SAND2022-2012R

Vector-Matrix Multiplication Engine for Neuromorphic Computation with a CBRAM Crossbar Array

Blayne Tolleson[1], Matthew Marinella[1], Christopher Bennett[2], Hugh Barnaby[1], Donald Wilson[1], and Jesse Short [1]
[1]Arizona State University, Tempe, AZ
[2]Sandia National Laboratories, Albuquerque, NM

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

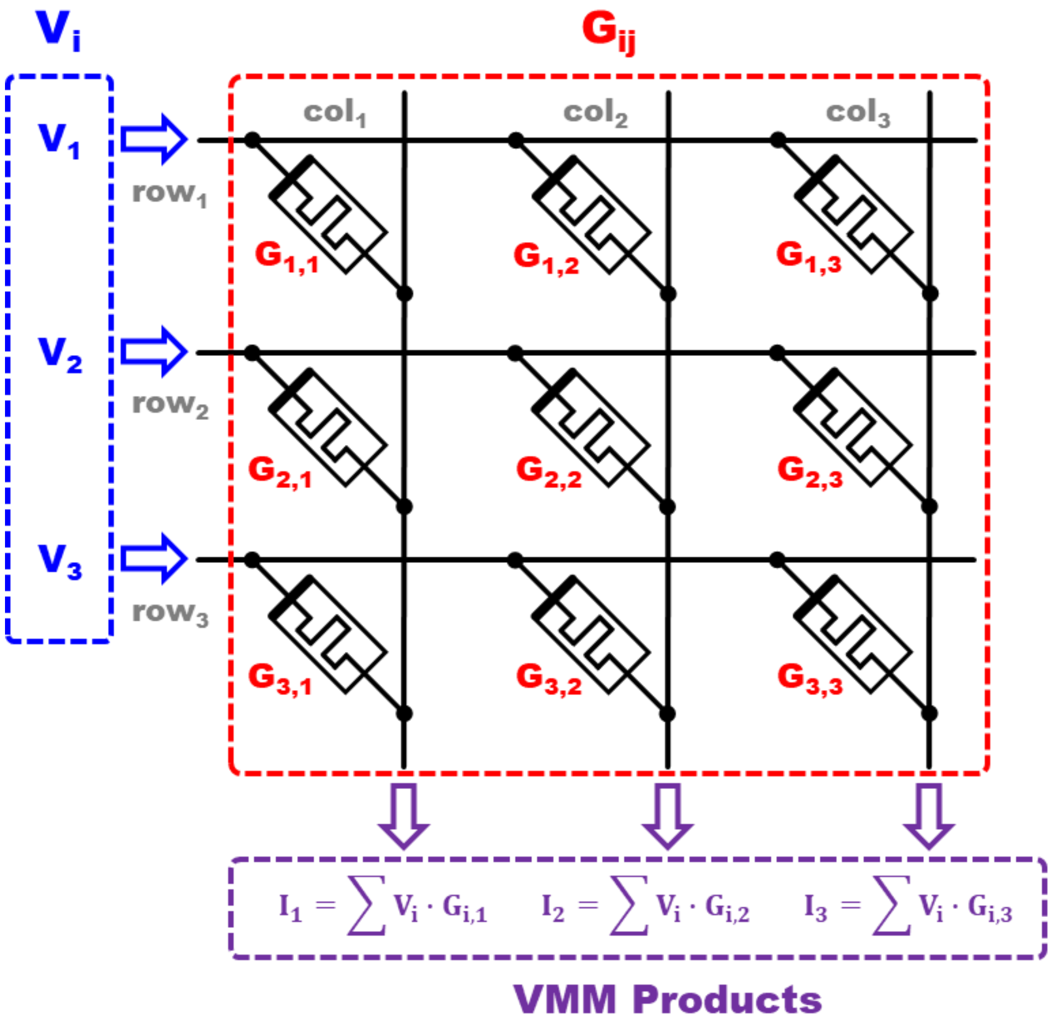
Acknowledgements

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This presentation describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

This work was funded in part by the Defense Threat Reduction Agency (DTRA) under grant no. HDTRA1-17-1-0038. The authors would like to thank Jacob Calkins of DTRA for his support of this work.

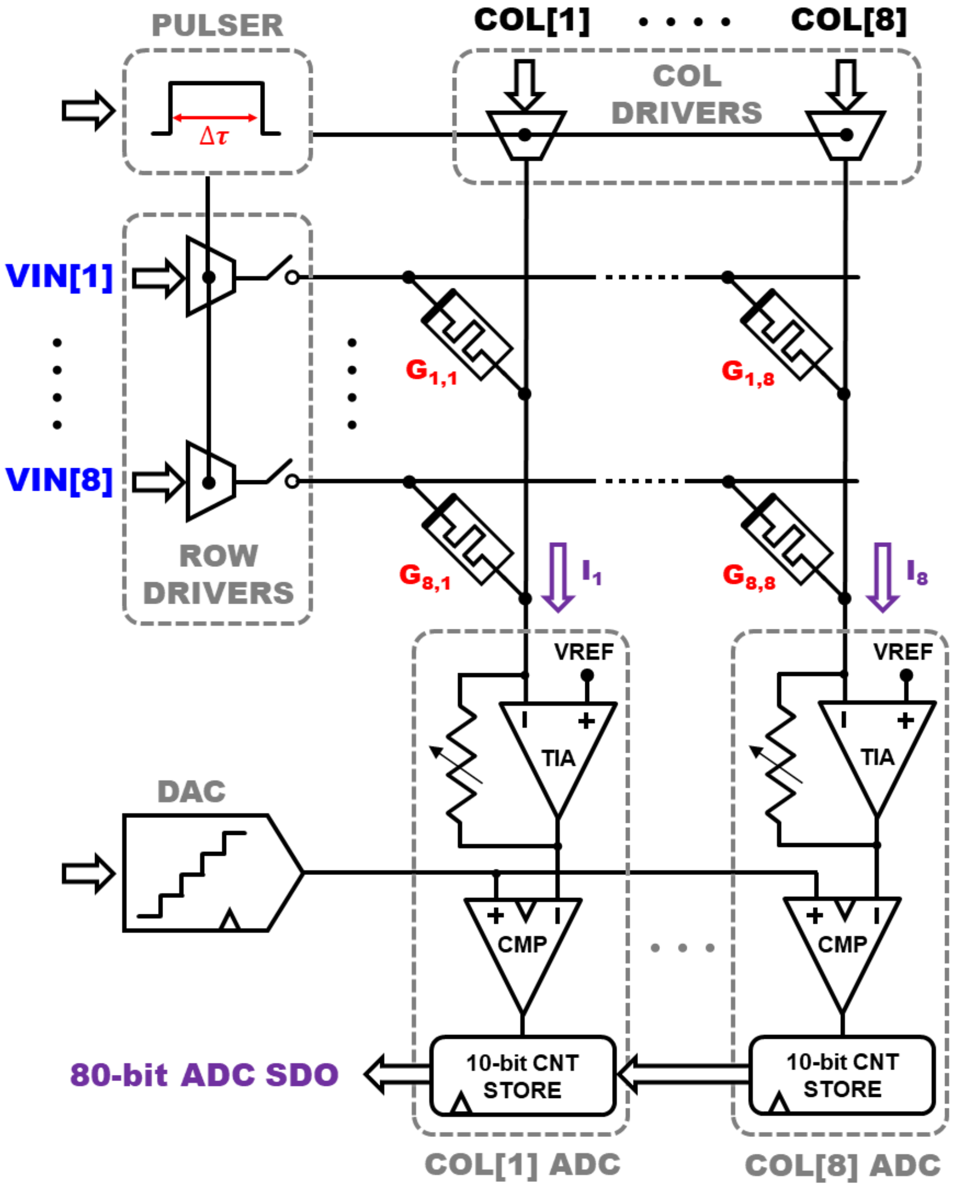
Introduction

- Core function of many neural network algorithms is the dot product, or vector matrix multiply (VMM) operation
- Crossbar arrays utilizing resistive memory elements can reduce computational energy in neural algorithms by up to five orders of magnitude compared to conventional CPUs [1].
- Moving data between a processor, SRAM, and DRAM dominates energy consumption [1].
- By utilizing analog operations to reduce data movement, resistive memory crossbars can enable processing of large amounts of data at lower energy than conventional memory architectures [1].

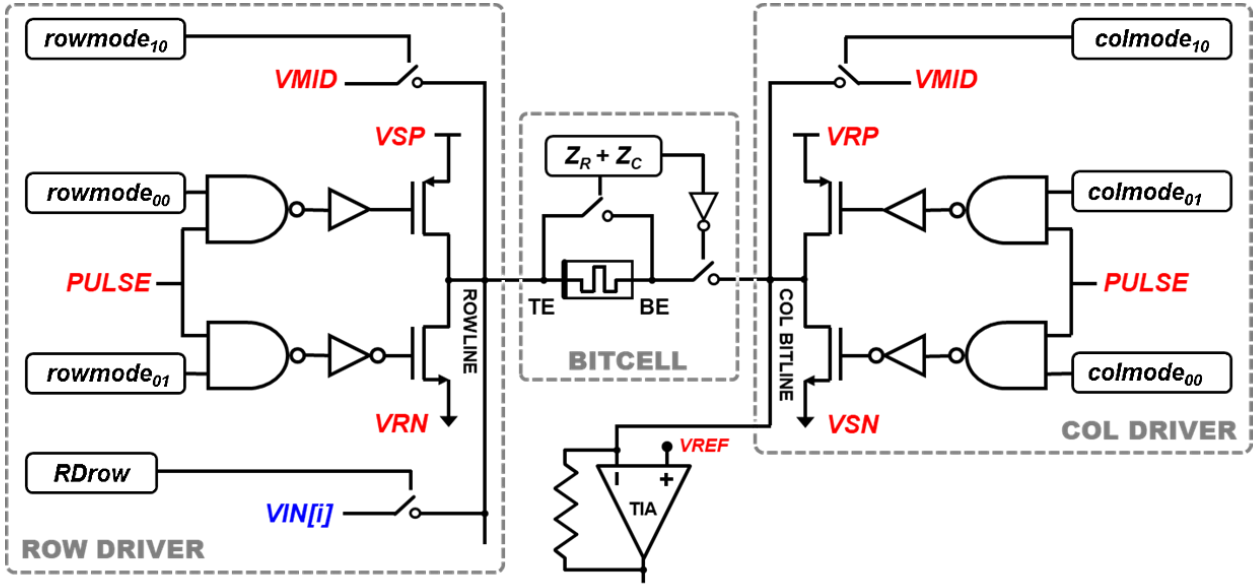


VMM ASIC Prototype Architecture

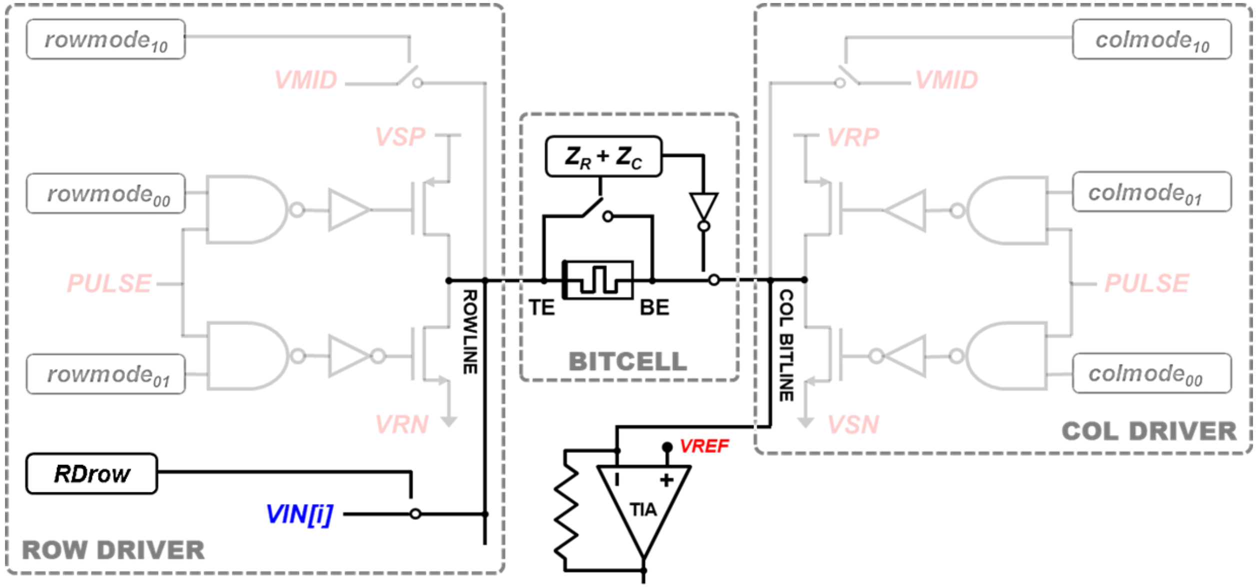
- **Nonvolatile Memory (NVM) Array** – Silver-doped chalcogenide based (Ag-Ge-Se) conductive-bridging random-access memory (CBRAM) post-processed onto CMOS wafer
- **Readout & ADC** – Transimpedance stage sums and converts row device currents for on-chip ADC operation. Typically, ADC area/power dominate circuit density/energy efficiency [2]. Prototype on-chip 10-bit ADC design amortizes area/power by using single comparator and counter per column with shared ramping DAC [2]
- **TIA** – 10 programmable gain settings enable scaling of dot product sums to span more LSBs of ADC.
- **Write & Voltage Pulser** – Weight/conductance changes achieved by setting modes in row/column drivers then triggering on-chip voltage pulser to turn on drivers for duration of pulse width. Test mode enables direct pulse width measurements
- **Interface Configuration** – 48 I/O pads used in design to maintain small areal footprint. A simple serial scan interface used to input ~68 internal configuration bits. Additionally, 80-bit serial data output from ADC arranged in long shift register for simple scan out of VMM values over 2-wire serial interface



Mode Configuration

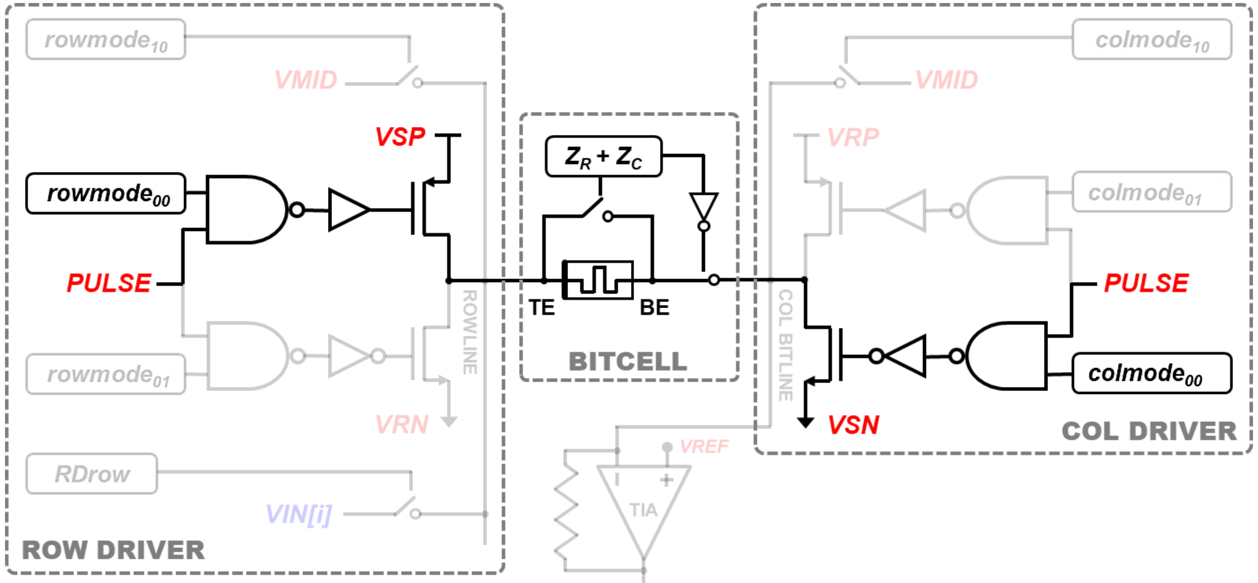


- Row/column drivers and array bitcells form an "H-bridge" structure. Mode control bits configure bitcell into different modes by turning branches of row/column drivers on or off

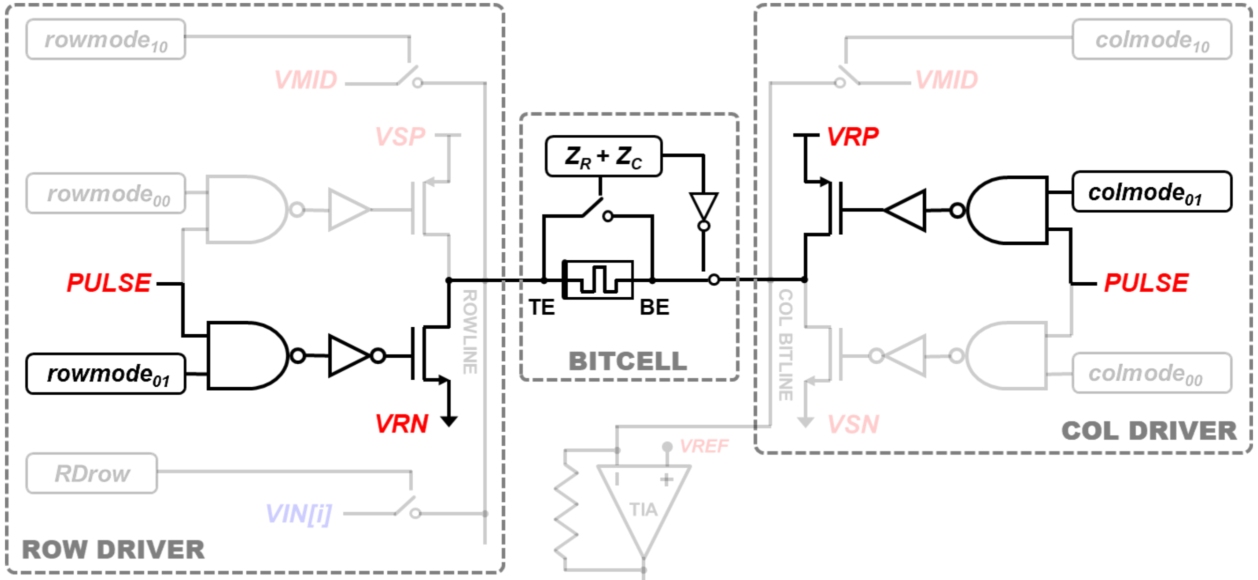


- **Read Mode** – An input VIN[i] voltage is driven onto rowline and dot product sum of the conductances and input vector results in a scaled voltage on each column's TIA output.

Mode Configuration

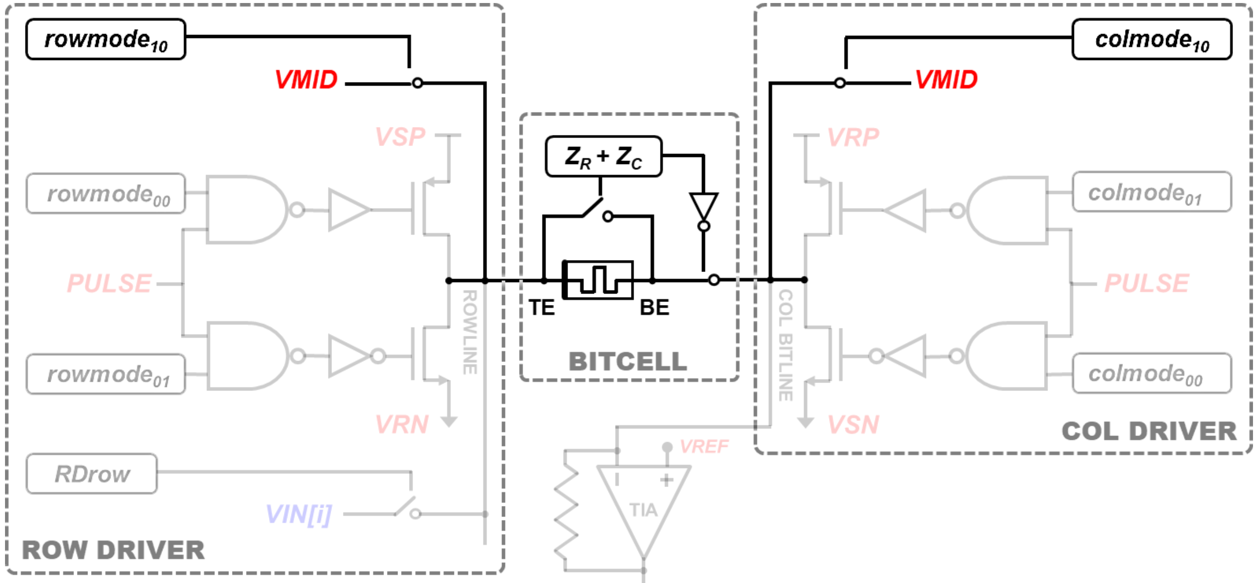


- **SET Mode** – Mode control bits enable top branch of the row driver bottom branch of column driver. External drive bias ~|VSP-VSN| applied from TE to BE of CBRAM element, <u>increasing</u> its weight value. Bias is applied for duration of PULSE output from on-chip voltage pulser

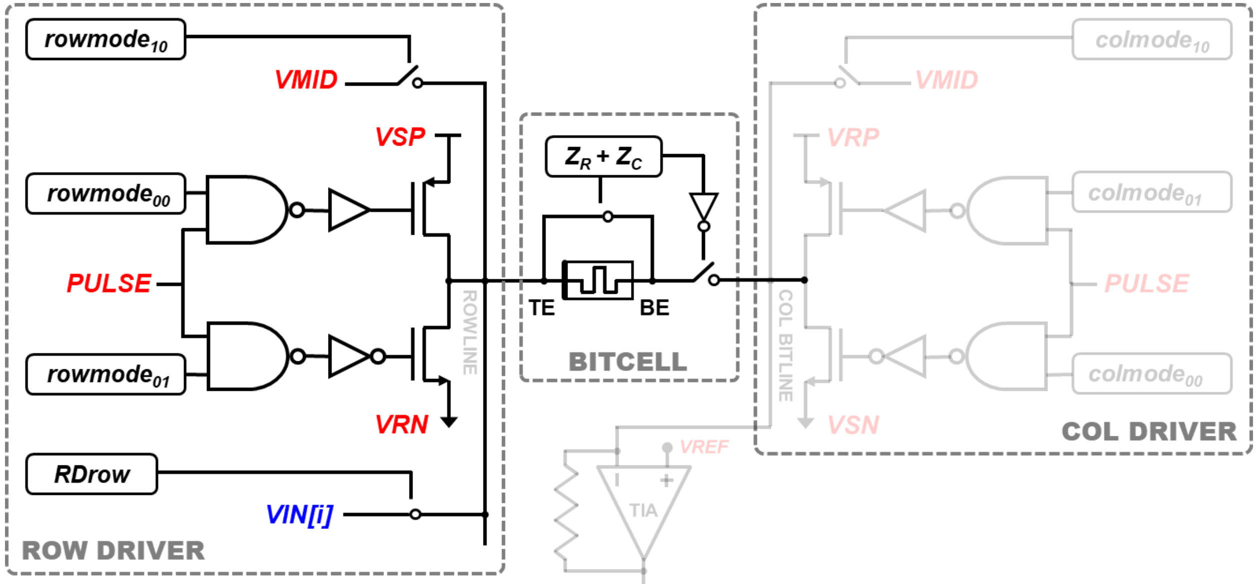


- **RST Mode** – Mode control bits enable bottom branch of the row driver and top branch of column driver. External drive bias ~|VRP-VRN| applied from BE to TE of CBRAM element, <u>decreasing</u> weight value. Bias is applied for duration of PULSE output from on-chip voltage pulser

Mode Configuration



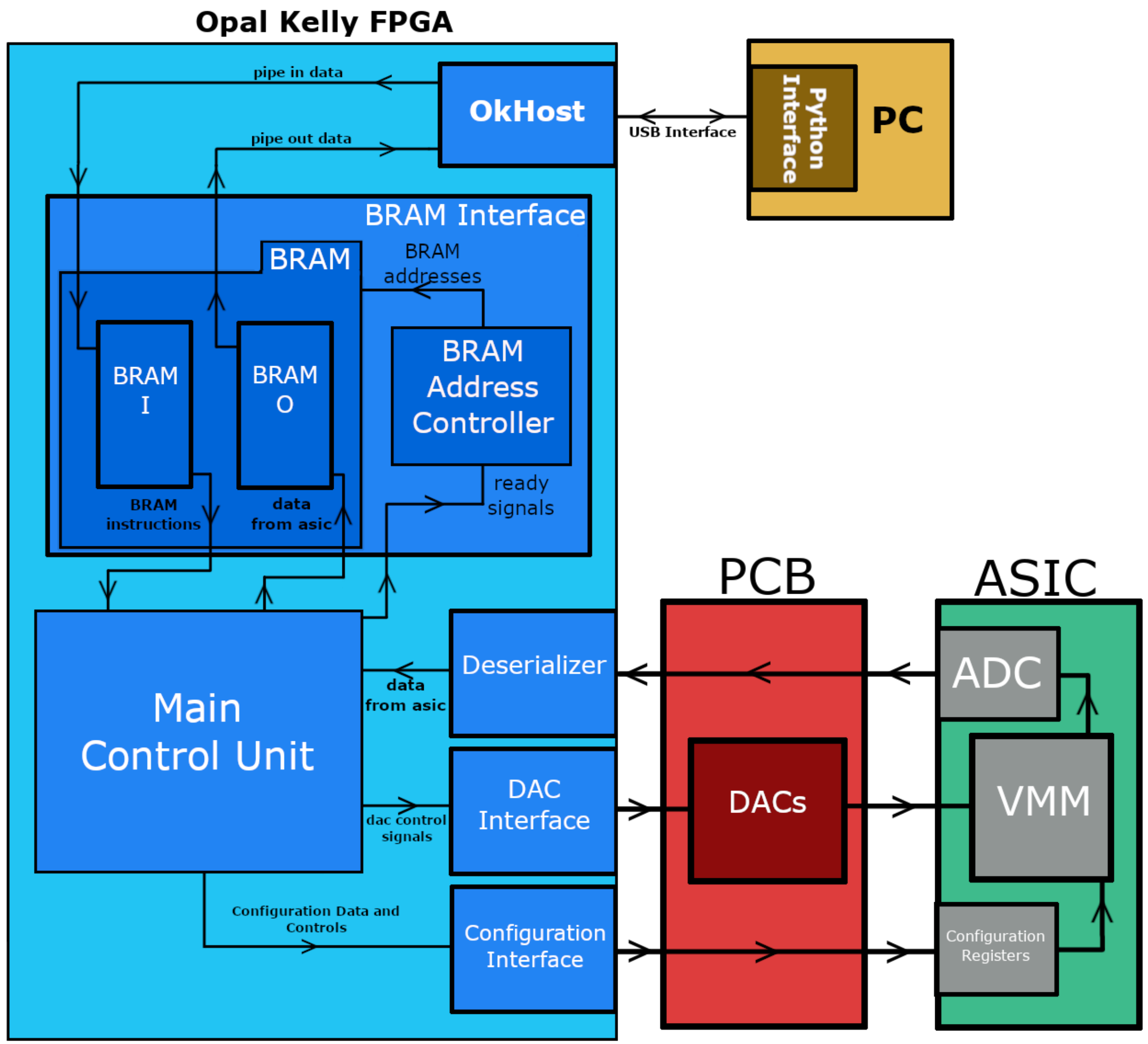
- **VMID Reference** – External VMID reference voltage exists for returning both TE and BE of CBRAM device to a known neutral voltage after each Write pulse. Ensures device does not dwell with a large voltage (due to parasitic capacitances) across it for long periods after Write pulse has finished



- **High-Impedance Mode** – Cells put in a disconnected mode where CBRAM device is disconnected from bitline. This ensures no participation in Read or Write actions impressed across the shared bitline. A parallel shunt switch is closed, so that any transient charging/discharging sees a low impedance path

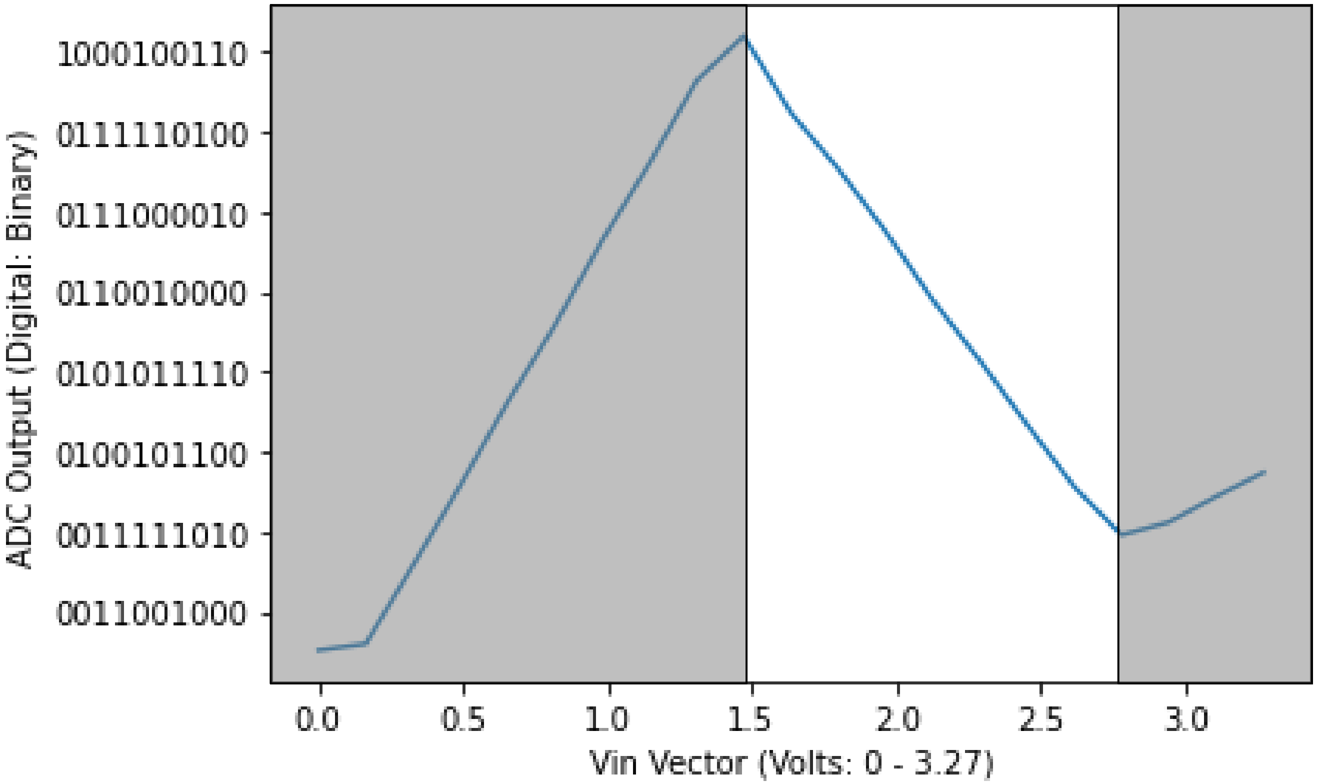
Python Interface & Dataflow

- Python interface created to send/receive VMM ASIC data
- Instructions encoded in Python interface on PC and decoded in Main Control Unit (MCU) of Opal Kelly FPGA
- Interface PCB designed to enable analog inputs to ASIC via board-level DAC ICs using SPI interface
- FPGA deserializer unit takes 80-bit serial data output (SDO) stream from ASIC into the MCU then stored in BRAM; ASIC output data then analyzed on a PC
- **Takeaway** – Python interface will be further developed to work in conjunction with neuromorphic simulation platform *CrossSim* to train network and implement Write-Verify routines [1]



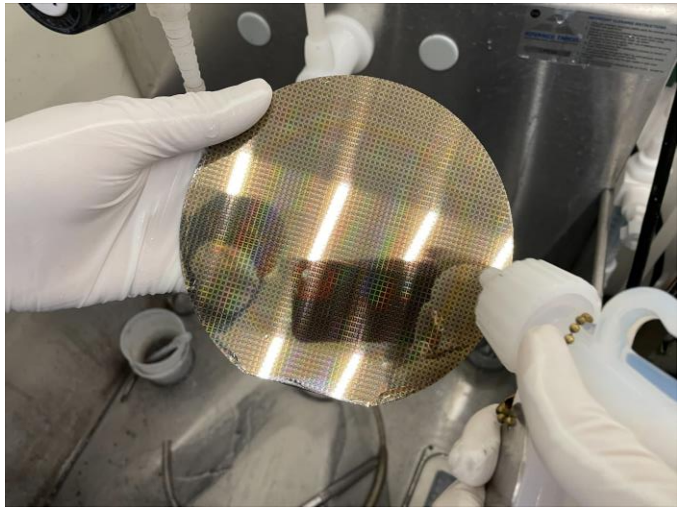
VMM ASIC Output Data

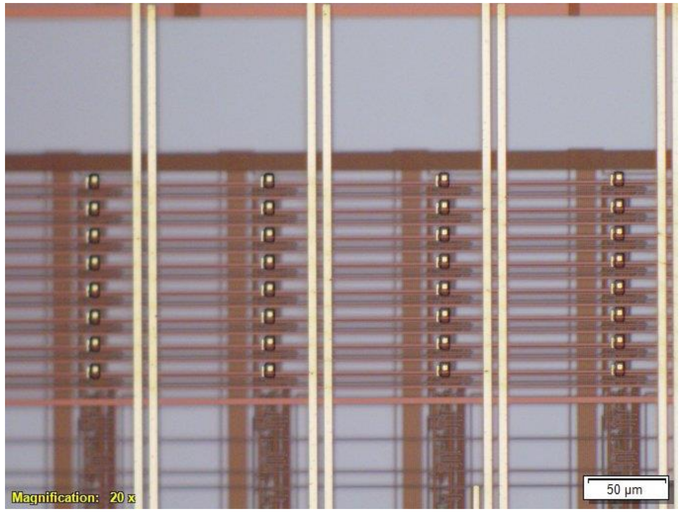
- Testing VMM Read/ADC operation prior to CBRAM device array integration, last array column populated with fixed resistors
- 10-bit digital counts output from on-chip ADC when the input voltage levels are swept
- Each digital output count corresponds to VMM sum at each input voltage (for single test column)
- Column TIAs have inverting gain. Higher column currents (i.e., higher input voltages or lower column device resistances) cause column comparator to trip quicker and result in a smaller digital output counts.

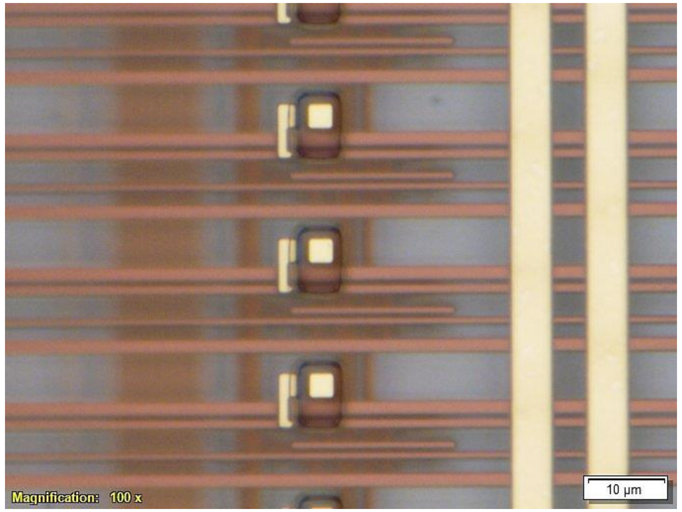


- Adjustments to references voltages in column TIA and Ramp DAC are being tested to try and increase range of outputs showing linear and negative slope (unshaded region in plot).

Ongoing Development







- **CBRAM Device Arrays** – Several wafers have now gone through a chemical mechanical polishing (CMP) process and dicing. The diced samples have been patterned to begin post-processing and integration of CBRAM arrays. Masks have been designed and ASU has a well established CBRAM process. In coming weeks, CBRAM device arrays will be integrated onto the ASIC and measured

- **Python Interface** – Continuing to refine and optimize Python interface to handle dataflow to and from VMM ASIC. Write-Verify routines being developed in anticipation of CBRAM device array integration. Goal is to adapt the existing interface to incorporate CrossSim for network training using measured device properties and variabilities

References

[1] S. Agarwal, S.J. Plimpton, R. Schiek, I. Richter, A. Hsia, D. Hughart, R. Jacobs-Gedrim, C. James, M. Marinella. (2017). CrossSim. [Online]. Available: http://cross-sim.sandia.gov

[2] M. J. Marinella et al., "Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 86-101, March 2018, doi: 10.1109/JETCAS.2018.2796379.