Arizona State University

Fully Parallel RRAM Synaptic Array for Implementing Binary Neural Network with (+1, -1) Weights and (+1, 0) Neurons

Xiaoyu Sun, Xiaochen Peng, Pai-Yu Chen, Rui Liu, Jae-sun Seo, and Shimeng Yu

Arizona State University, Tempe, AZ, 85281, USA

Email: pchen72@asu.edu

http://faculty.engineering.asu.edu/shimengyu/

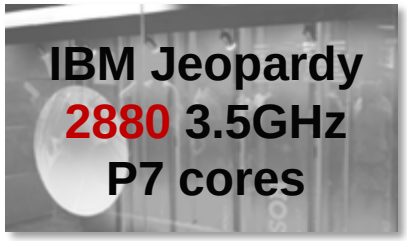
School of Electrical, Computer and Energy Engineering (ECEE)

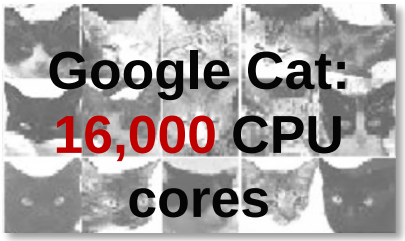
Outline

- **Challenges of On-chip Implementation of Deep Neural Networks and Demand on Binary Neural Networks**
- **RRAM Based Parallel-BNN Accelerator Design**
- **Comparison Between Conventional Serial-BNN and Proposed Parallel-BNN**
- **Summary**

Demands for On-chip Implementation of DNNs

- Deep learning in the cloud: expensive computation, huge training data, **low energy efficiency**, high precision





- Edge computing needs novel **hardware** and **algorithms**
  - Local to the sensor, **low power**, **small area**

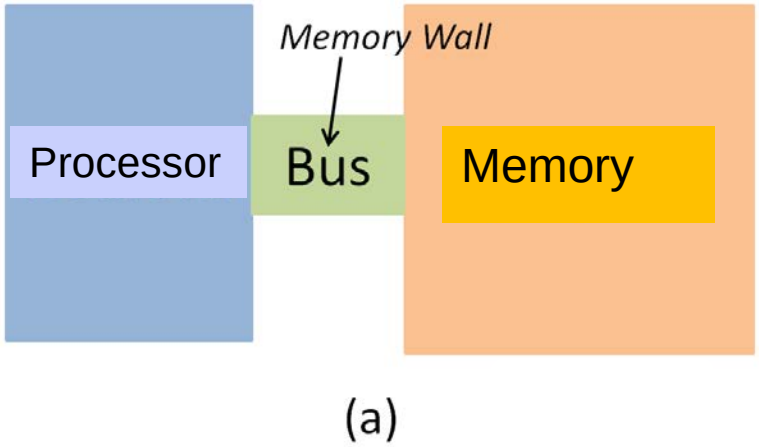
Why Binary Neural Networks?

- When the neural network gets deeper, the high demands on **memory capacity** and **computational power** make it unsuitable for on-chip implementation.
  - ResNet-50 has 25.5M parameters and requires 3.9G high precision operations to classify one image.
- Binary Neural Networks are able to achieve satisfying classification accuracy with significant savings on memory usage and computational resources.

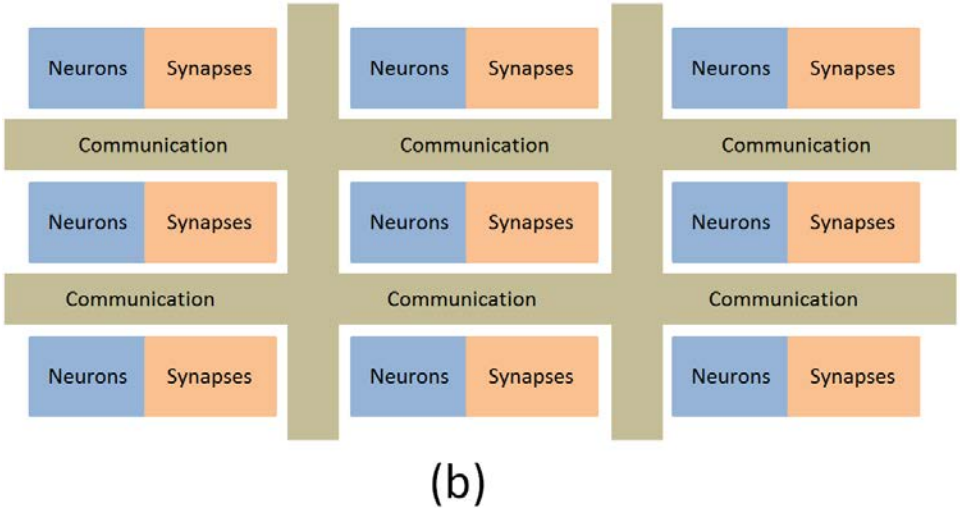
Dataset	FL Precision	Binary Precision
MNIST	98.72%	98.54%
CIFAR-10	89.98%	88.47%

A Shift in Computing Paradigm

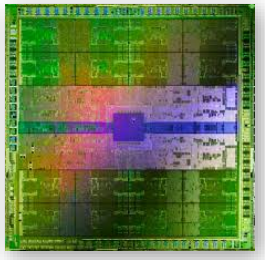
von-Neumann architecture



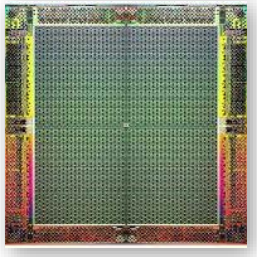
Neuro-inspired architecture



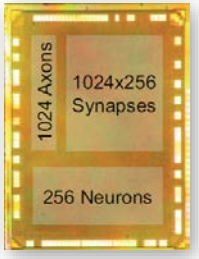
Hardware Acceleration Platforms



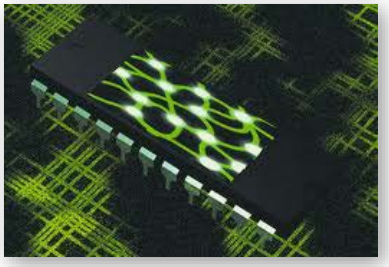
GPU
10 – 30 X



FPGA
10 – 50 X



CMOS ASIC
10^2 – 10^3 X



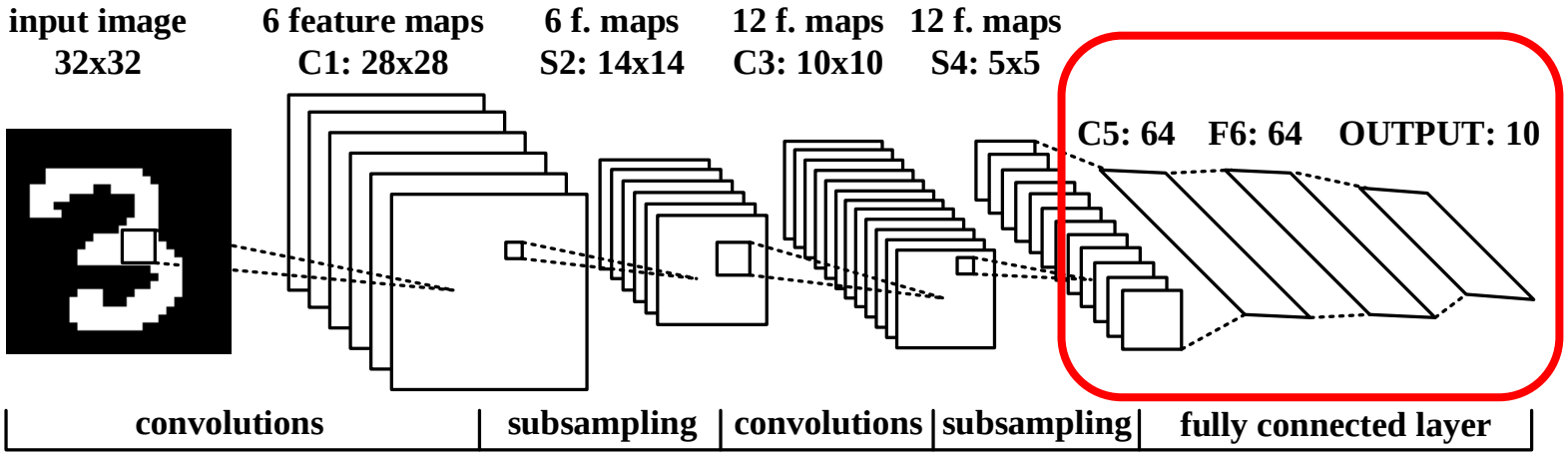
Beyond CMOS
>10^3 X

- **Solution**: beyond CMOS with emerging non-volatile memory
  - Maximizing the **parallel operation** in hardware, load more weights on-chip as emerging NVMs are much smaller than SRAM
  - Loading more weights on-chip eliminate the off-chip memory access, thereby saving latency and energy consumption

Outline

- **Challenges of On-chip Implementation of Deep Neural Networks and Demand on Binary Neural Networks**
- **RRAM Based Parallel-BNN Accelerator Design**
- **Comparison Between Conventional Serial-BNN and Proposed Parallel-BNN**
- **Summary**

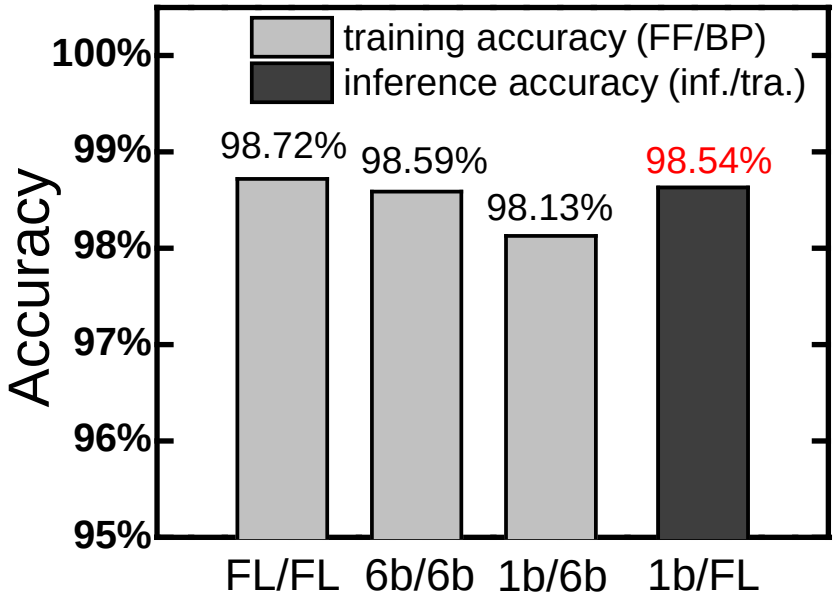
The CNN Topology Used for Evaluation



- Similar to LeNet, 2 convolutional stages, followed by a 64-64-10 MLP.
  - Kernel size: 5x5
  - Dataset: MNIST, binarized to black and white
  - Only the MLP are evaluated on the proposed architecture

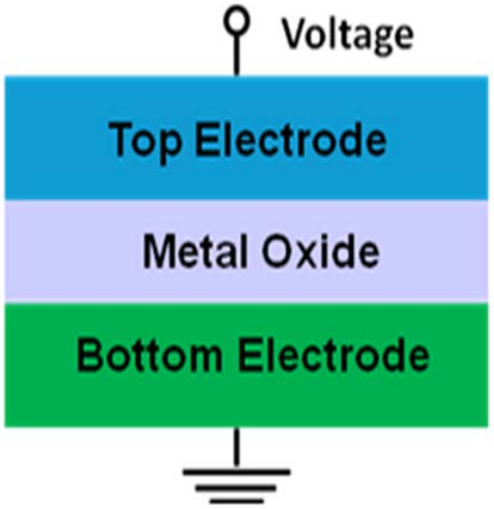
Truncating Weights and Neurons to Binary

- We investigated the impact of precision truncation on the saturated training accuracy.
  - 98.72% with floating point precision in both feed forward and back propagation.
  - **6-bit precision** is necessary for back propagation to maintain a good training accuracy.

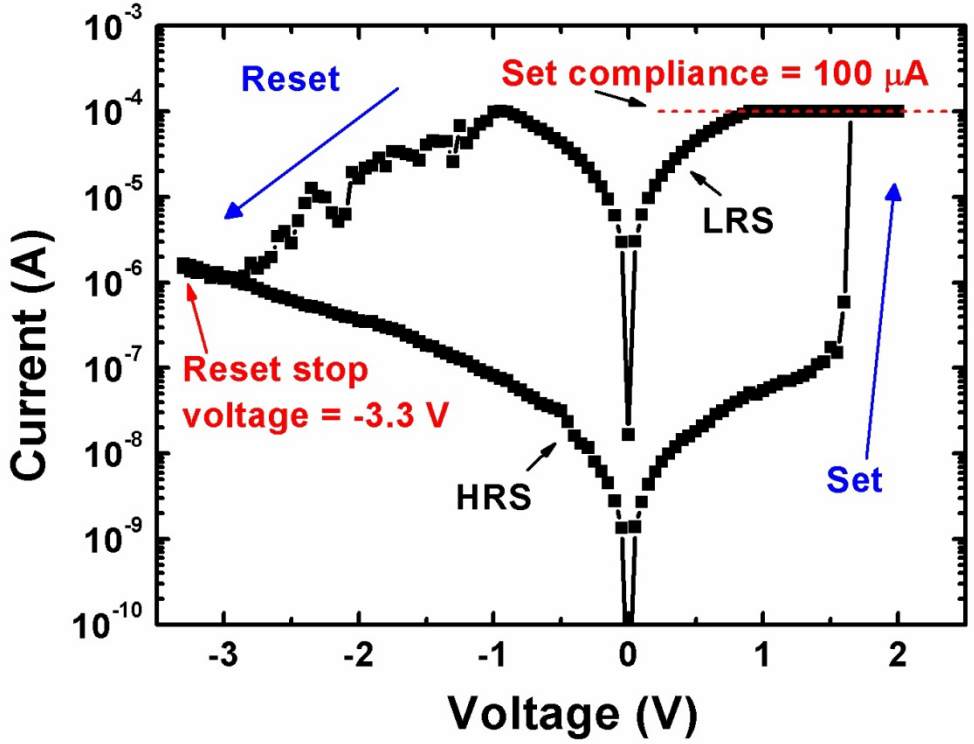


- The baseline accuracy is **98.54%**,where the precision is truncated to 1-bit for inference after the network is well-trained with floating point precision.

What Is Oxide RRAM?



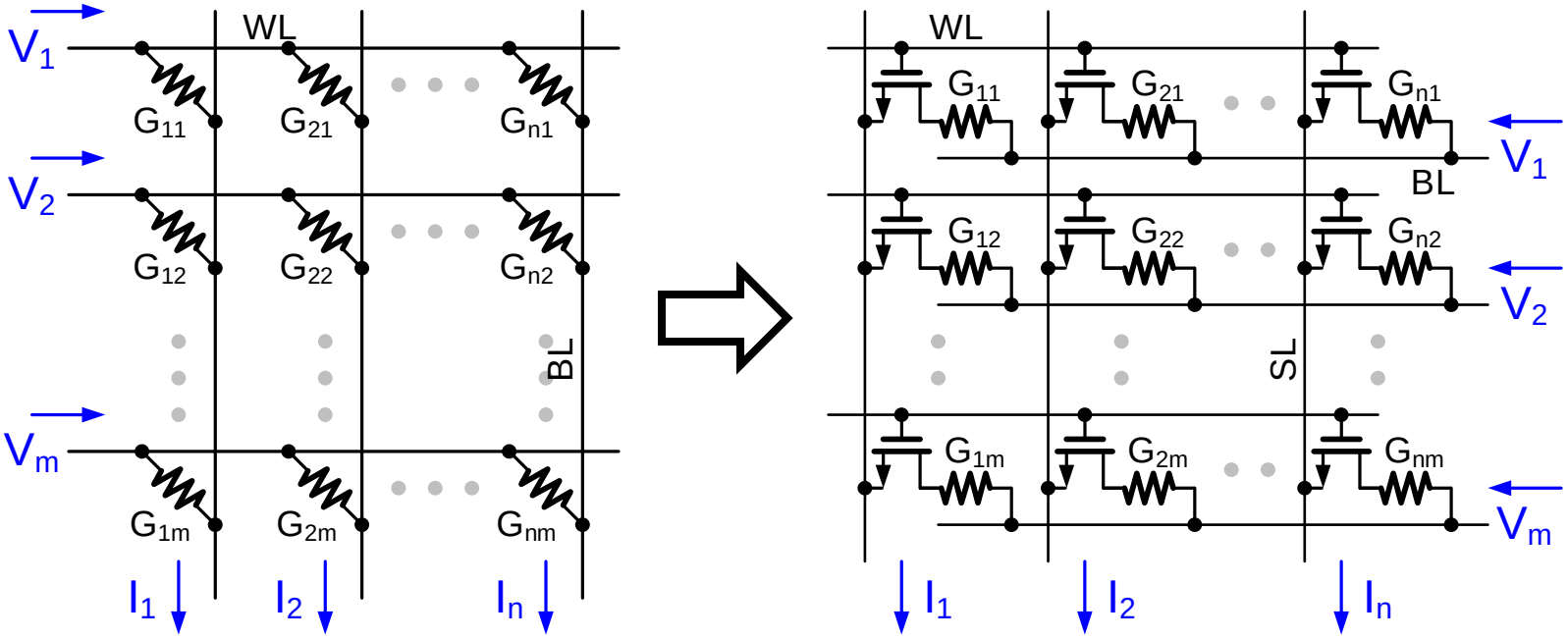
Typical I-V Curve



- ❑ **“0” : High Resistance State (HRS, OFF-state)**
- ❑ **“1” : Low Resistance State (LRS, ON-state)**
- ❑ **HRS→LRS: SET**
- ❑ **LRS→HRS: RESET**

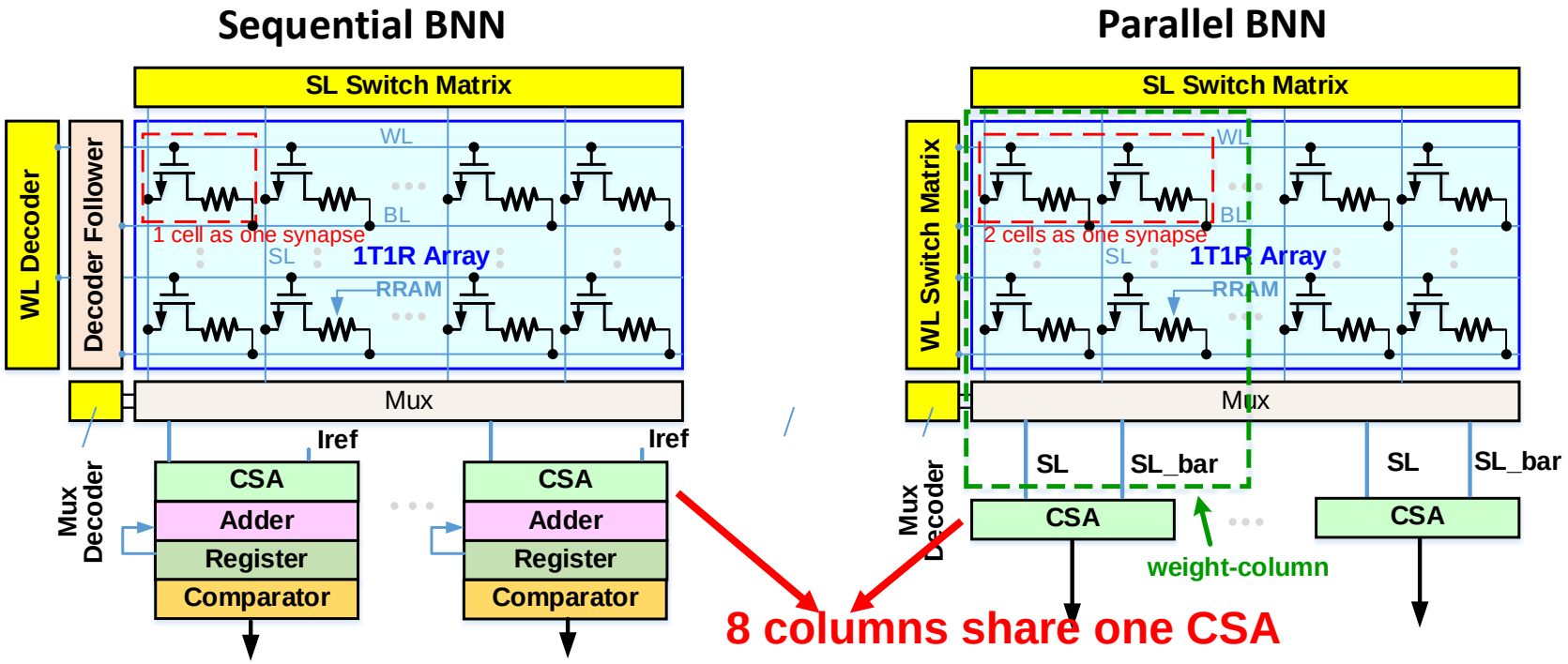
RRAM Arrays for Matrix-Vector Multiplication

- **1T1R vs. crossbar array:**
  - ✓ Less weight update energy
  - ✓ No write disturbance issue
  - ✗ Larger cell area
  - ✗ IR drop on transistor



Crossbar Array **1T1R (Pseudo-crossbar) Array**

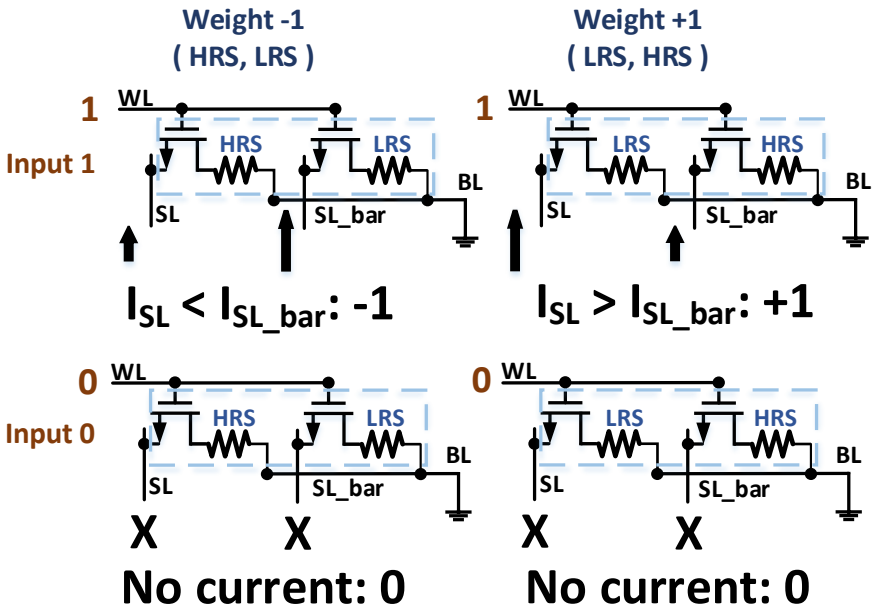
RRAM Based Parallel-BNN Accelerator



- Conventional RRAM synaptic array with **row-by-row read-out**
- 1 cell as one synapse
- Lots of peripheral circuits (sense amplifier, adder, register, etc.) to generate 1-bit neuron output

- Parallel RRAM synaptic array with **parallel read-out,** multiple WLs are activated simultaneously
- 2 cells as one synapse
- A single CSA behaves as the neuron circuit

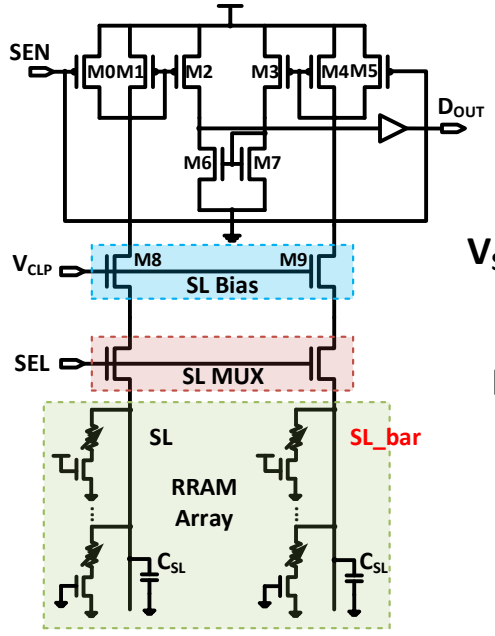
RRAM Cell Configuration Enabling Parallel Operation

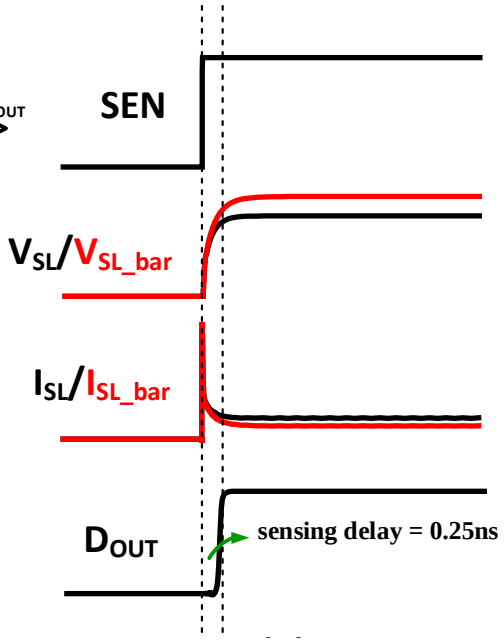


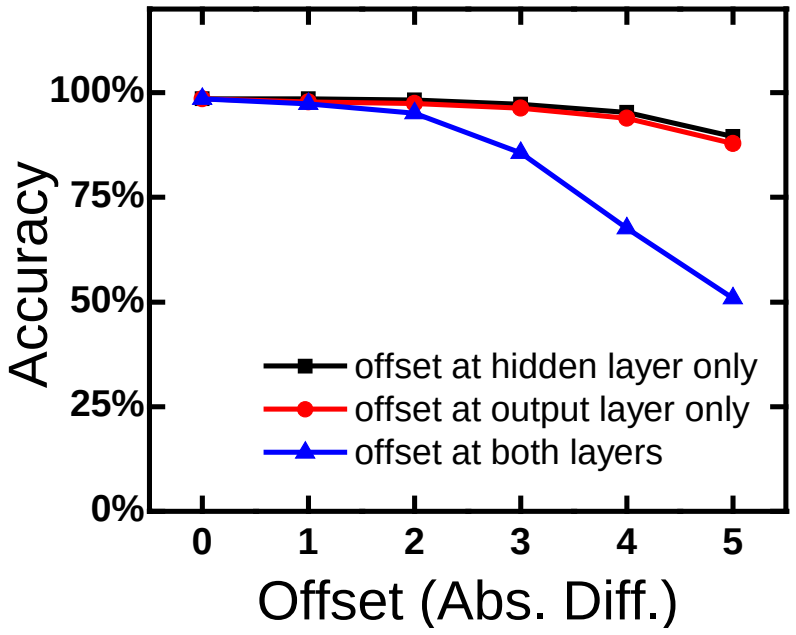
	Weight = -1	Weight = +1
Input = 1	-1	+1
Input = 0	0	0

- We use (+1, -1) for weights, (+1, 0) for neurons
- 2 cells with complimentary resistance states are grouped as 1 synapse
- When input is “1”, WL is activated ➔ **current difference between SL and SL_bar**
- When input is “0”, WL is off ➔ no current
- Key idea: the difference of the number of LRS-cells between two SLs in one weight-column is equal to the absolute value of the weighted sum

Current Mode Sense Amplifier and Its Impact on Classification Accuracy

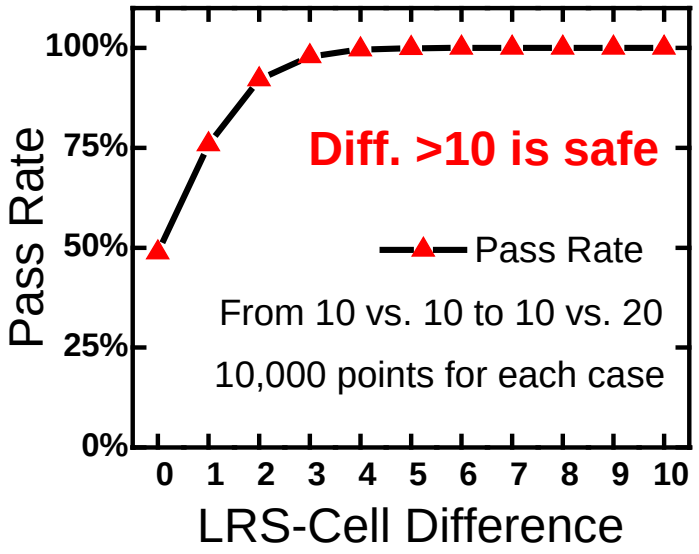






- CSA behaves as the neuron circuit ➔ **Heaviside function**
- CSA offset may cause wrong output when difference between I_{SL} and I_{SL_bar} is too small
- Taken offset into consideration during software simulation, the classification accuracy drops quickly when offset is larger than 2.

Current Mode Sense Amplifier and Its Impact on Classification Accuracy



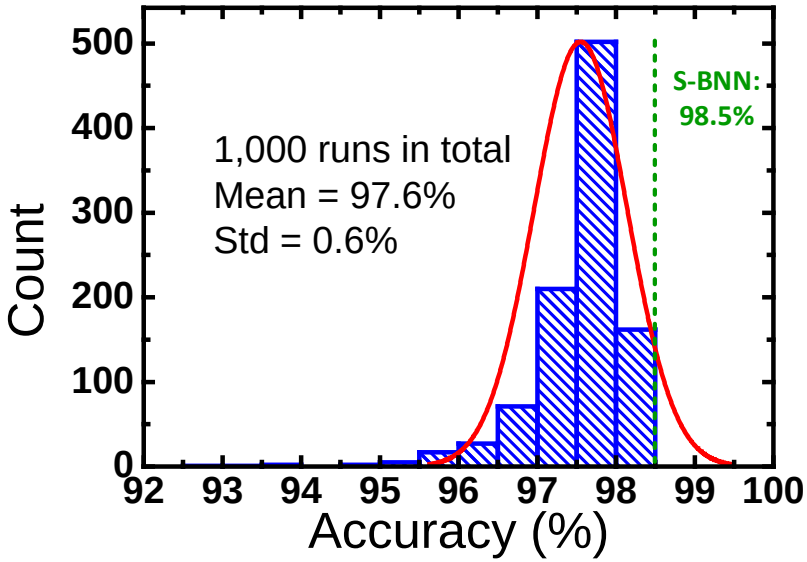
Diff.	# of LRS-cell on SL						
	5	6	7	8	9	10	…
1~3	/	/	/	97.6	96.8	95.8	…
4~6	/	/	100	100	100	100	…
7~9	100	100	100	100	100	100	…

- Instead of the assumption that all the CSAs have a uniform offset pattern, the practical offset pattern is generated from Monte Carlo simulation using TSMC65GP PDK.
- To save time, when the difference is in [1, 3], [4, 6], and [7, 9], we used the offset information of 2, 5, and 8 respectively. 35 sets of MC simulations to cover all the cases.

Current Mode Sense Amplifier and Its Impact on Classification Accuracy

- **Inference with generated offset pattern:** The average accuracy is **~97.6%**, close to the accuracy of S-BNN ~98.5%. The standard deviation is ~0.6%, which means ~68% of runs could achieve >97% (<1.5% loss) and ~95% of runs could achieve >96.4% (<2.1% loss).

- As the CSA used in this work is a common design, the accuracy loss could be effectively reduced by employing a more advanced CSA with offset-cancellation techniques.

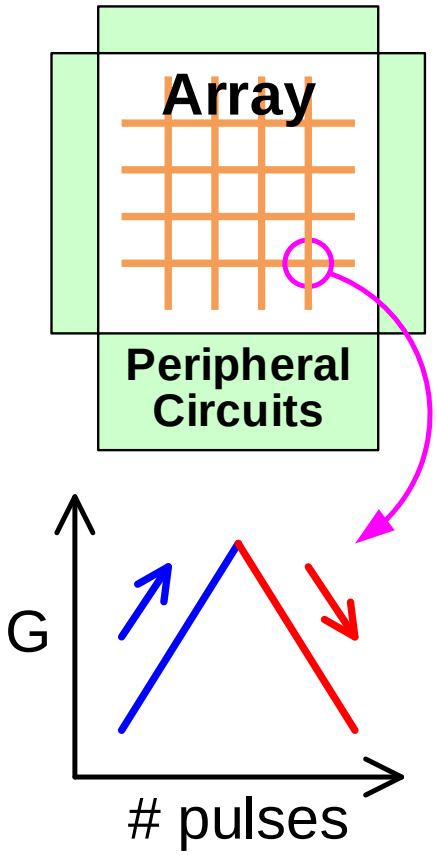


Outline

- **Challenges of On-chip Implementation of Deep Neural Networks and Demand on Binary Neural Networks**
- **RRAM Based Parallel-BNN Accelerator Design**
- **Comparison Between Conventional Serial-BNN and Proposed Parallel-BNN**
- **Summary**

NeuroSim: A Circuit-level Macro Model

- NeuroSim is a circuit-level benchmark simulator developed in C++ that can estimate the area, latency, energy and leakage for neuro-inspired computing with the supporting peripheral circuits to facilitate the design exploration.
- **NeuroSim has flexible design options:**
  - Array architecture type and array size
  - HP/LSTP transistors with technology nodes from 130 nm to 7 nm
  - Memory cell parameters such as eNVM resistance and read/write voltage
  - System parameters such as the activity factor in weighted sum and weight update operation



Comparison on Latency, Energy, Area, and Accuracy

Performance	S-BNN	P-BNN	Improvement
Accuracy	~98.5%	~97.6% (Mean)	**-0.9%**
Area ($\mu m^{2)}$	7913.97	6665.89	**16%**
Latency (ns)	2376.00	10.42	**228X**
Energy (pJ)	710.06	34.48	**20X**
TOPS/W	6.67	137.35	**20X**

- Area overhead is reduced by ~16% even with doubled array size, mainly due to the **elimination of MAC peripheral circuits**. (adder, register, etc.)
- P-BNN could achieve a high energy efficiency of **137.35TOPS/W**, improved by **20X** compared to conventional S-BNN, with only **~0.9%** accuracy degradation.

Outline

- **Challenges of On-chip Implementation of Deep Neural Networks and Demand on Binary Neural Networks**
- **RRAM Based Parallel-BNN Accelerator Design**
- **Comparison Between Conventional Serial-BNN and Proposed Parallel-BNN**
- **Summary**

Summary

- Binary neural network is a promising solution for on-chip implementation of DNNs due to significant reduction in memory size and computation load.
- A RRAM based parallel-BNN hardware accelerator is proposed, aiming to achieve a better energy-efficiency than conventional approach.
- The impact of CSA offset on classification accuracy is well analyzed through Monte Carlo simulation using TSMC65 PDK.
- The proposed architecture could achieve 137.35TOPS/W, improved by 20X, with only 0.9% accuracy degradation.
- Next: scalability on larger networks and datasets

Acknowledgement