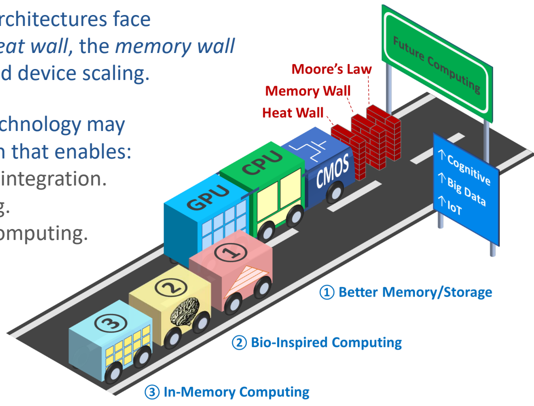
RRAM fabric for neuromorphic and in-memory computing applications

Mohammed Zidan and Wei Lu

University of Michigan
Electrical Engineering and Computer Science

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.

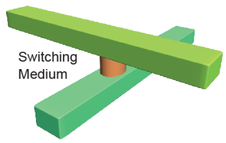


M. A. Zidan, J. P. Strachan, and W. D. Lu, Nature Electronics 1: 22–29 (2018)

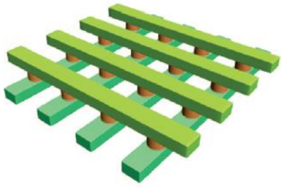
Lu Group
U. Michigan

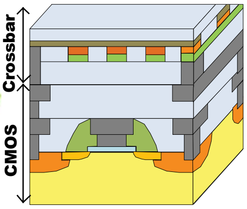
Two-Terminal Memory Devices and Crossbar Arrays

Single-cell structure



Crossbar Structure

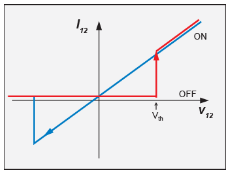




Resistive memory (RRAM), *memory + resistor* (memristor)

Hysteretic resistive switches and crossbar structures

- 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
- Memory, logic/neuromorphic applications

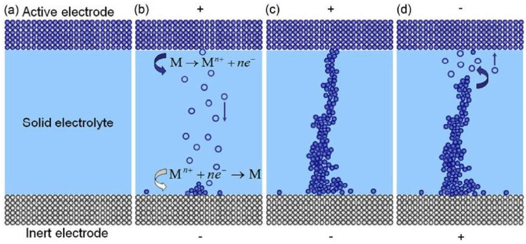


Lu group

Lu Group
U. Michigan

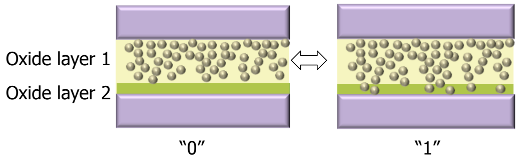
Physically reconfigurable materials and devices: Resistive Memory

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)

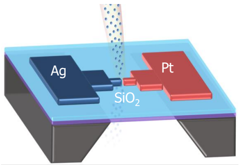


- ***Modulating exiting material properties***
- **Filament based on oxygen exchange between two oxide layers**
- **Electrode plays minor role**

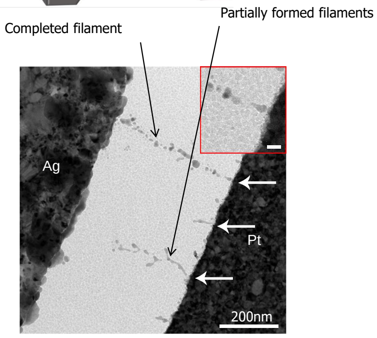
Yuchao Yang and Wei Lu, Nanoscale, 5, 10076 (2013)

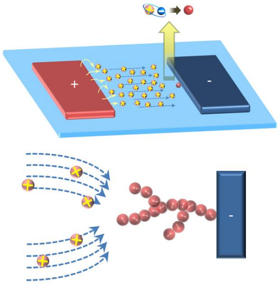
Lu Group
U. Michigan

Visualization of Filament



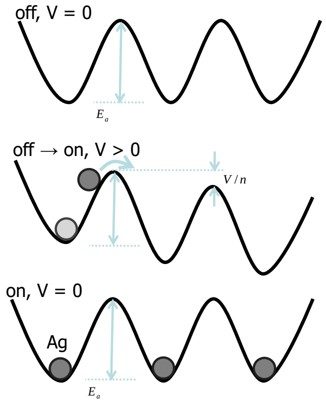
- •Ag/SiO_2/Pt structure, sputtered SiO_2 film
- •The filament grows from the IE backwards toward the AE
- •Branched structures were observed with wider branches pointing to the AE





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

Driving Ions with Electric Field

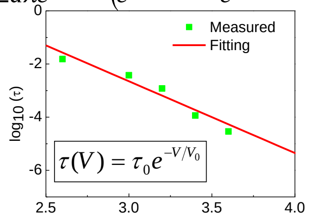


$$\Gamma = 1/\tau = \upsilon e^{-E_a^{'}(V)/k_BT}$$

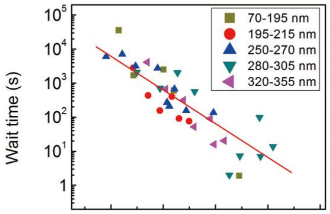
$$E_a^{'}(V) = E_a - V/2n$$

Jo, Kim, Lu *Nano Lett.* 9, 496-500 (2009).

$$v = 2d\lambda e^{-E_a/k_BT}\left(e^{qEd/2k_BT} - e^{-qEd/2k_BT}\right)$$

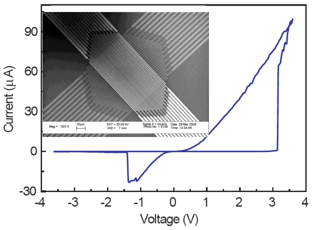


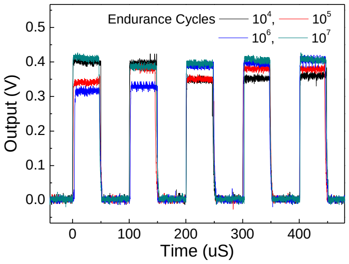
$$\tau(V) = \tau_0 e^{-V/V_0}$$

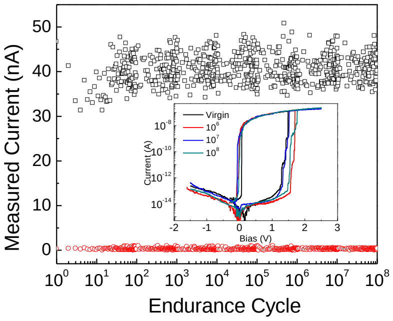


- •Filament formation is a thermally activated process.
- •Activation energy reduced by applied bias.
- •Speed is a ca. *exponential* function of voltage.

Resistance Switching Characteristics





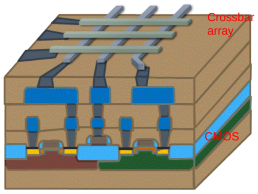


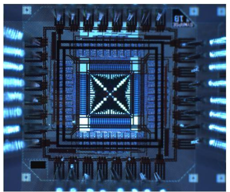
- ➢ 1e6 on/off
- ➢ 1e8 W/E endurance
- ➢ Switching speed ~10ns

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

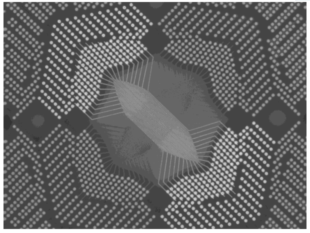
Kim, Jo, W. Lu, *Appl. Phys. Lett.* 96, 053106 (2010)

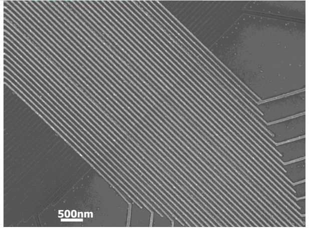
Integrated RRAM Crossbar/CMOS System





- •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 n^2 memory cells

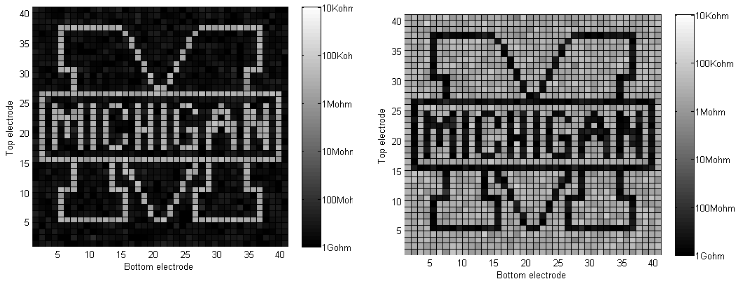




Kim, Gaba, Wheeler, Cruz-Albrecht, Srivinara, W. Lu *Nano Lett.*, 12, 389–395 (2012).

Integrated Crossbar Array/CMOS System

- **Crossbar array operation, array written followed by read**
- **Programming and reading through integrated CMOS address decoders**
- **Each bit written with a single pulse**

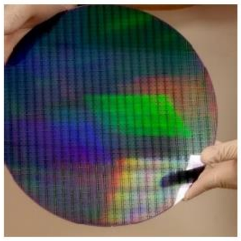


Results from a 40x40 crossbar array integrated on CMOS

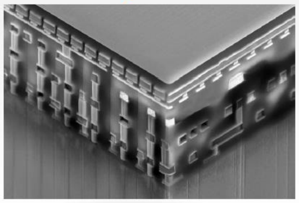
Kim, Gaba, Wheeler, Cruz-Albrecht, Srivinara, W. Lu *Nano Lett.*, 12, 389–395 (2012).

From Lab to Fab - Crossbar RRAM Technology

- **CMOS** Compatible
- **3D** Stackable, Scalable Architecture – Low thermal budget process
- **Architectures** proven include multiple Via schemes and Subtractive etching
- **Crossbar Inc** founded in 2010, $100M VC funding to date
- **Commercial Products** offered in 2016 based on 40nm CMOS

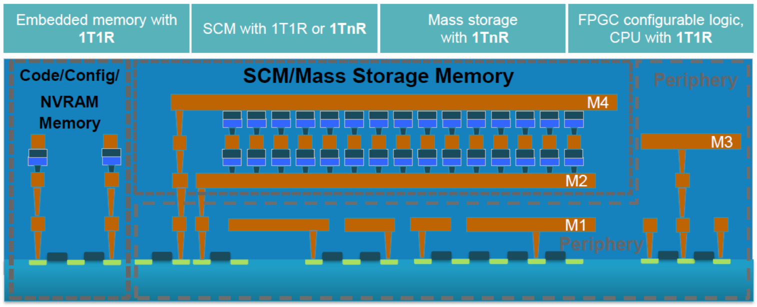


Simple CMOS Integration



Hybrid Integration of Memory with Logic

1T1R and 1TnR 3D stackable memory arrays



- Monolithic logic/memory integration
- Different memory components integrated on the same chip
- Flexibility of speed/density/cost

Improving Computing Efficiency using RRAM Arrays

Different approaches for improving computing efficiency (depending on the application):

- **Bring memory as close to logic as possible, still largely based on conventional architecture**
- **Neuromorphic computing in artificial neural networks**
- **More bio-inspired, taking advantage of the internal ionic dynamics at different time scales**
- **Other compute applications based on vector-matrix multiplications**

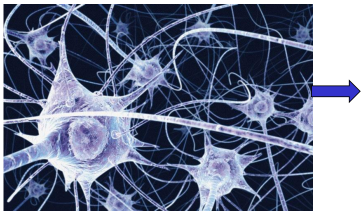
Towards a general in-memory computing fabric based on a common physical substrate

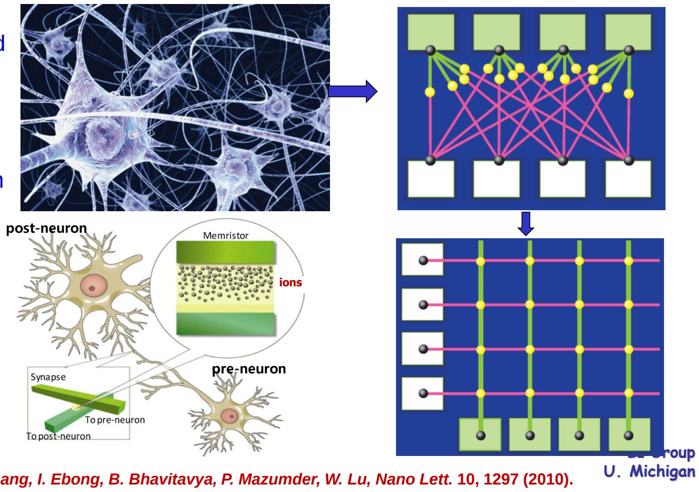
RRAM Based Neural Network Hardware

Synapse – reconfigurable two-terminal resistive switches

Co-located memory-compute

High parallelism





S. H. Jo, T. Chang, I. Ebong, B. Bhavitavya, P. Mazumder, W. Lu, Nano Lett. 10, 1297 (2010).

Neuromorphic Computing with RRAM Arrays

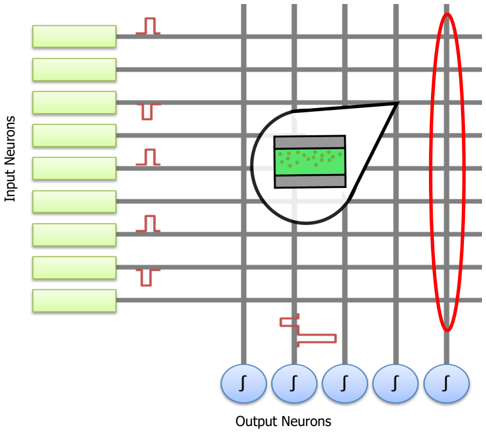
RRAM perform learning and inference functions

DARPA UPSIDE program

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

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

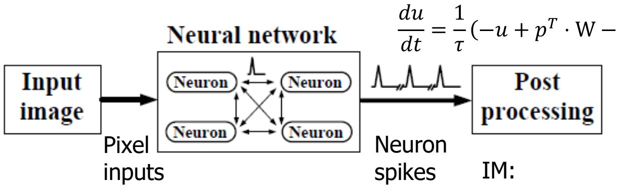
- Integrate and fire neurons
- Learning achieved by backpropagating spikes



Lu Group
U. Michigan

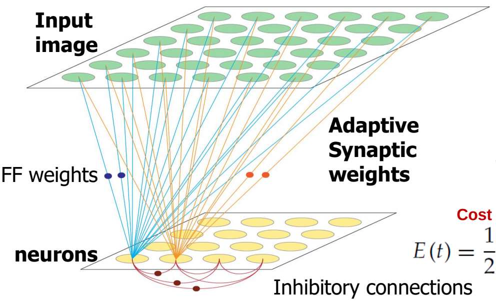
Neural Network for Image Processing based on Sparse Coding

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



IM:

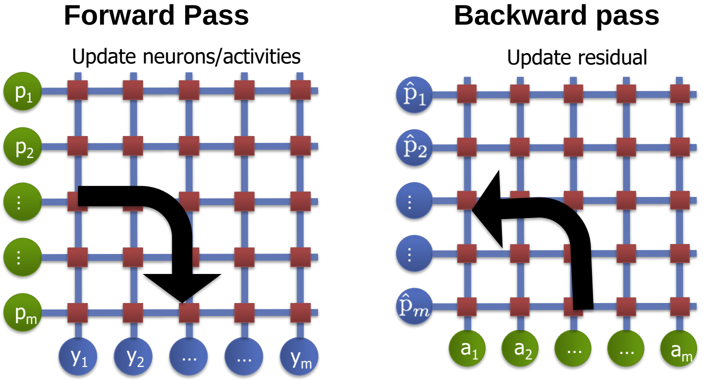
1. Network adapt during training following local plasticity rules
2. FF weights form neuron receptive fields (dictionary elements)
3. Output as neuron firing rates



Cost Function:

$$E(t) = \frac{1}{2}\left|\left|s(t) - \hat{s}(t)\right|\right|^2 + \lambda \sum_m C(a_m(t)) \,.$$

Sparse Coding Implementation in RRAM Array



$$y = p^\top W$$

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

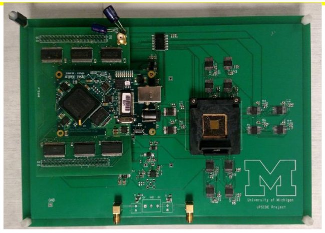
Neuron membrane potential

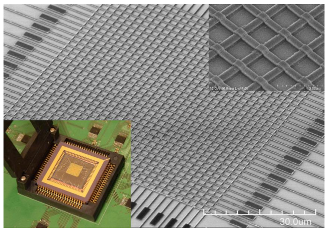
$$\frac{du}{dt} = \frac{1}{\tau}\left(-u + p^T \cdot \mathrm{W} - a \cdot (\mathrm{W}^\top \mathrm{W} - I)\right)$$

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

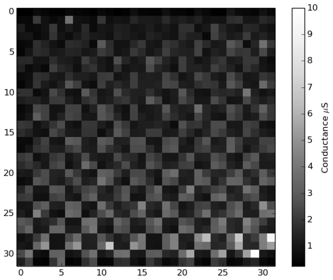
Sheridan et al., *Nature Nanotechnology*, 12, 784–789 (2017)

Hardware Implementation





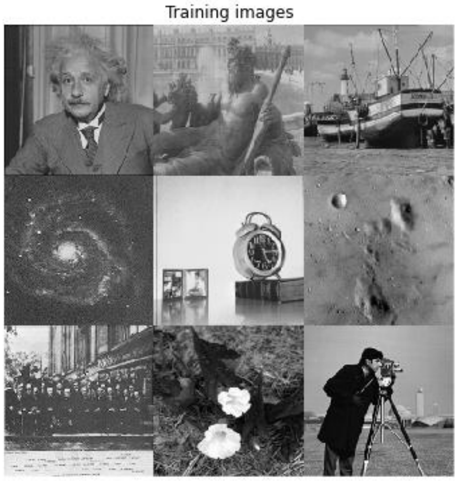
32x32 memristor array



- Checkerboard pattern
- 32 x 32 array
- Direct storage and read out
- No read-verify or re-programming

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

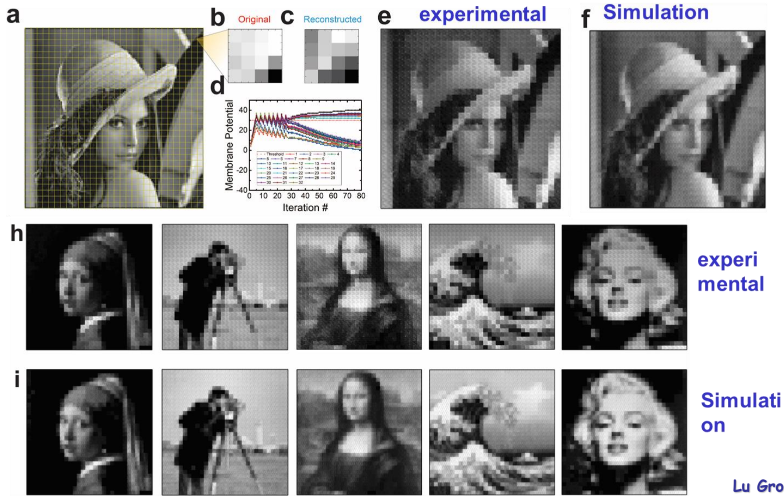
Training



9 Training Images
128x128px
4x4 patches
Trained in random order

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

Image Reconstruction with RRAM Crossbar



experi mental

Simulati on

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

Improving Computing Efficiency using RRAM Arrays

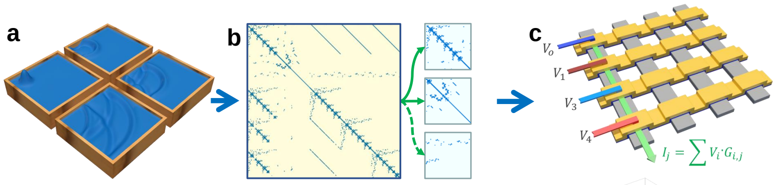
Different approaches for improving computing efficiency (depending on the application):

- Bring memory as close to logic as possible, still largely based on conventional architecture
- Neuromorphic computing in artificial neural networks
- More bio-inspired, taking advantage of the internal ionic dynamics at different time scales
- **Other compute applications based on vector-matrix multiplications**

Towards a general in-memory computing fabric based on a common physical substrate

Arithmetic Applications: Numerical Simulation

Solving partial-differential equations (PDEs)

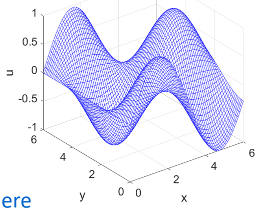


Solving an A·x=b problem in matrix form

- A second order Poisson equation as a toy example,

$$\nabla^2 u = -2 \cdot \sin(x) \cdot \cos(y)$$

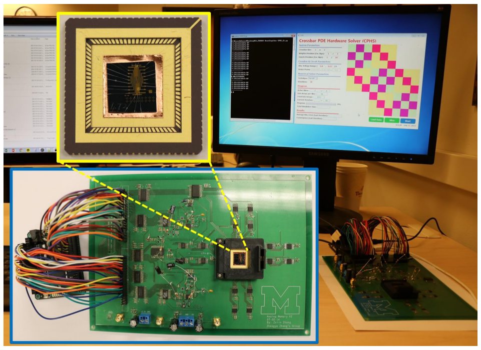
- The problem is solved using finite difference (FD), where matrix can be sliced into a set of few similar slices.



Hardware Prototyping

» Hardware Test bench

- The test board consists of: **(i)** RRAM crossbar, **(ii)** DACs to control the input signals, **(iii)** sense amplifiers and ADCs to sample the output current, **(iv)** MUXs to route the signals, and **(v)** FPGA to enables the software interface and control.

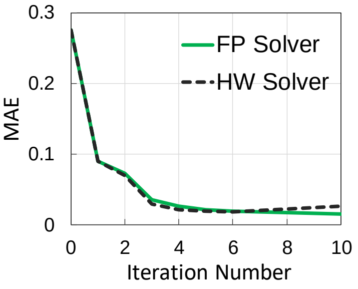


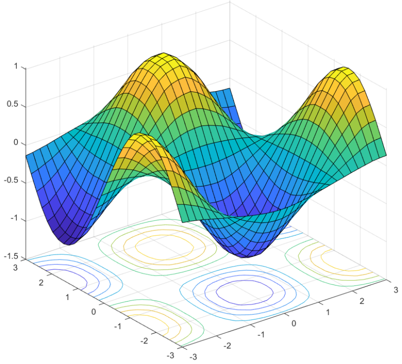
M. A. Zidan, Y.J. Jeong, J. Lee, B. Chen, S. Huang, M. J. Kushner, & W. D. Lu, Nature Electronics, **1, 411–420 (2018)**

Hardware Prototyping

» Solving a toy example

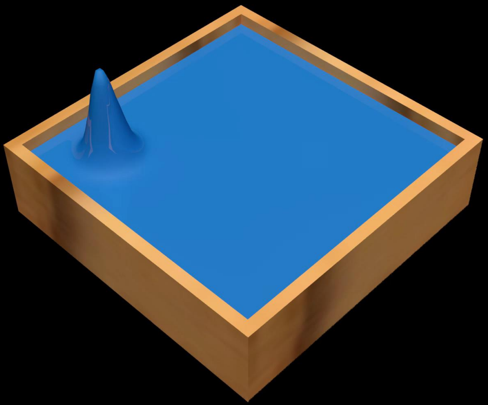
Measured Results for the toy Example





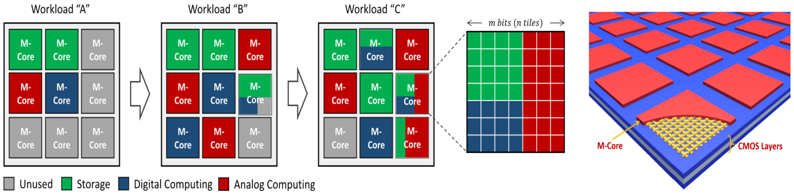
M. A. Zidan, Y.J. Jeong, J. Lee, B. Chen, S. Huang, M. J. Kushner, & W. D. Lu, Nature Electronics, **1, 411–420 (2018)**

» Results Reconstructed as a 3D Animation



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

General In-Memory Computing Fabric



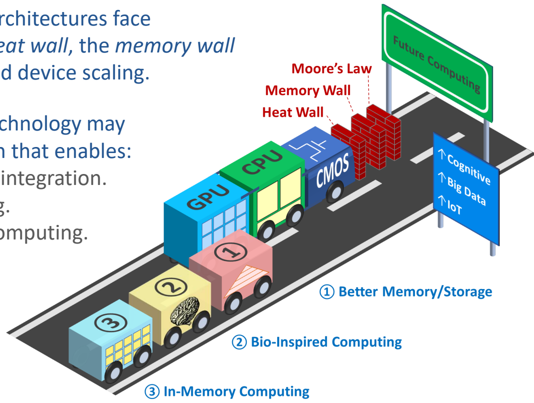
- **Memory-Computing Unit (MPU)**
- "General" purpose by design: the same hardware supports different tasks – low precision or high precision. Not just an neuromorphic accelerator
- Dense local connection, sparse global connection
- Run-time, dynamically reconfigurable. Function defined by software.

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)

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

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.



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

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)

Summary

Different approaches for improving computing efficiency (depending on the application):

- **Bring memory as close to logic as possible, still largely based on conventional architecture**
- **Neuromorphic computing in artificial neural networks**
- **More bio-inspired, taking advantage of the internal ionic dynamics at different time scales**
- **Other tasks based on vector-matrix multiplications**

⇨ **Towards a general in-memory computing fabric based on a common physical substrate**