Design Methodologies for Building a CTT-based Analog in-Memory Computing (iMC) Engine

Steven Moran, UCLA CHIPS (steven.moran@g.ucla.edu)

Introduction

Background

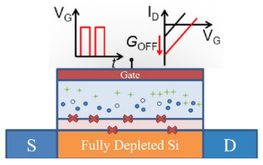
- CTT **analog** synapses inspired by CTT-based digital eNVM solutions

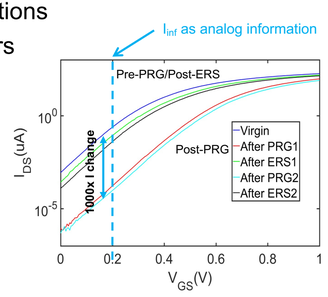
Milestones

- Demonstrated analog dot-product computation in CTT array with < 3% error [1]
- Benchmarked CTT with other state-of-the-art analog devices for NN applications
- Developed CIRCUS: tool for characterizing the effect of circuit-induced errors and device non-idealities on computation

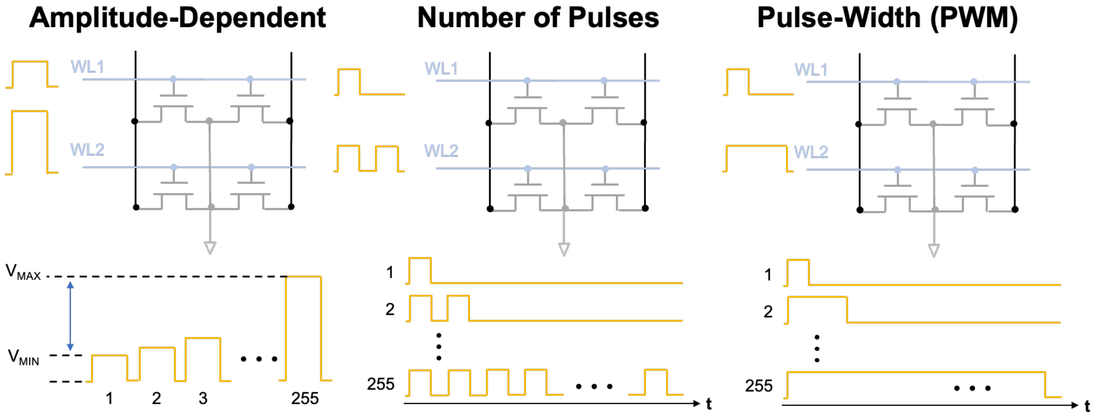
On-going & Future Work

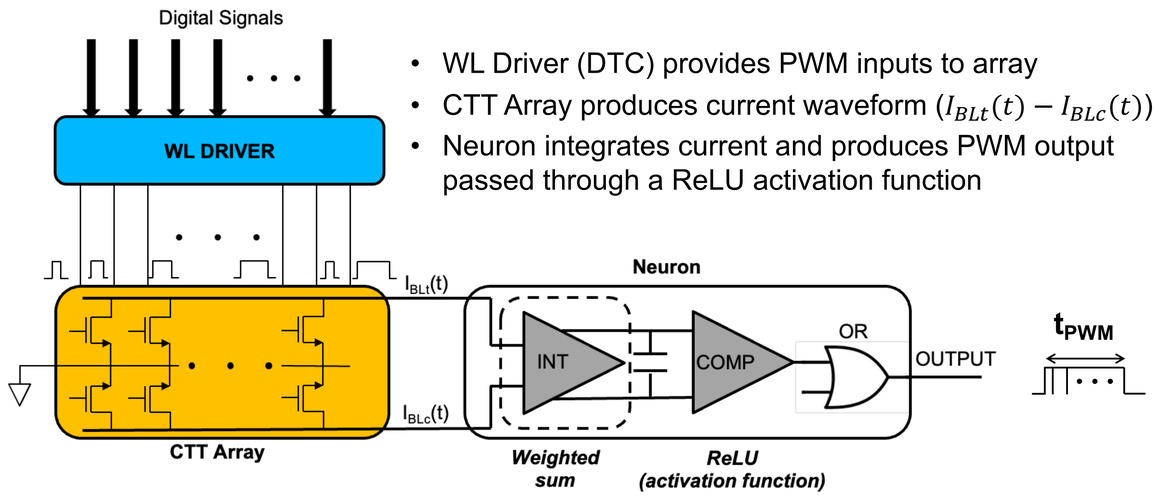
- Taped out NeuroCTT 0.3 (Hardware expected ~December)
- Finalized Daughterboard + Motherboard Testing Design
- CIRCUS post-silicon validation



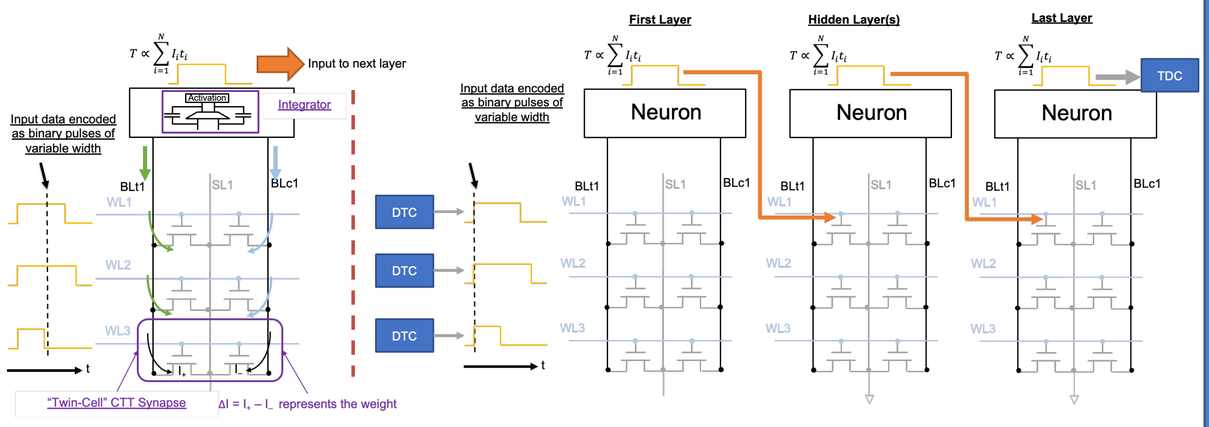


Input & System Architecture



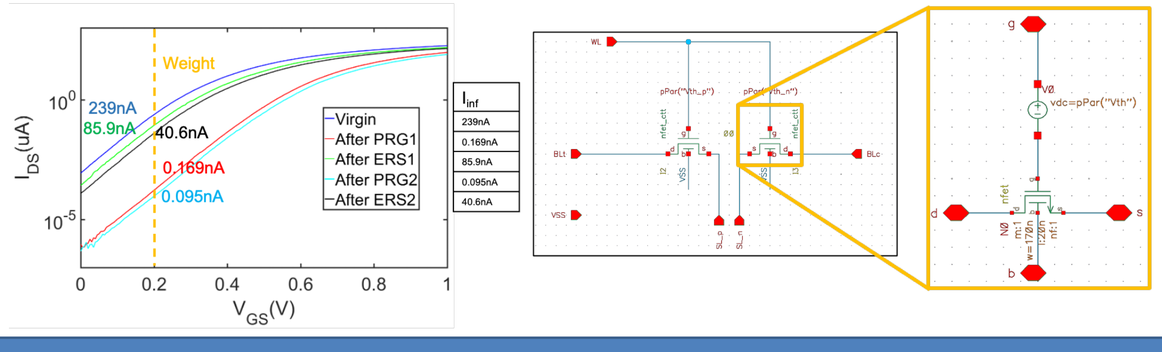


- WL Driver (DTC) provides PWM inputs to array
- CTT Array produces current waveform $(I_{BLt}(t) - I_{BLc}(t))$
- Neuron integrates current and produces PWM output passed through a ReLU activation function

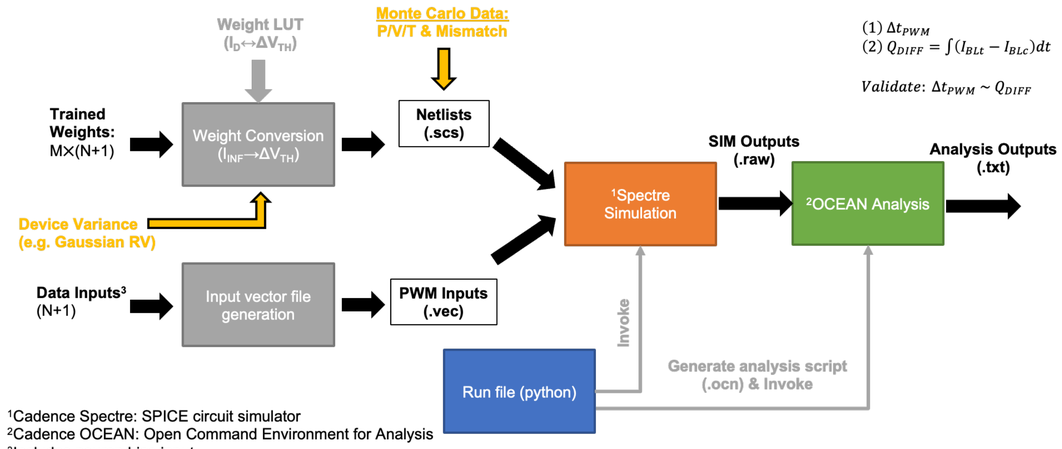


CTT-Hardware-based Inference Realistic Circuit Universal Simulator (CIRCUS)

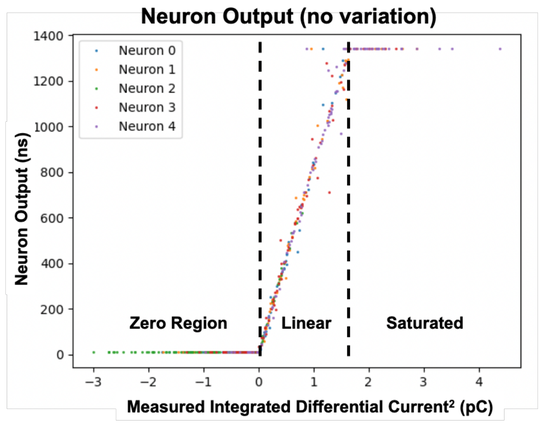
- Weights encoded as ΔV_{TH} parameters

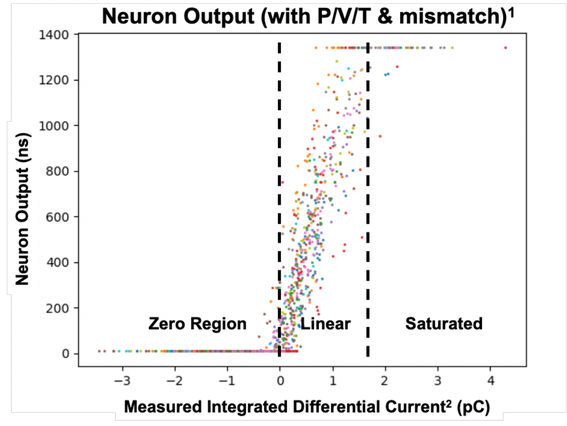


CIRCUS Flow Overview



[1]Cadence Spectre: SPICE circuit simulator
[2]Cadence OCEAN: Open Command Environment for Analysis
[3]Includes neuron bias input

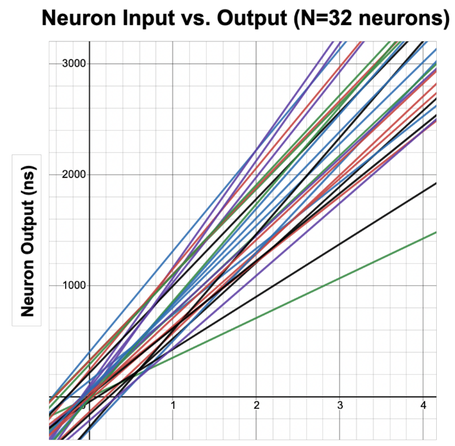




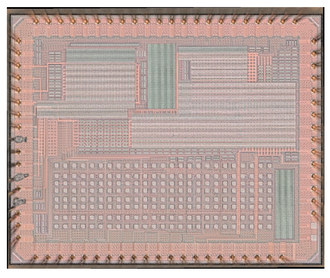
[1]Post-calibration
[2]$N_j = \int \left(I_{BLc_j} - I_{BLc_j}\right) dt$

Net Takeways:

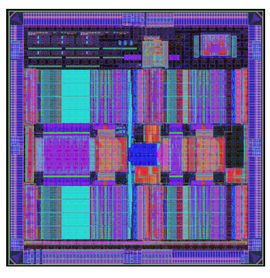
- Each individual neuron has different input/output characteristics
- Encode effective weights / column
  - Original Method: set weight, I_D, per device
  - New Method:
    1. **Slope correction:** set weight s.t. it produces a particular neuron output (e.g. t_{PWM})
    2. **Offset correction:** Utilize extra WLs to compensate for input offset



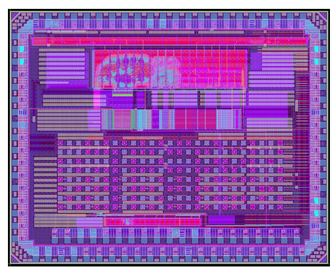
NeuroCTT 0.1



NeuroCTT 0.2



NeuroCTT 0.3



Conclusions

- CIRCUS: tool for characterizing the effect of circuit-induced errors and device non-idealities on computation
- Taped-out NeuroCTT 0.3 Design (Delivery: ~Dec. 2021)
- Validate CIRCUS Simulations with silicon data & release Abstracted Error Model for inference & training simulations

This work was supported in part by the UCLA CHIPS consortium, the Defense Advanced Research Projects Agency (DARPA), and UC Multicampus Research Programs and Initiatives (UC-MRPI). We would like to thank GlobalFoundries and ON-Semiconductor for their support.

References

1. X. Gu, *et. al*, "Charge-Trap Transistors for CMOS-Only Analog Memory," in IEEE Transactions on Electron Devices, Oct. 2019.

UCLA Samueli School of Engineering

2021 UCLA CHIPS Workshop
Poster Session

CHIPS CENTER FOR HETEROGENEOUS INTEGRATION AND PERFORMANCE SCALING