Energy Efficient Implementation of MVM Operations Using Filament-free Bulk RRAM Arrays

Ashwani Kumar*, J. Park, Y. Zhou, J. Kim, S. Jain, C. D. Schuman, G. Cauwenberghs, D. Kuzum*

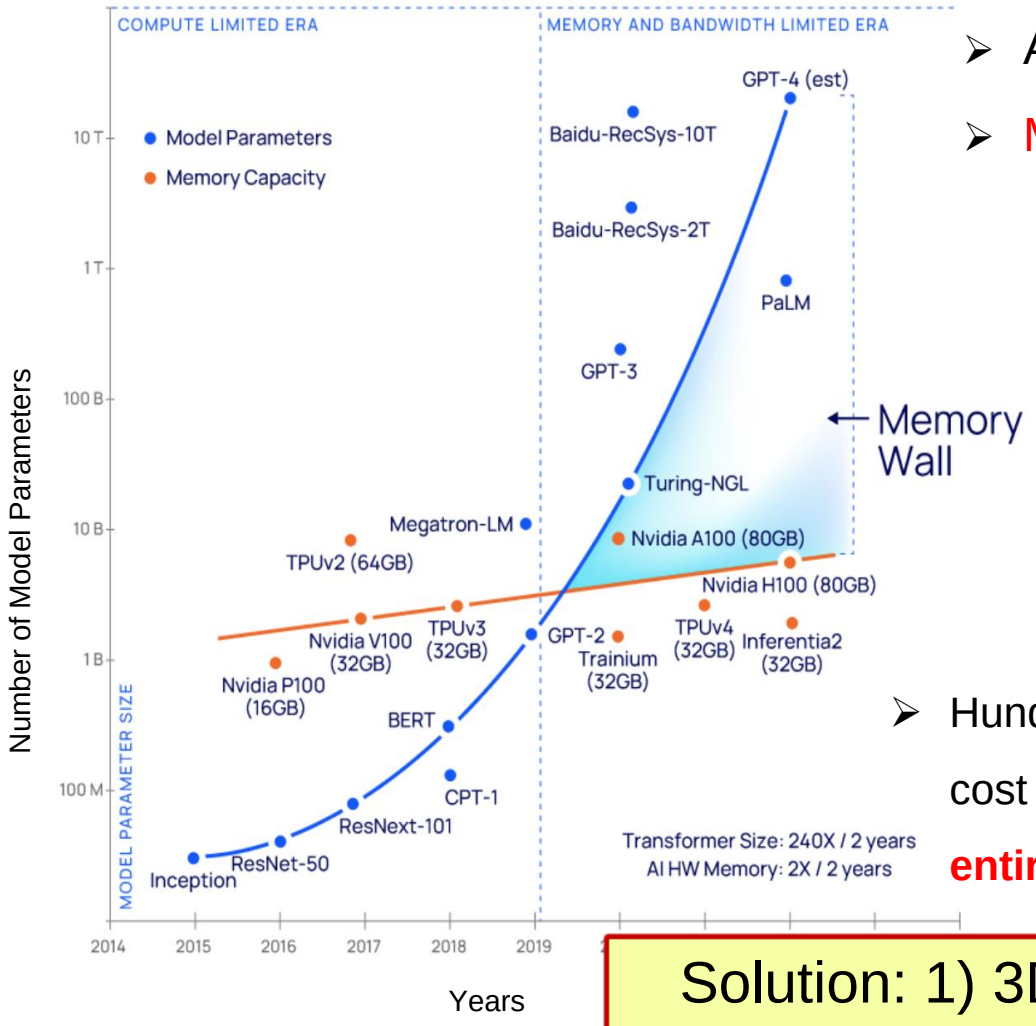
*(ask010@ucsd.edu, dkuzum@ucsd.edu)

Neuroelectronics Lab
Department of Electrical and Computer Engineering
University of California San Diego, CA, USA

Outline

- Filament-Free Bulk RRAM Fabrication and Characterization
- Weight Mapping on RRAM Crossbar
- SNN Implementation with our Bulk RRAM Crossbars
- Conclusion

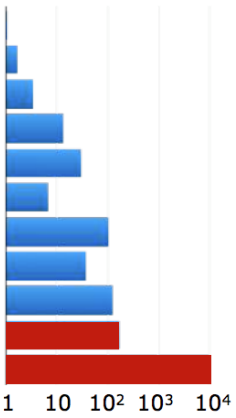
Memory Access



- AI operations require high parallelism
- Memory access is the bottleneck

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

Relative Energy Cost



[**Horowitz**, *ISSCC* 2014]

- Hundreds of millions of queries on ChatGPT cost **~1 GWh** daily, **enough to power an entire city for one day** (~33,000 U.S. homes)!

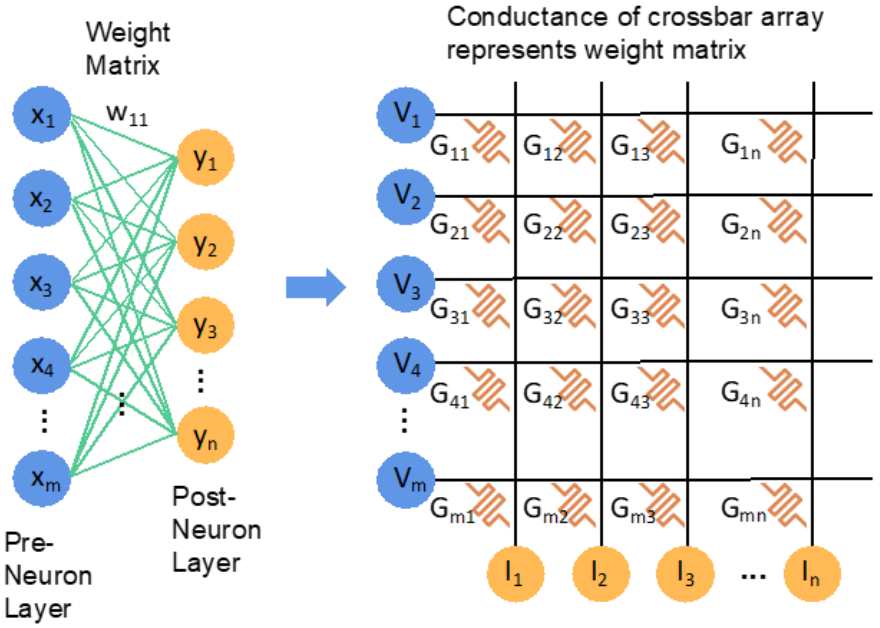
Solution: 1) 3D stacking of memory, 2)Reduce the data movement by performing **Compute in Memory**

Analog Compute-in-Memory (CIM)

- Challenging requirements by today's AI Models:
  - Massive training and inference exercises require large amount of energy
  - MAC operations (as MVM - matrix vector multiplication) contribute 70- 90% of the total operational cost of neural network implementation.
- Need energy efficient MVM operations:

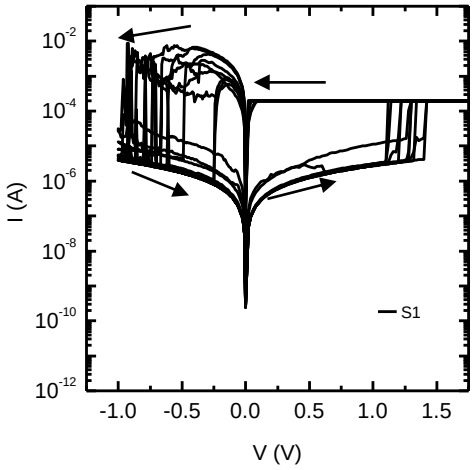
TASK	OPERATION
$Y = W \cdot X$	$I_i = \sum_j G_{ij} \cdot V_j$

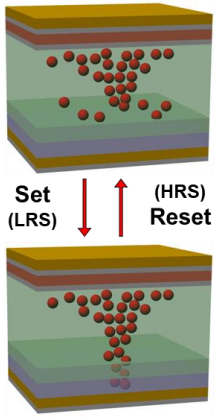
- CMOS compatible RRAM crossbar array for MVM using CIM.



Optimizing RRAM Technology for CIM

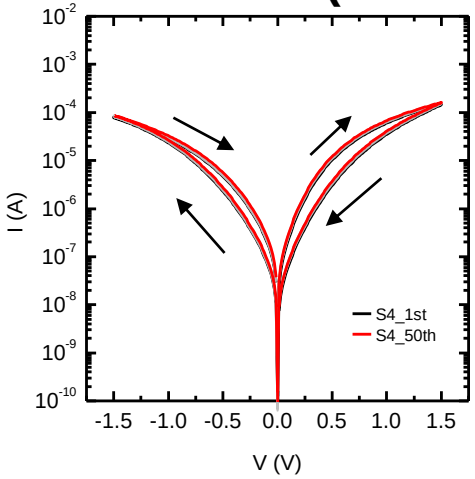
❑ Filamentary RRAM

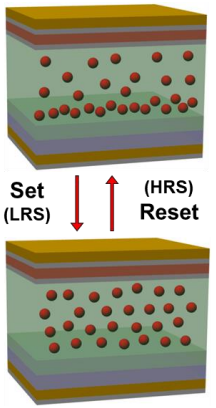




1. High voltage forming → not compatible with advanced CMOS, requires additional peripheral circuitry to support this operation.
2. Abrupt resistive switching → Variations and noise, accuracy loss, many iterations of read/verify cycles.
3. Low ON state resistance (~kΩ) → increases power consumption, limits the arrays size and parallel MAC operations.
4. Limited number of states or binary operations → not suitable for on-chip learning.

Solution: Bulk (Area-type) RRAM

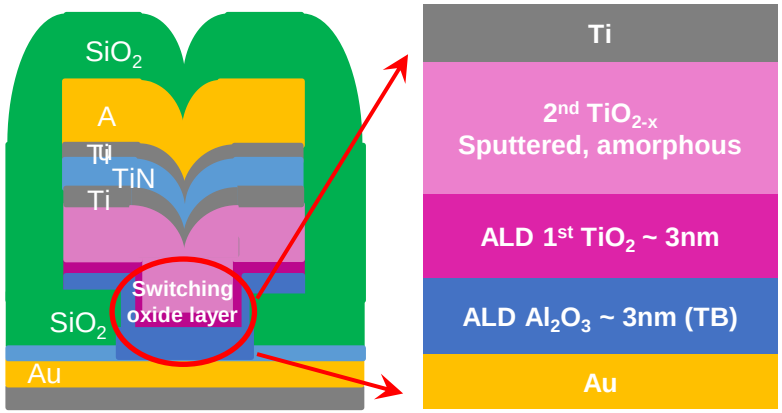


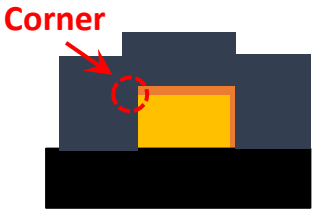


1. Forming-free operation, no filaments
2. Area-type switching, uniform switching with no compliance current
3. MΩ level resistance enables large size arrays and parallel read, reduced array level energy consumption
4. Multi-level gradual switching for on-chip learning

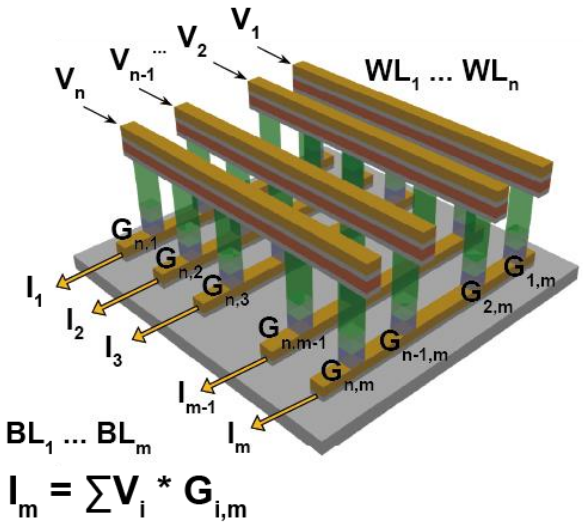
Fabrication of Trilayer Bulk RRAM

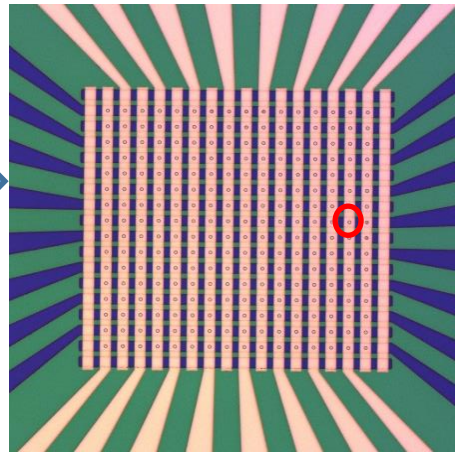
- Trilayer bulk RRAM stack:
  - Al_2O_3(3nm) / TiO_2(3nm) / TiO_x (40nm)
  - Tunnel barrier from Al_2O_3, high oxygen vacancy concentration in TiO_x , separated by ALD deposited TiO_2
  - Crossbar with via-hole structured RRAM

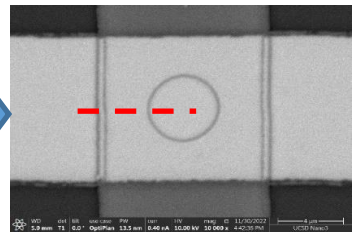


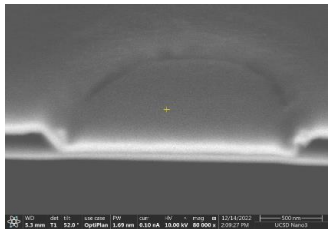


*Eliminates the edge effects due to high-field corners or sidewalls



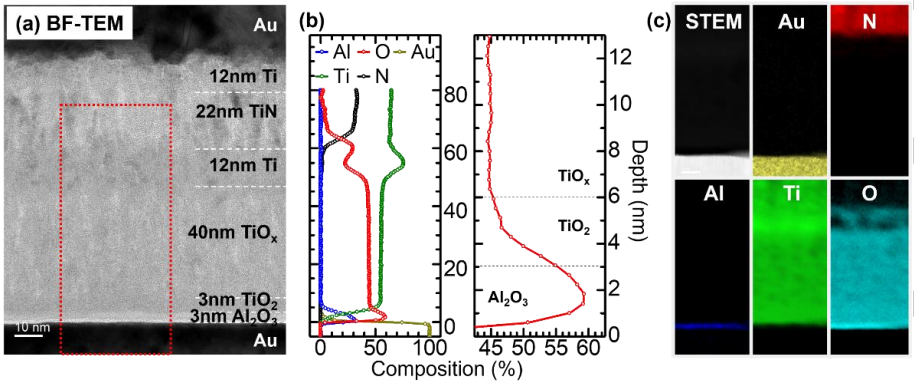






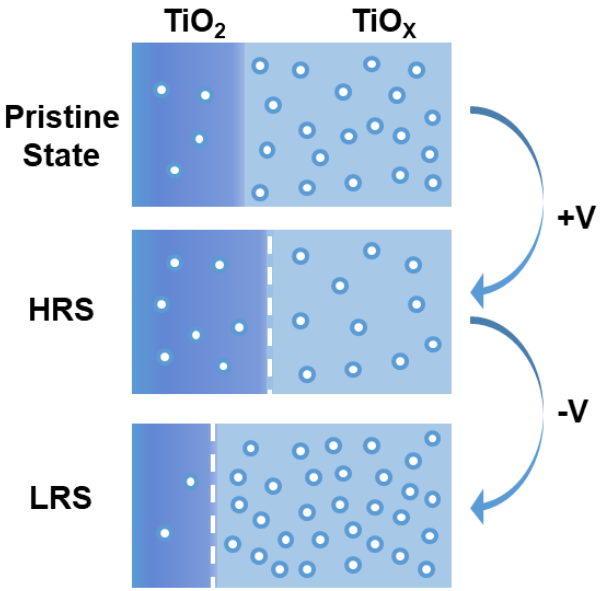
Park, J., Kumar, A., Zhou, Y. *et al. Nat Commun* **15**, 3492 (2024).

RRAM Structure and Resistive Switching

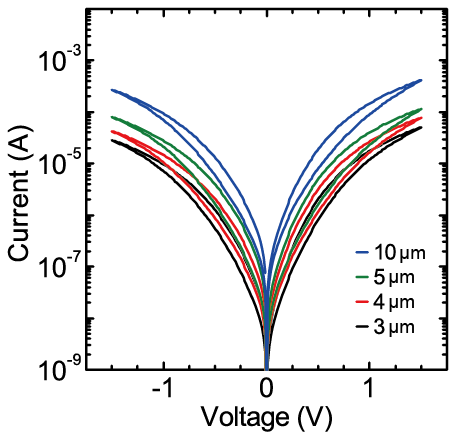


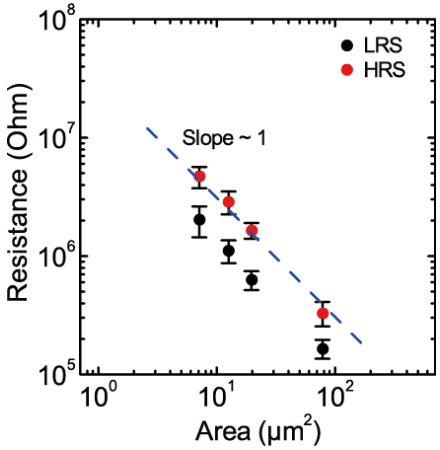
- Darker contrast of ALD TiO_2 confirms higher atomic density than sputtered 40nm TiO_x
- STEM-EELS line-scan profile also shows lower oxygen concentration in 40nm TiO_x
- STEM-EELS composition map (red-dotted (a)) shows nm-scale dark areas pointing to a porous structure.

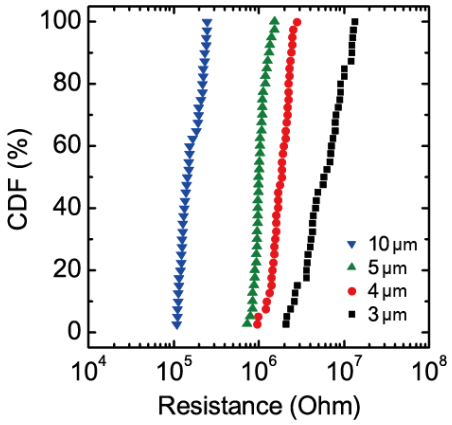
- Bulk RRAM's resistive switching mechanism:
  - Distribution of oxygen vacancies (V_O) is modulated between TiO_x and TiO_2 layers by applying a field across the device.
  - Migration of oxygen vacancies near the TiO_2/TiO_x interface either extend or reduce the effective thickness of oxygen vacancy rich TiO_x layer to switch the device in LRS or HRS.

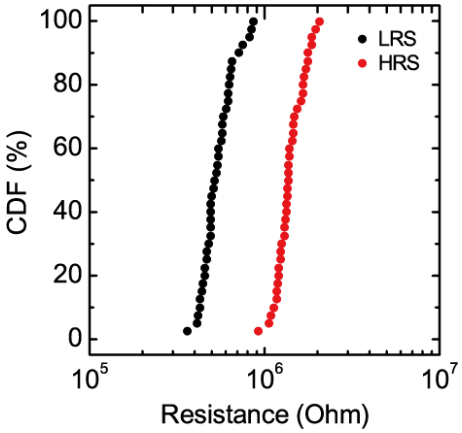


Bulk RRAM DC Switching Characterization

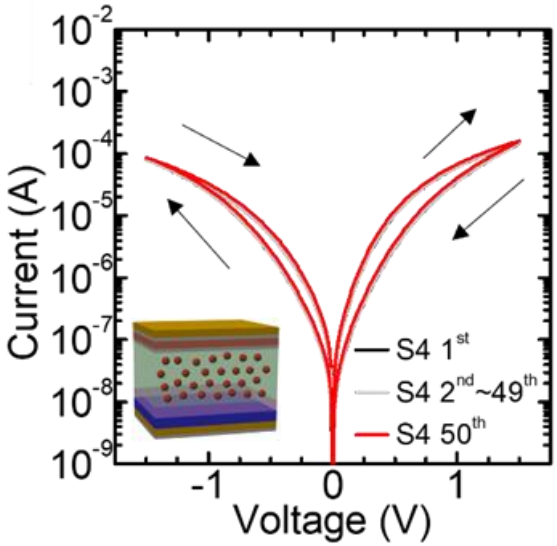






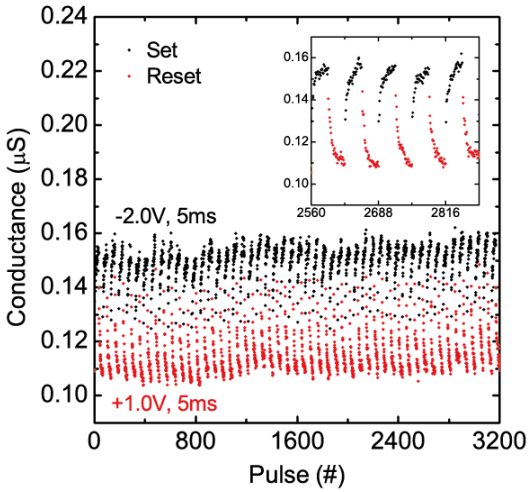


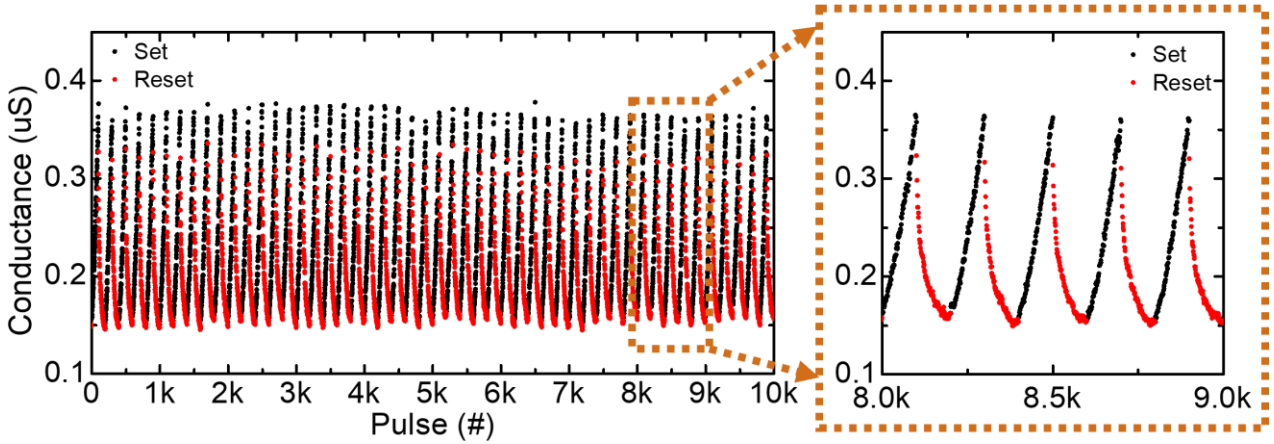
- Bulk (area-type) Switching: Resistance scales with area for both HRS and LRS.
- MΩ level bulk switching
- Low device-to-device and cycle-to-cycle variations



Park, J., Kumar, A., Zhou, Y. *et al. Nat Commun* **15**, 3492 (2024).

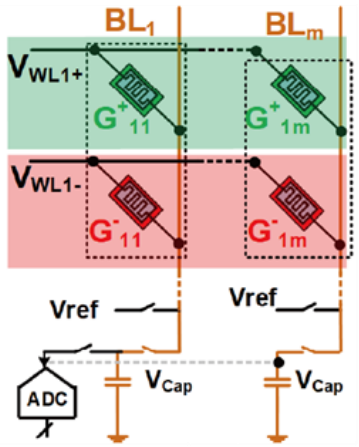
Bulk RRAM Pulse Switching Characterization





- Achieved Multilevel States at MΩ Range:
  1. Identical pulse programming scheme → same pulse amplitude
  2. Incremental pulse programming scheme → pulse amplitude increases with 20mV step.

Row Differential Weight Scheme in Crossbar

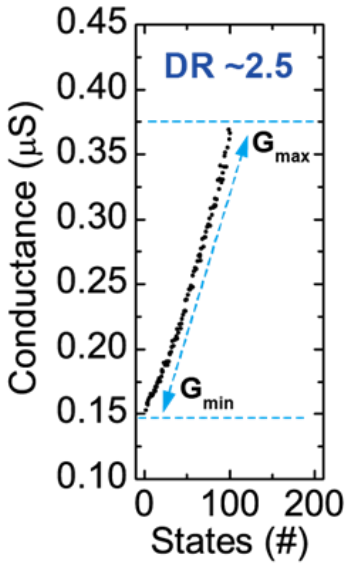


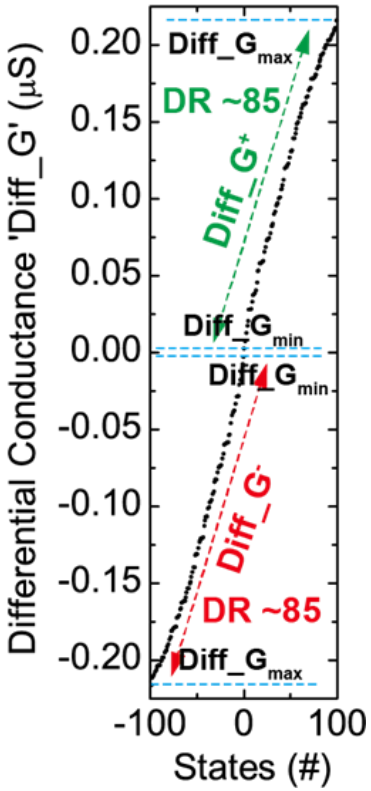
$DR = G_{max}/G_{min}$

$Diff\_G = G^{+} - G^{-}$;

Effective DR

$2(G_{max}-G_{min})/|Diff\_G_{min}|$

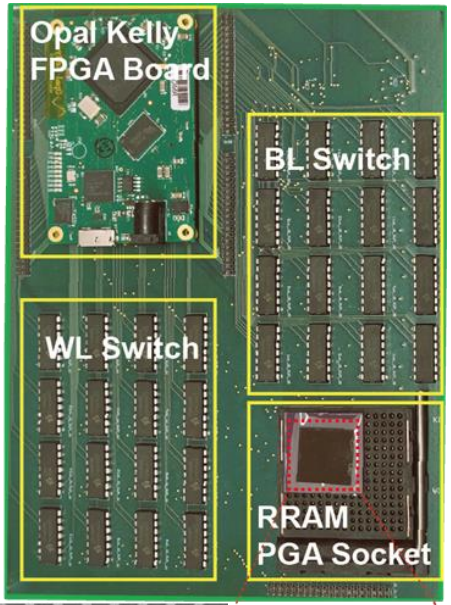


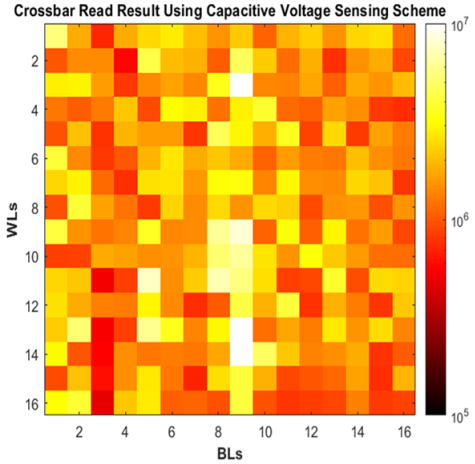


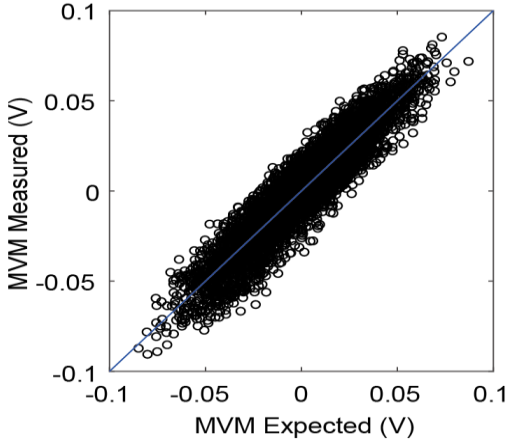
- Row Differential Crossbar Schematic

- Signed weight implementation with voltage-sensing scheme.
- Increased effective switching dynamic range (~170) while observing the many (100) conductance levels.
- Enables mapping of a wide range of real-valued weights

MVM Operations with Bulk RRAM Crossbars

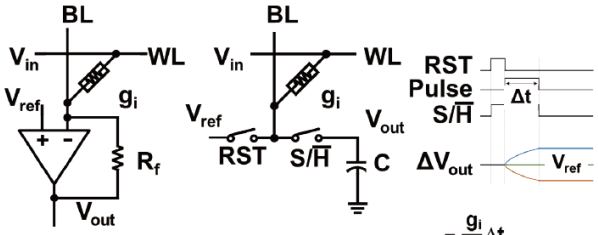






- A neuromorphic CIM platform utilizing a switched capacitor voltage sensing
  - Packaged crossbar array tested using neuromorphic-board developed with on board energy efficient voltage sensing.
  - A representative resistance map of 16x16 bulk RRAM crossbar read by using voltage sensing scheme.
  - Measured MVM and expected MVM result show good linearity (low error) for differential mapping scheme.

Current Sensing



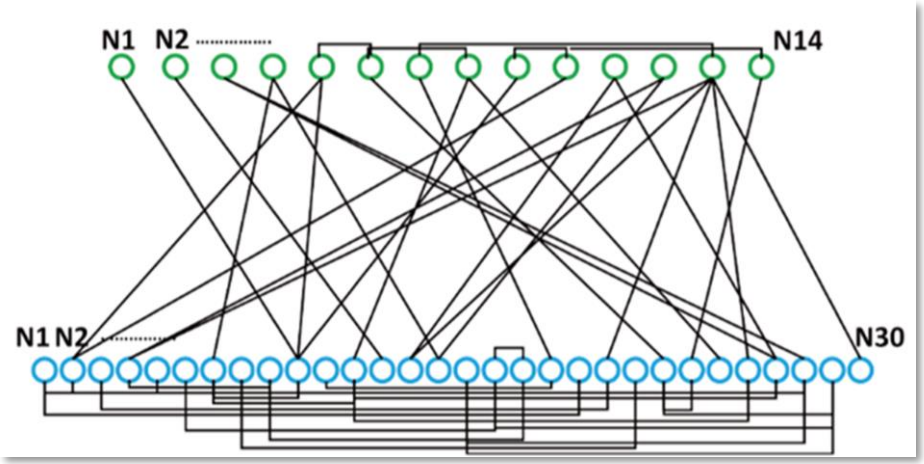
$\Delta V_{out} = -R_f g_i \Delta V_{in}$

Voltage Sensing

$\Delta V_{out} = (1 - e^{-\frac{g_i}{C}\Delta t})\Delta V_{in}$

In collaboration with Prof. G. Cauwenberghs @UCSD
Jain et al. IEEE ISCAS, 2023.

SNN Implementation: F-1 racetrack navigation

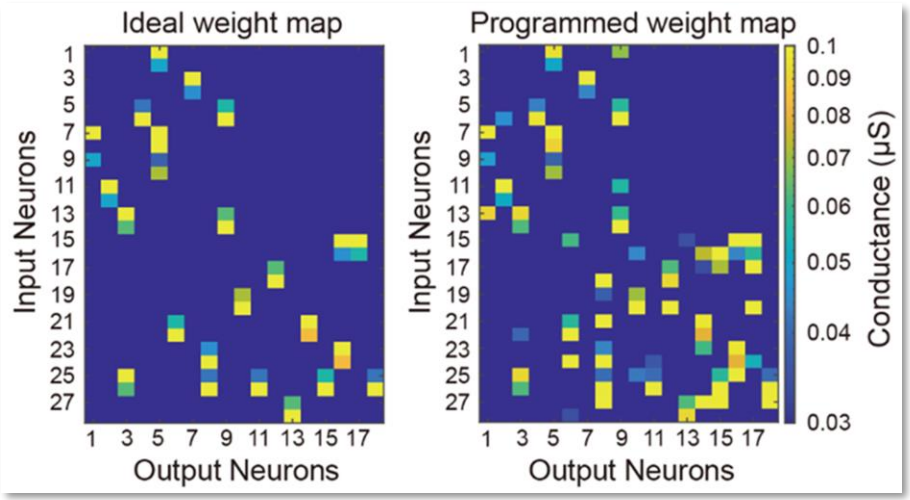


- SNN Model Using Evolutionary Optimization for Neuromorphic Systems (EONS):
  - SNN is optimized and trained for small-scale autonomous racing task (representative tracks).
  - Trained on 5 F-1 tracks and tested on an additional 15 tracks.
  - Pruned SNN consists of 14 input neurons and 30 output neurons.

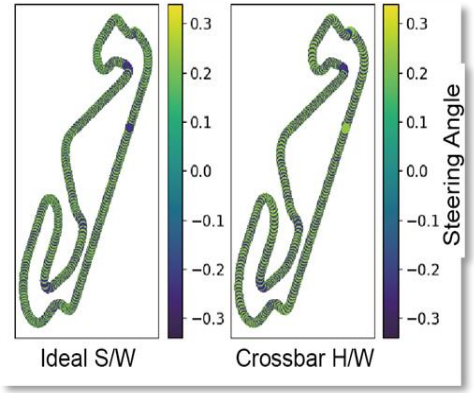
In collaboration with Prof. C. Schuman @UTK

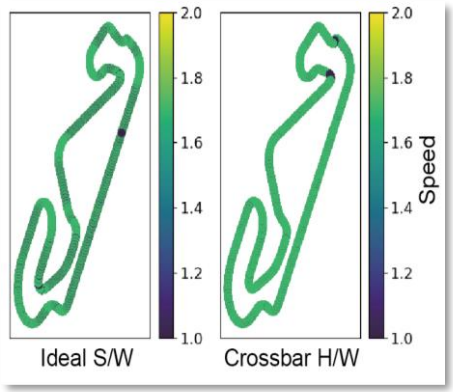
J. S. Plank, et al., IEEE Letters of the Computer Society, 2018.
C. D. Schuman, et al., NICE, Workshop, 2020.

SNN Weight Implementation on Bulk RRAM Crossbars



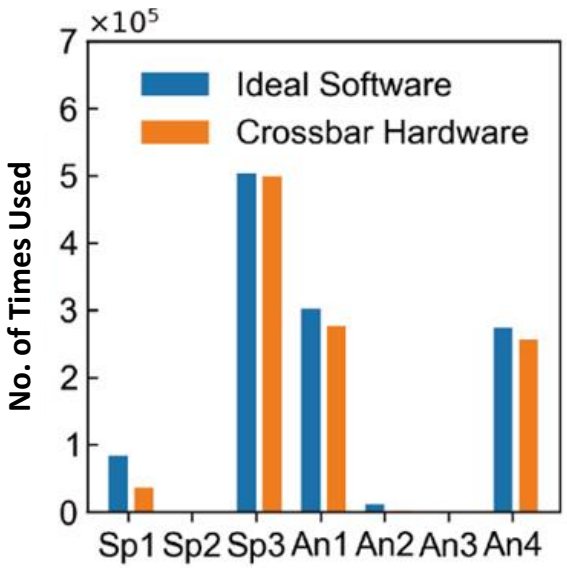
- Two 16x16 crossbars for all encoded weights
  - SNN's signed 4-bit weights were encoded into differential conductance (G+ and G−) using row differential scheme and programmed in crossbars.

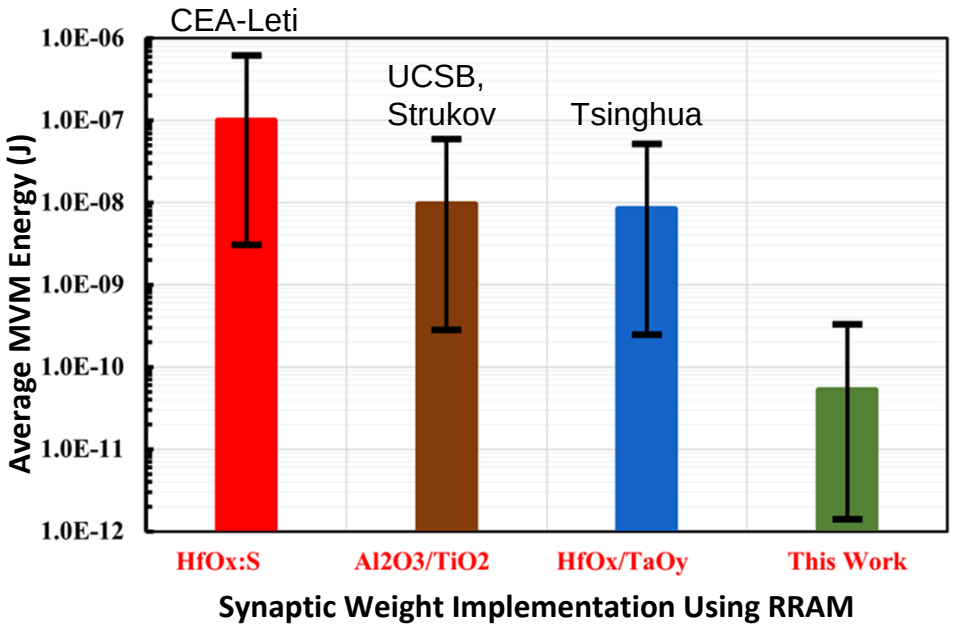




- Ideal (software) verses programmed (hardware) weight map in RRAM crossbars.
- Network outputs: steering angle and speed

SNN Hardware Implementation: Results





- Network Performance and Energy Comparison:
  - Speed and steering angle computations across navigation through all 15-racetracks show highly consistent results between software and hardware implementations.
  - Average energy consumed for MVM operations across all 15 tracks shows that our trilayer bulk RRAM substantially (more than two orders of magnitude) reduces energy consumption compared to other filamentary RRAM technologies.

[CEA-Leti] L. Grenouillet et al., IEEE International Memory Workshop (IMW), 2021.
[UCSB, Strukov] H. Kim, et al., Nature communications, 2021.
[Tsinghua] W. Wan et al., Nature, 2022.

Conclusion

- Developed a novel trilayer filament-free bulk RRAM crossbar technology
- Proposed row-differential weight mapping to achieve higher dynamic range for mapping of a wide range of real-valued weights in bulk RRAM crossbars.
- Performed highly linearized MVM operation in an energy efficient way using in-house design neuromorphic CIM hardware platform.
- Presented SNN implementation using our bulk RRAM crossbars for autonomous navigation tasks for scaled F1-tracks and showed great agreement with ideal software and hardware results.
- Our bulk RRAM crossbars for at edge neuromorphic computing application substantially reduced energy consumption compared to other filamentary RRAM technologies.
- Presented bulk RRAM crossbar technology with capability of multilevel switching in MΩ regime and CMOS-BEOL compatibility addresses several challenges and offer great potential for energy and area efficient computing.

Acknowledgements

Collaborators:
-Prof. Gert Cauwenberghs (UC San Diego)
-Prof. Catherine Schuman (UT Knoxville)
-Prof. Ivan Schuller (UC San Diego)

Students:

Yuhan Shi

Sangheon Oh

Madison Wilson

Mehrdad Ramezani

Yucheng Zhou

Yuyi Zhang

Shaan Shah

Fengyi Sun

Postdocs:

Ashwani Kumar

Yue Zhou

Visitors:

Seonghyun Kim (SK Hynix)

Neuroelectronics Lab

Thank You!

References

1. J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean, and G. S. Rose, ”The TENNLab exploratory neuromorphic computing framework,” IEEE Letters of the Computer Society, 1, 2, 17-20, 2018.
2. C. D. Schuman, J. P. Mitchell, R. M. Patton, T. E. Potok, and J. S. Plank, ”Evolutionary optimization for neuromorphic systems,” Annual NeuroInspired Computational Elements Workshop, pp. 1-9. 2020.
3. https://github.com/f1tenth/f1tenth racetracks.
4. L. Grenouillet et al., ”16kbit 1T1R OxRAM arrays embedded in 28nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors,” IEEE International Memory Workshop (IMW), 1-4, 2021.
5. H. Kim, M. Mahmoodi, H. Nili, D. B. Strukov, ”4K-memristor analog grade passive crossbar circuit,” Nature communications 12, 5198, 2021.
6. W. Wan et al., ”A compute-in-memory chip based on resistive random access memory,” Nature 608, 504-512, 2022.
7. B. Fleischer et al., "A Scalable Multi- TeraOPS Deep Learning Processor Core for AI Trainina and Inference," 2018 IEEE Symposium on VLSI Circuits, Honolulu, HI, USA, 2018.