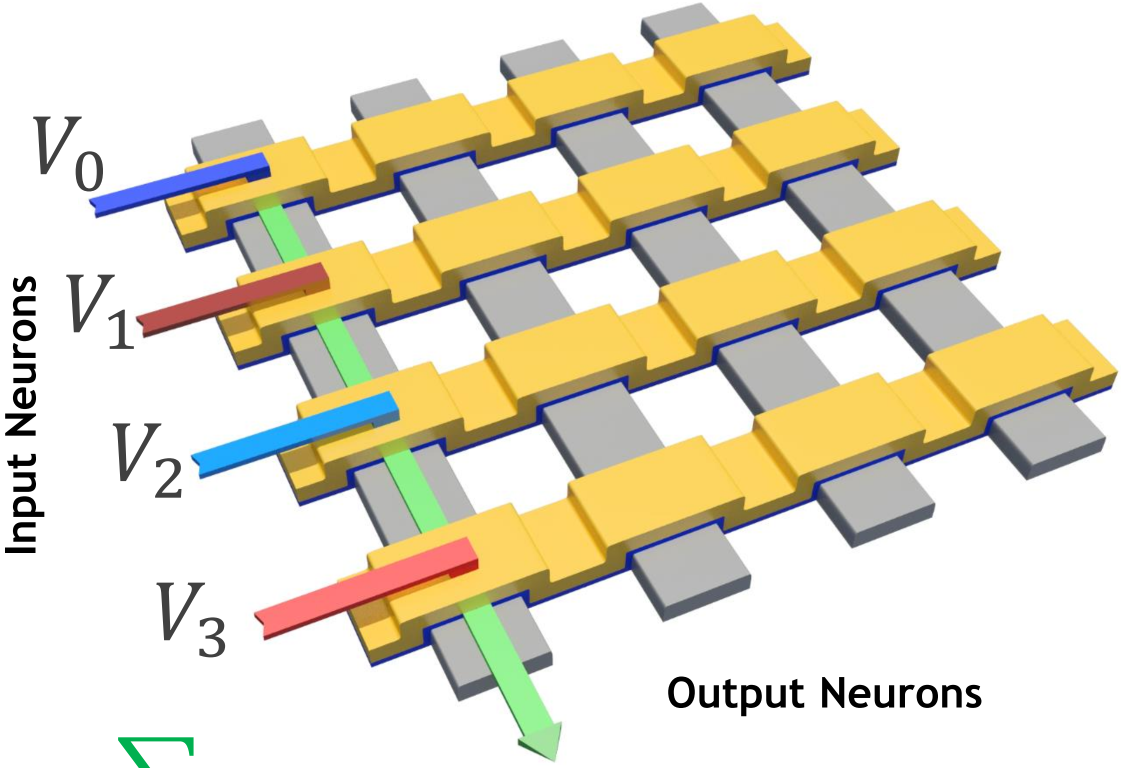
» RRAM perform learning and inference functions

- RRAM weights form dictionary elements (features)
- Image input, pixel intensity represented by widths of pulses
- RRAM array natively performs matrix operation:

$$\vec{I} = \vec{v} \cdot \overleftrightarrow{\Phi}$$

- Integrate and fire neurons
- Learning achieved by backpropagating spikes

M. A. Zidan, J. P. Strachan, and W. D. Lu, Nature Electronics 1: 22–29 (2018)



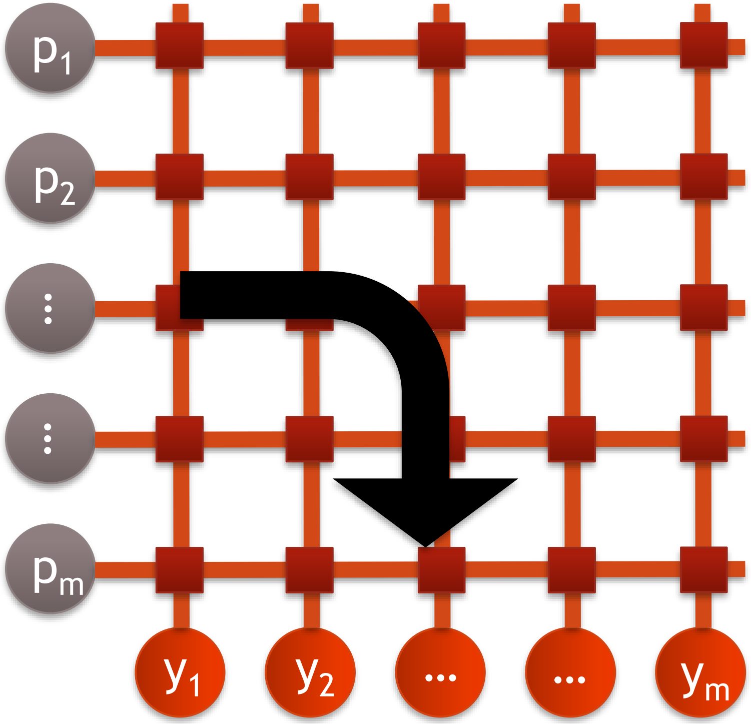
$$I_j = \sum V_i \cdot G_{i,j}$$

$$= V_0 \cdot G_{0,j} + V_1 \cdot G_{1,j} + V_2 \cdot G_{2,j} + V_3 \cdot G_{3,j} + \ldots$$

Forward and Backward Data Flow in RRAM Array

Forward Pass

Update neurons/activities

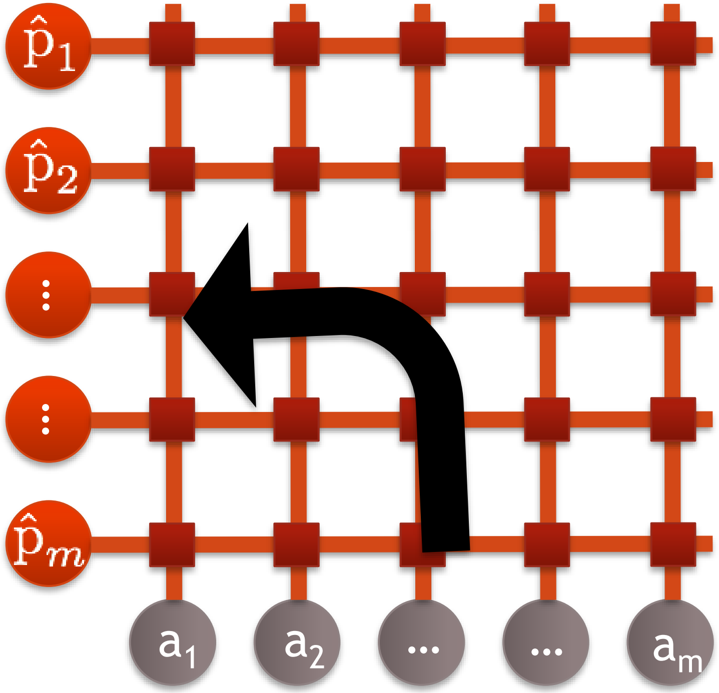


$$y = p^{\intercal} W$$

$$\frac{du}{dt} = \frac{1}{\tau}(-u + p^T \cdot \mathrm{W} - a \cdot (\mathrm{W}^{\intercal}\mathrm{W} - I)\,)$$

Backward pass

Update residual



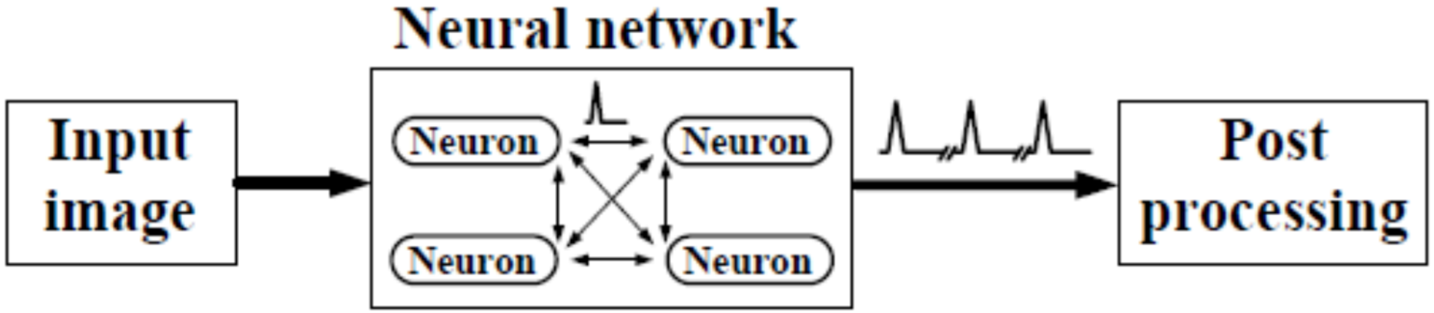
Sheridan et al., Nature Nanotechnology, 12, 784–789 (2017)

$$\hat{p} = aW^{\intercal}$$

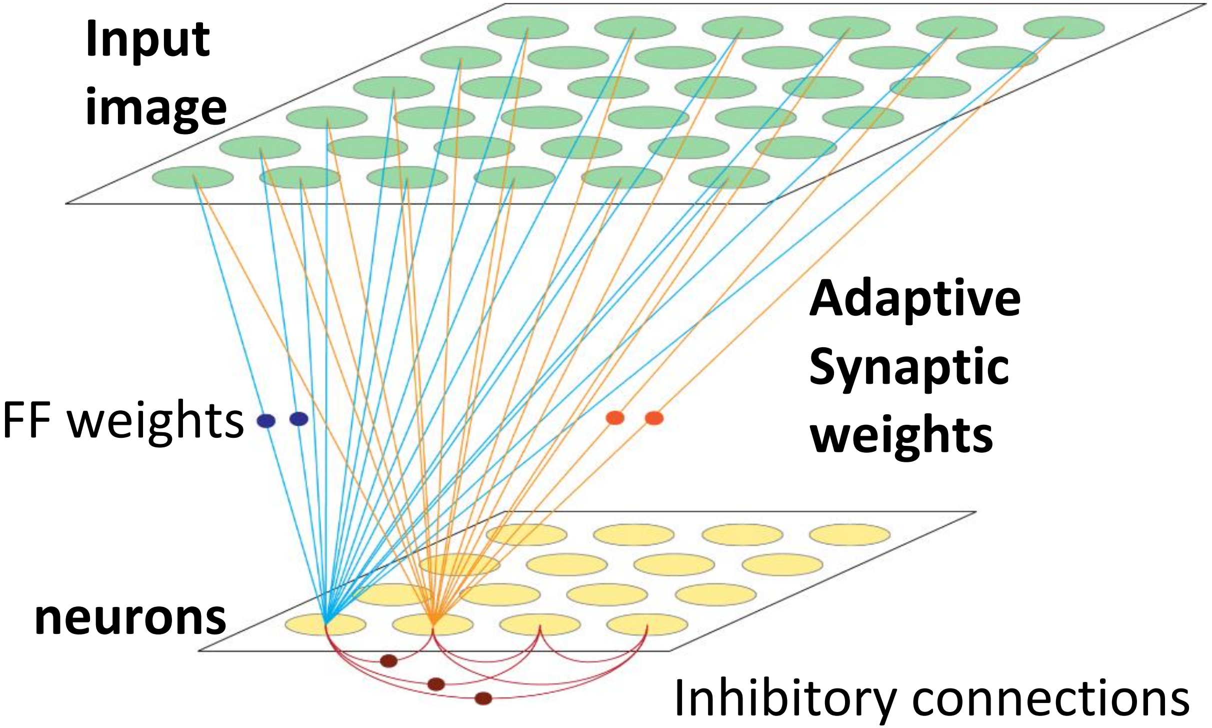
Neuron membrane potential

$$\frac{du}{dt} = \frac{1}{\tau}(-u + (p - \hat{p})^{\intercal} W + a)$$

RRAM Network for Image Processing

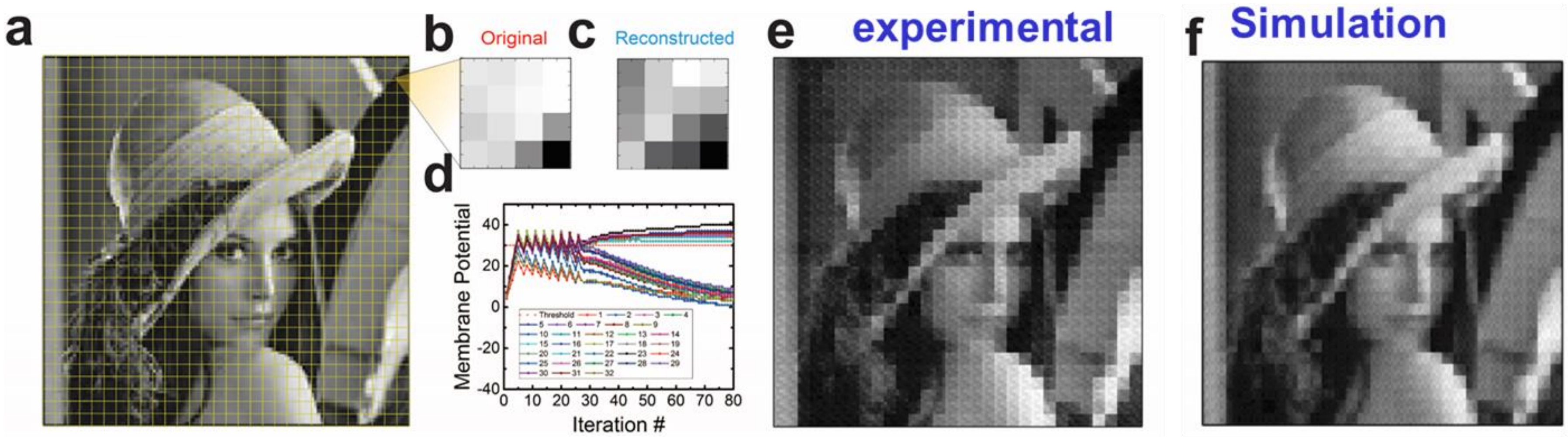


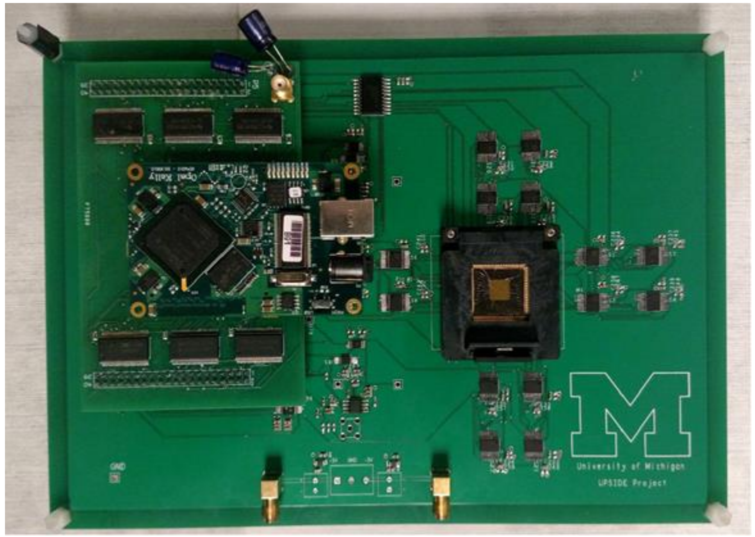
$$\frac{du}{dt} = \frac{1}{\tau}(-u + p^T \cdot \mathrm{W} - a \cdot (\mathrm{W}^\top \mathrm{W} - I)\,)$$

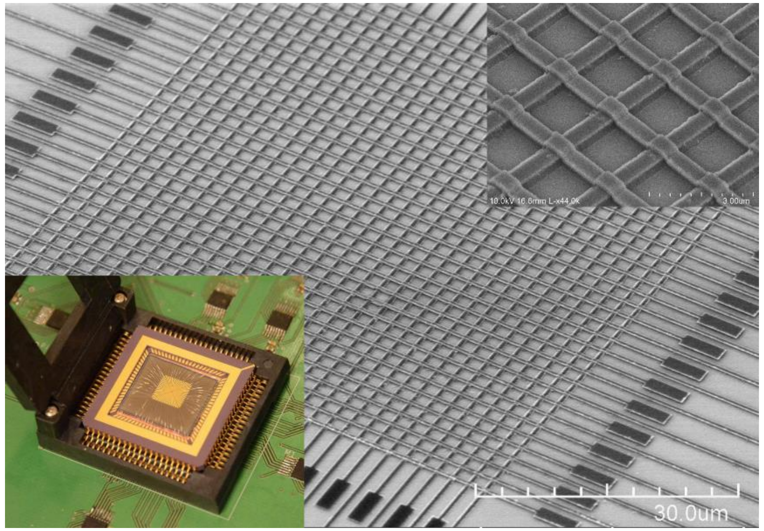


- Network adapt during training following local plasticity rules
- FF weights form neuron receptive fields (dictionary elements)
- Output as neuron firing rates
- Firing neurons determined by FF convolution and lateral inhibition.

Sparse Coding with RRAM Crossbar



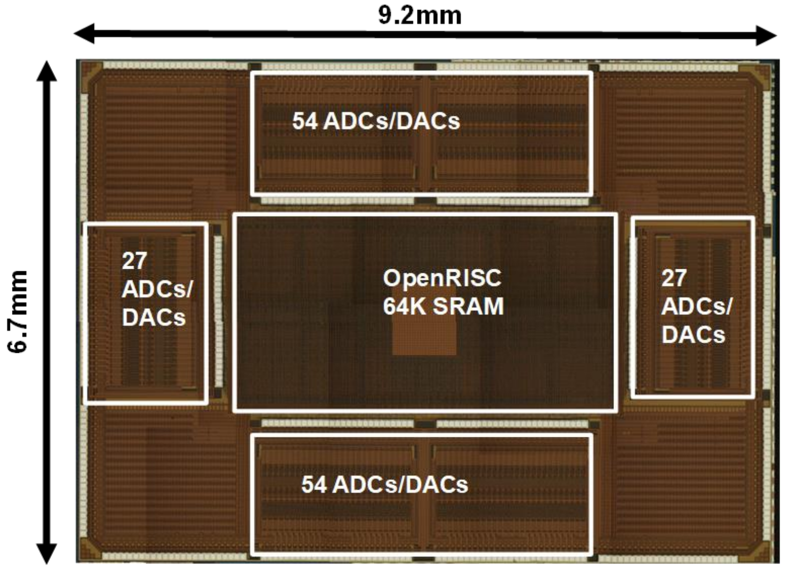


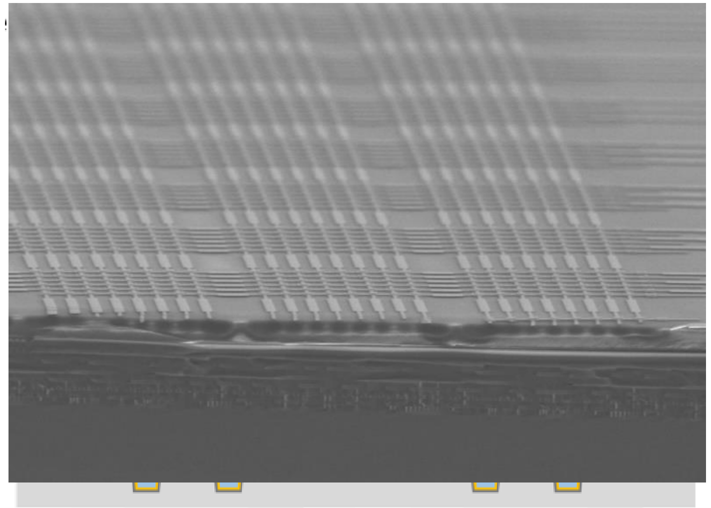


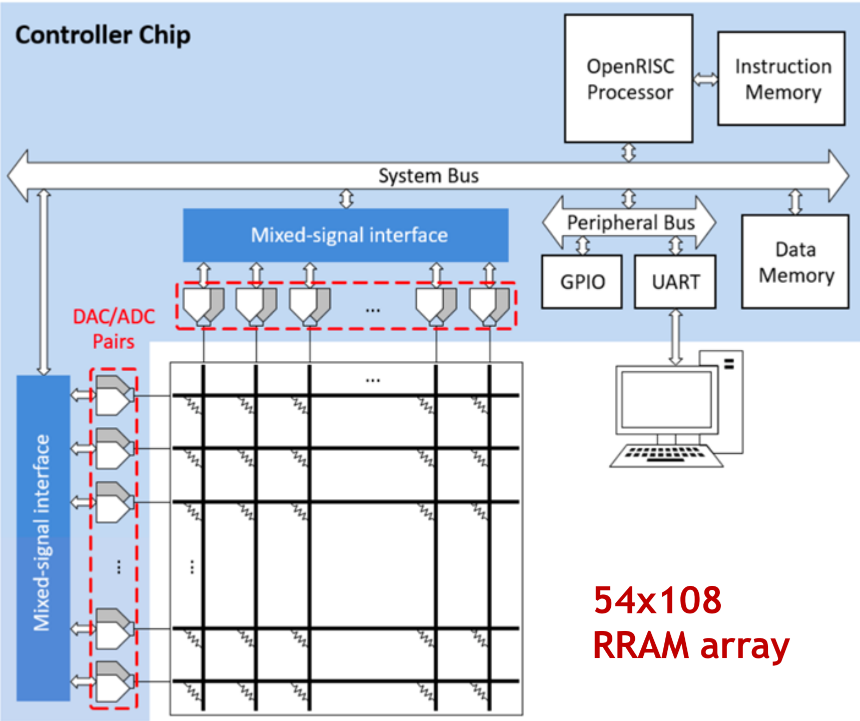
32x32 RRAM array

Sheridan et al., *Nature Nanotechnology*, 12, 784-789 (2017)

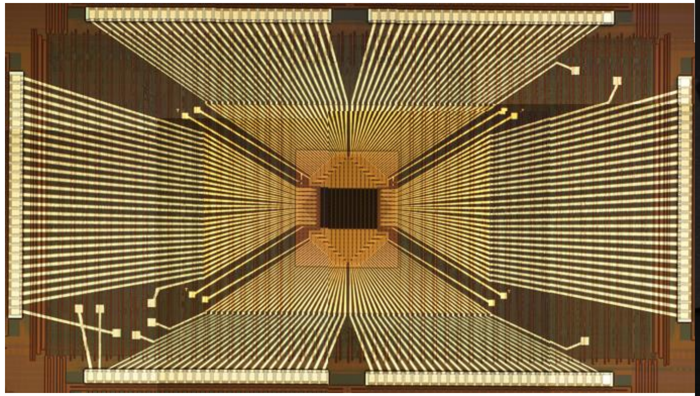
Fully integrated RRAM/CMOS chip

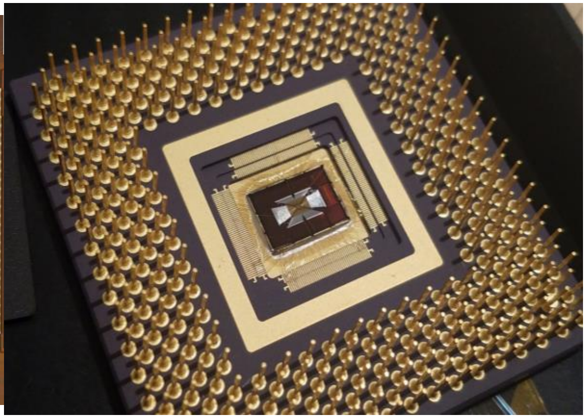






54x108 RRAM array

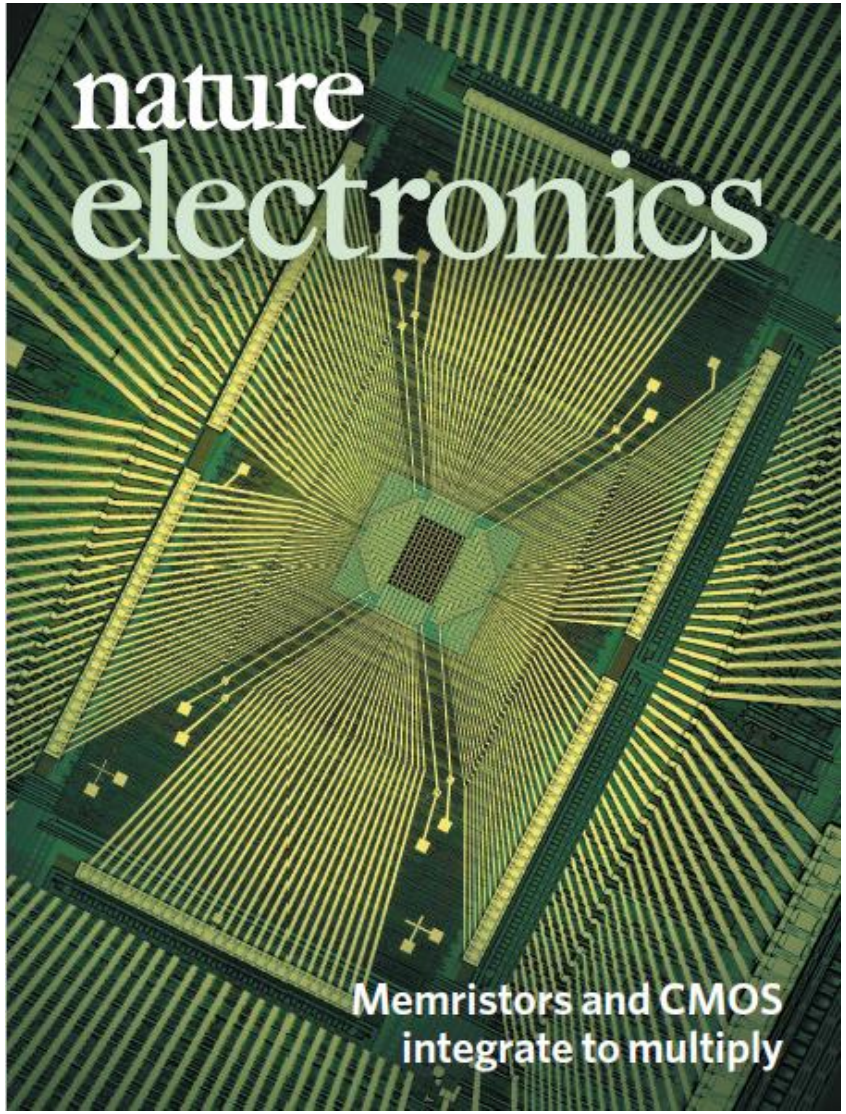


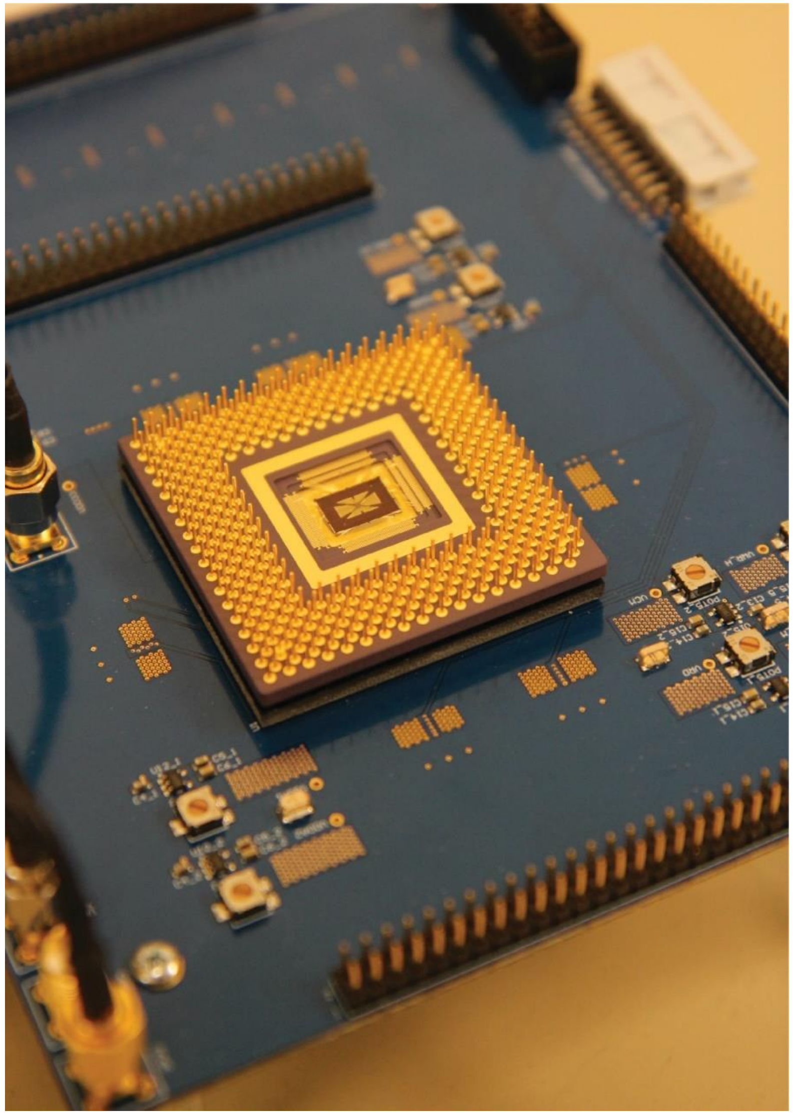


Fully integrated chip with all required ADCs, DACs, digital buses, and an on-chip OpenRISC Processor

Cai et al. Nature Electronics, DOI: 10.1038/s41928-019-0270-x

Fully integrated RRAM/CMOS chip





Fully integrated chip with all required ADCs, DACs, digital buses, and an on-chip OpenRISC Processor

Can run simple ML models end-to-end

Can be reprogrammed to run different ML models

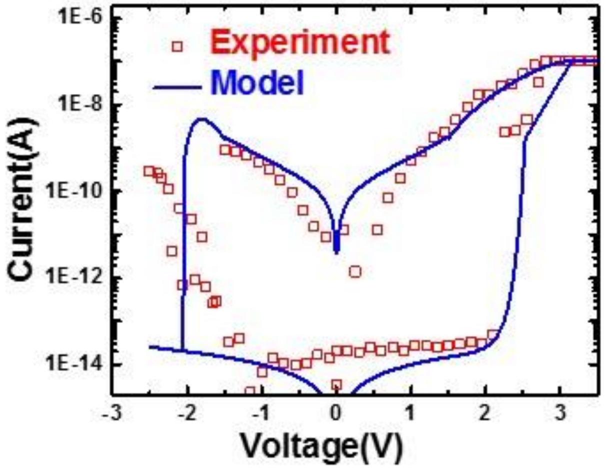
Cai et al. Nature Electronics, DOI: 10.1038/s41928-019-0270-x

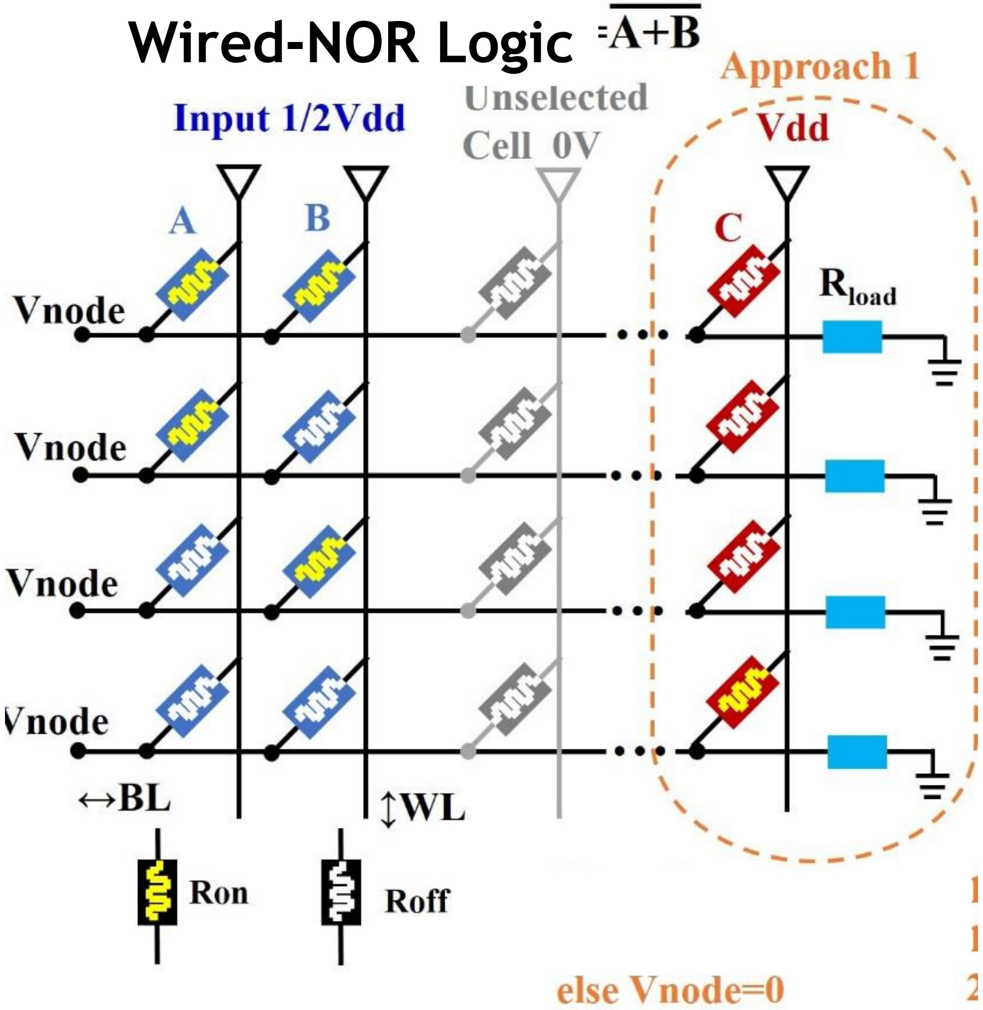
Outline

- Introduction - RRAM Devices
- Improving Computing Efficiency using RRAM Arrays
  - Bring memory as close to logic as possible
  - Neuromorphic computing in artificial neural networks
  - More bio-inspired networks, taking advantage of the internal ionic dynamics
  - In-memory computing for logic and arithmetic operations
- Future - reconfigurable systems based on a common physical fabric

In-memory Arithmetic Computing

- Parallel write enables efficient programming/ storage for new functions
- "computation" involves simple read operation
- (binary) RRAM device with low power (Ion < 100nA) and high on/off is used for the arithmetic operations.

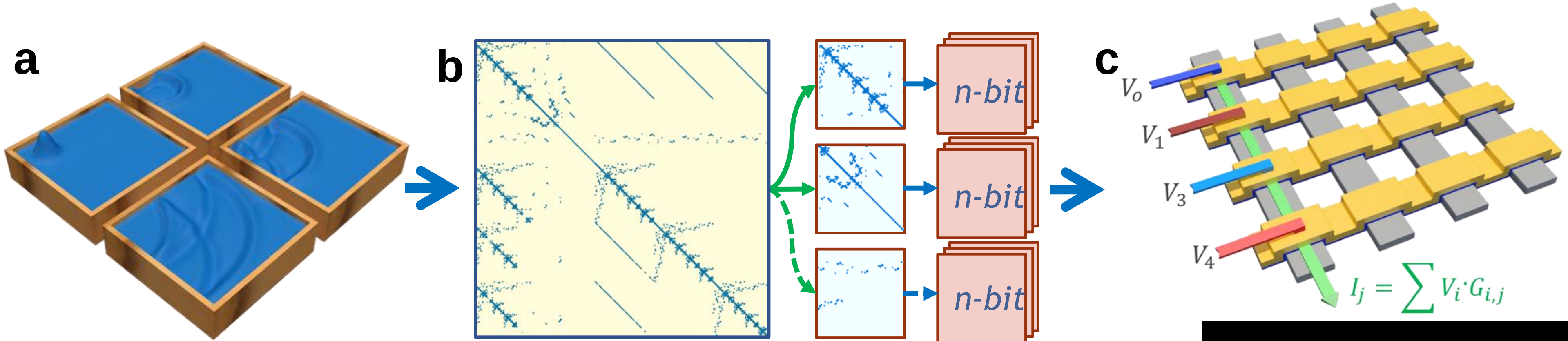




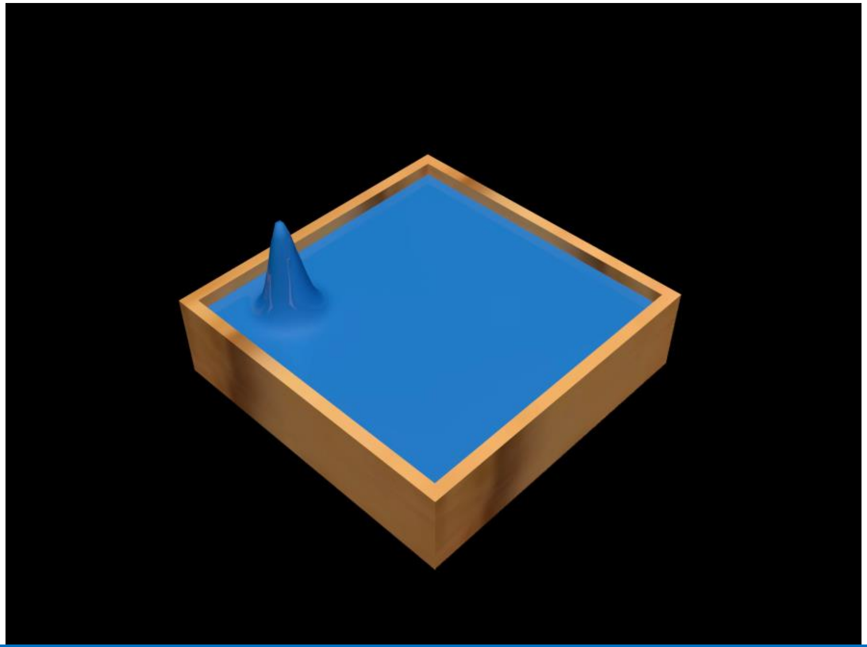
C. Bing, F. Cai, W. Ma, P. Sheridan, W. Lu, IEDM 2015

High Precision Arithmetic Computing

Solving partial-differential equations (PDEs)



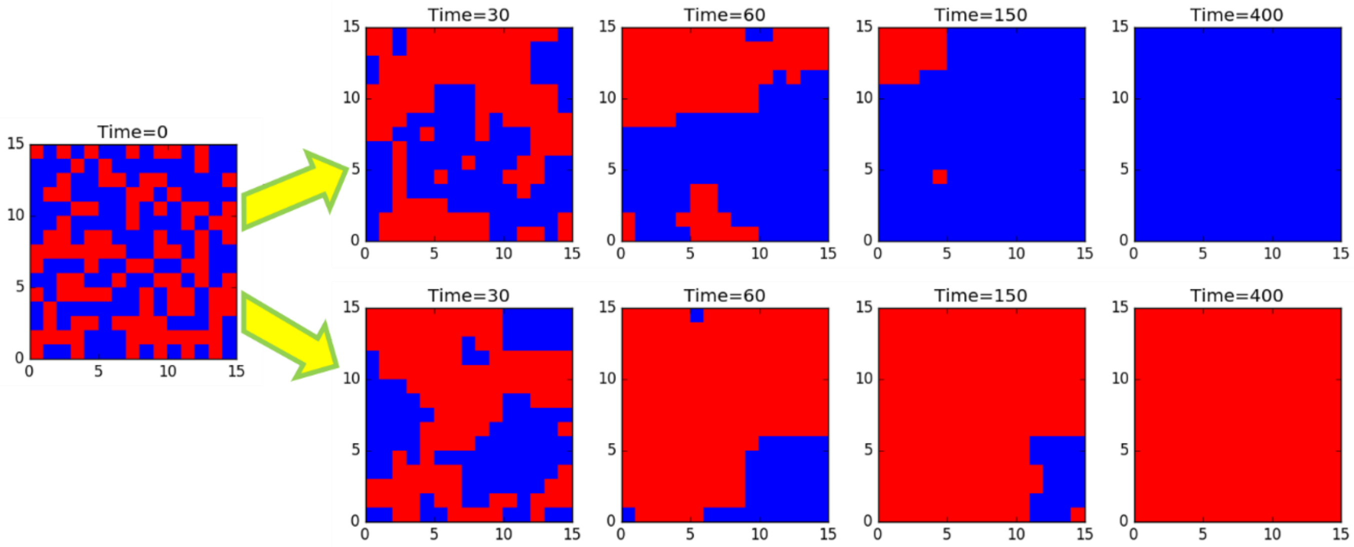
Solving an A·x=b problem in matrix form

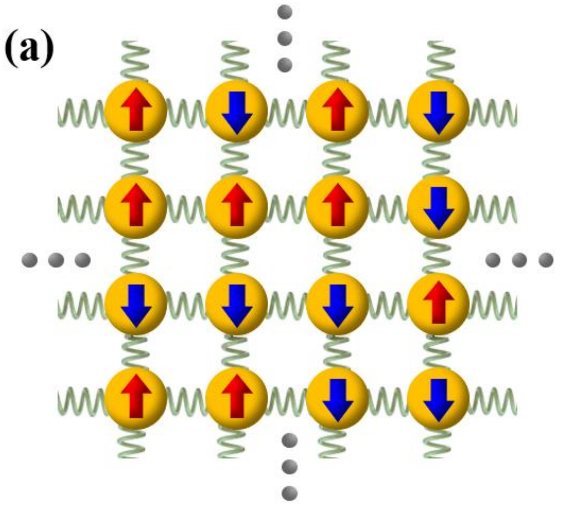


- Requires high precision and accurate solutions vs. neural networks which can tolerate low precision and inaccurate solutions
- Numerical simulation of water drop in a shallow pool

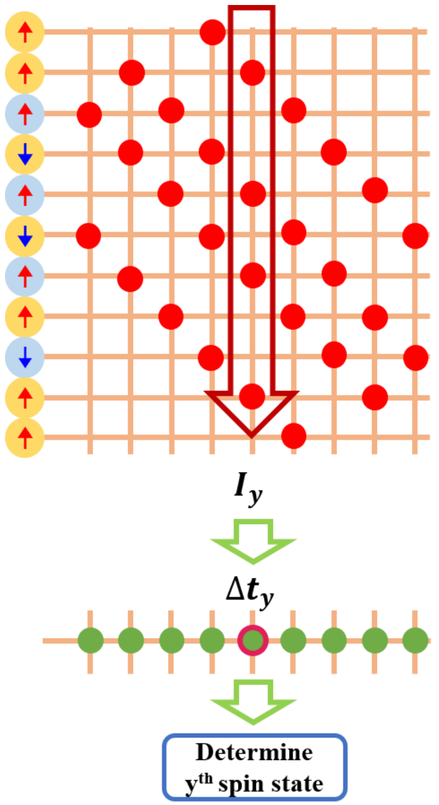
M. A. Zidan, Y.J. Jeong, J. Lee, B. Chen, S. Huang, M. J. Kushner, & W. D. Lu, *Nature Electronics*, 1, 411-420 (2018)

Hardware Acceleration of Simulated Annealing

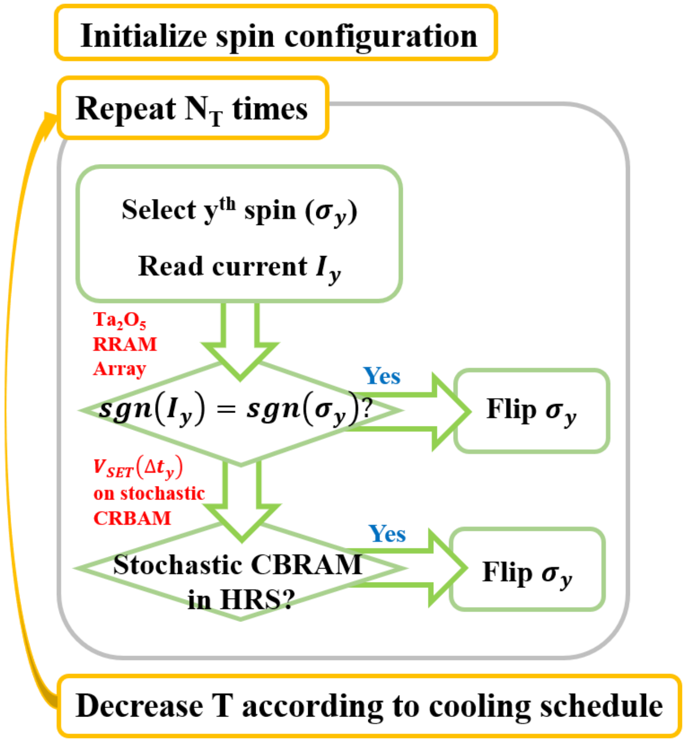




Experimentally implemented Simulated Annealing of a spin-glass problem using RRAM arrays



Determine y^{th} spin state



Initialize spin configuration

Repeat N_T times

Select y^{th} spin (σ_y)

Read current I_y

Ta_2O_5 RRAM Array

$sgn(I_y) = sgn(\sigma_y)$?

Yes

Flip σ_y

$V_{SET}(\Delta t_y)$ on stochastic CRBAM

Stochastic CBRAM in HRS?

Yes

Flip σ_y

Decrease T according to cooling schedule

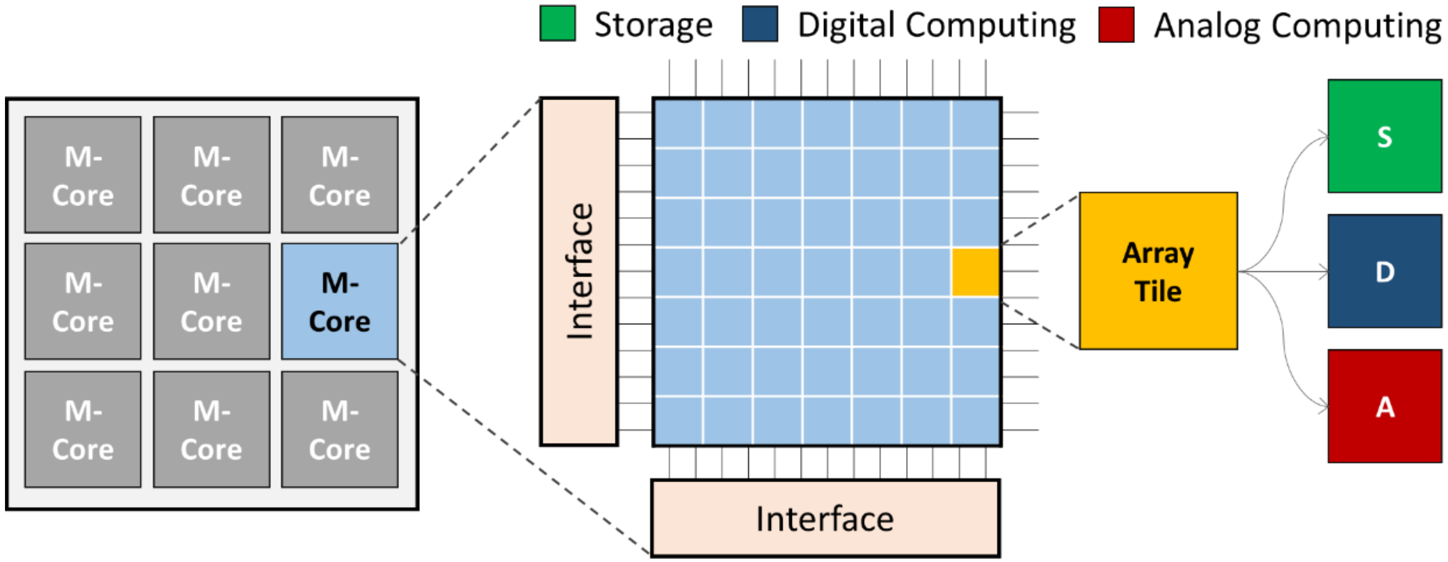
J. Shin, et al. IEDM 2018

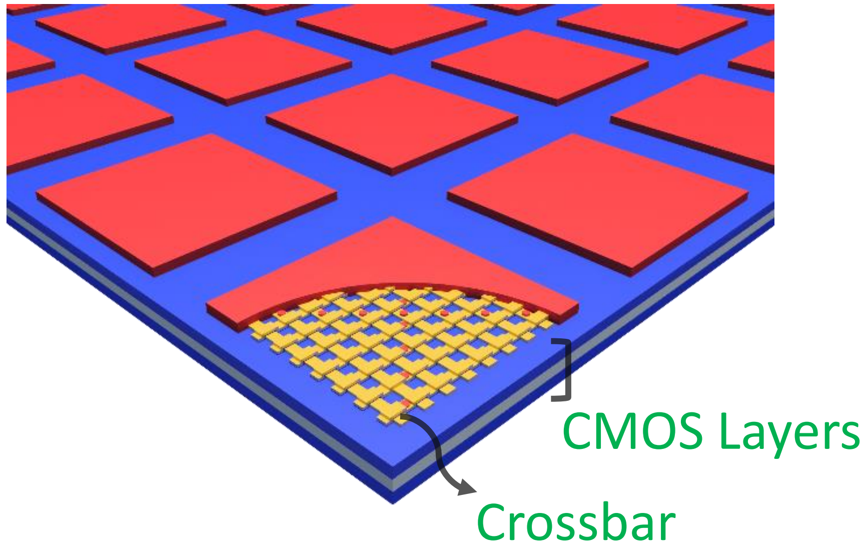
Outline

- Introduction - RRAM Devices
- Improving Computing Efficiency using RRAM Arrays
  - Bring memory as close to logic as possible
  - Neuromorphic computing in artificial neural networks
  - More bio-inspired networks, taking advantage of the internal ionic dynamics
  - In-memory computing for logic and arithmetic operations
- Future - reconfigurable systems based on a common physical fabric

Dynamically reconfigurable Computing Fabric

» A reconfigurable hardware system with modular reconfigurable blocks

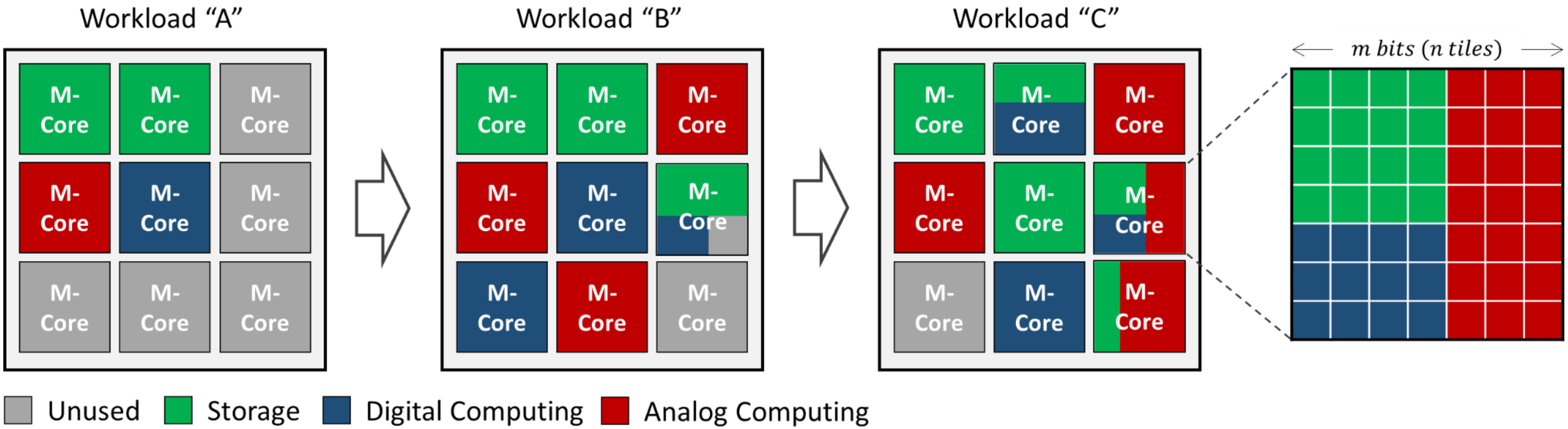




- Hierarchically structured interconnects: locally dense connection + globally asynchronous serial link
- Reconfigurable computing modules at both fine-grained and coarse-grained levels

M. Zidan, Y. Jeong, J. H. Shin, C. Du, Z. Zhang, and W. D. Lu, IEEE Trans Multi-Scale Comp Sys, DOI 10.1109/TMSCS.2017.2721160 (2017)

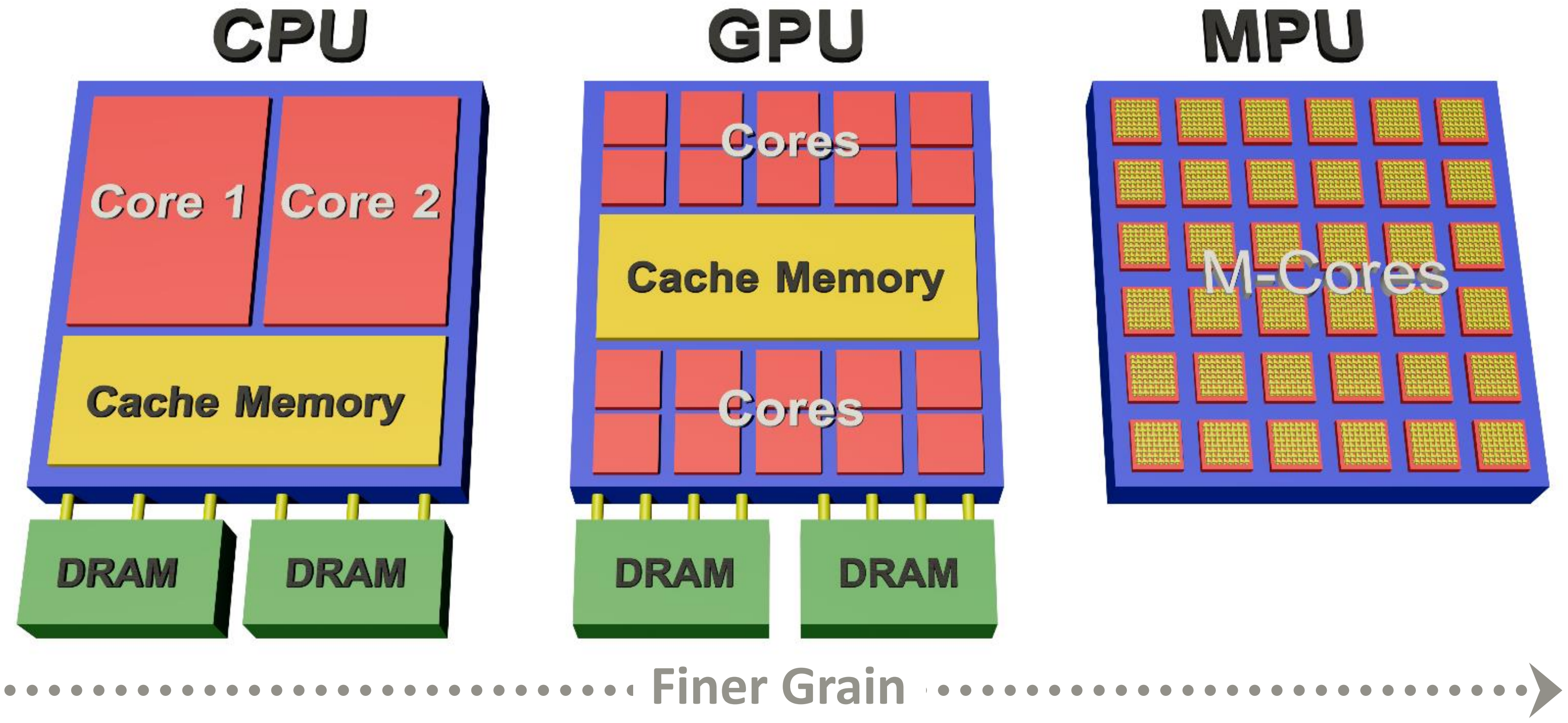
Dynamically reconfigurable Computing Fabric



- "General" purpose by design: the same hardware supports different tasks – image, video, speech, …
- Dense local connection, sparse global connection
- Run-time, dynamically reconfigurable. Function defined by software.

M. Zidan, Y. Jeong, J. H. Shin, C. Du, Z. Zhang, and W. D. Lu, IEEE Trans Multi-Scale Comp Sys, DOI 10.1109/TMSCS.2017.2721160 (2017)

Possible evolutions



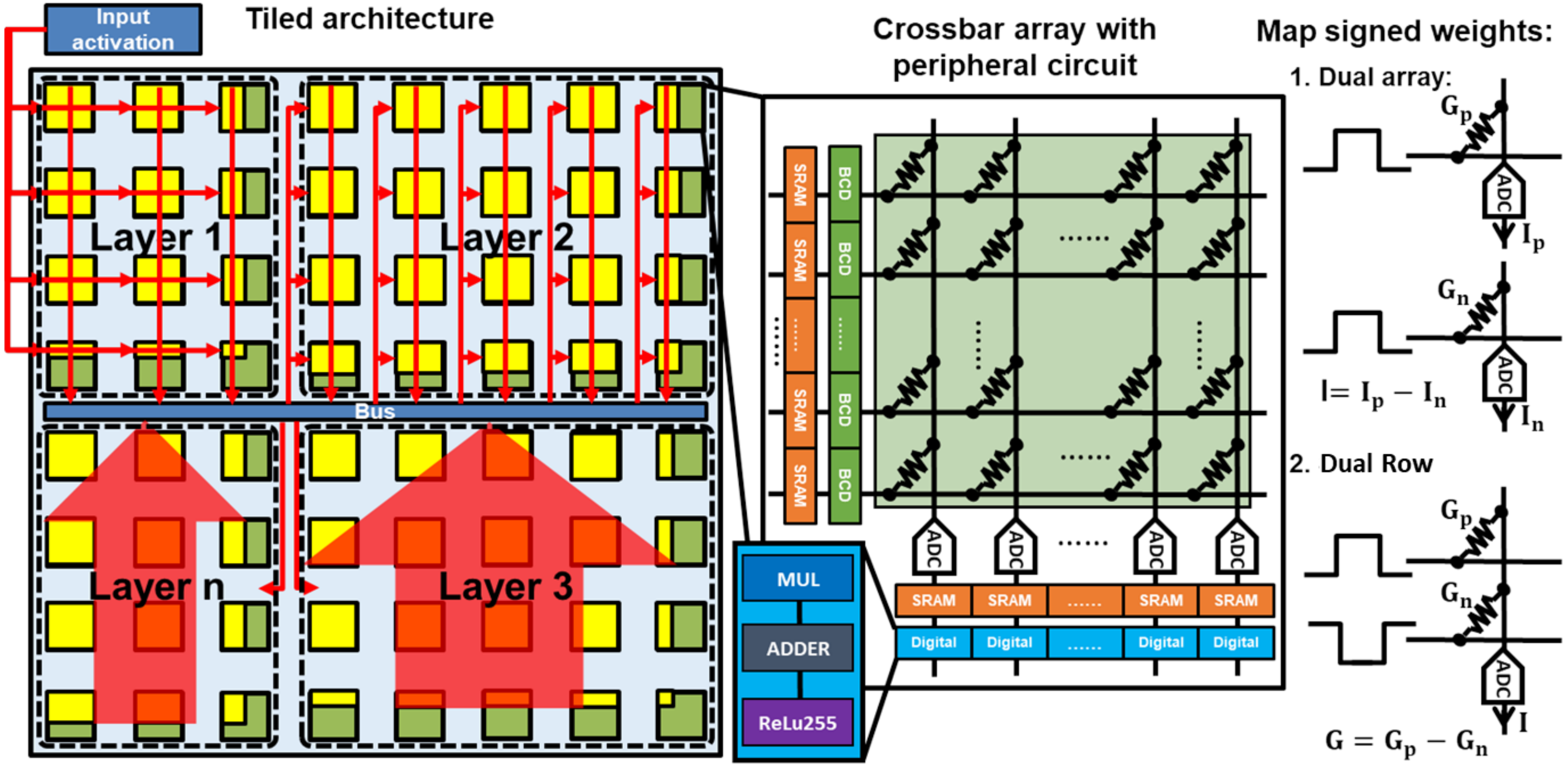
- Course grain cores
- Memory Bottleneck

- Finer grain cores
- Faster memory access

- Device level computing
- In-memory Computing

M. A. Zidan, J. P. Strachan, and W. D. Lu, Nature Electronics 1: 22–29 (2018)

Implementing Large Networks: Modular Systems



X. Wang

Q. Wang

Tiled architecture for practical model implementation:

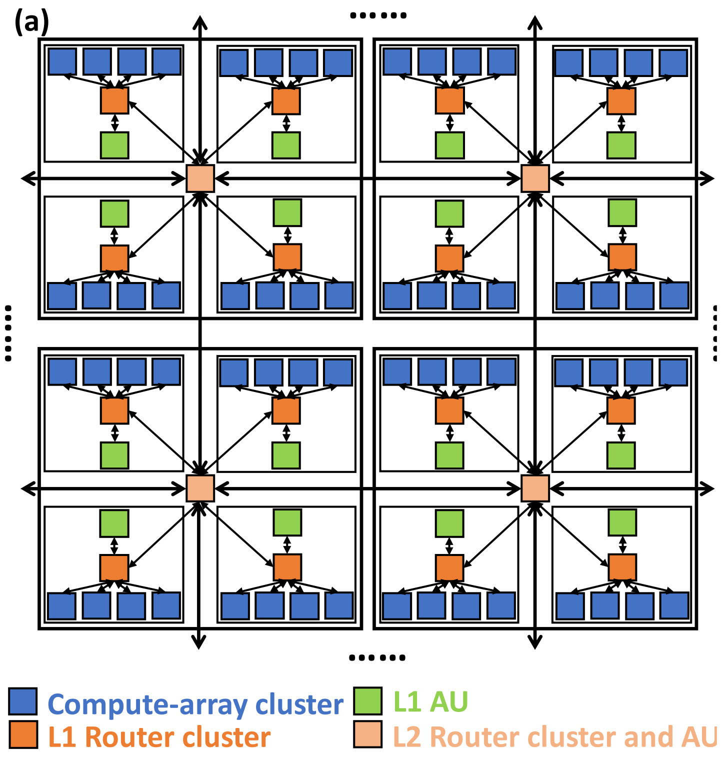
- Weight mapping
- ADC quantization, partial products
- Device and circuit nonideality

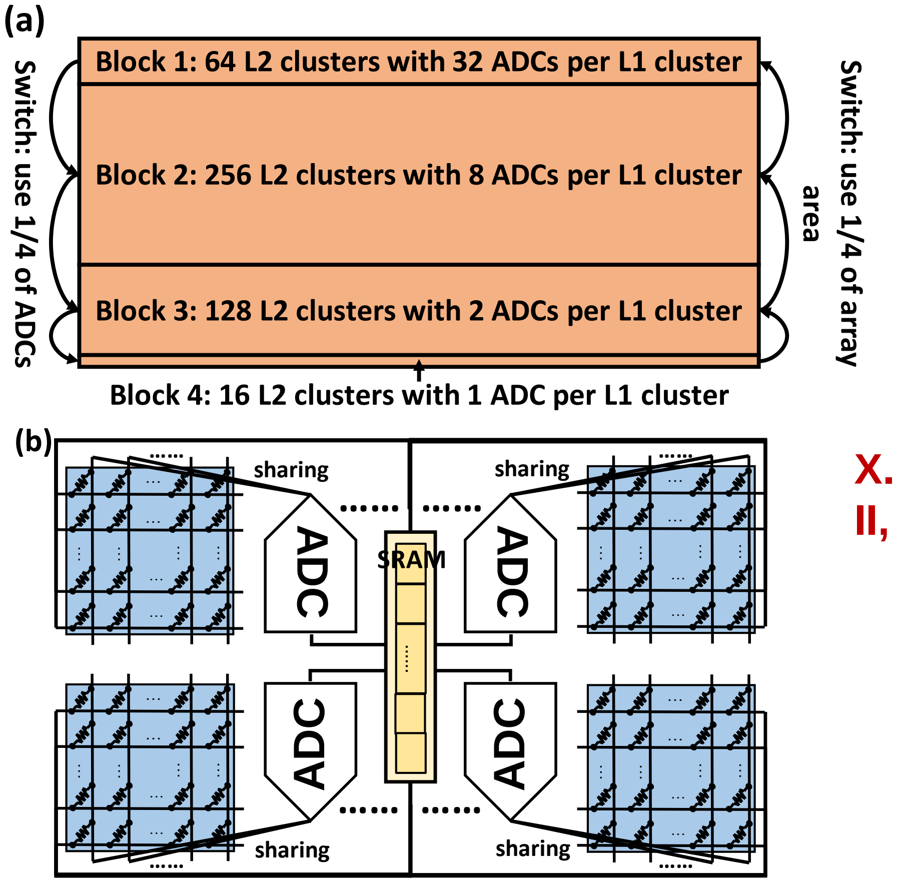
Wang et al. IEDM 14.4 (2019)

TAICHI: General RRAM IMC Chip Design

- The chip design should be compatible with a wide range of models.
- A general RRAM IMC chip based on analog RRAM tiles and a heterogeneous NoC structure
- Optimally designed blocks based on four types of compute arras work well for different popular models.

X. Wang





Block 1: 64 L2 clusters with 32 ADCs per L1 cluster

Block 2: 256 L2 clusters with 8 ADCs per L1 cluster

Block 3: 128 L2 clusters with 2 ADCs per L1 cluster

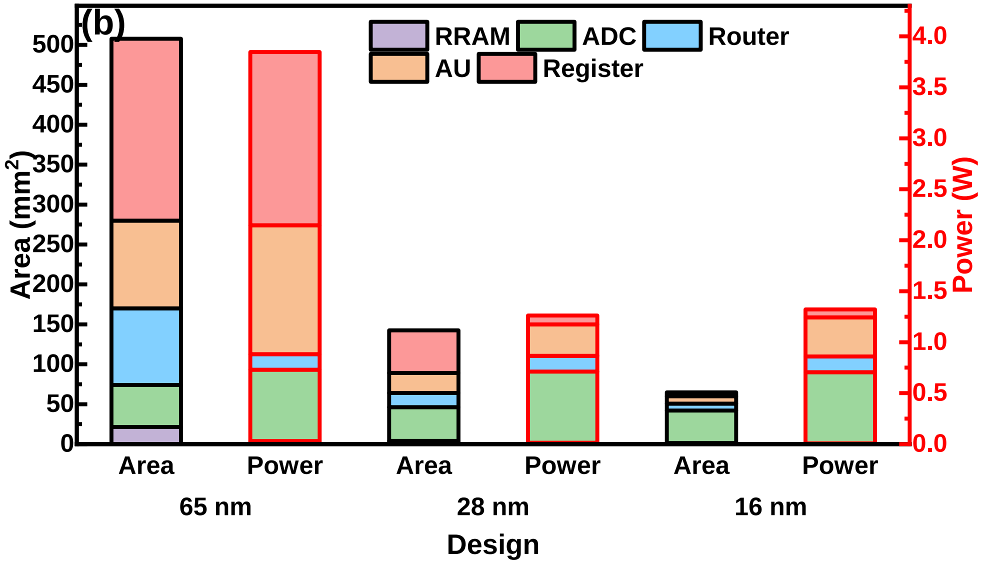
Block 4: 16 L2 clusters with 1 ADC per L1 cluster

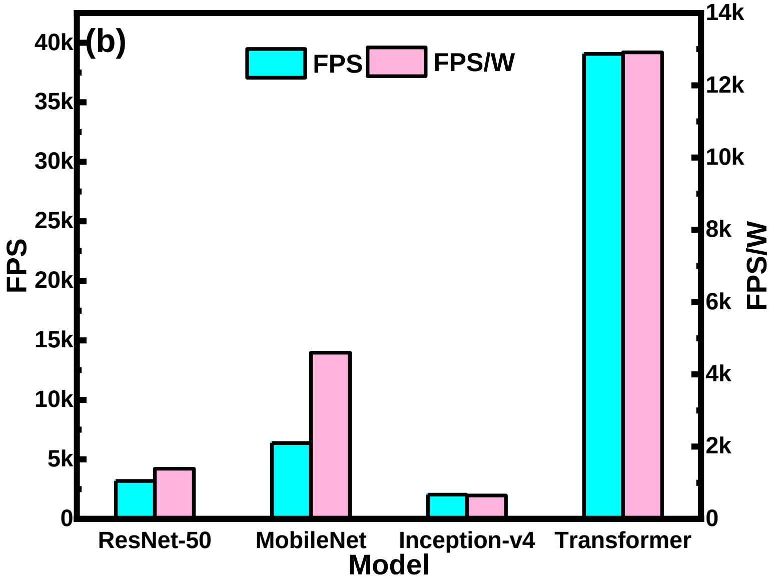
X. Wang, et al. IEEE TCAS II, 69(2), 559–563. (2022)

TAICHI: Chip Performance Analysis

- Register and ADC dominate the chip area and power.
- 70TOPS/W (int-8) estimated at 28nm.
- High throughput and energy efficiency for common models based on the single chip design: 1391 FPS/W (ResNet-50), 4602 FPS/W (MobileNet), 646 FPS/W (Inception-v4) and 12911 FPS/W (Transformer).

X. Wang



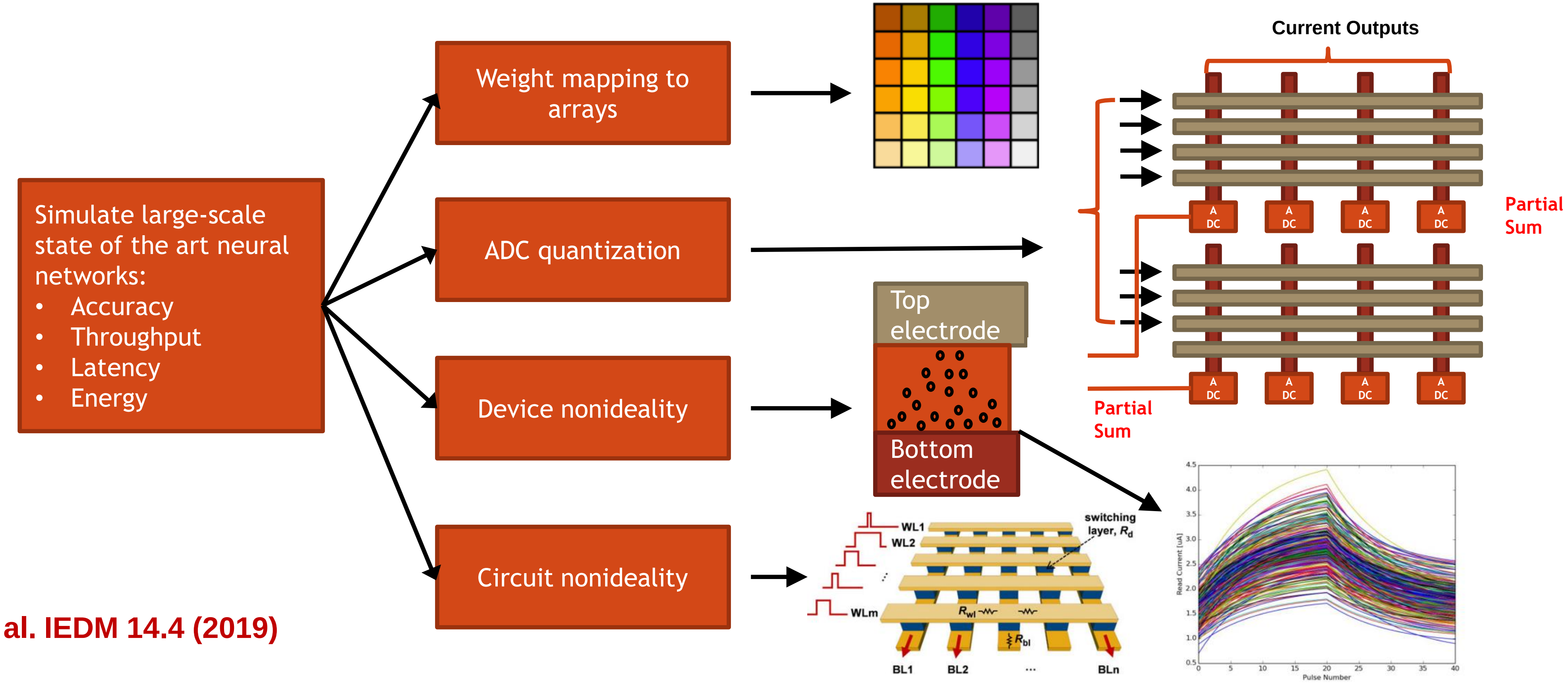


X. Wang, IEEE TCAS II, 69(2), 559–563. (2022)

Effects of Device and Architecture Non-idealities

X. Wang

Q. Wang



Simulate large-scale state of the art neural networks:

- Accuracy
- Throughput
- Latency
- Energy

Weight mapping to arrays

ADC quantization

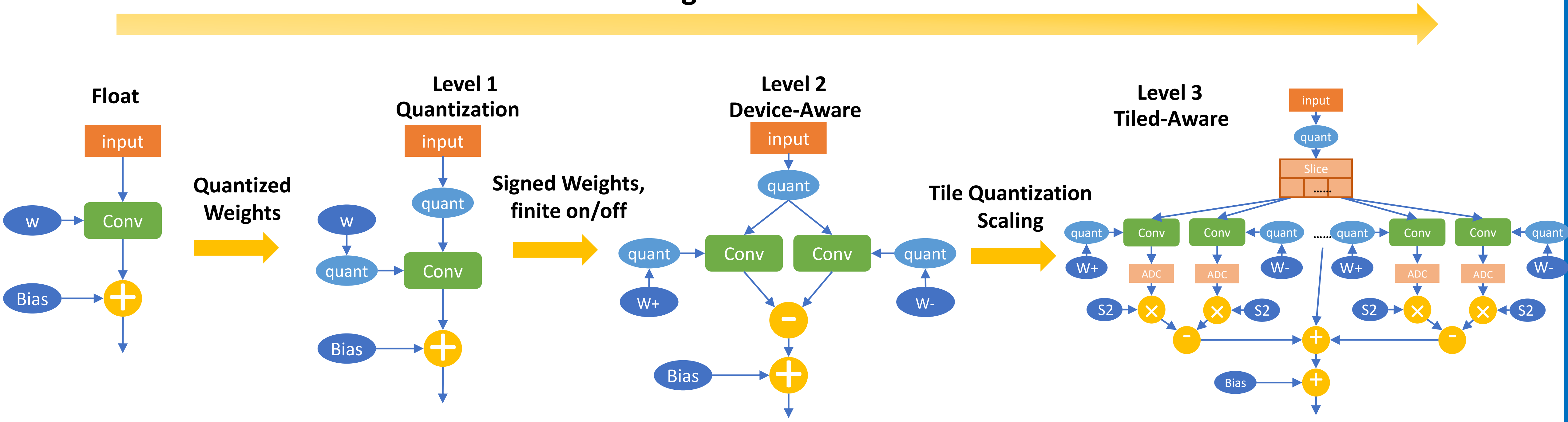
Device nonideality

Circuit nonideality

Wang et al. IEDM 14.4 (2019)

Architecture-Aware Training Topology

Adding more architecture details



Float

Quantized Weights

Level 1 Quantization

Signed Weights, finite on/off

Level 2 Device-Aware

Tile Quantization Scaling

Level 3 Tiled-Aware

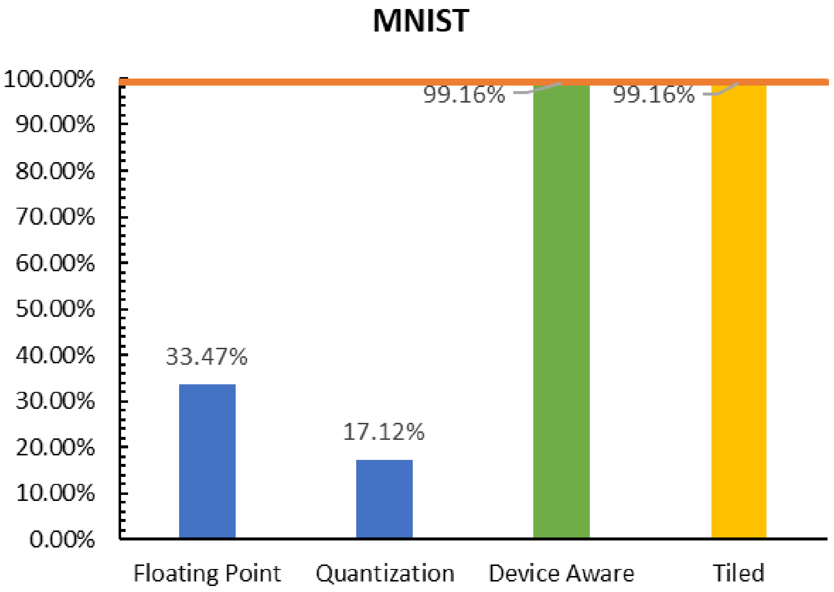
- Architecture details need to be included during training processes to produce accurate inference results

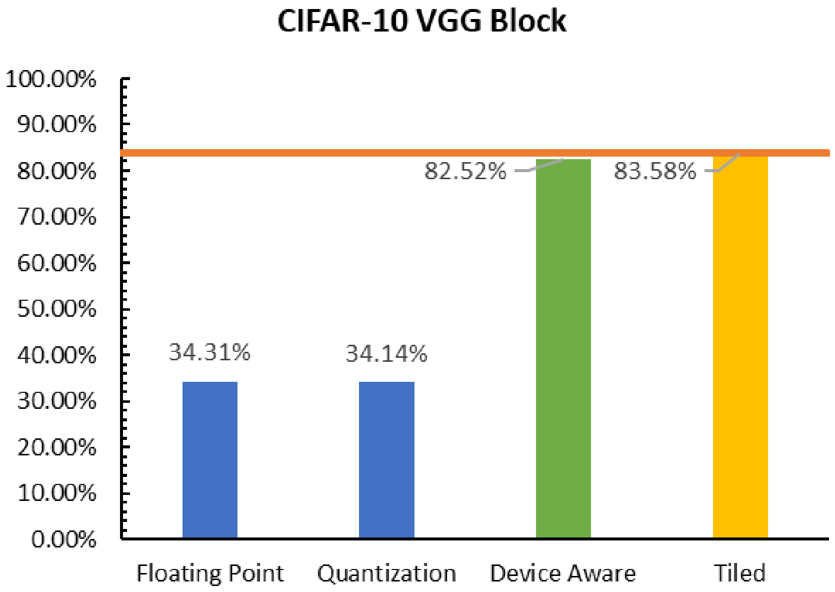
Q. Wang, Y. Park, and W. D. Lu, "Device Non-Ideality Effects and Architecture-Aware Training in RRAM In-Memory Computing Modules," ISCAS, 2021

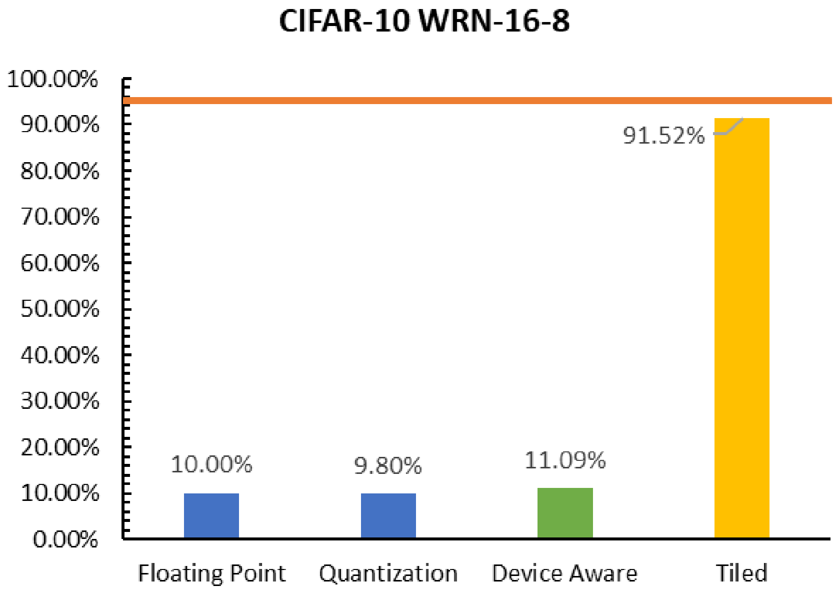
Inference Accuracy

- Training: Levels 0-3
- Inference: Tiled Architecture (Level-3)
- Weight Precision: 4bits
- On/off Ratio: 10

- Array size: 256x64
- ADC: 8bit


MNIST

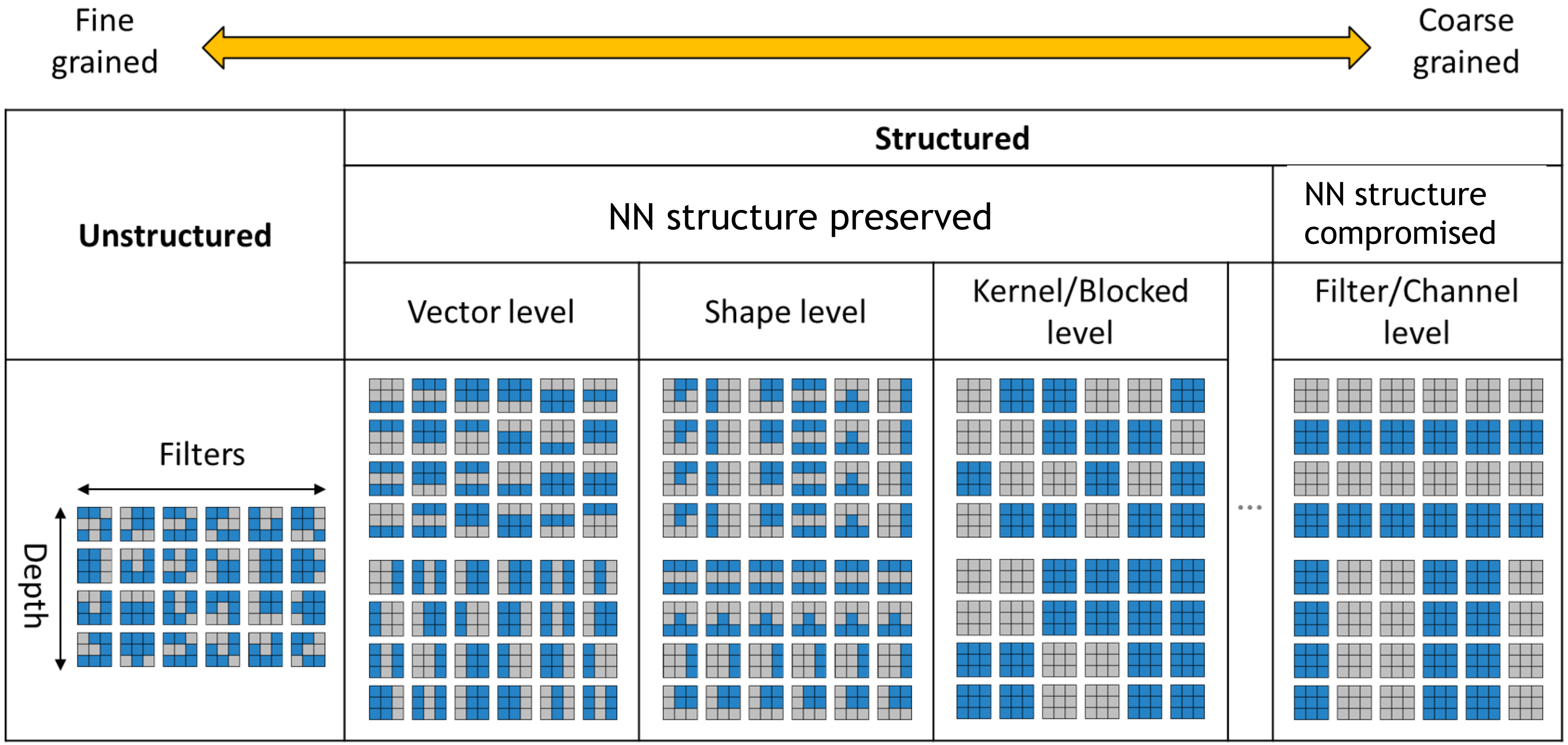

CIFAR-10 VGG Block


CIFAR-10 WRN-16-8

- In relatively complex models, the tiled architecture has to be accounted for during training to achieve acceptable accuracy

Q. Wang, Y. Park, and W. D. Lu, “Device Non-Ideality Effects and Architecture-Aware Training in RRAM In-Memory Computing Modules,” ISCAS, 2021

Structured Pruning to Fit Larger Models on Chip



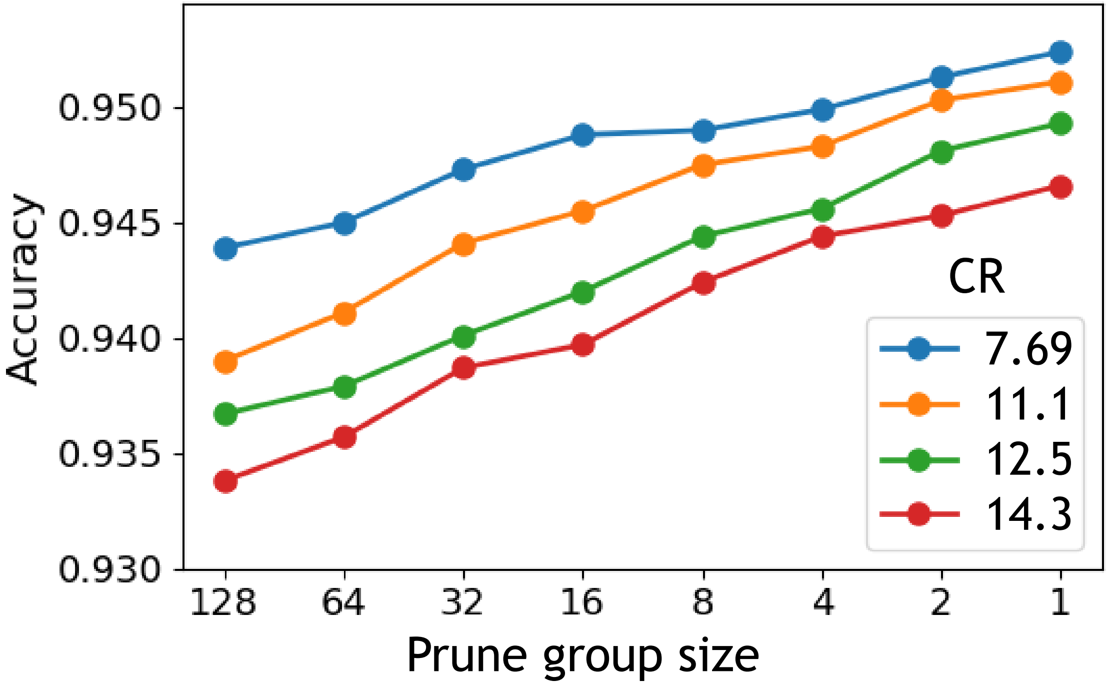
- Memory capacity is fixed on CIM chips after fabrication, but model sizes keep increasing

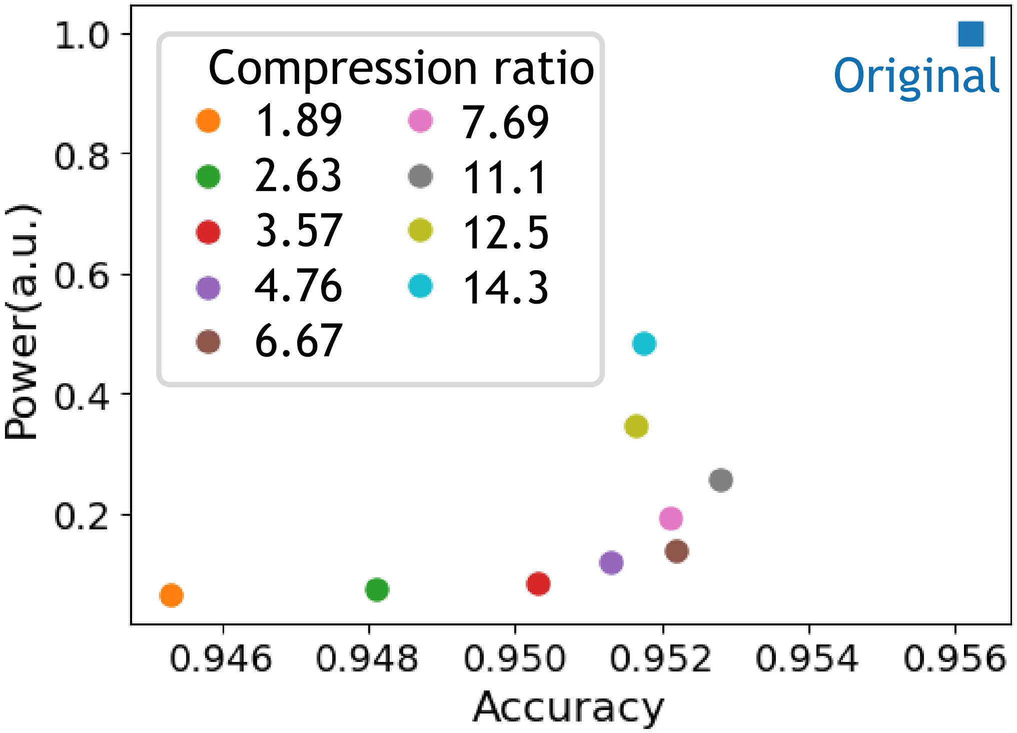
- Structured pruning can allow mapping of larger models – tradeoff of compression ratio, pruning granularity, and accuracy

F. Meng et a. unpublished

Fine-Grained Structured Pruning

Cifar 10

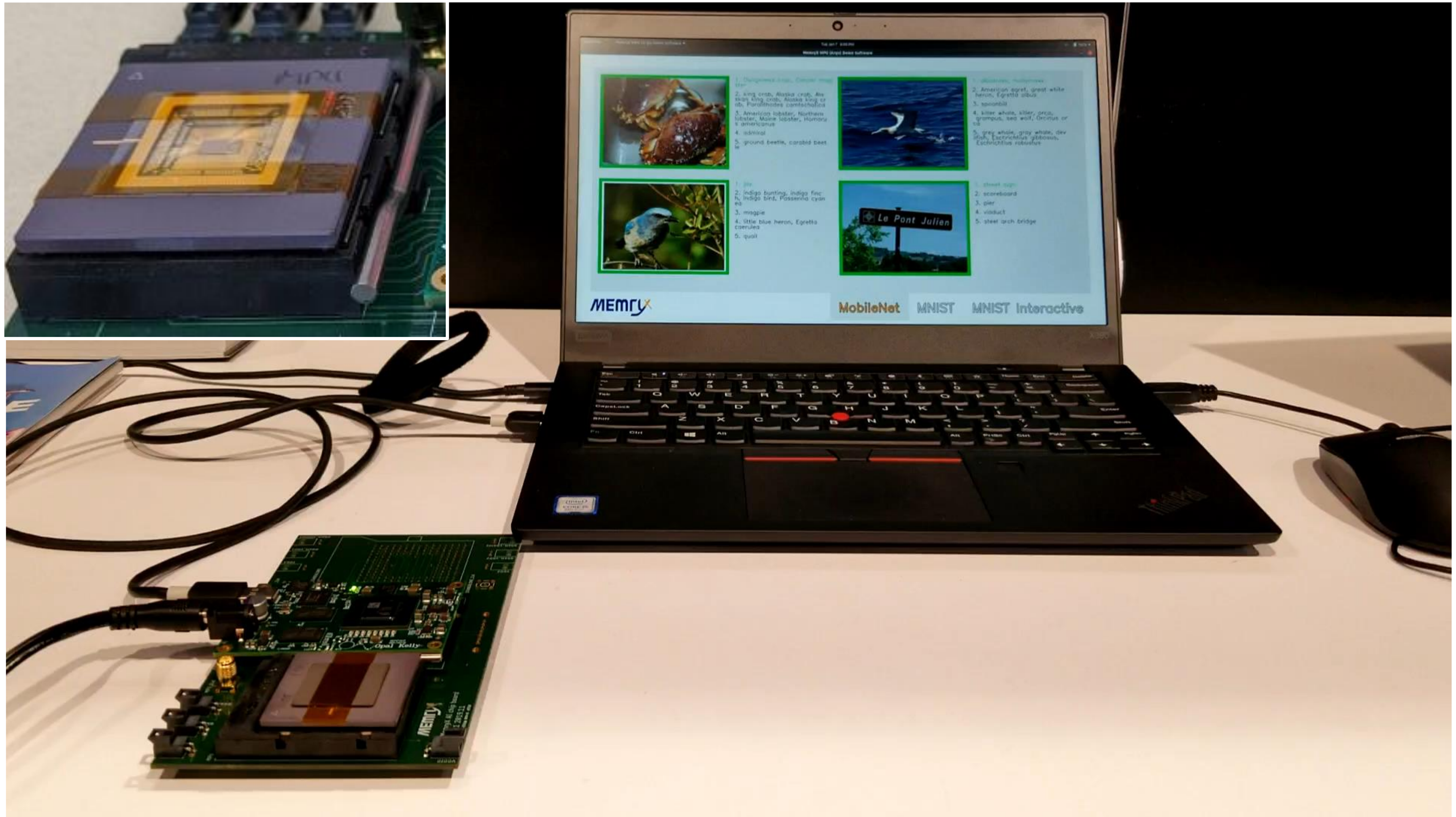




- The proposed fine-grained structured pruning improves accuracy and allows compression ratio up to 10x, enabling the mapping of larger models

F. Meng et a. unpublished

Larger Scale Implementations (> 10M devices)



Fully mapped MobileNet v2 on RRAM chip (no external DRAM)

Streaming images in, streaming classification out (batch = 1)

Nonvolatile – instant on, no data lost during power interrupts

Conclusions

- Significant progress has been made in RRAM-enabled AI accelerators
  - At the module level and at the system architecture level.
  - on the cusp of commercialization
- Challenges for large scale implementation can be mitigated through multiple approaches
  - Tiled-architecture implementation
  - Architecture-aware training
  - Fine-grained structure pruning

wluee@umich.edu

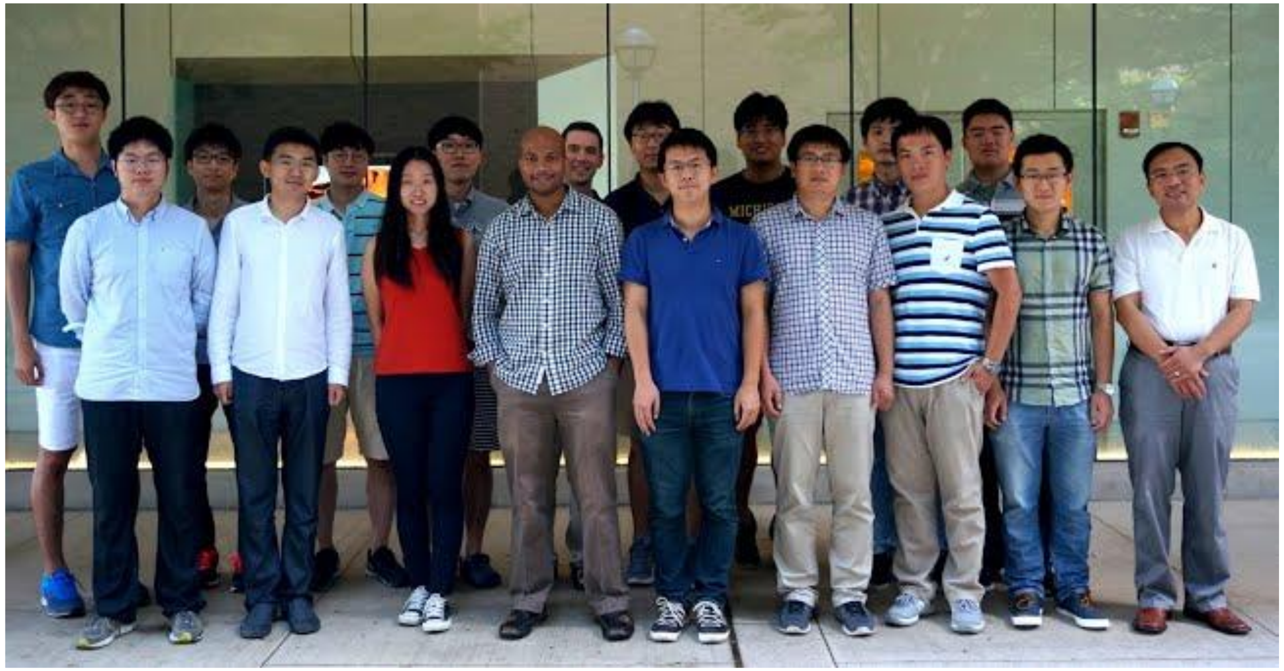
Acknowledgements

Grad students:

- **Sung-Hyun Jo*
- **Kuk-Hwan Kim*
- **Siddharth Gaba*
- **Ting Chang*
- **Patrick Sheridan*
- **ShinHyun Choi*
- **Jiantao Zhou*
- **Chao Du*
- *Jihang Lee*
- *Wen Ma,*
- *Fuxi Cai*
- *Yeonjoo Jeong*
- *Jong Hong Shin*
- *John Moon*
- *Billy Schell*
- *Qiwen Wang*
- **Eric Dattoli*
- **Wayne Fung*
- *Lin Chen*
- **Seok-Youl Choi*
- **Woo Hyung Lee*

PostDocs:

- *Dr. Mohammed Zidan*
- *Dr. Xiaojian Zhu*
- **Dr. Yuchao Yang*
- **Dr. Sungho Kim*
- **Dr. Bing Chen*
- **Dr. Taeho Moon*
- **Dr. Zhongqing Ji*
- ** Dr. Qing Wan*



- *Prof. Z. Zhang, Prof. M. Flynn, UM*
- *Dr. G. Kenyon, LANL*
- *Prof. C. Teuscher, PSU*
- *Prof. D. Strukov, UCSB*
- *Prof. J. Hasler, GeorgiaTech*
- *Dr. I. Valov, Prof. R. Waser*

Funding:

- *ADA Center through the SRC/DARPA JUMP Program*
- *DARPA UPSIDE program, DARPA ACCESS program*
- *National Science Foundation (ECS-0601478, CCF-0621823, ECCS-0804863, CNS-0949667, ECCS-0954621).*
- *DARPA SyNAPSE program*
- *Air Force HyNano MURI program, Air Force q-2DEG program*

References:

M. A. Zidan, J. P. Strachan, and W. D. Lu, Nature Electronics 1: 22–29 (2018)
Y. Yang and W. Lu, Nanoscale, 5, 10076 (2013)
Y. Yang, Gao, Chang, Gaba, Pan, and W. Lu, Nature Communications, 3, 732, (2012)
Y. Yang, Gao, Li, Pan, Tappertzhofen, Choi, Waser, Valov, W. Lu, Nature Communications 5, 4232 , (2014)
S. H. Jo, K.-H. Kim, W. Lu Nano Lett. 9, 496-500 (2009)
S. H. Jo, Kim, W. Lu, Nano Lett., 8, 392 (2008)
K.-H. Kim, S. H. Jo, W. Lu, Appl. Phys. Lett. 96, 053106 (2010)
K.-H. Kim, S. Gaba, D. Wheeler, J. Cruz-Albrecht, N. Srivinara, W. Lu Nano Lett., 12, 389–395 (2012)
S. H. Jo, T. Kumar, S. Narayanan, W. D. Lu, H. Nazarian, 6.7, IEDM 2014
S. Kim, S. Choi, W. Lu, ACS Nano, 8, 2369–2376 (2014)
P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, W. D. Lu, Nature Nanotechnology, 12, 784–789 (2017)
S. Choi, P. Sheridan, J. Shin, W. Lu, Nano Lett. 17, 3113–3118 (2017)
S. H. Jo, T. Chang, I. Ebong, B. Bhavitavya, P. Mazumder, W. Lu, Nano Lett. 10, 1297 (2010)
C. Du, W. Ma, T. Chang, P. Sheridan, W. D. Lu, Adv. Func. Mater., 25, 4290, (2015)
S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, W.D. Lu, Nano Lett, 15, 2203 (2015)
C. Du, F. Cai, M. Zidan, W. Ma, W. Lu, Nature Communications, 8: 2204, (2017)
B. Chen, F. Cai, W. Ma, P. Sheridan, W. Lu, 17.5, IEDM 2015
M. A. Zidan, Y.J. Jeong, J. Lee, B. Chen, S. Huang, M. J. Kushner, & W. D. Lu, Nature Electronics, 1, 411–420 (2018)
J. H. Shin, Y. J. Jeong, M. A. Zidan, Q. Wang W. D. Lu, 3.3, IEDM 2018.
J. Moon, W. Ma, J. H. Shin, F. Cai, C. Du, S. H. Lee, W. D. Lu, *Nature Electronics* https://doi.org/10.1038/s41928-019-0313-3
M. Zidan, Y. Jeong, J. H. Shin, C. Du, Z. Zhang, and W. D. Lu, IEEE Trans Multi-Scale Comp Sys, 4, 698-710 (2017)
X. Wang, Q. Wang, S. H. Lee, F. S. Meng, W. D. Lu, IEDM 2019