

Exceptional service in the national interest



CrossSim: GPU-Accelerated Simulation of Analog Neural Networks

T. Patrick Xiao, Christopher H. Bennett, Ben Feinberg, Matthew Marinella, Sapan Agarwal

Sandia National Laboratories, Albuquerque, NM
txiao@sandia.gov



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.

Deep learning inside memory arrays

Matrix-vector
multiplication:

$$\mathbf{Ax}$$

Mathematical

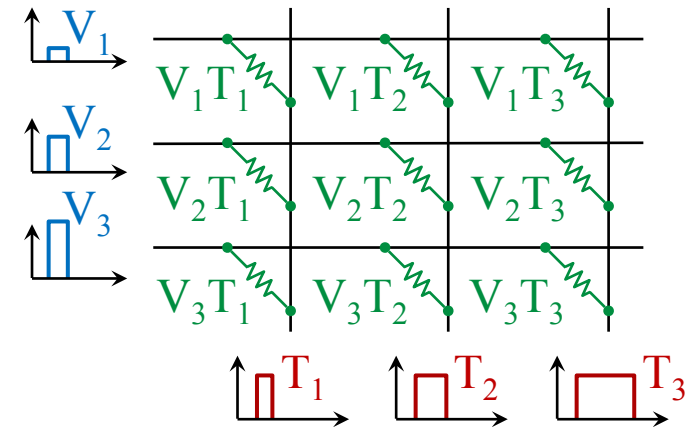
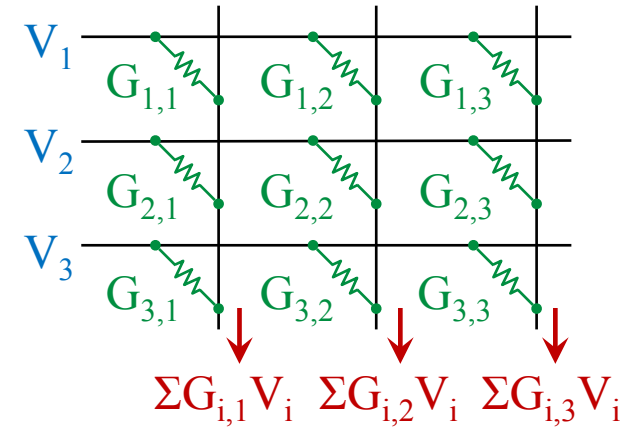
$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^T \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix} = \begin{bmatrix} \sum A_{i,1} x_i & \sum A_{i,2} x_i & \sum A_{i,3} x_i \end{bmatrix}$$

Outer product
update:

$$\mathbf{x}\delta^T$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \otimes \begin{bmatrix} x_1\delta_1 & x_1\delta_2 & x_1\delta_3 \\ x_2\delta_1 & x_2\delta_2 & x_2\delta_3 \\ x_3\delta_1 & x_3\delta_2 & x_3\delta_3 \end{bmatrix} = \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 \end{bmatrix}$$

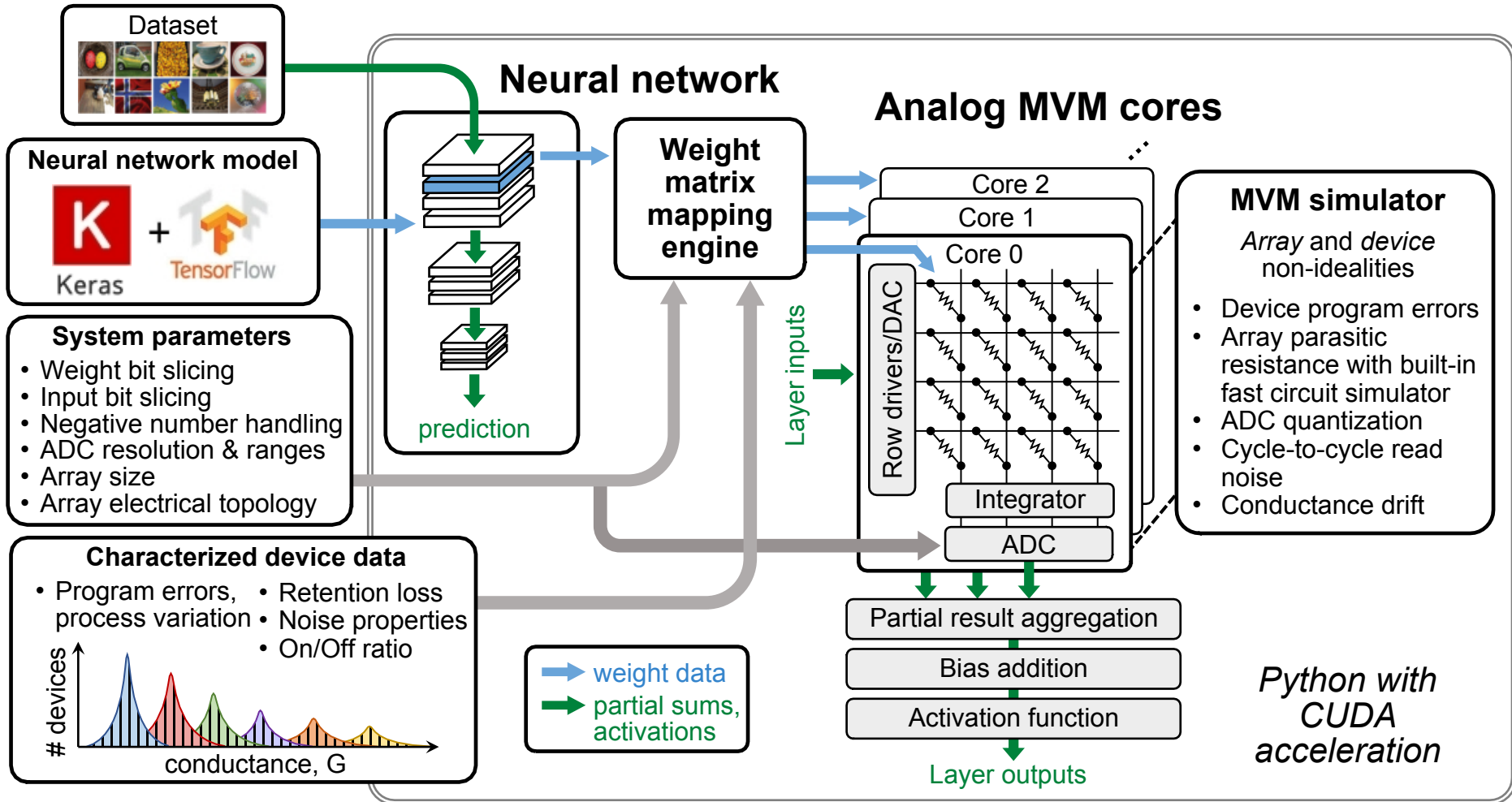
Electrical



Highly energy-efficient, *but is it accurate enough?*

#ROSS SIM Inference

Inputs to CrossSim



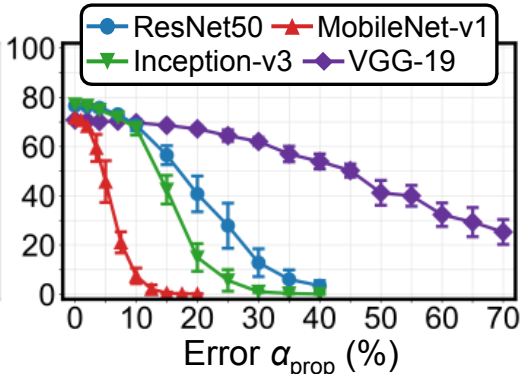
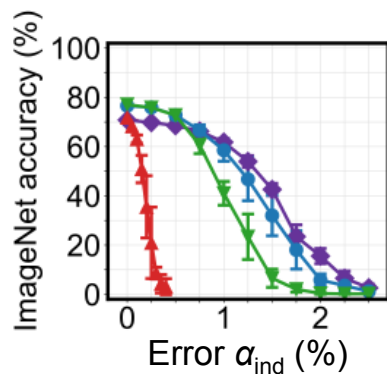
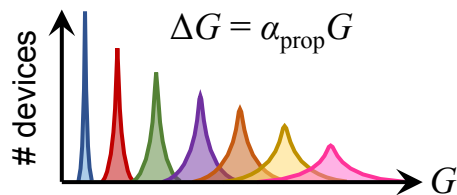
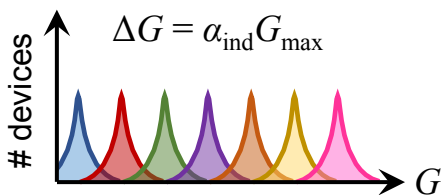
To be released soon! Check cross-sim.sandia.gov

Multi-scale modeling of inference accuracy

Device properties affect accuracy

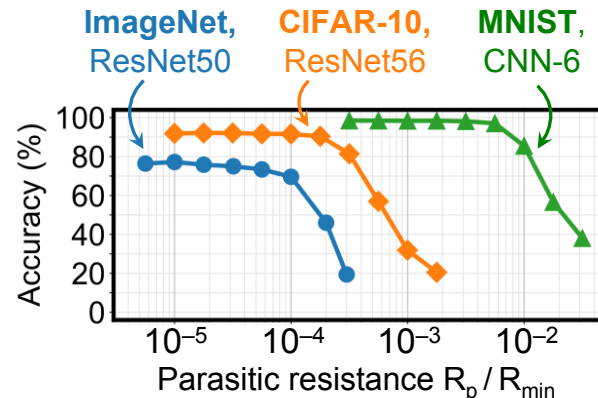
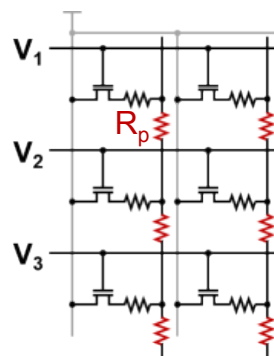
State-independent programming error

State-proportional programming error



Array design affects accuracy

CrossSim's fast built-in circuit simulator



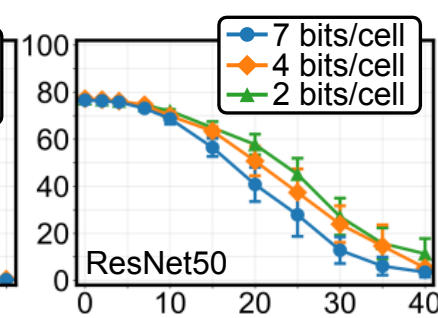
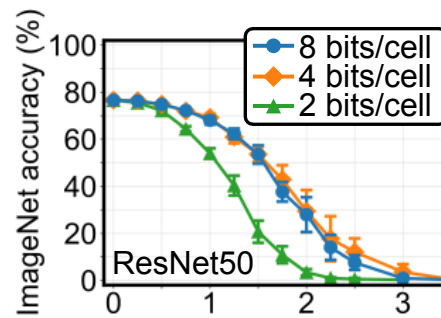
System architecture affects accuracy

Offset subtraction

Differential cells

$$W_{ij} \sim G_{ij} - G_{\text{offset}}$$

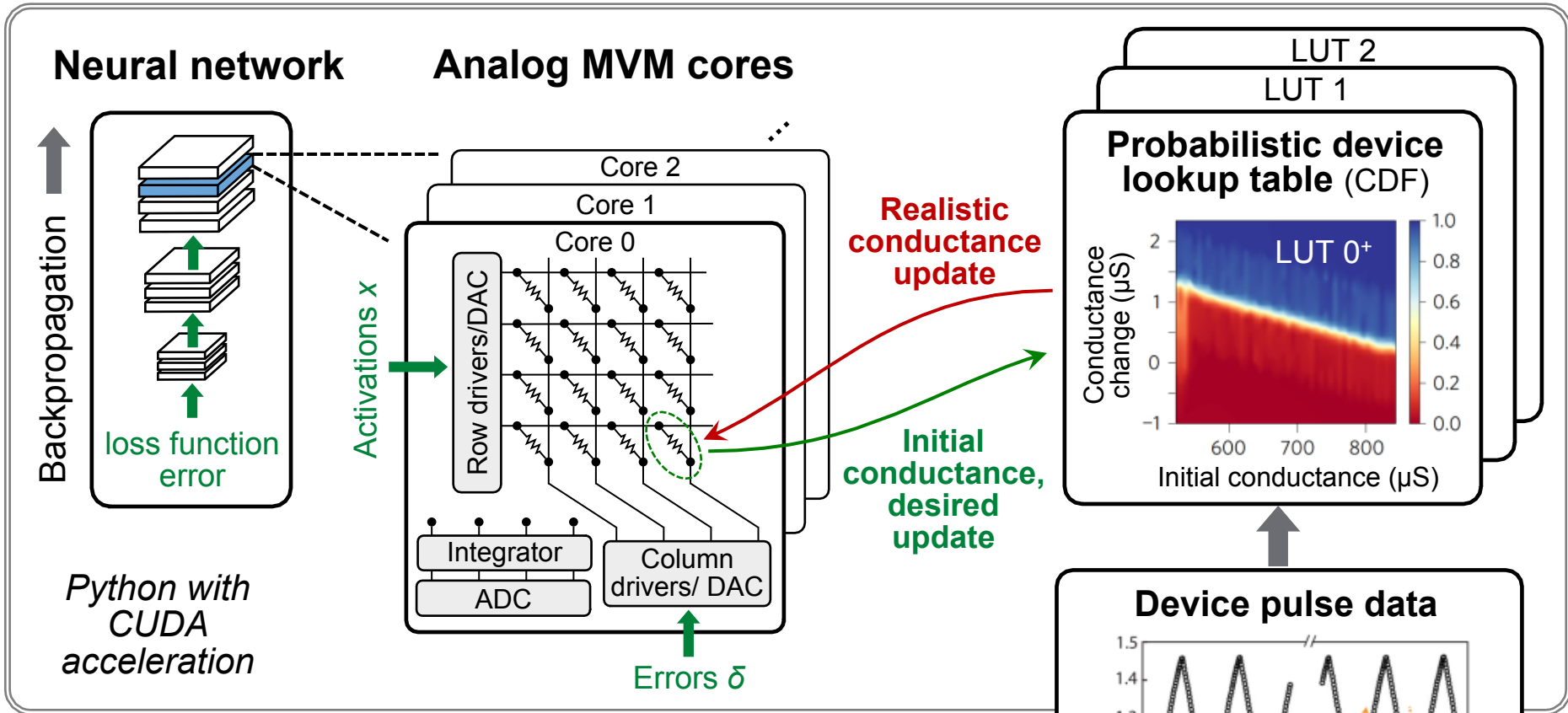
$$W_{ij} \sim G_{ij}^+ - G_{ij}^-$$



Xiao et al, *arXiv:2109.01262*, 2021

Xiao et al, *Semi Sci Tech*, Accepted (in press), 2021

#ROSS SIM Training



Lookup tables can model:

- Arbitrary device update **nonlinearity** and **asymmetry** properties, not describable by analytical equations
- Cycle-to-cycle **write noise**
- Device-to-device **variation**

Fuller et al, *Science* 2019
 Bennett et al, *IRPS* 2019

From device measurements to accuracy

Device

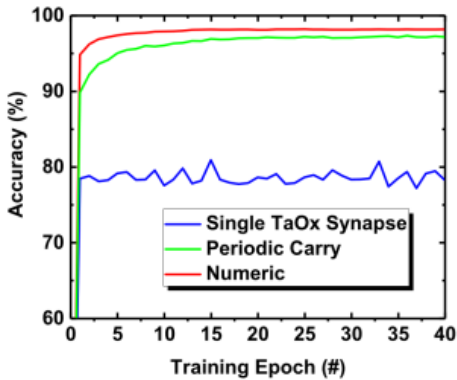
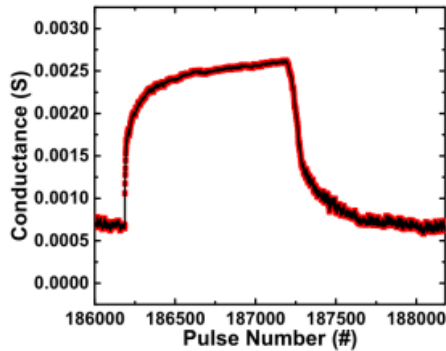
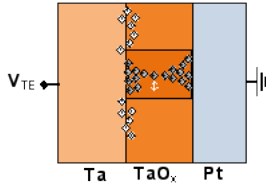


Pulse data



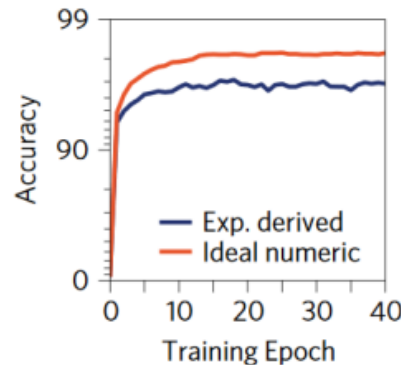
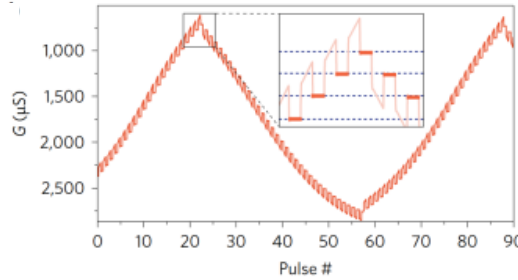
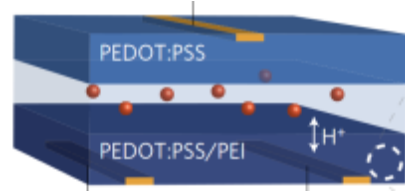
MNIST accuracy (2-layer MLP)

TaO_x ReRAM



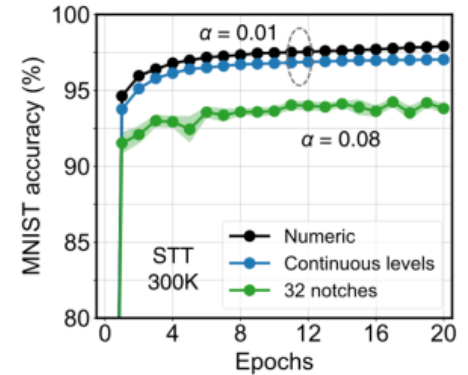
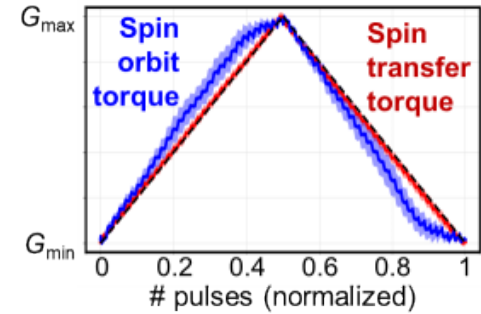
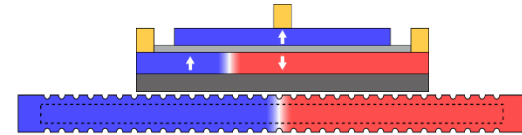
Marinella, Agarwal et al, *JETCAS* 2018

Electrochemical RAM



Van der Burgt et al, *Nature Materials* 2017

Domain wall magnetic tunnel junction



Liu et al, *Appl Phys Lett*, 2021

GPU-Accelerated Simulation of Analog Neural Networks

Inference

Training

