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.

Two-Terminal Memory Devices and Crossbar Arrays

- 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 4F²
    - Terabit potential
  - Large connectivity
  - Memory, logic/neuromorphic applications

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

Ag/SiO₂/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

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


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

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/cm²
• CMOS units – larger but fewer units needed. 2^n CMOS cells control n² memory cells


Resistance Switching Characteristics
• 1e6 on/off
• 1e8 W/E endurance
• Switching speed ~10ns


Driving Ions with Electric Field
• Filament formation is a thermally activated process.
• Activation energy reduced by applied bias.
• Speed is a ca. exponential function of voltage.
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


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

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

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
**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
  - Memristor array natively performs matrix operation
    \[ \mathbf{t} = \mathbf{v} \cdot \mathbf{W} \]
  - Integrate & fire neurons
  - Learning achieved by backpropagating spikes

**DARPA UPSIDE program**

**Neural Network for Image Processing based on Sparse Coding**

- Input image
- **Neural network**
  - Forward Pass
  - Update neurons/activities
  - Backward pass
  - Update residual
  - Neuron membrane potential

**Sparse Coding Implementation in RRAM Array**

- Cost Function:
  \[ E(t) = \frac{1}{2} \| s(t) - \hat{s}(t) \|^2 + \lambda \sum_n C(a_n(t)) \]
  \[ \frac{du}{dt} = \frac{1}{\tau} \left( -u + p^T \cdot W - a \cdot (W^T W - t) \right) \]
  \[ \frac{dv}{dt} = \frac{1}{\tau} \left( -u + (p - \hat{p})^T W + a \right) \]


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

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


Image Reconstruction with RRAM Crossbar

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

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

Solving partial-differential equations (PDEs)

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

- A second order Poisson equation as a toy example,
- The problem is solved using finite difference (FD), where matrix can be sliced into a set of few similar slices.

Solving an $A \cdot x = b$ problem in matrix form

- 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 enable the software interface and control.

Measured Results for the toy Example

Results Reconstructed as a 3D Animation
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.

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.

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