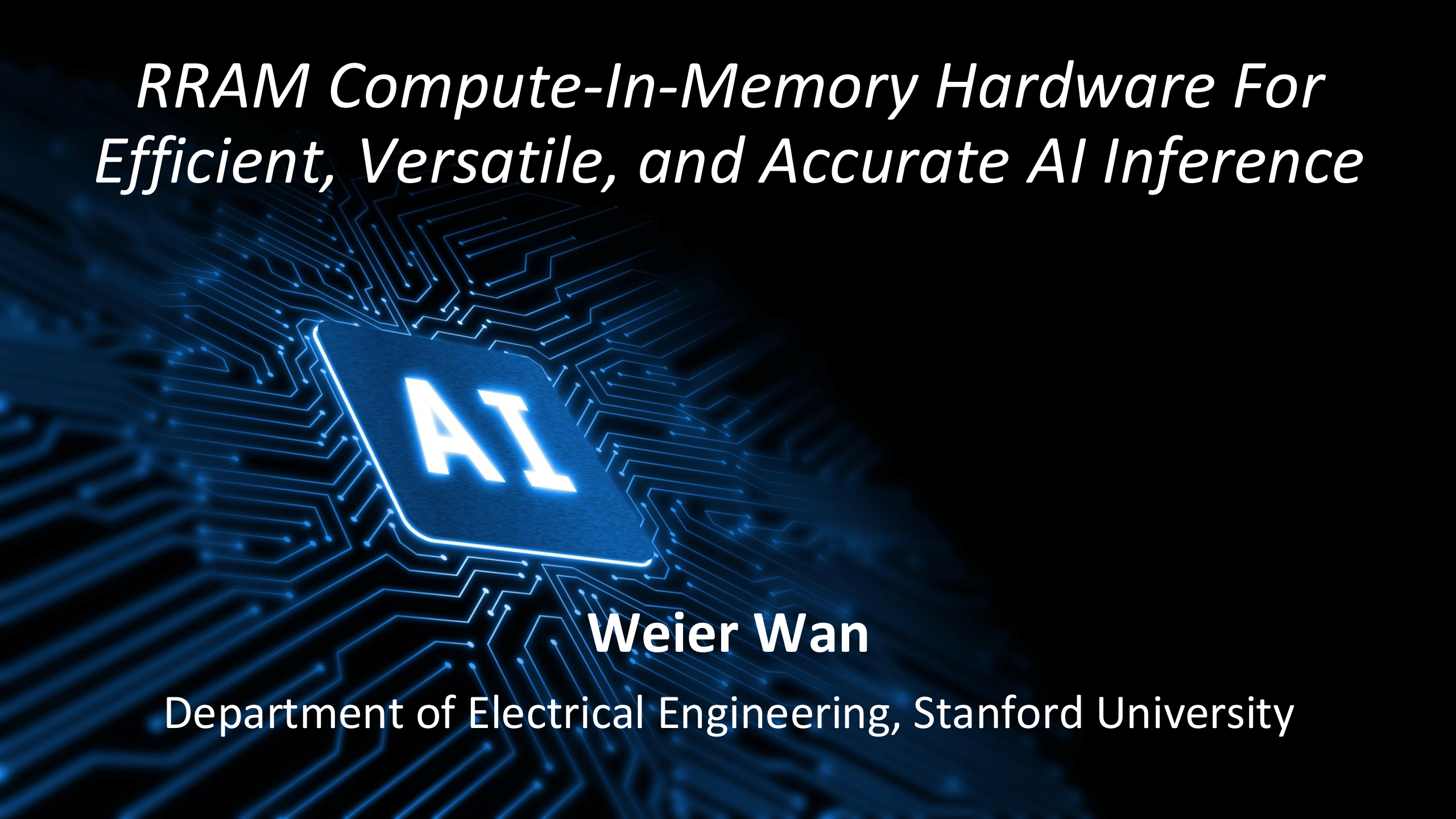
RRAM Compute-In-Memory Hardware For Efficient, Versatile, and Accurate AI Inference

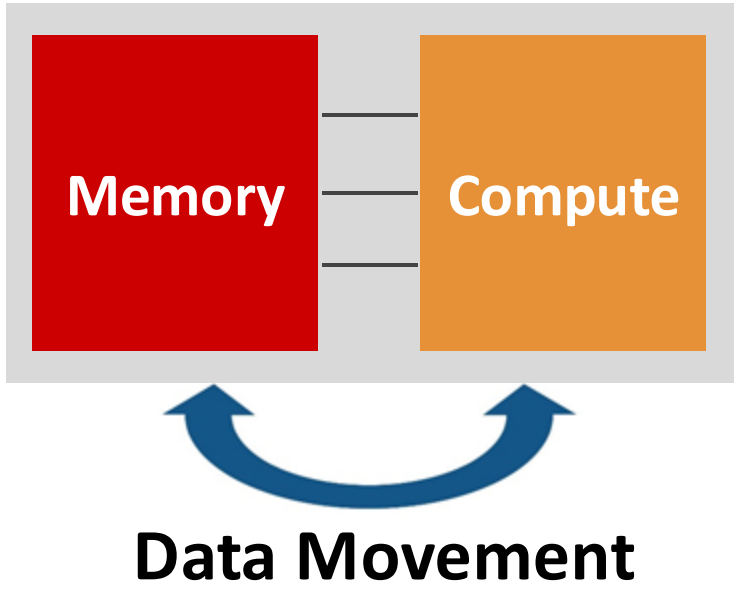


Weier Wan

Department of Electrical Engineering, Stanford University

TOWARDS LESS DATA MOVEMENT

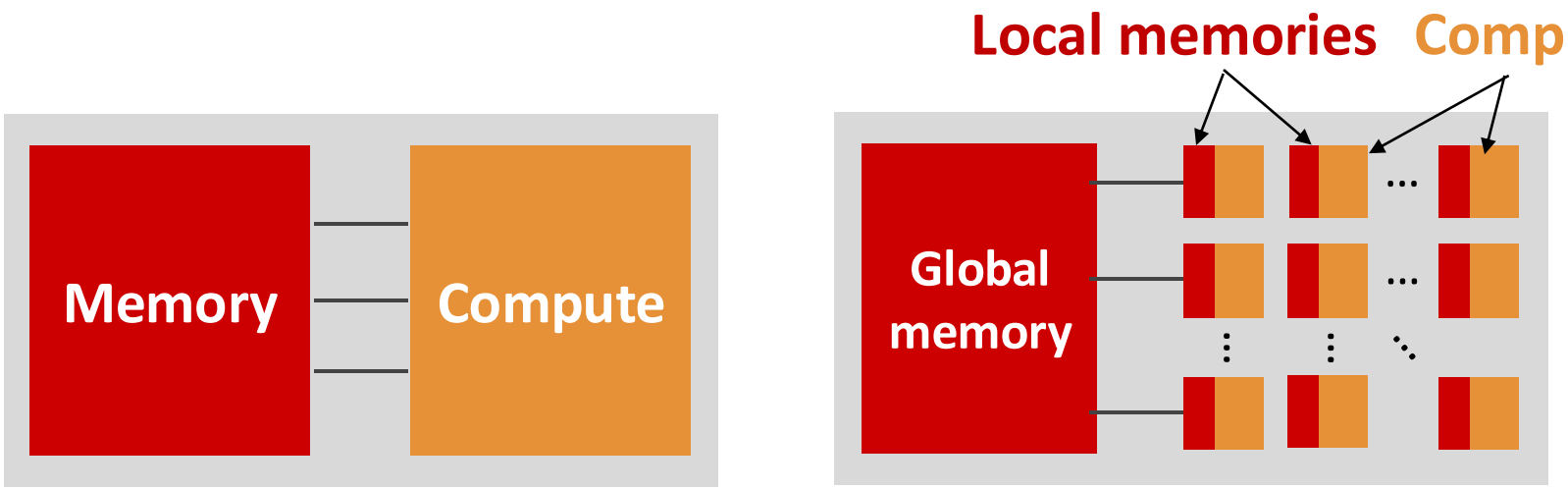
SEPARATE MEMORY AND COMPUTE UNITS



Less Data Movement

TOWARDS LESS DATA MOVEMENT

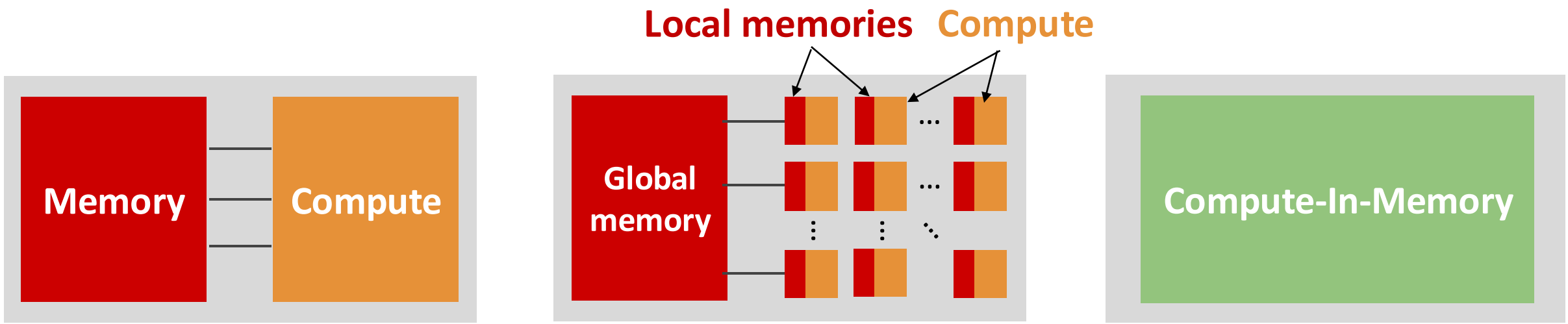
FINE-GRAINED, DISTRIBUTED MEMORY HIERARCHY



Less Data Movement

TOWARDS LESS DATA MOVEMENT

COMPUTE-IN-MEMORY (CIM)



Data processed at where they are stored

Less Data Movement

TOWARDS LESS DATA MOVEMENT

COMPUTE-IN-MEMORY (CIM)

What it takes to get there?

- *High-density memory offering enough capacity for model weights*
- *Memory and logic devices integrated monolithically on a single chip*
- *Support energy-efficient in-memory computation*

RRAM BASICS

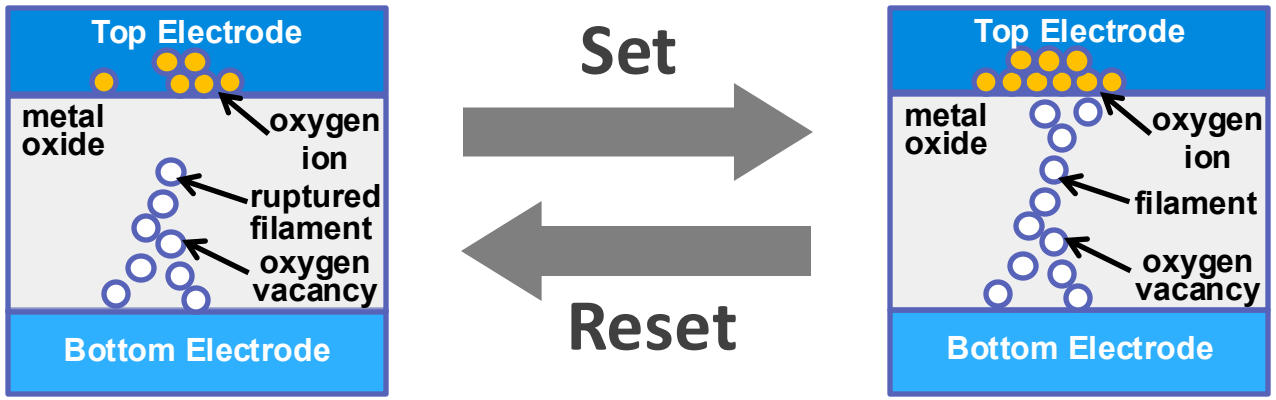
Resistive Switching Random-Access Memory

“0”

High resistance

“1”

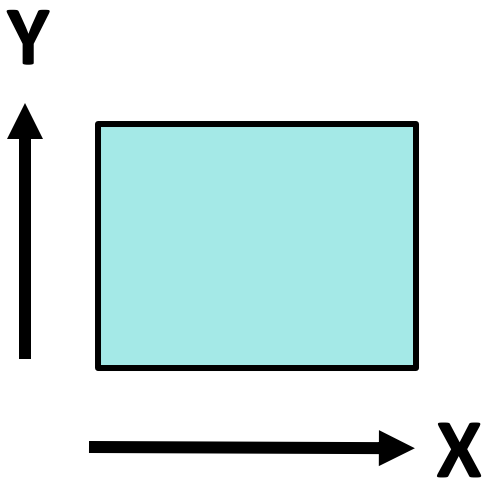
Low resistance

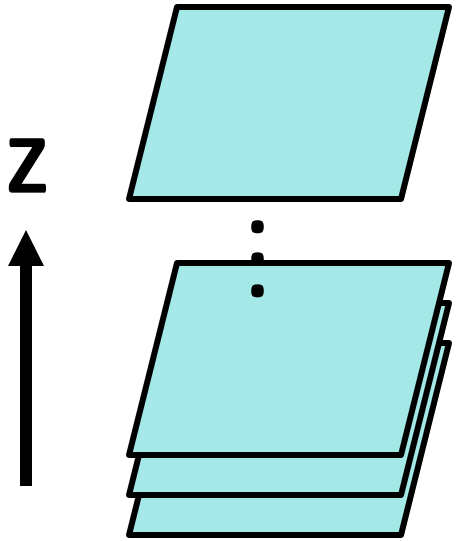


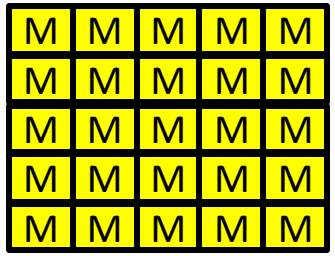
H.-S. P. Wong, S. Salahuddin, *Nature Nanotech.*, 2015

LARGE MEMORY CAPACITY:

X, Y, Z, M





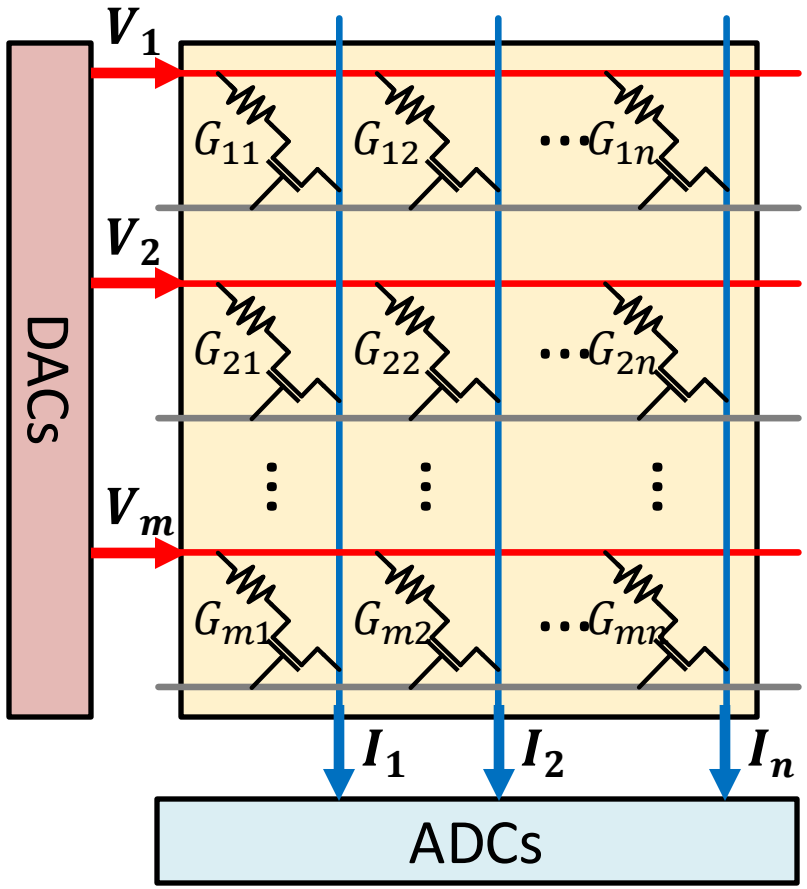


- Two dimensional down scaling
- 3D layers
- Store multiple bits per memory cell

Inspired by Jian Chen (Western Digital), *IEDM Short Course*, 2019

COMPUTE-IN-MEMORY (CIM) USING RRAM

INTERNALLY ANALOG, EXTERNALLY DIGITAL



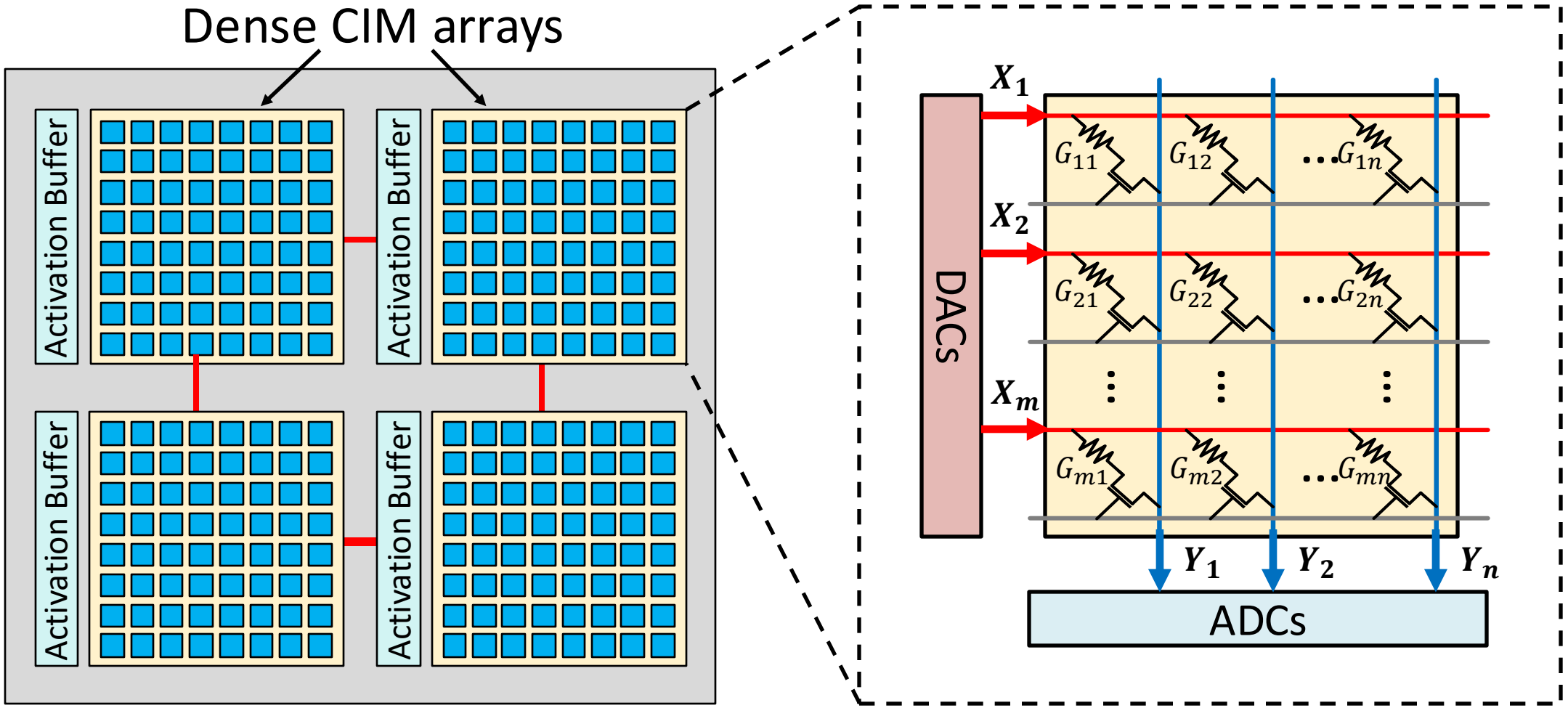
Kirchhoff's law matrix-vector multiplication

$$(V_1 \quad \cdots \quad V_m) \begin{pmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \ddots & \vdots \\ G_{m1} & \cdots & G_{mn} \end{pmatrix} = (I_1 \quad \cdots \quad I_n)$$

→ Basis of AI models' inference and training

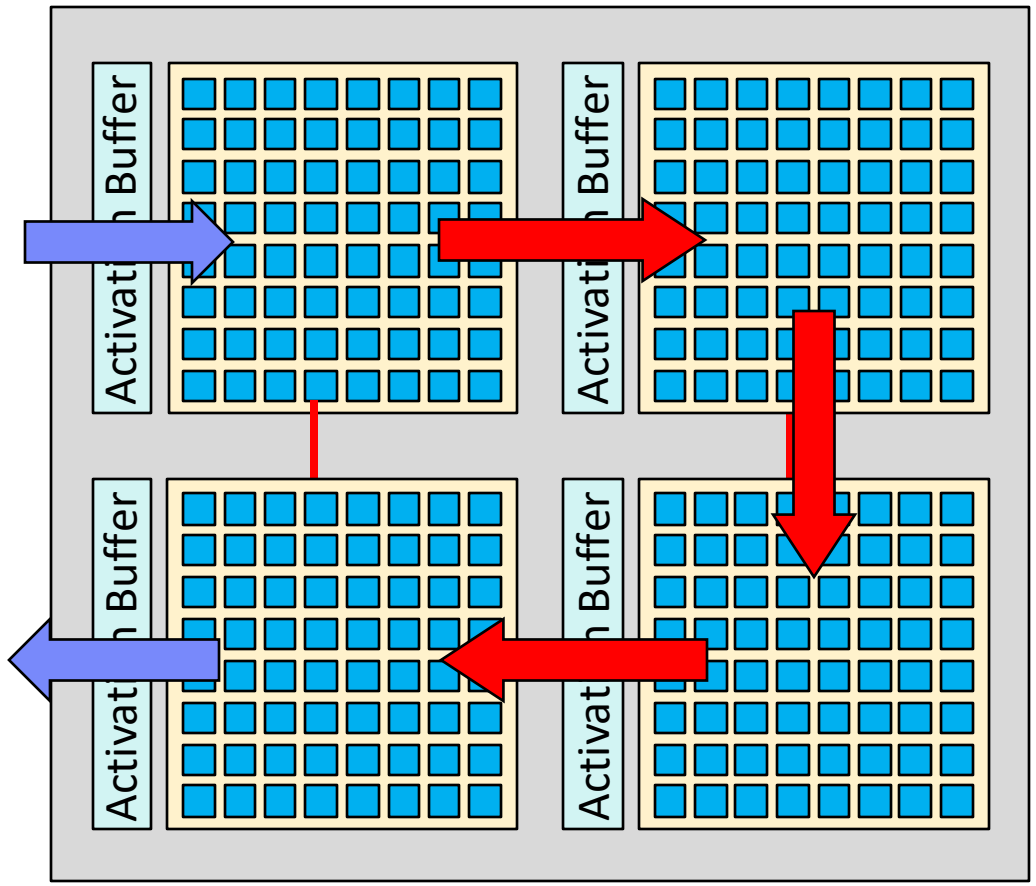
WEIGHT-STATIONARY AI INFERENCE

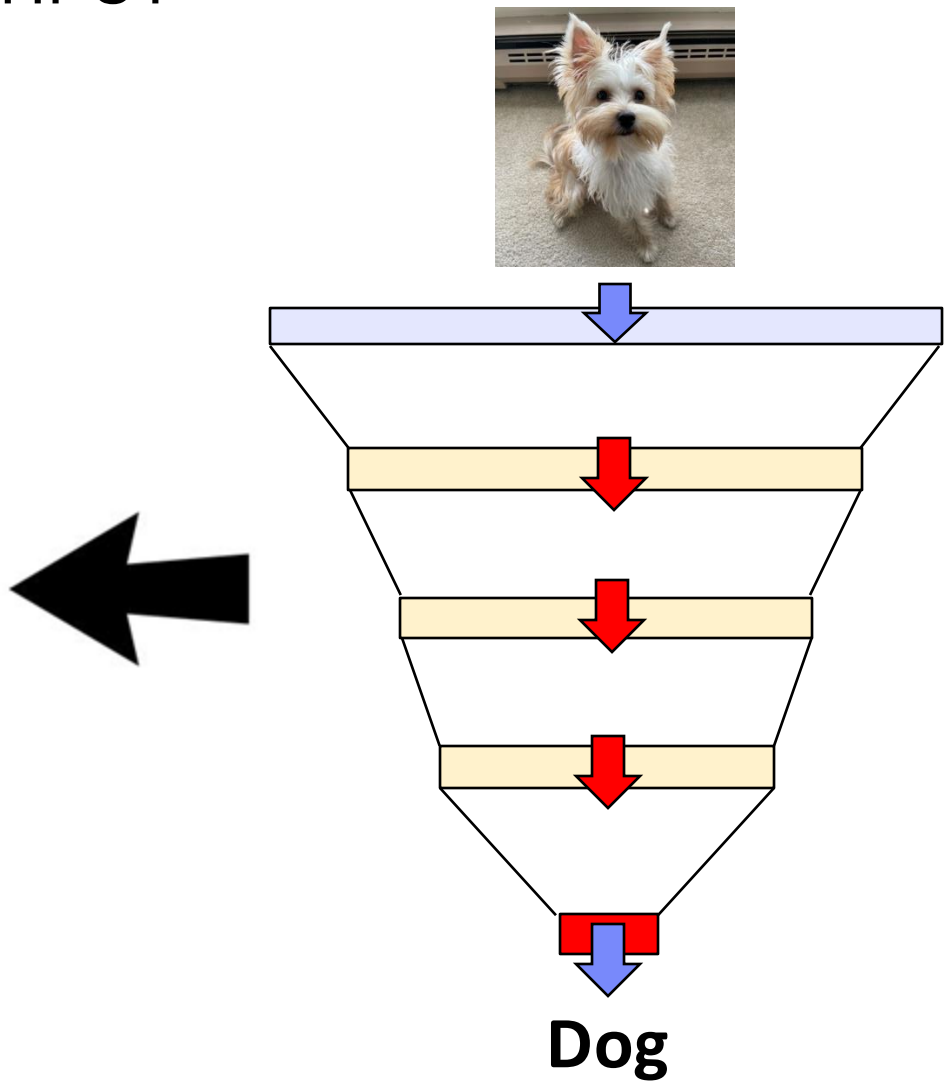
ELIMINATES WEIGHT MOVEMENT



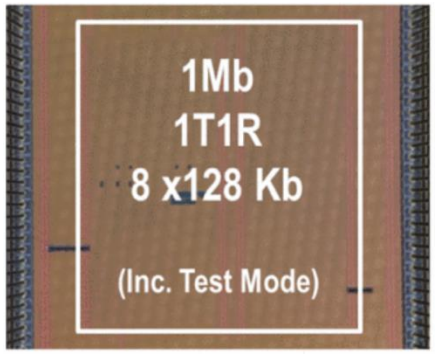
MULTI-CORE PIPELINED DNN INFERENCE

DENSE COMPUTE, HIGH THROUGHPUT

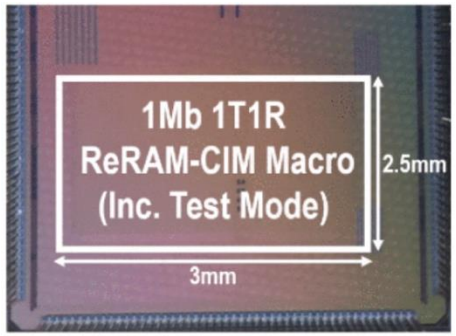




LOTS OF PROGRESS, BUT…

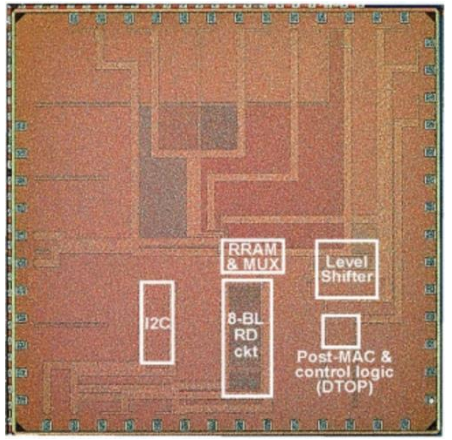


W.-H. Chen *et al.*, ISSCC 18’



C.-X. Xue *et al.*, ISSCC 19’

Q. Liu *et al.*, ISSCC 20’



J.-H. Yoon *et al.*, ISSCC 21’

…

- Few work simultaneously addressed ***efficiency***, ***versatility***, and ***accuracy***
- Results were ***partially simulated*** (rather than fully measured)
- Demonstrated tasks lacked ***complexity*** and ***diversity***

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

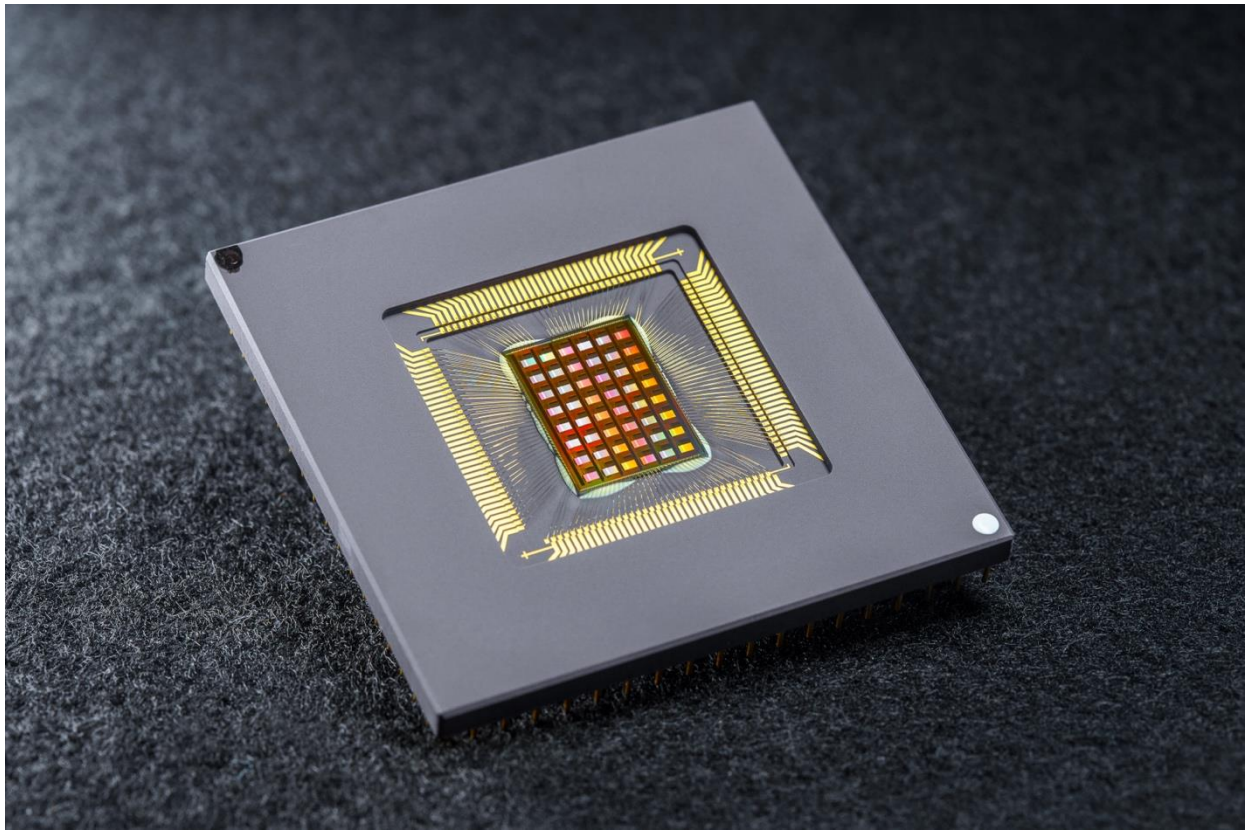
nature › articles › article

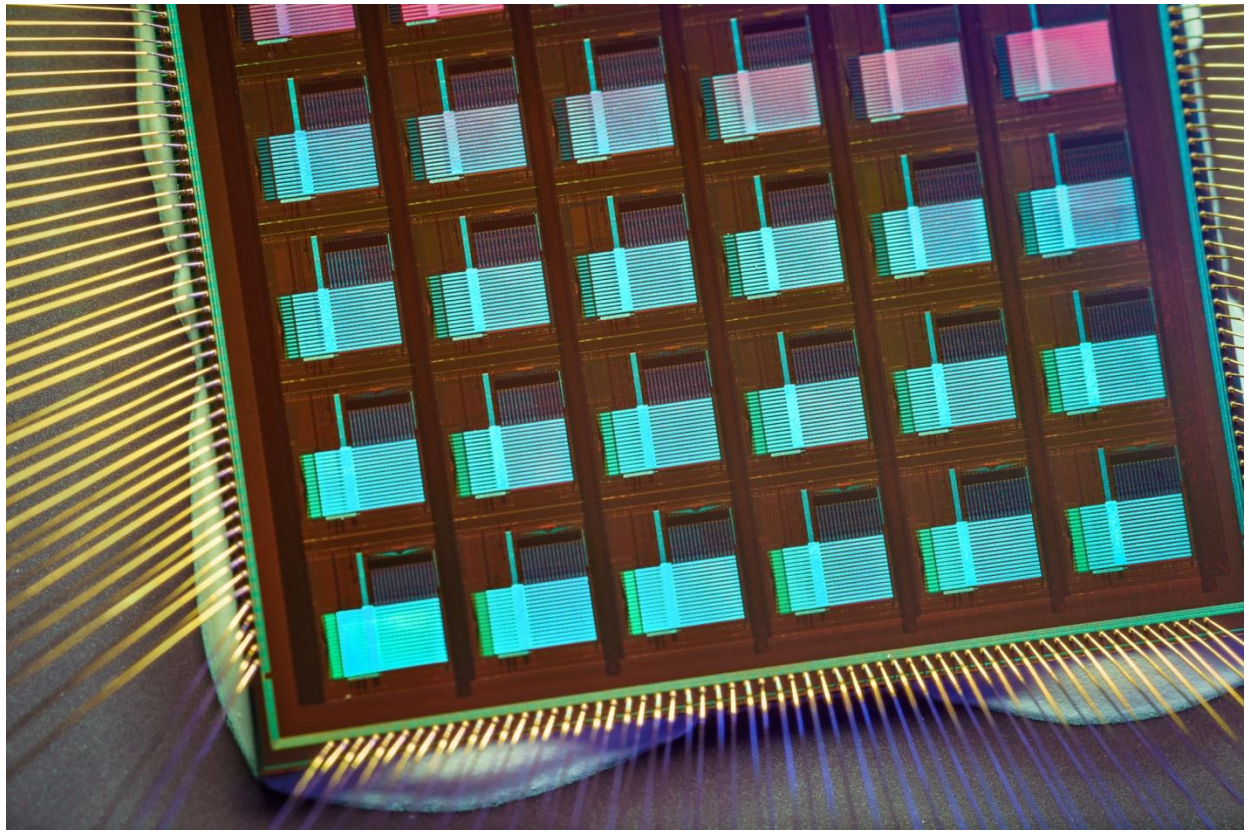
Article | Open Access | Published: 17 August 2022

A compute-in-memory chip based on resistive random-access memory

Weier Wan ✉, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao ✉, Siddharth Joshi ✉, Huaqiang Wu ✉, H.-S. Philip Wong ✉ & Gert Cauwenberghs ✉

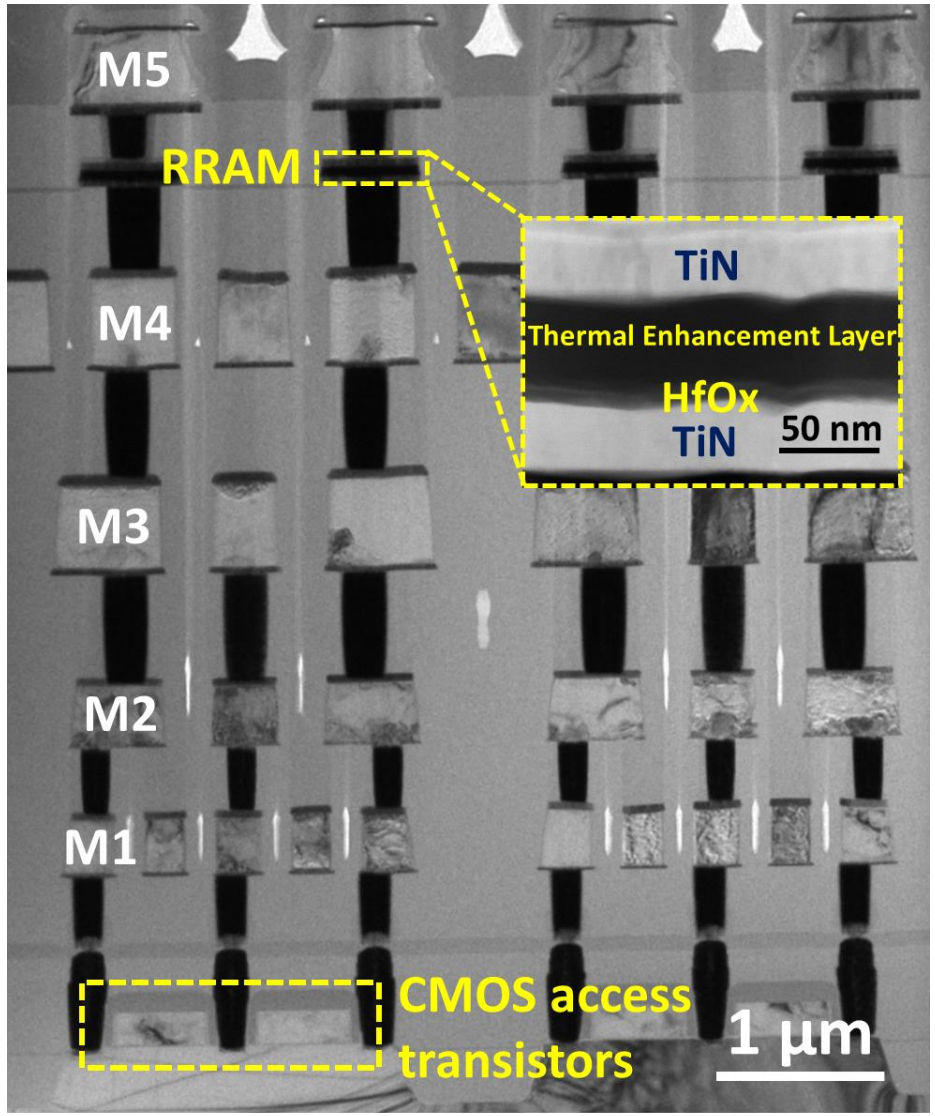
NEURRAM: 48-CORE RRAM-CIM CHIP





Photos by David Baillot/University of California San Diego

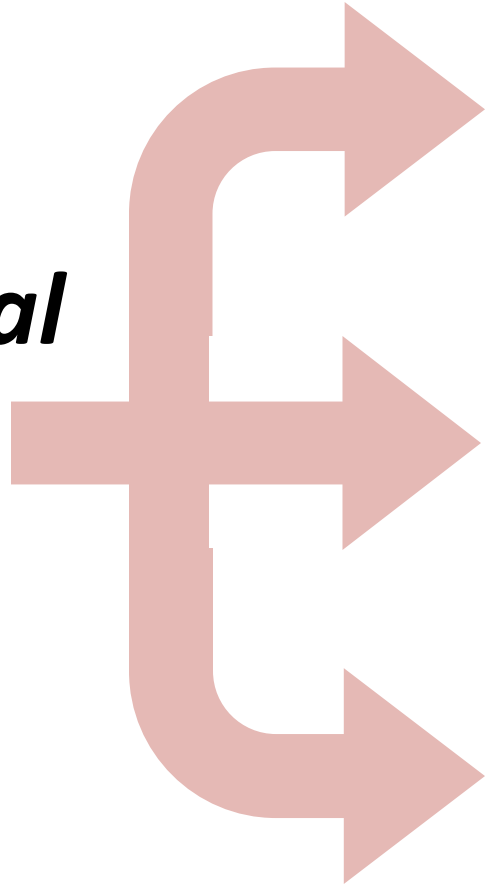
RRAM-CMOS MONOLITHIC INTEGRATION



- CMOS & M1-M4 fabricated using a commercial foundry 130 nm process
- RRAM & M5 integrated using a research lab process

CHALLENGES FOR RRAM-CIM CHIP DESIGN

Fundamental Design Trade-off



ENERGY-EFFICIENCY

Limited by peripheral circuits (e.g. ADCs)

RECONFIGURABILITY

Re-programming weights is expensive

INFERENCE ACCURACY

Analog computation susceptible to device & circuit imperfections

FULL-STACK CO-DESIGN

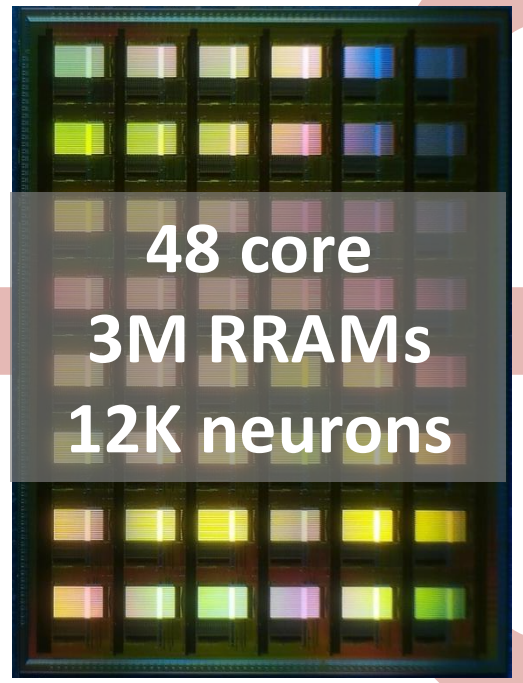
Algorithm

System

Architecture

Circuit

Technology



W. Wan *et al.*, *Nature* 2022

ENERGY-EFFICIENCY

Highest among RRAM CIM chips

RECONFIGURABILITY

CNN, MLP, LSTM, RBM, etc.

INFERENCE ACCURACY

Comparable to software model with 4-bit weights

ENERGY-EFFICIENCY

Open-circuit voltage-mode sensing scheme

RECONFIGURABILITY

Dataflow reconfigurability realized by Transposable Neurosynaptic Array

ACCURACY

Non-ideality-aware DNN training & fine-tuning

ENERGY-EFFICIENCY

Open-circuit voltage-mode sensing scheme

RECONFIGURABILITY

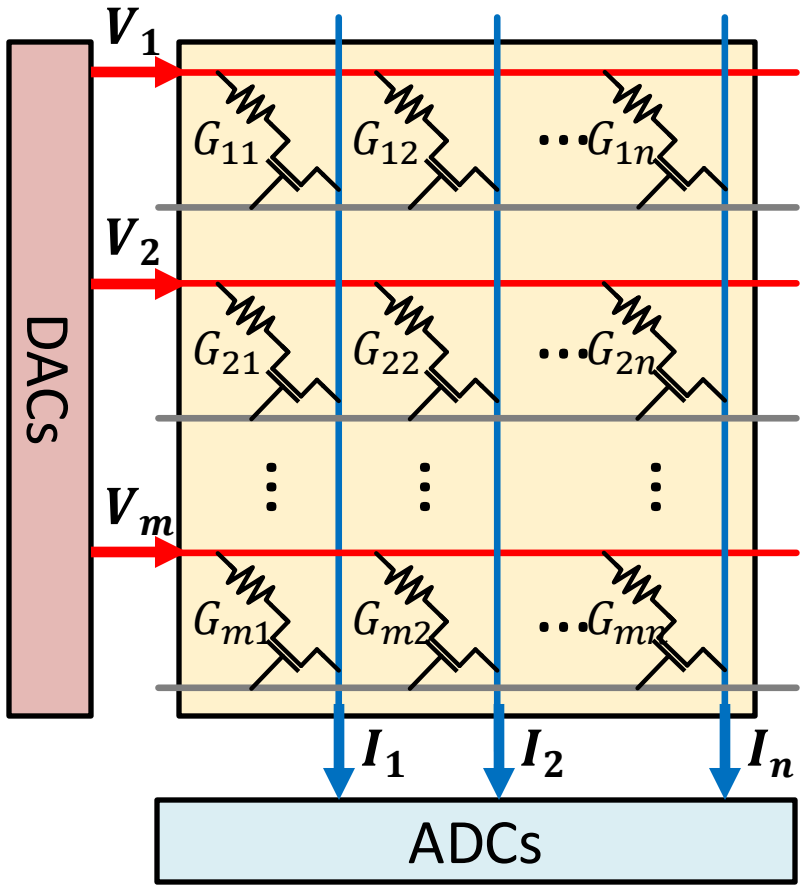
Dataflow reconfigurability realized by Transposable Neurosynaptic Array

ACCURACY

Non-ideality-aware DNN training & fine-tuning

OUTPUT SENSING SCHEME

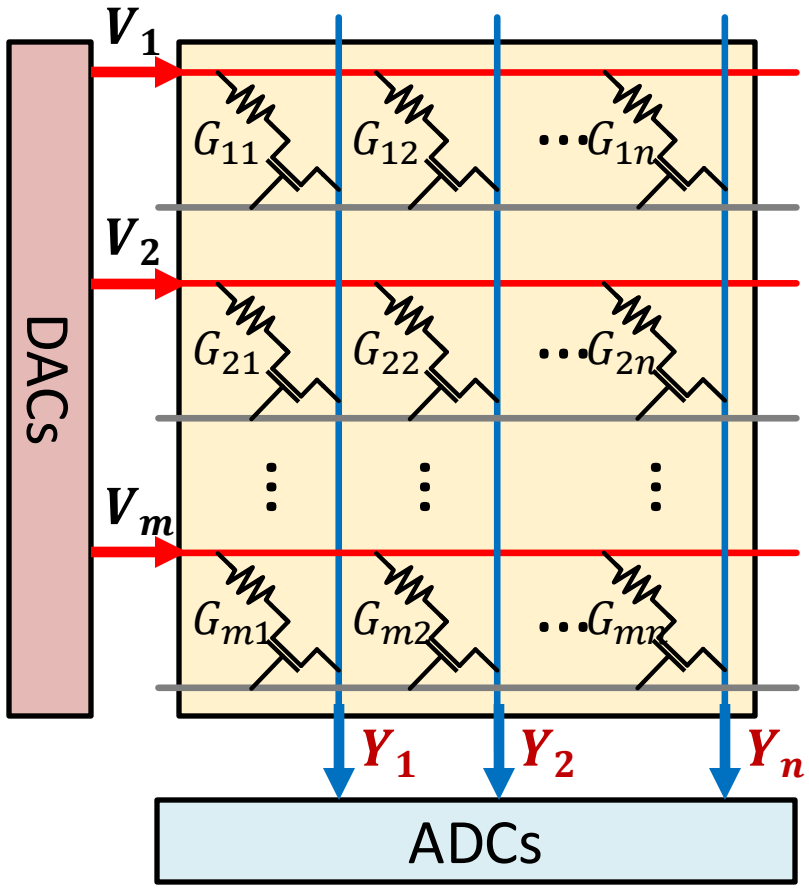
KEY TO ENERGY-EFFICIENCY



$$(V_1 \quad \cdots \quad V_m) \begin{pmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \ddots & \vdots \\ G_{m1} & \cdots & G_{mn} \end{pmatrix} = (I_1 \quad \cdots \quad I_n)$$

OUTPUT SENSING SCHEME

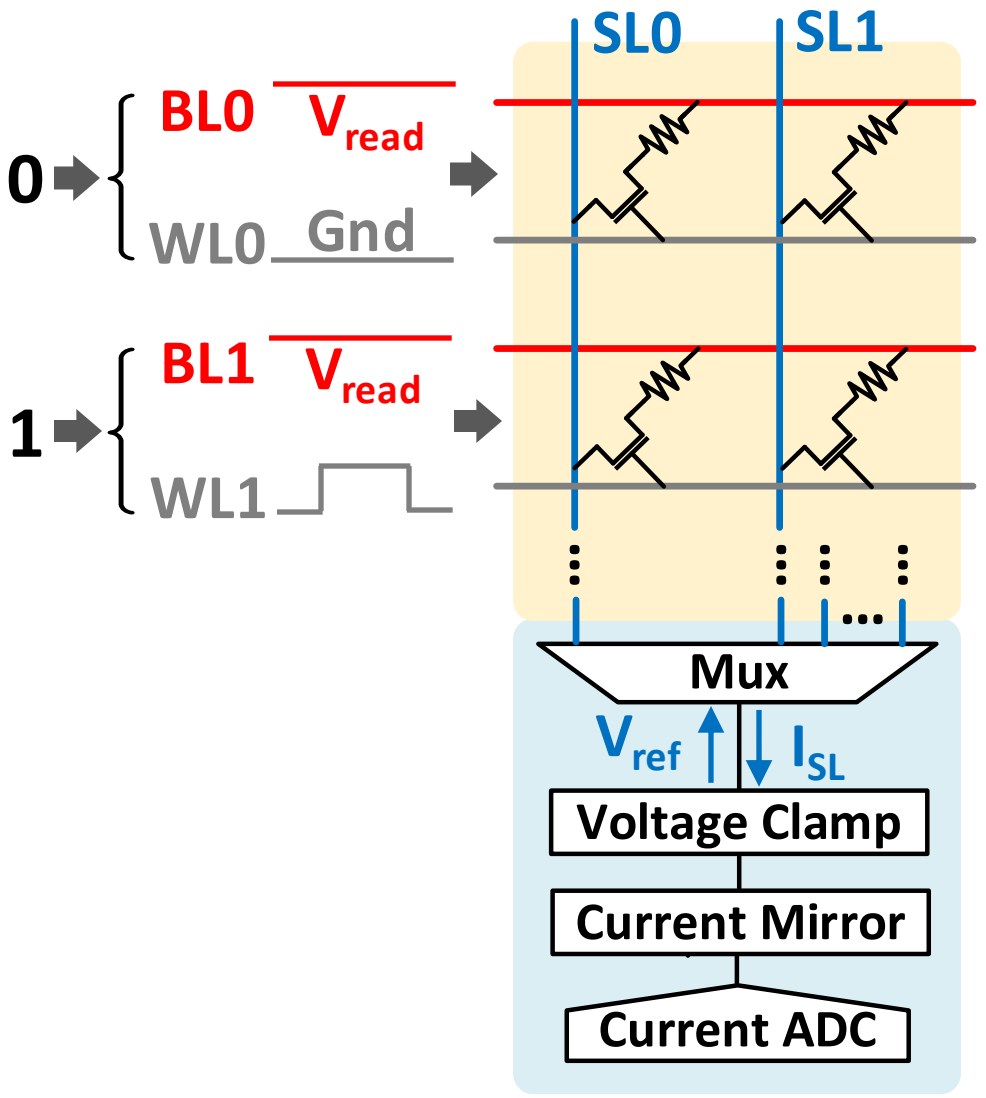
KEY TO ENERGY-EFFICIENCY



$$(V_1 \quad \cdots \quad V_m) \begin{pmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \ddots & \vdots \\ G_{m1} & \cdots & G_{mn} \end{pmatrix} = (Y_1 \quad \cdots \quad Y_n)$$

Y_i can be current or voltage depending on the **sensing scheme**

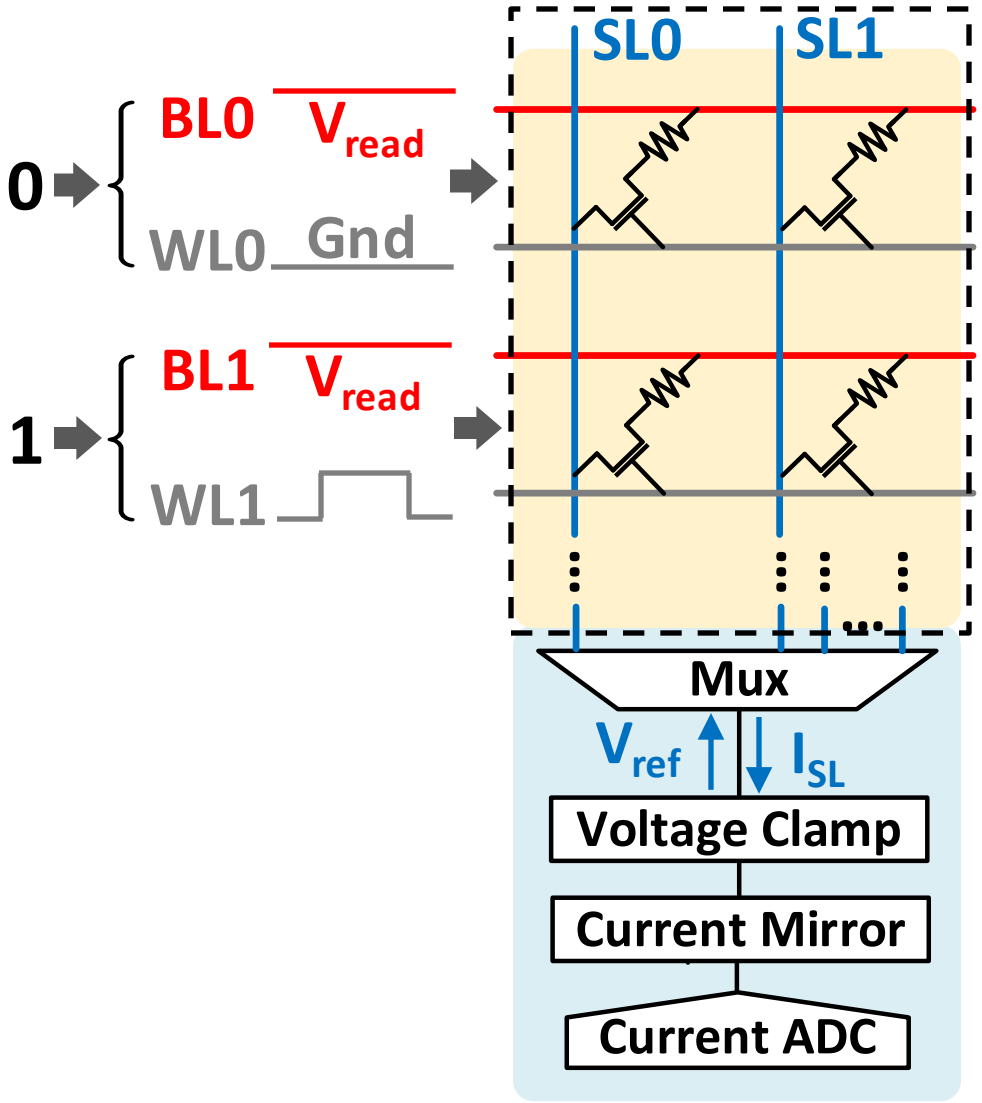
CONVENTIONAL CURRENT-MODE SENSING



- Drive all BLs to V_{read}
- Clamp SLs to V_{ref}
- Activate WLs with "1" input
- Sense SL current as output of MVM

$$I_{SL} = (V_{read} - V_{ref}) \sum_i X_i G_i$$

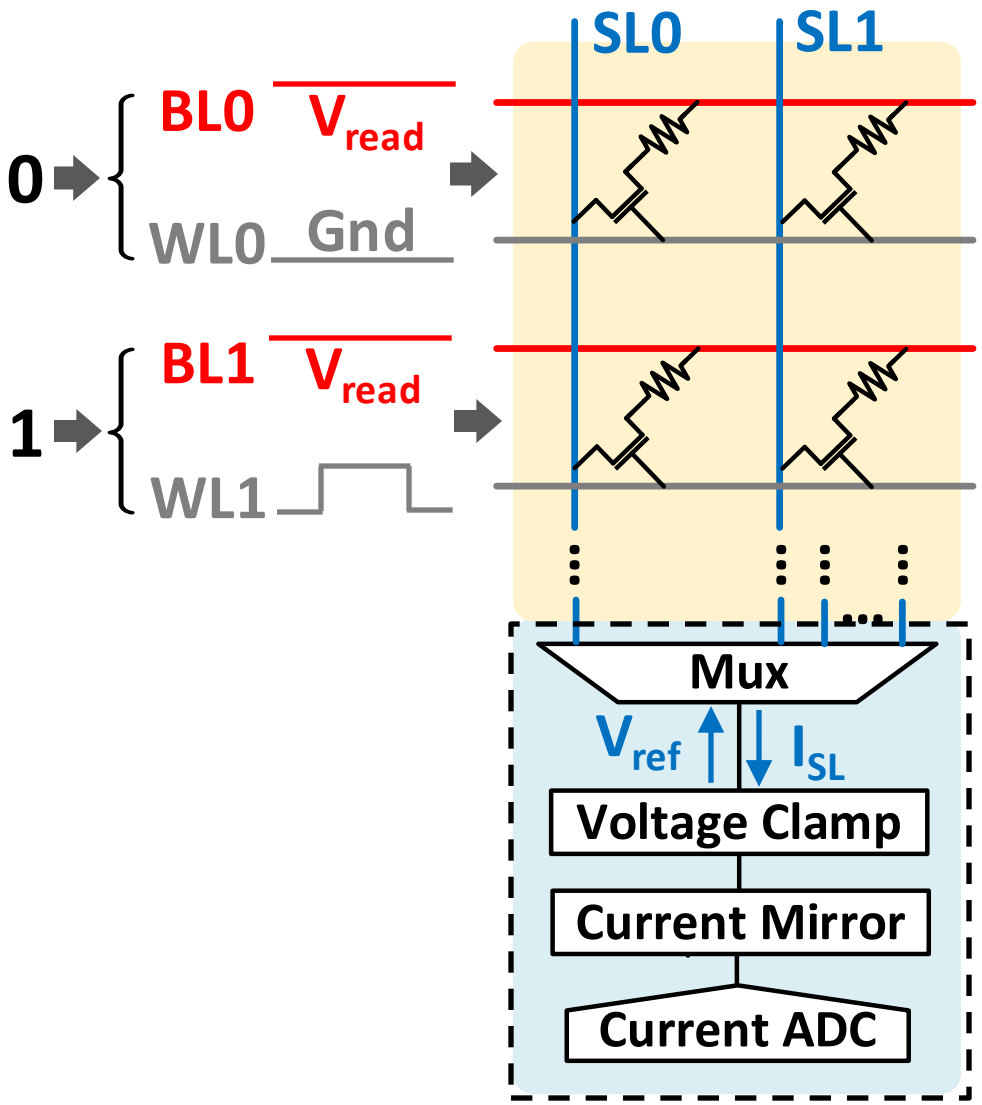
CONVENTIONAL CURRENT-MODE SENSING



- Current flows in array during A/D conversion
- ADC: latency grows with precision
- NN models require > 4-bit activation precision

→ ***High energy consumption***

CONVENTIONAL CURRENT-MODE SENSING



- Activate many rows in a cycle ("row-parallelism")

  → Large I_{SL}

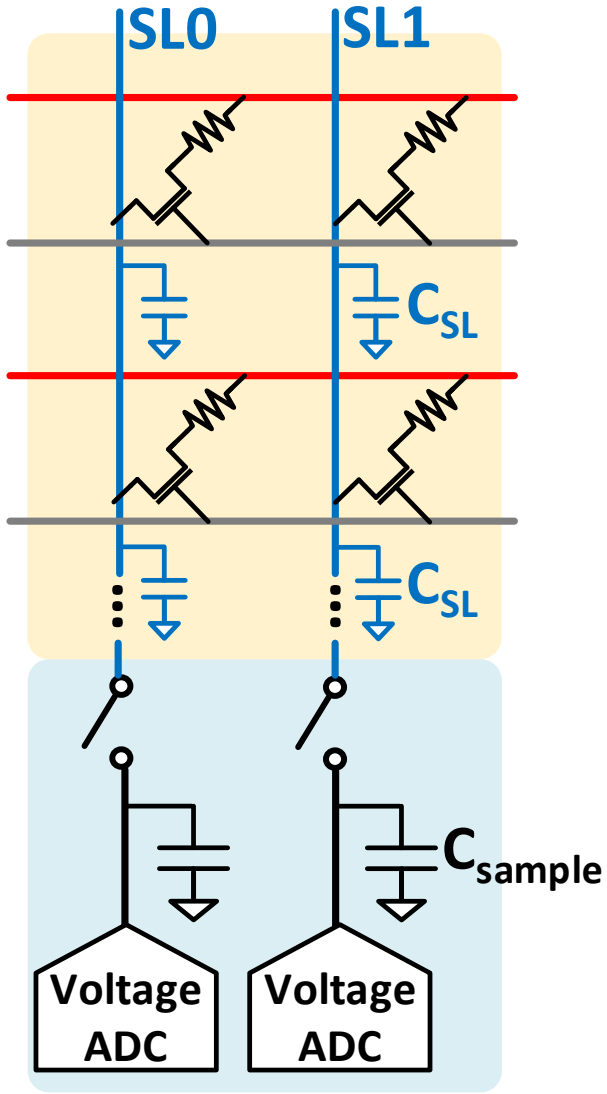
  → Peripheral circuits need **large transistors**

- To pitch match with dense RRAM array, peripherals time-multiplexed by many columns

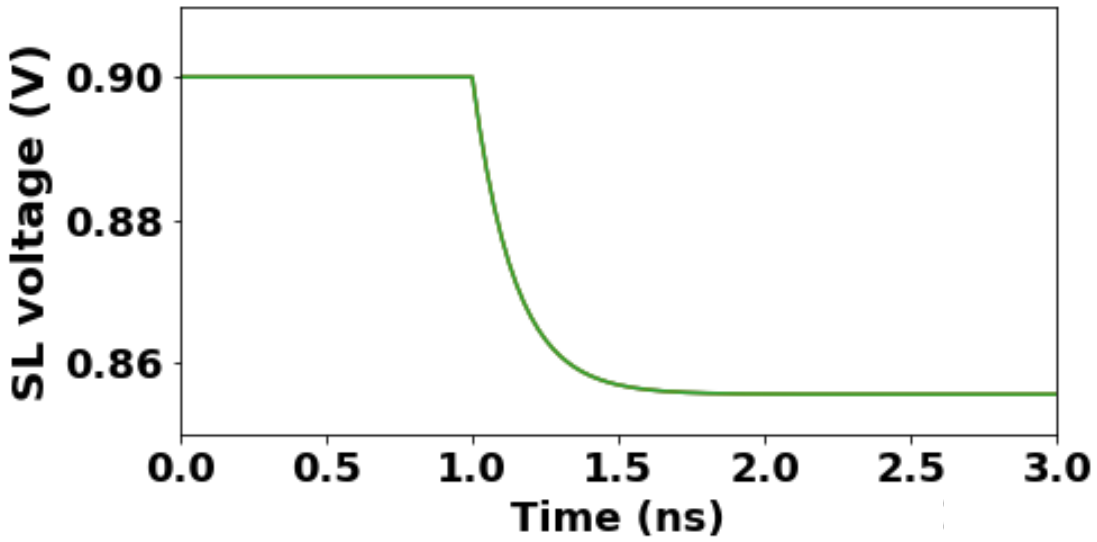
  → limits "column-parallelism"

→ ***Limited throughput***

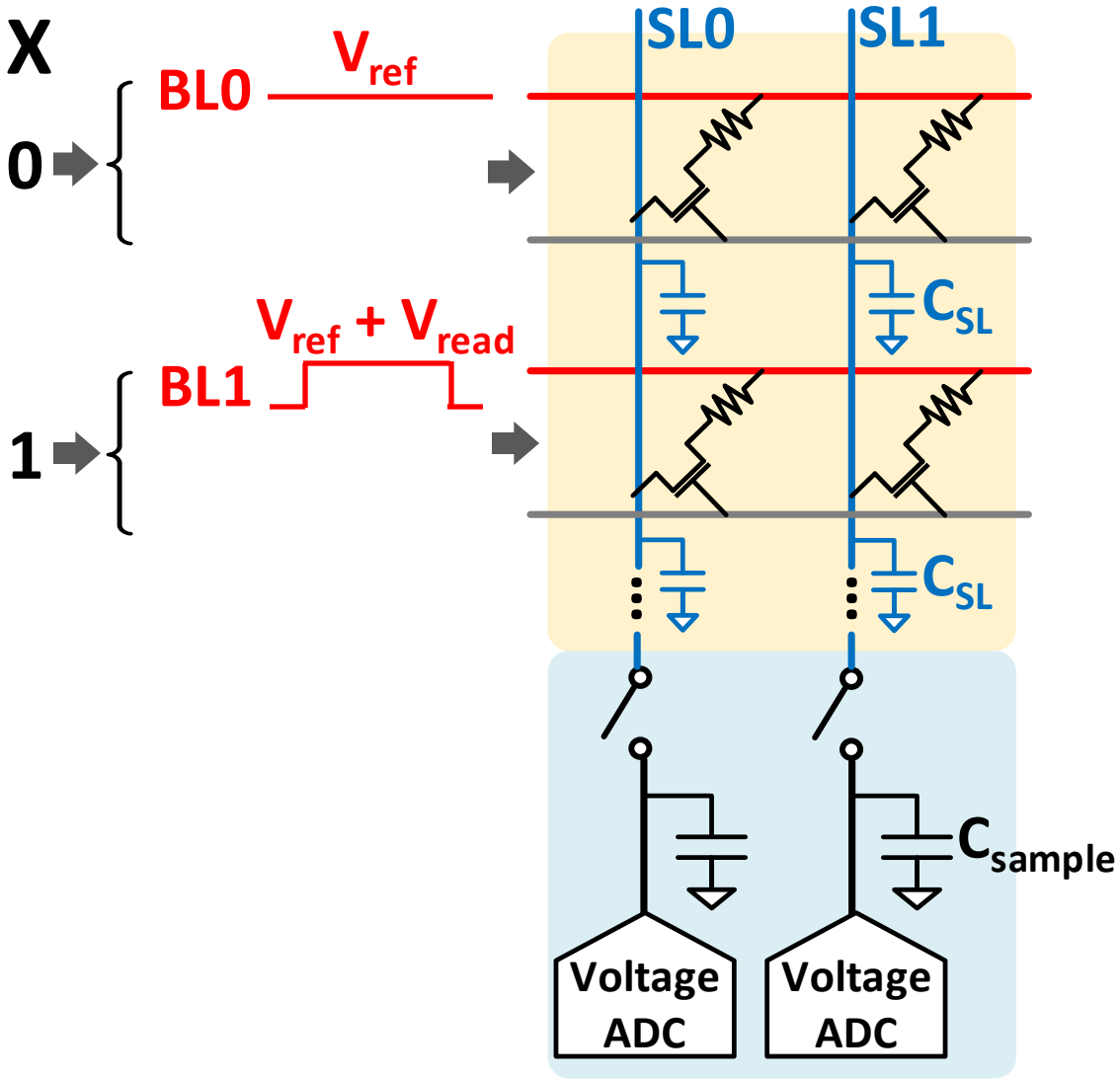
OPEN-CIRCUIT VOLTAGE-MODE SENSING



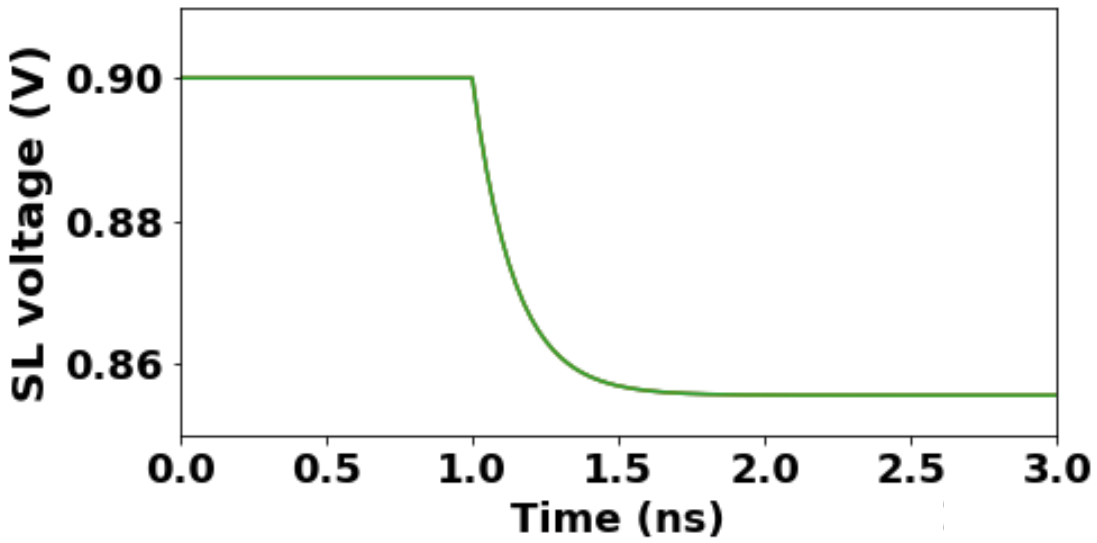
- Pre-charge SLs to V_{ref}



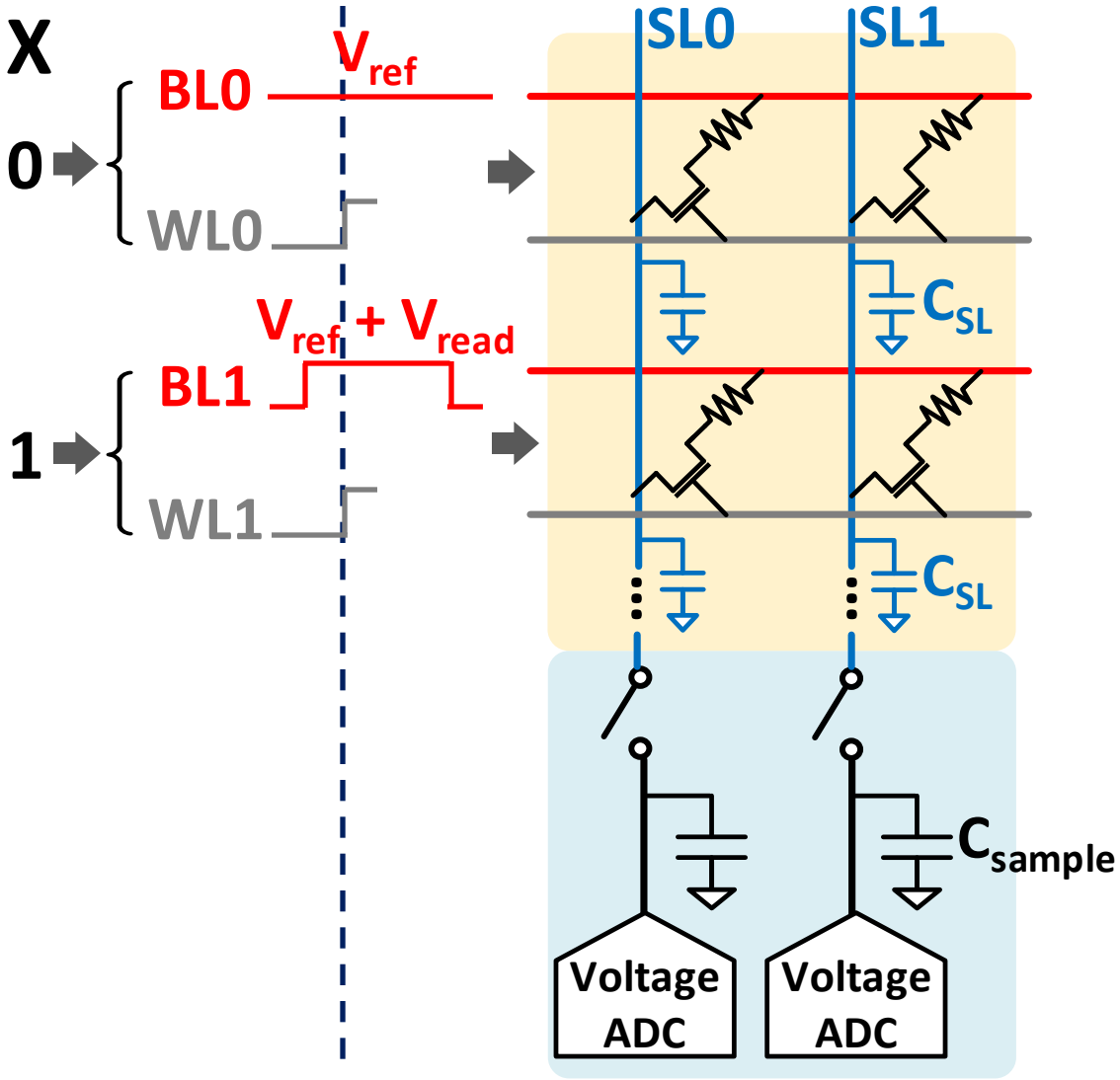
OPEN-CIRCUIT VOLTAGE-MODE SENSING



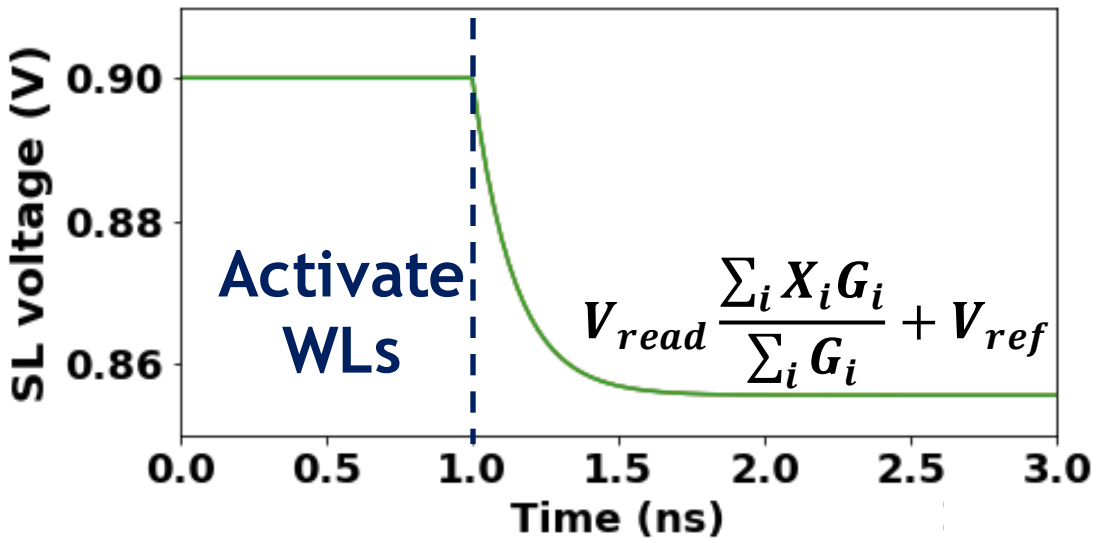
- Pre-charge SLs to V_{ref}
- Drive BLs to $V_{ref} + X_i \times V_{read}$



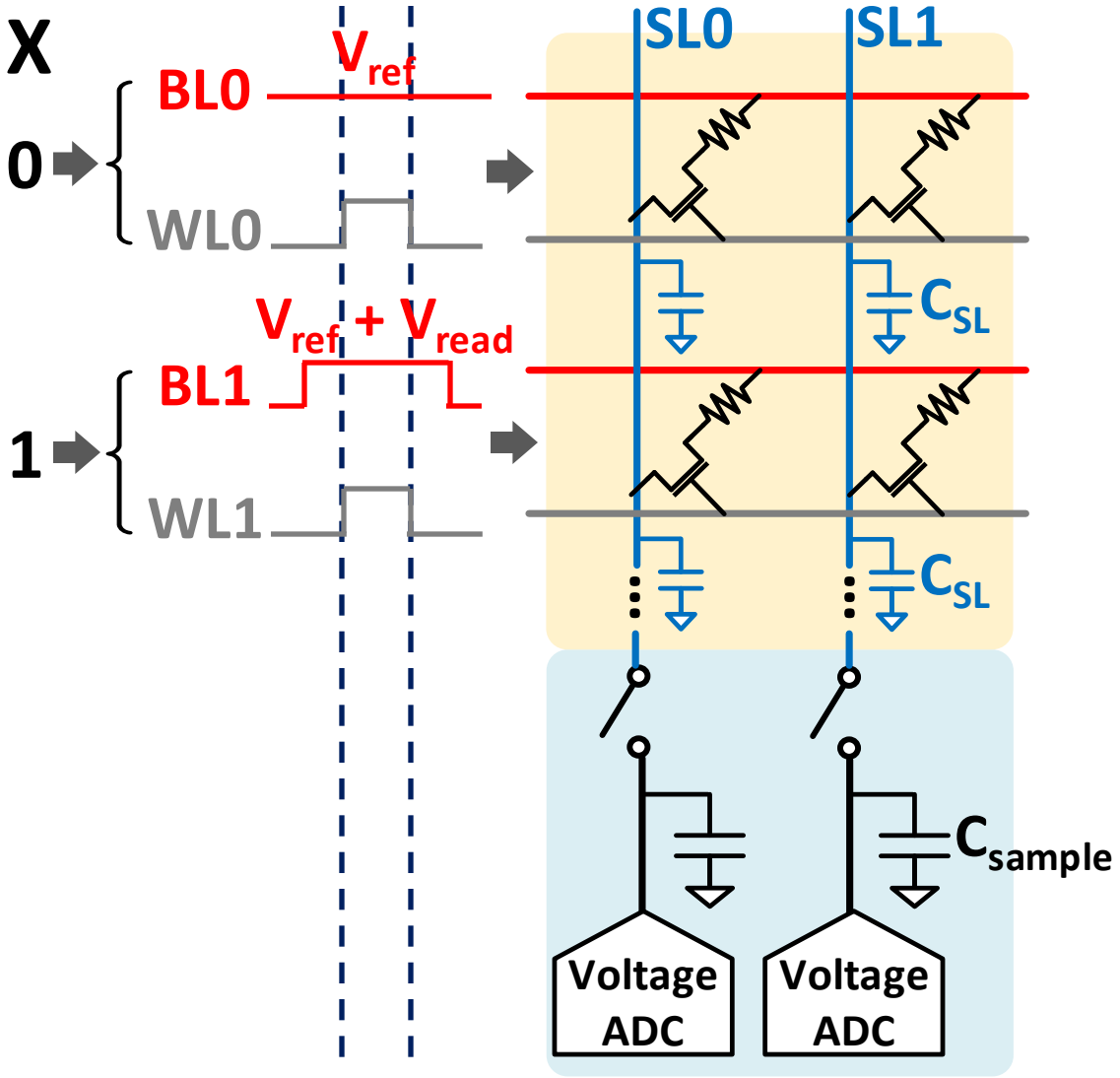
OPEN-CIRCUIT VOLTAGE-MODE SENSING



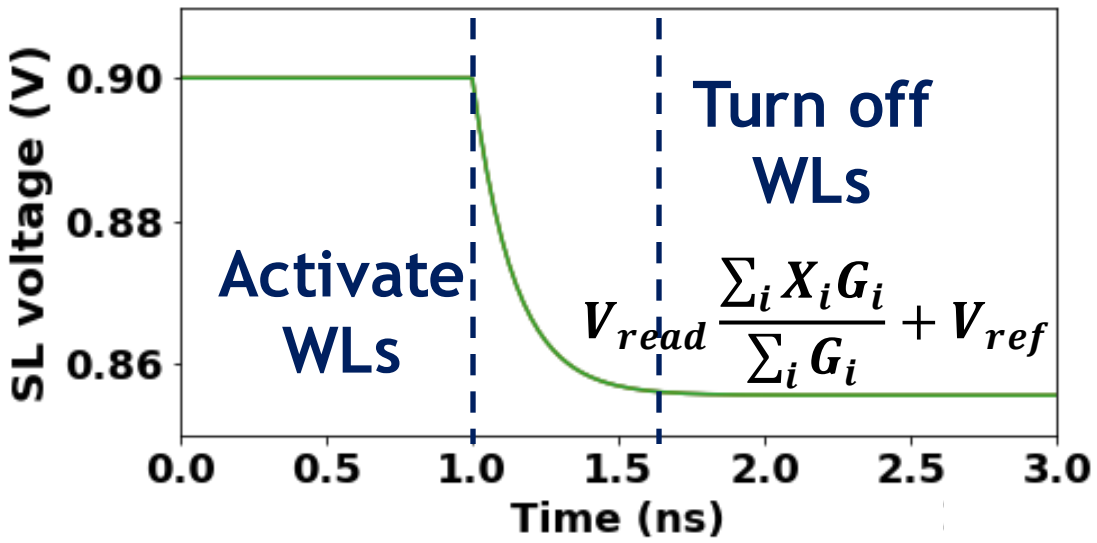
- Pre-charge SLs to V_{ref}
- Drive BLs to $V_{ref} + X_i \times V_{read}$
- Activate all WLs and keep SLs floating



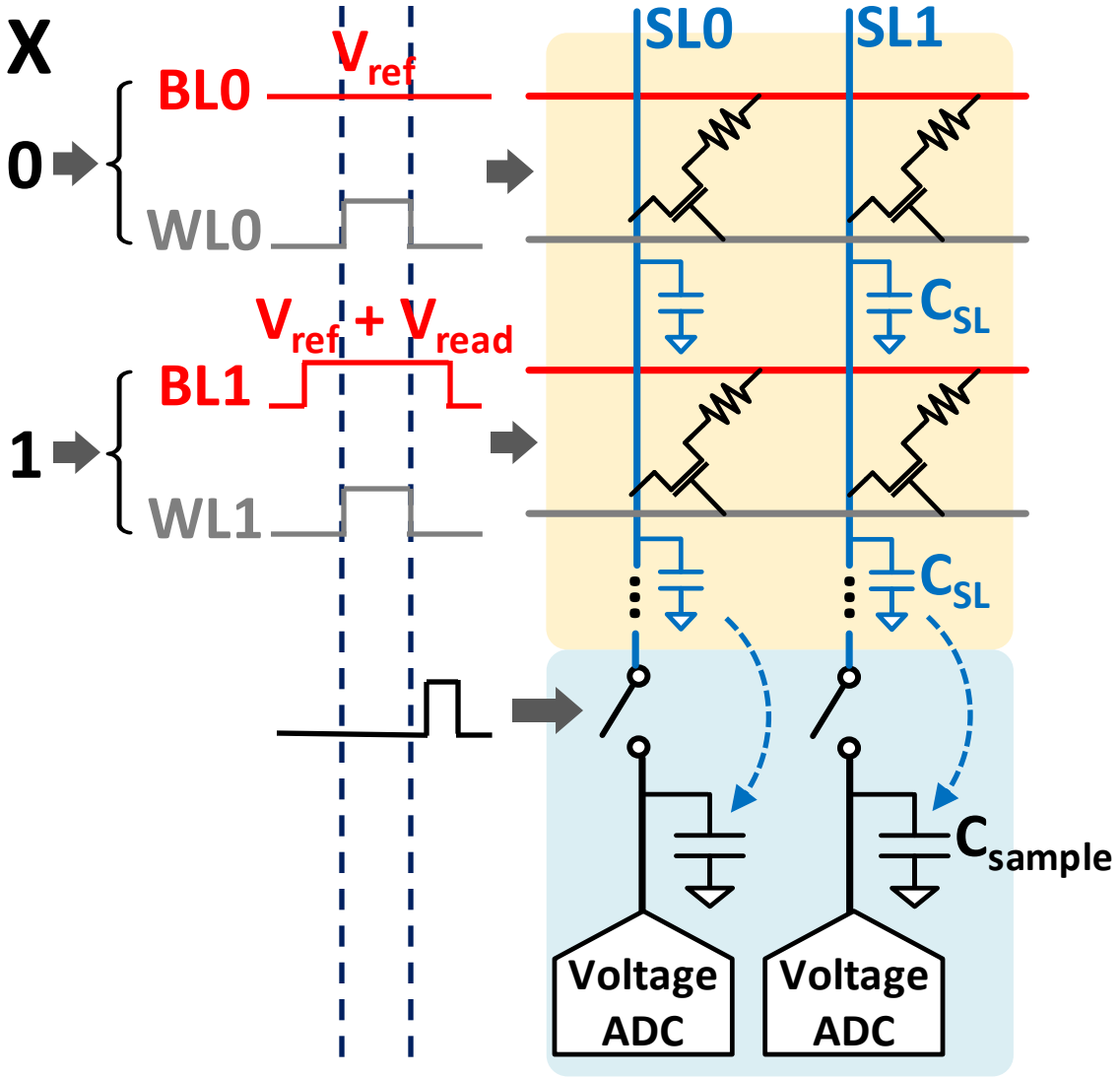
OPEN-CIRCUIT VOLTAGE-MODE SENSING



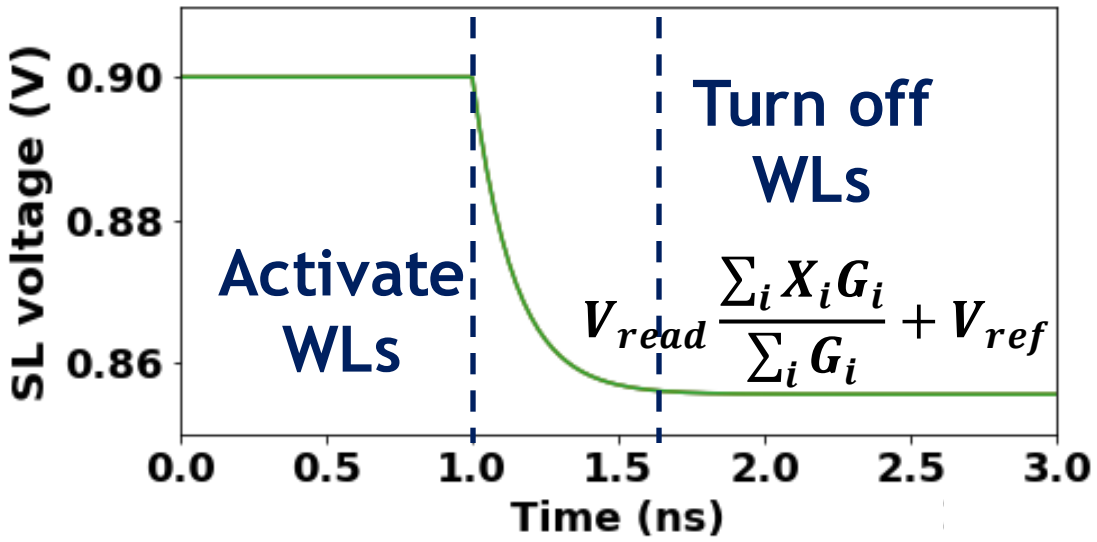
- Pre-charge SLs to V_{ref}
- Drive BLs to $V_{ref} + X_i \times V_{read}$
- Activate all WLs and keep SLs floating
- Turn off WLs when SL voltage settles



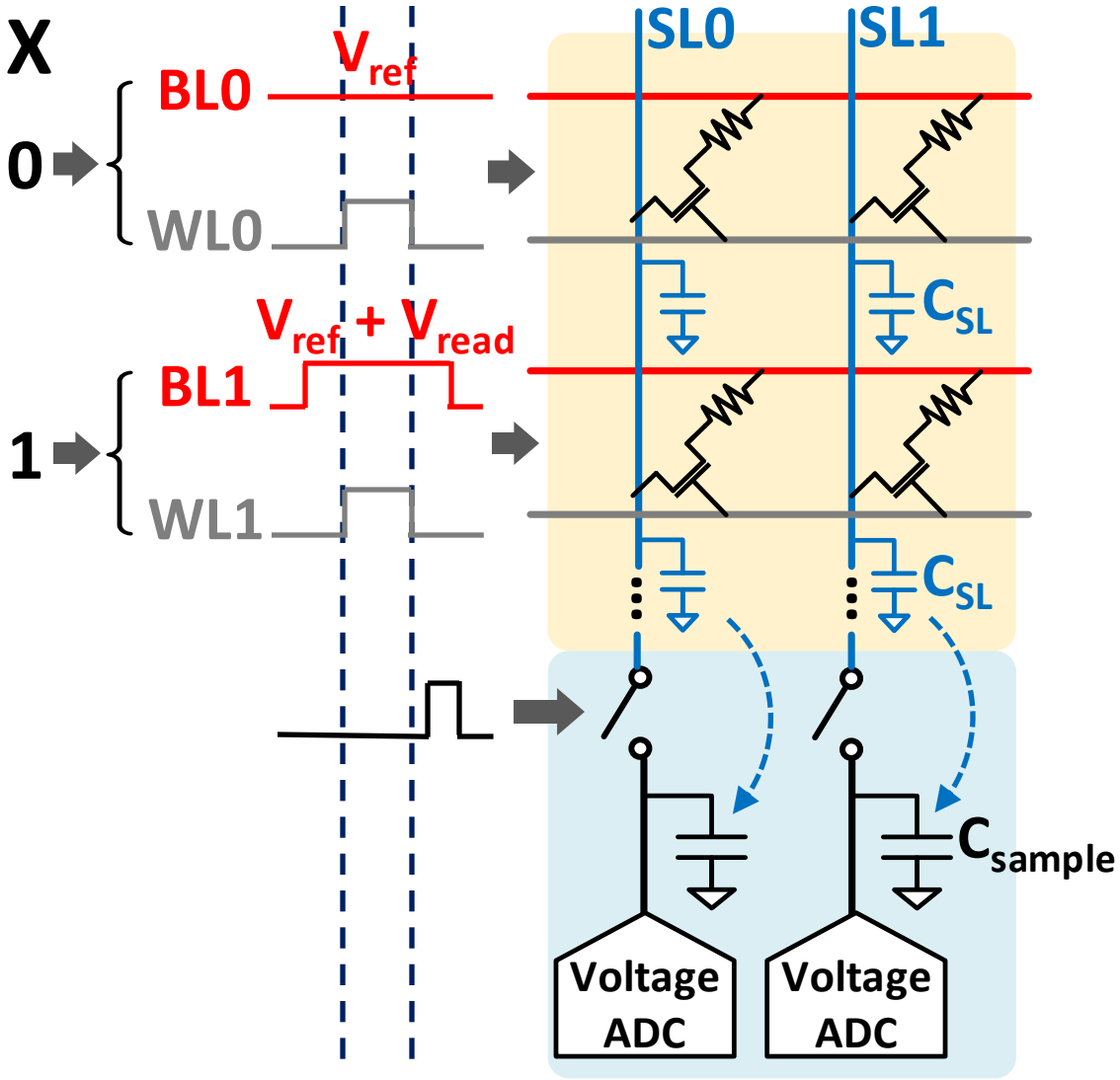
OPEN-CIRCUIT VOLTAGE-MODE SENSING



- Pre-charge SLs to V_{ref}
- Drive BLs to $V_{ref} + X_i \times V_{read}$
- Activate all WLs and keep SLs floating
- Turn off WLs when SL voltage settles
- Sample charges from C_{SL} to C_{sample}



OPEN-CIRCUIT VOLTAGE-MODE SENSING



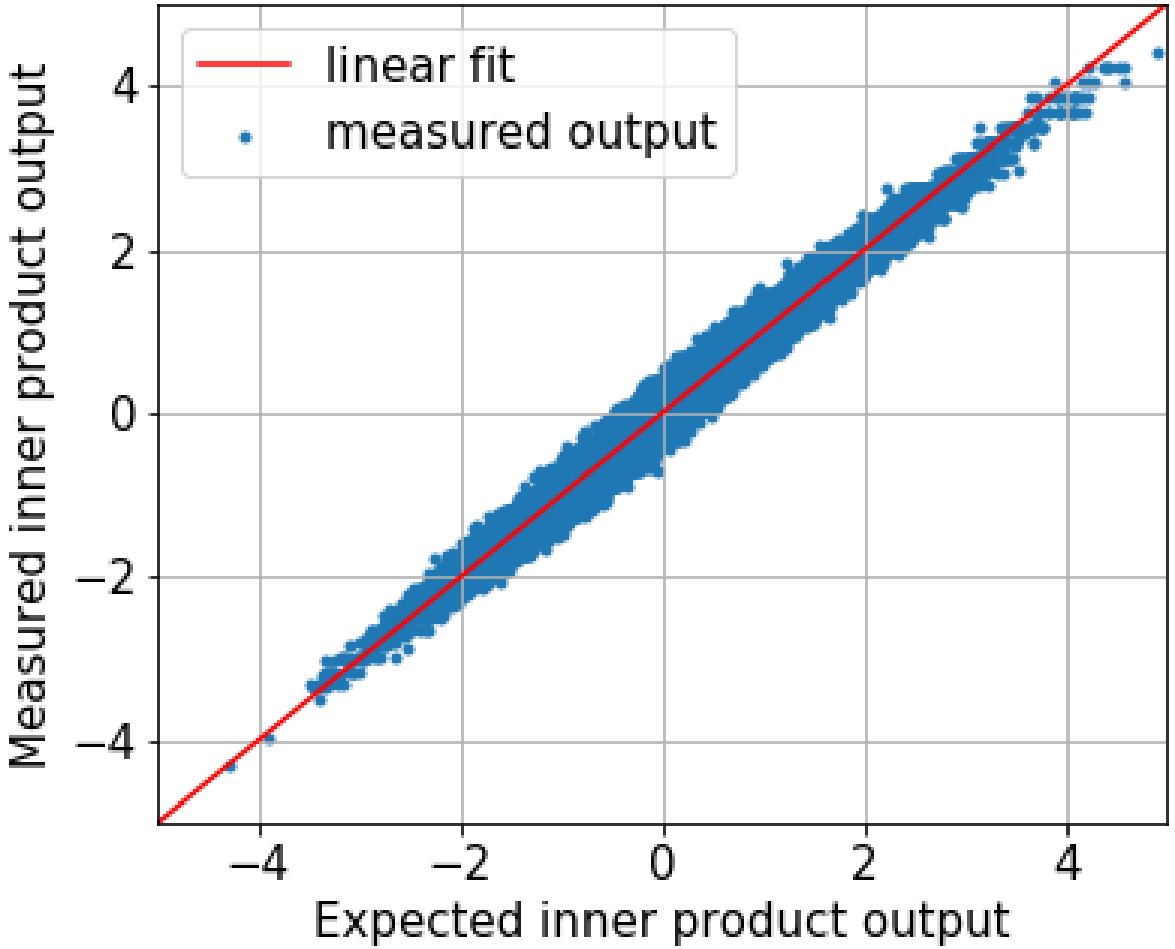
- Pre-charge SLs to V_{ref}
- Drive BLs to $V_{ref} + X_i \times V_{read}$
- Activate all WLs and keep SLs floating
- Turn off WLs when SL voltage settles
- Sample charges from C_{SL} to C_{sample}

$$V_{SL} = V_{read} \frac{\sum_i X_i G_i}{\sum_i G_i} + V_{ref}$$

Total conductance along a SL

Pre-computed once and multiplied back after ADC

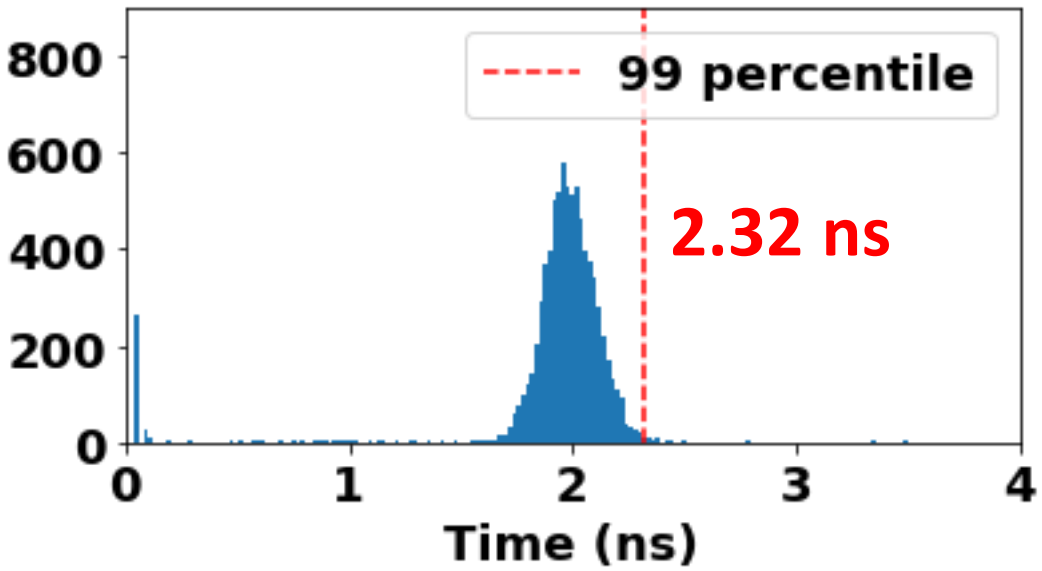
VOLTAGE-MODE MATRIX-VECTOR MULTIPLICATION (MVM) CHIP MEASUREMENTS



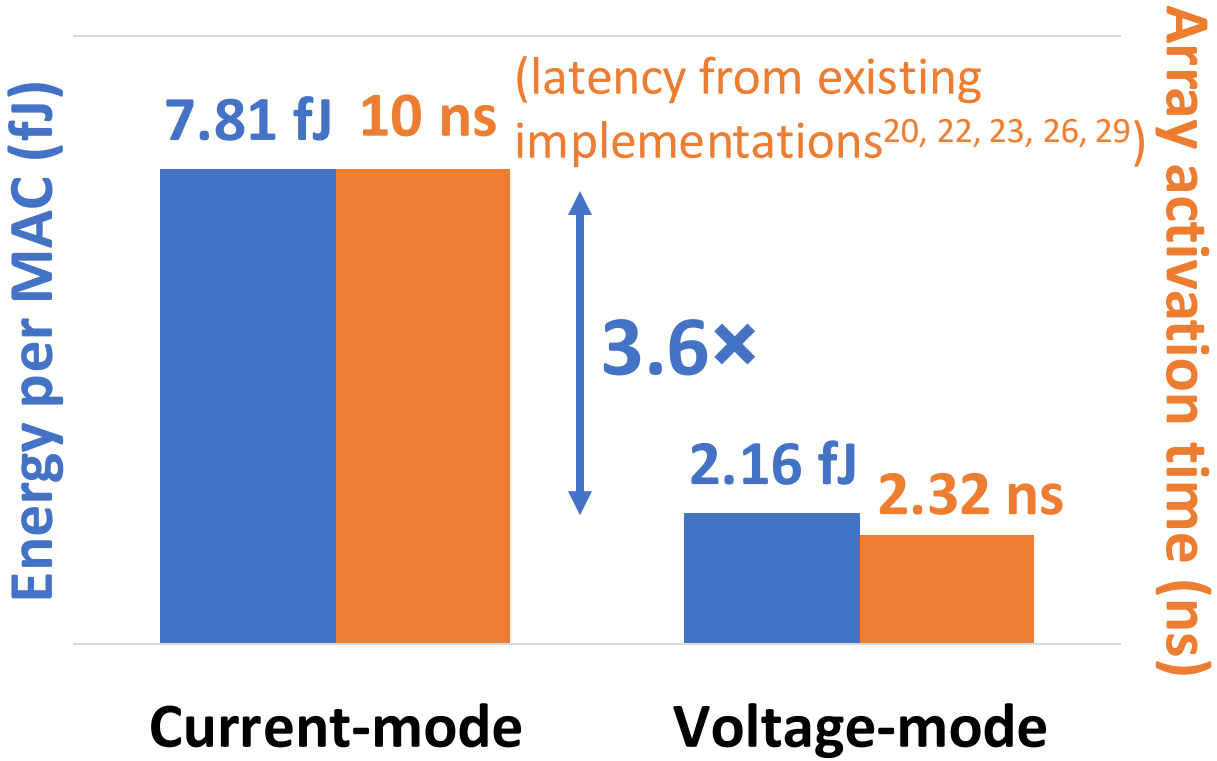
ENERGY-EFFICIENCY BENEFIT

DUE TO SHORTER ARRAY ACTIVATION TIME

SLs settling time distribution

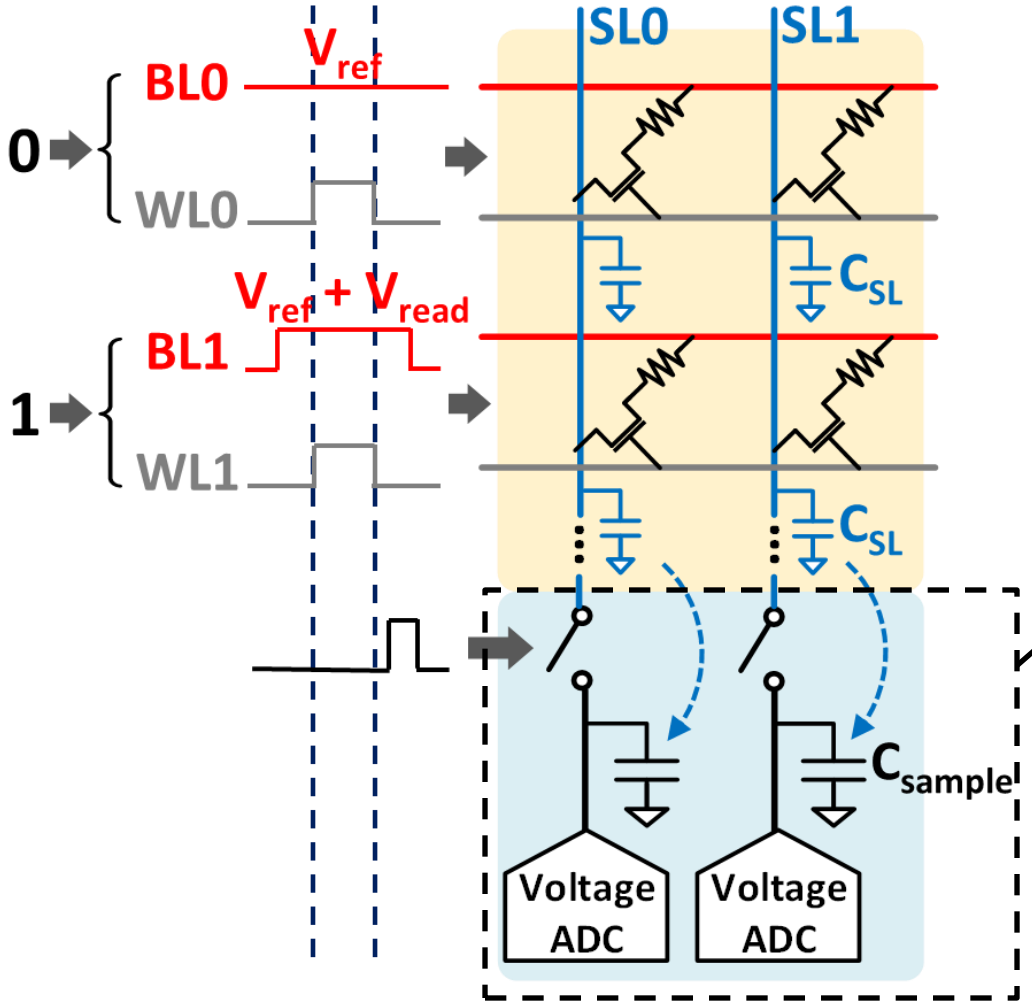


*** Includes only energy from RRAM array**



THROUGHPUT BENEFIT

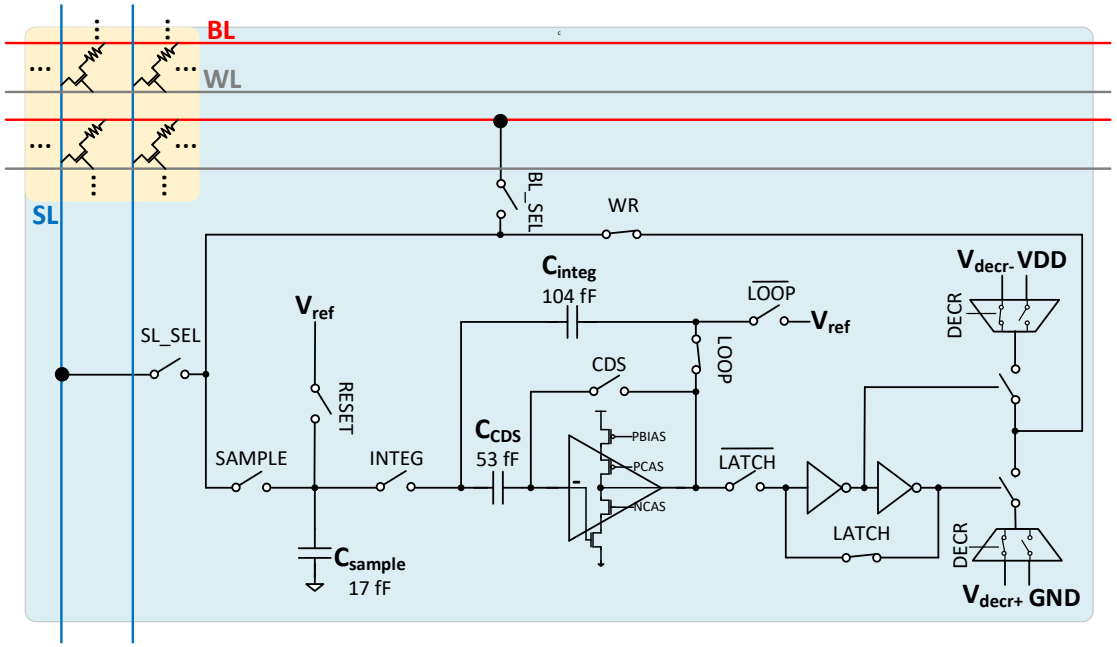
ROW PARALLELISM + COLUMN PARALLELISM

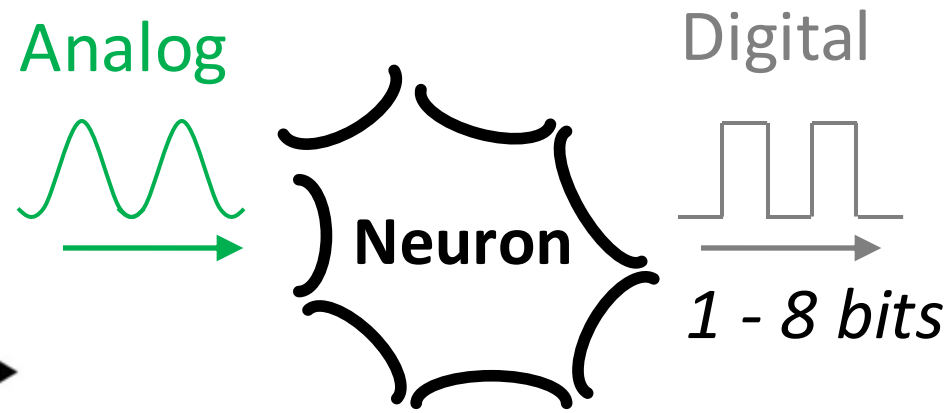


No need for large transistors or time-multiplexing

RECONFIGURABLE VOLTAGE-MODE NEURON

VARIABLE-PRECISION ADC + VARIOUS ACTIVATION FUNCTIONS

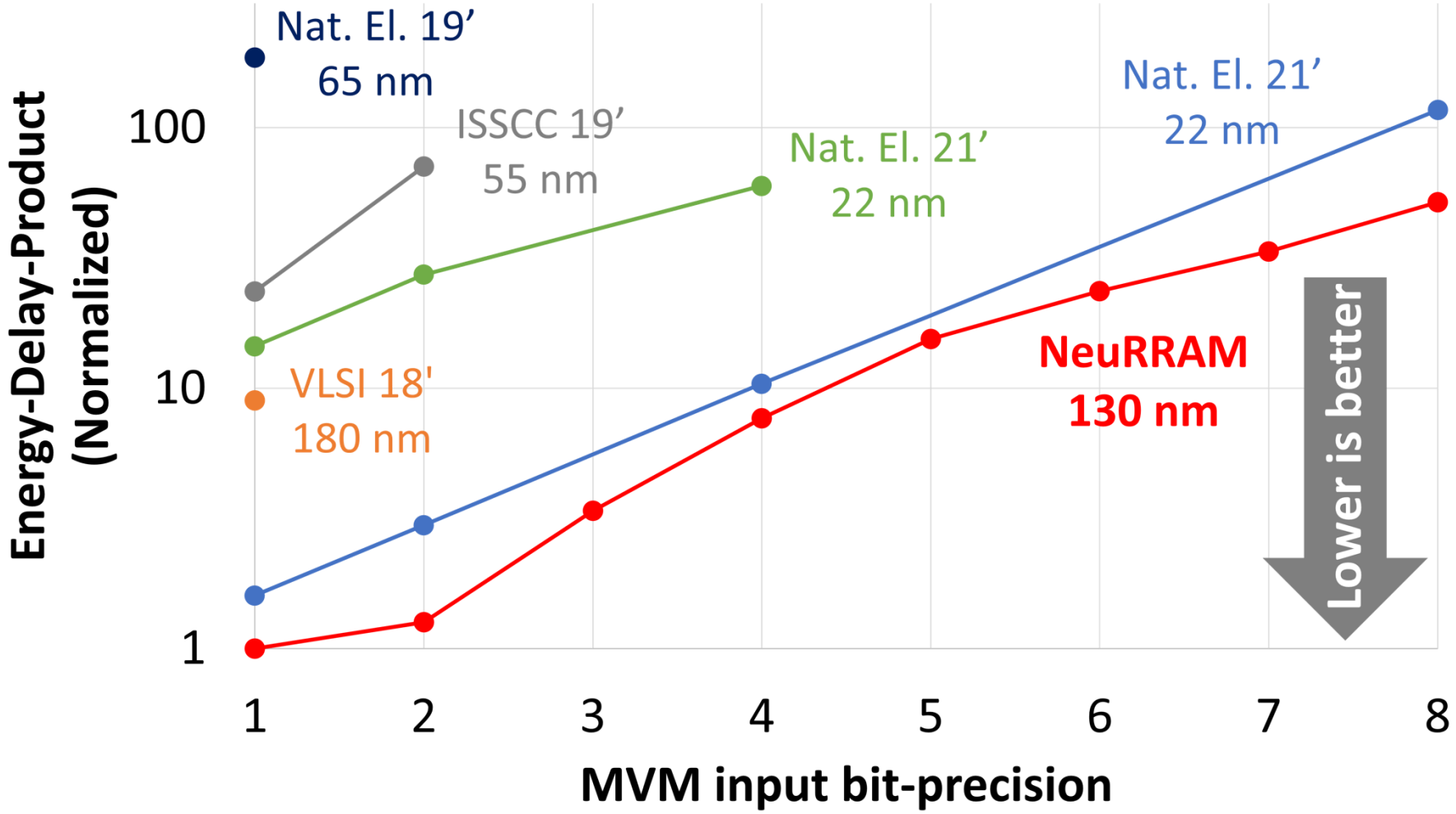




ReLU
Sigmoid
Tanh
Step
...

RECORD ENERGY-DELAY-PRODUCT (EDP)

MEASURED ACROSS VARIOUS COMPUTATIONAL BIT-PRECISIONS



W. Wan … , P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H.-S. P. Wong, *Nature*, 2022

ENERGY-EFFICIENCY

Open-circuit voltage-mode sensing scheme

RECONFIGURABILITY

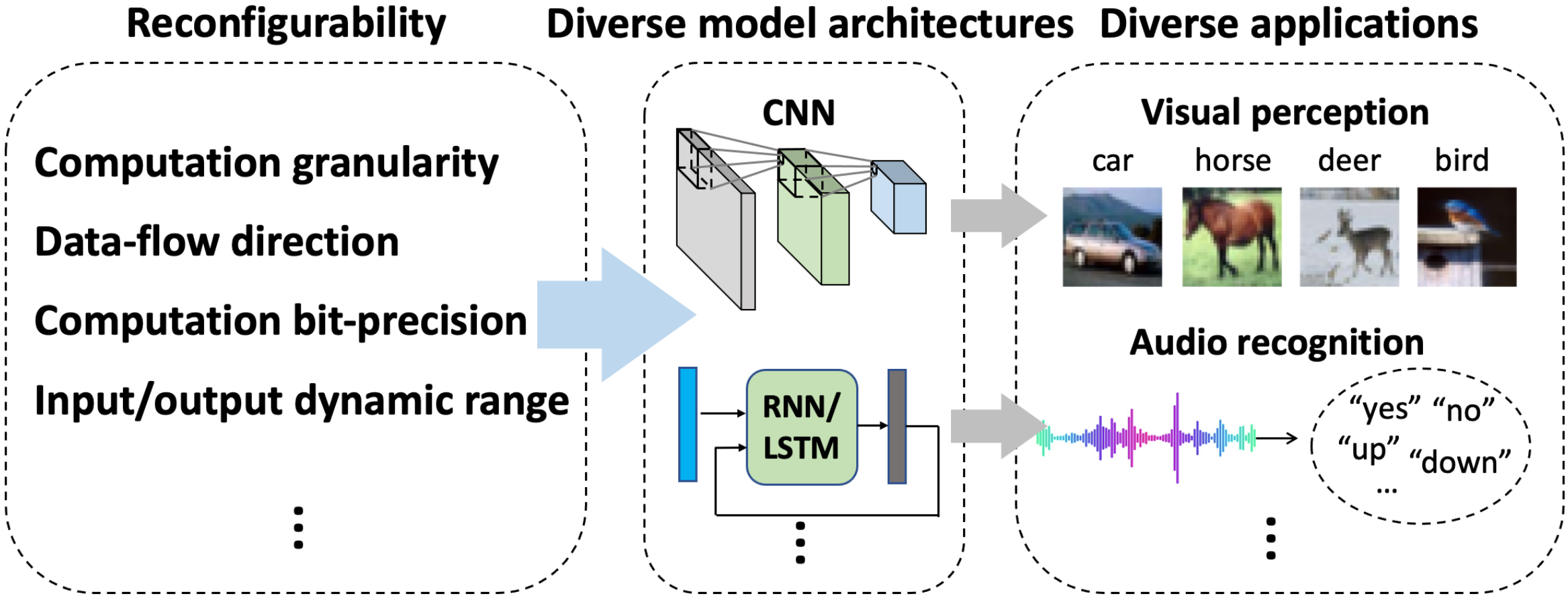
Dataflow reconfigurability realized by Transposable Neurosynaptic Array

ACCURACY

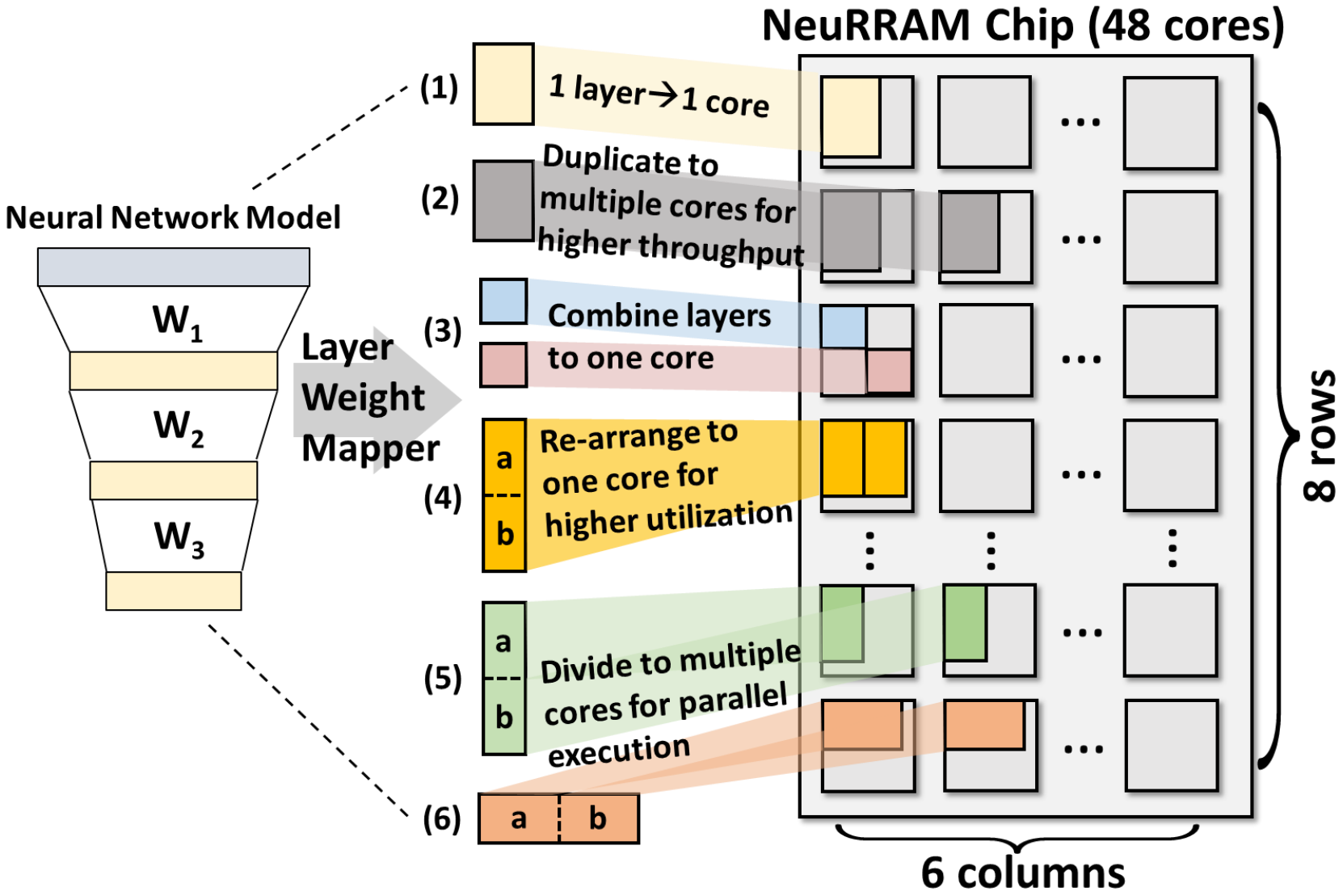
Non-ideality-aware DNN training & fine-tuning

RECONFIGURABLE FOR DIVERSE AI MODELS

RECONFIGURABLE FEATURES ACROSS FULL STACK OF THE DESIGN



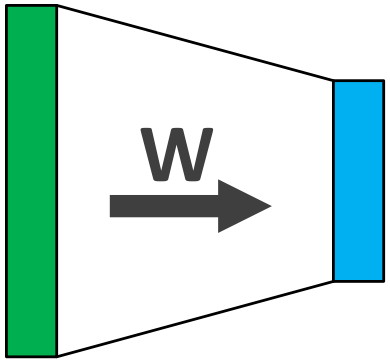
MULTI-CORE WEIGHT MAPPING STRATEGY



RECONFIGURABLE DATAFLOW DIRECTIONS

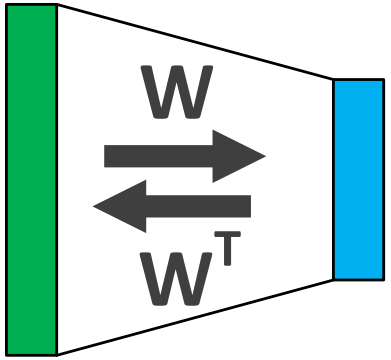
Forward

MLP, CNN



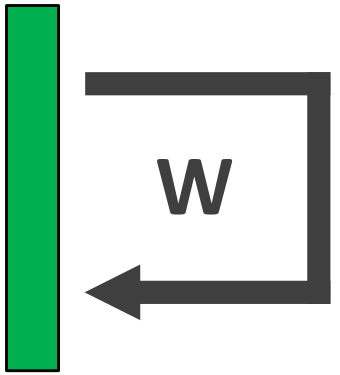
Bi-directional

Back-prop, RBM



Recurrent

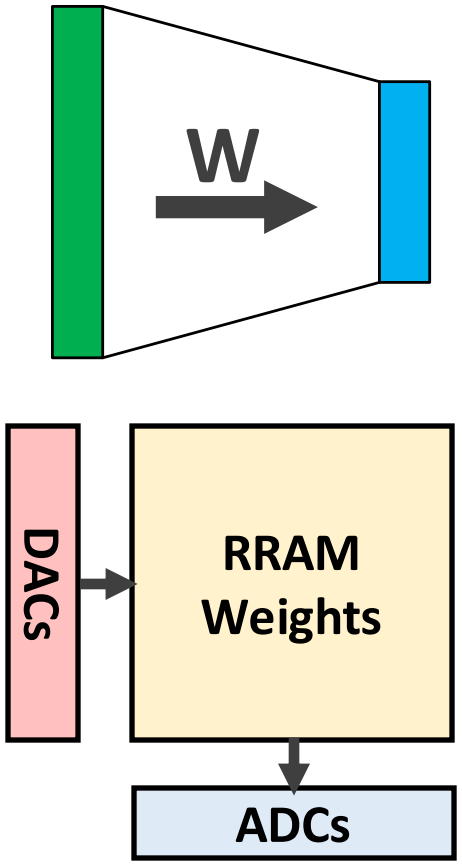
RNN, LSTM



CIM IMPLEMENTATION W/ LARGE OVERHEAD

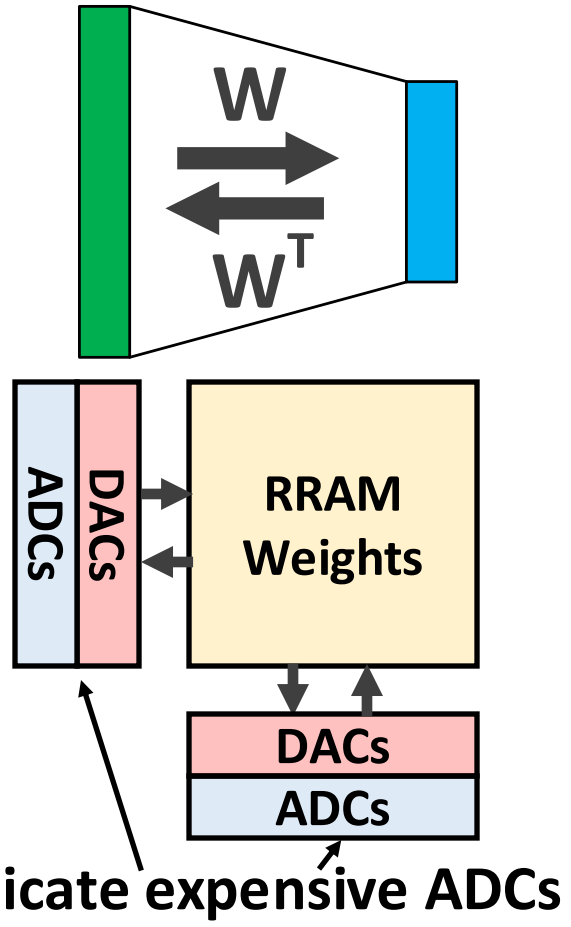
Forward

MLP, CNN



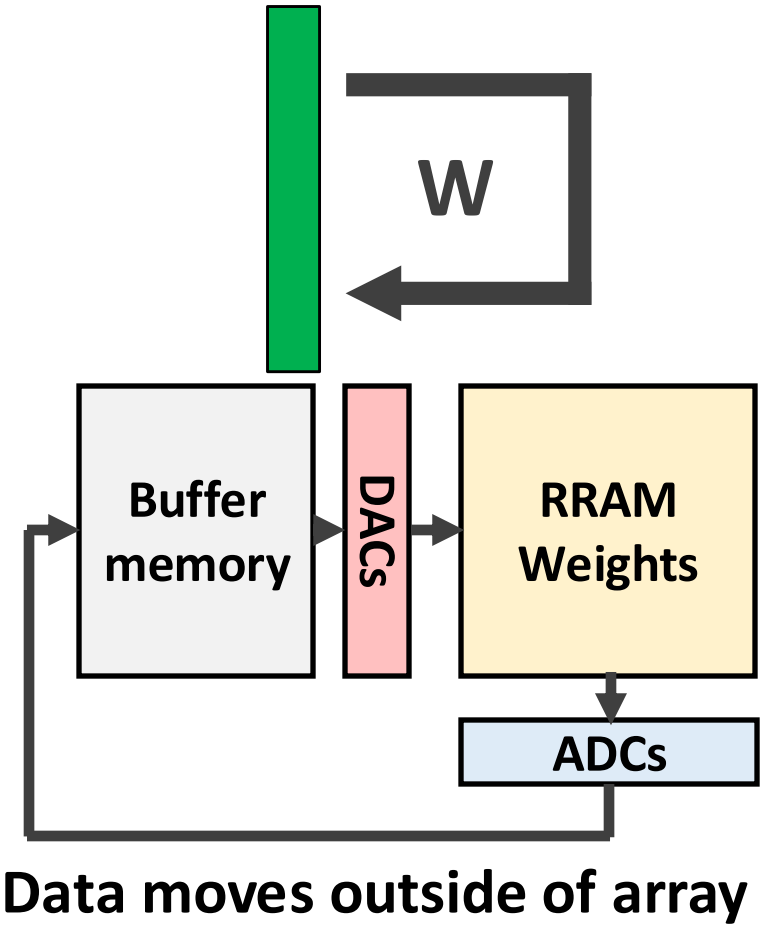
Bi-directional

Back-prop, RBM



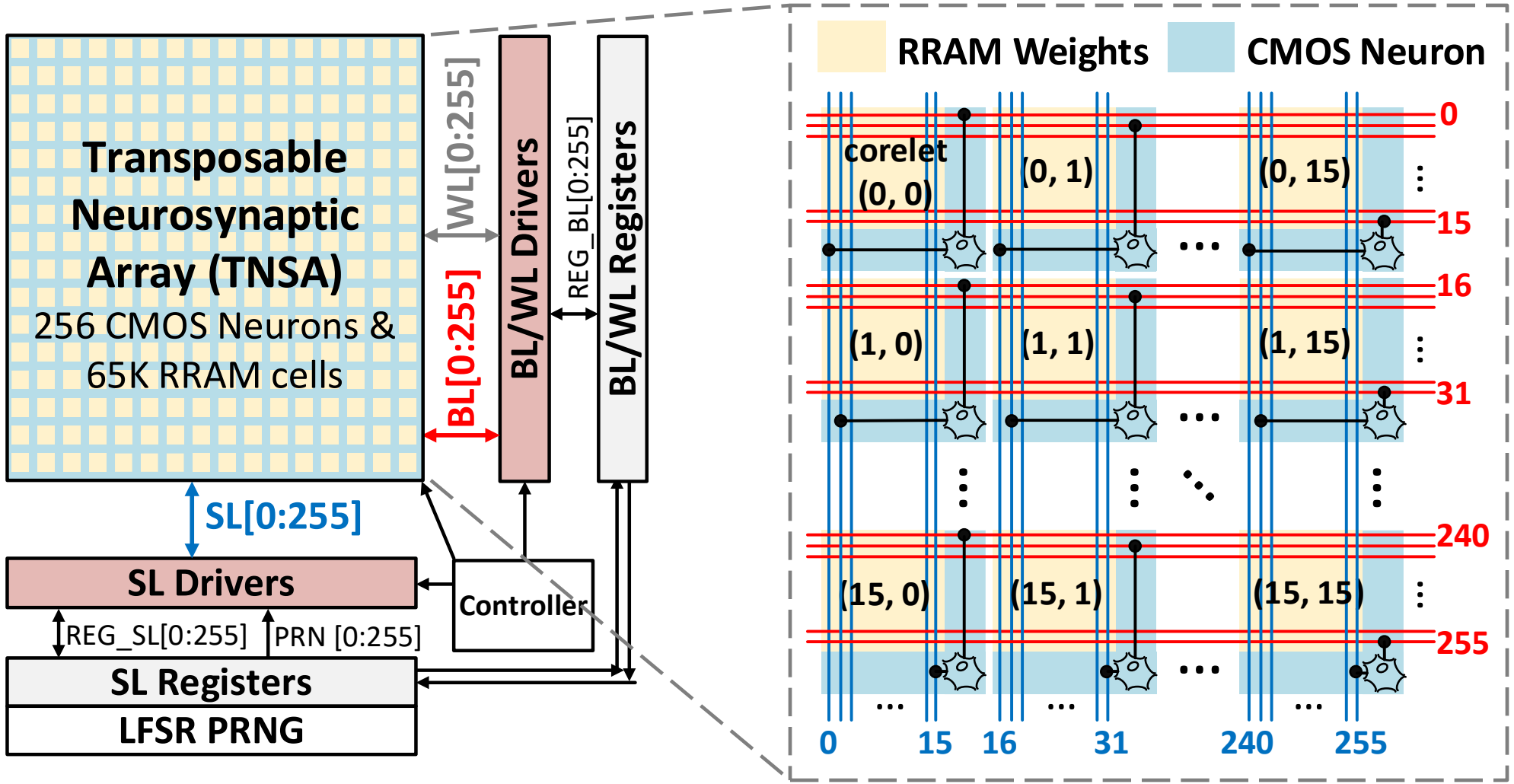
Recurrent

RNN, LSTM



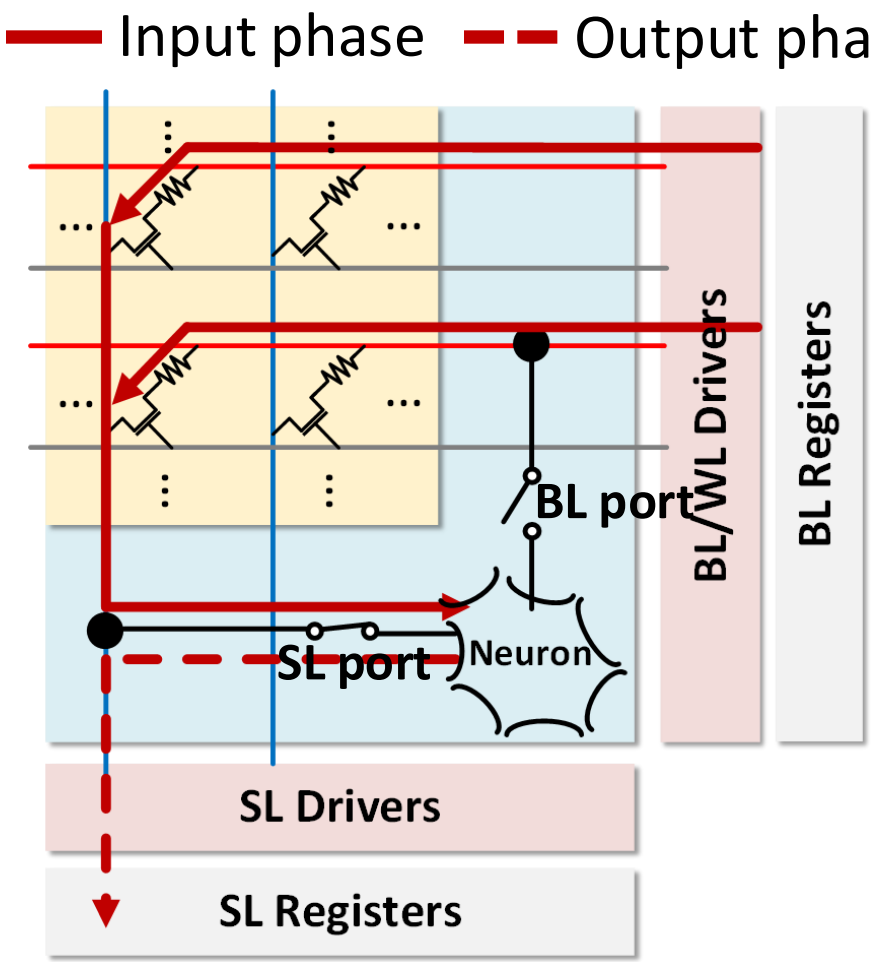
TRANSPOSABLE NEUROSYNAPTIC ARRAY (TNSA)

PHYSICALLY INTERLEAVES RRAM WEIGHTS AND CMOS NEURONS



RECONFIGURABLE MVM DIRECTIONS

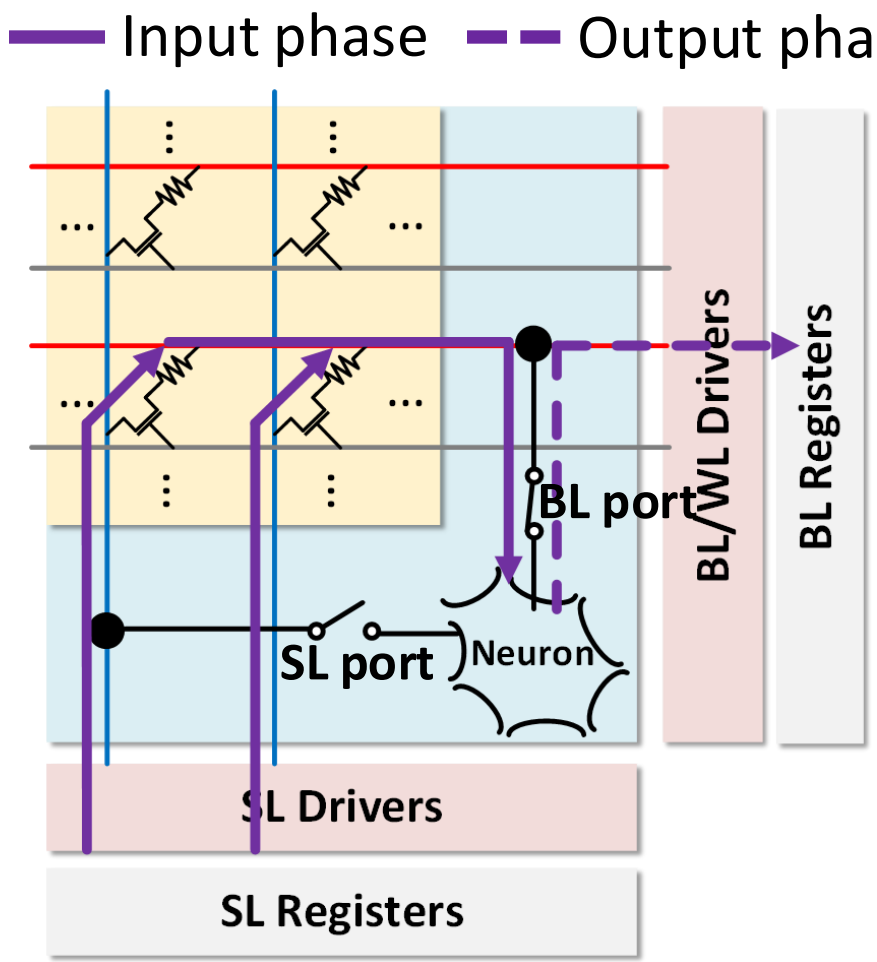
FORWARD (BL→SL)



pulse input	neuron input port	neuron output port	Operation
BL	SL	SL	Forward (BL→SL) MVM
SL	BL	BL	Backward (SL→BL) MVM
BL	SL	BL	BL Recurrent MVM
SL	BL	SL	SL Recurrent MVM
BL	BL		Directly driving neurons
SL	SL		

RECONFIGURABLE MVM DIRECTIONS

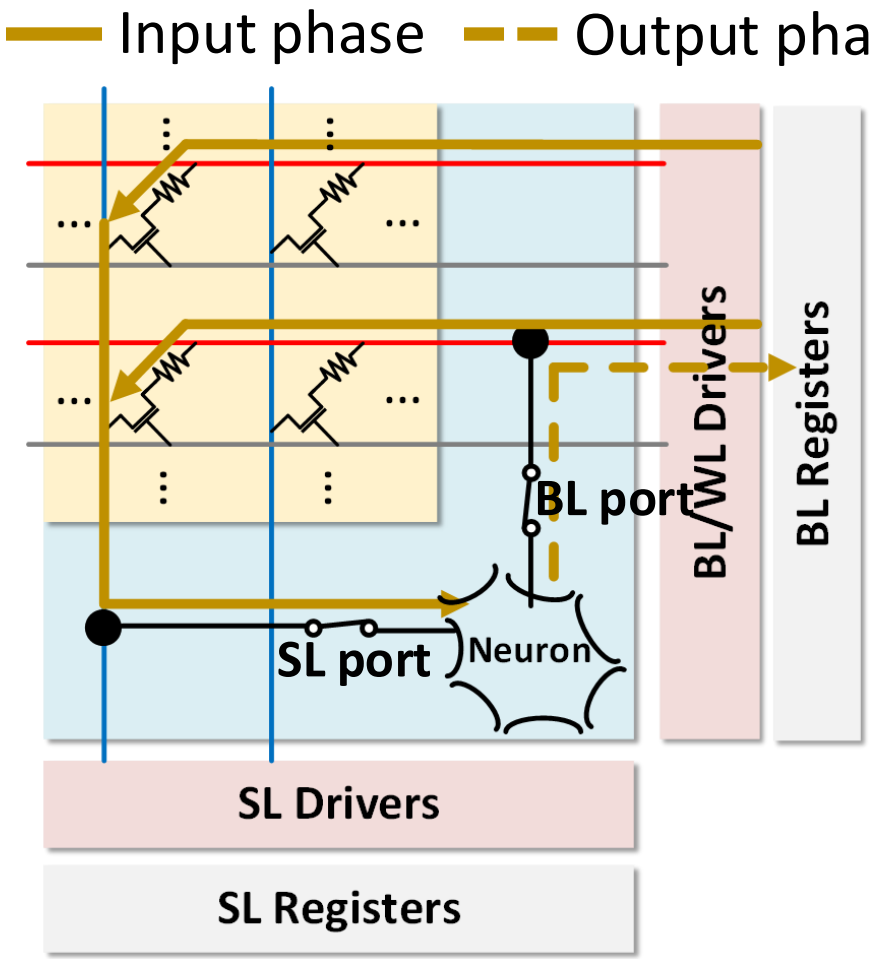
BACKWARD (SL→BL)



pulse input	neuron input port	neuron output port	Operation
BL	SL	SL	Forward (BL→SL) MVM
SL	BL	BL	Backward (SL→BL) MVM
BL	SL	BL	BL Recurrent MVM
SL	BL	SL	SL Recurrent MVM
BL	BL		Directly driving neurons
SL	SL		

RECONFIGURABLE MVM DIRECTIONS

RECURRENT (BL→BL)



pulse input	neuron input port	neuron output port	Operation
BL	SL	SL	Forward (BL→SL) MVM
SL	BL	BL	Backward (SL→BL) MVM
BL	SL	BL	BL Recurrent MVM
SL	BL	SL	SL Recurrent MVM
BL	BL		Directly driving neurons
SL	SL		

ENERGY-EFFICIENCY

Open-circuit voltage-mode sensing scheme

RECONFIGURABILITY

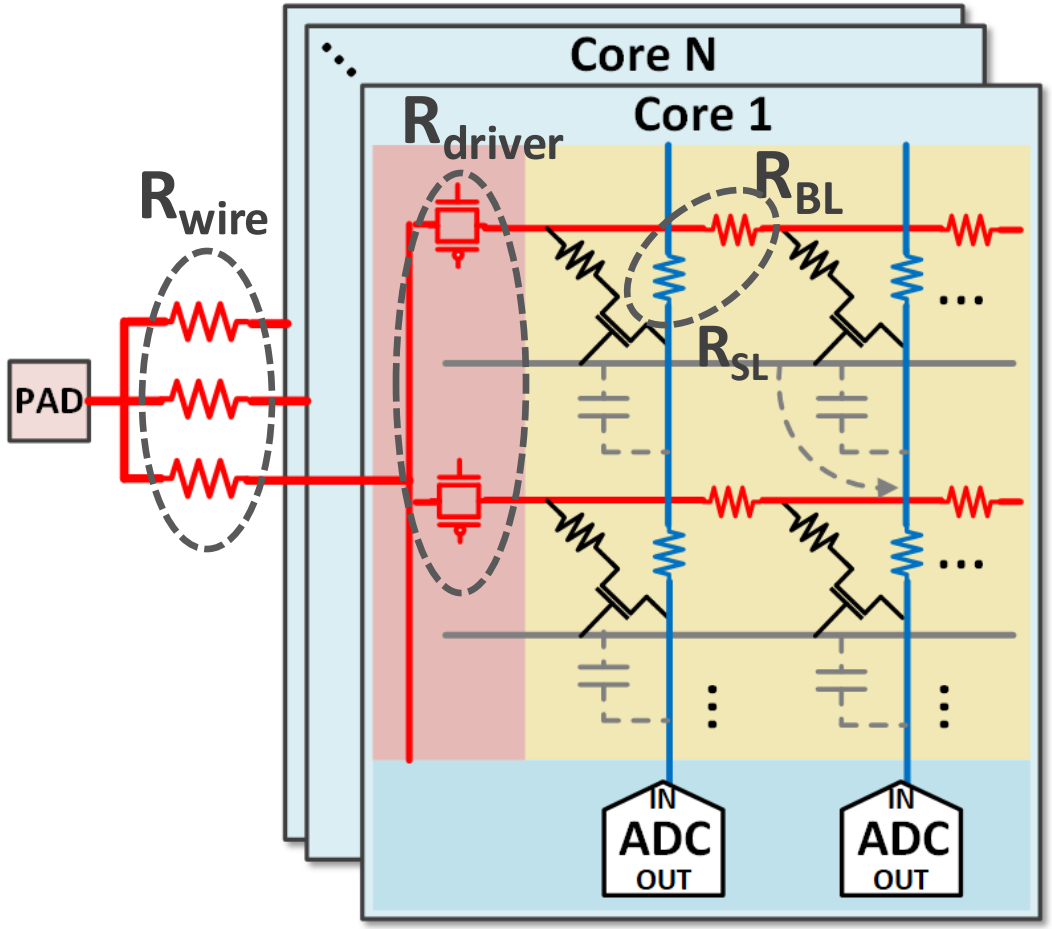
Dataflow reconfigurability realized by Transposable Neurosynaptic Array

ACCURACY

Non-ideality-aware DNN training & fine-tuning

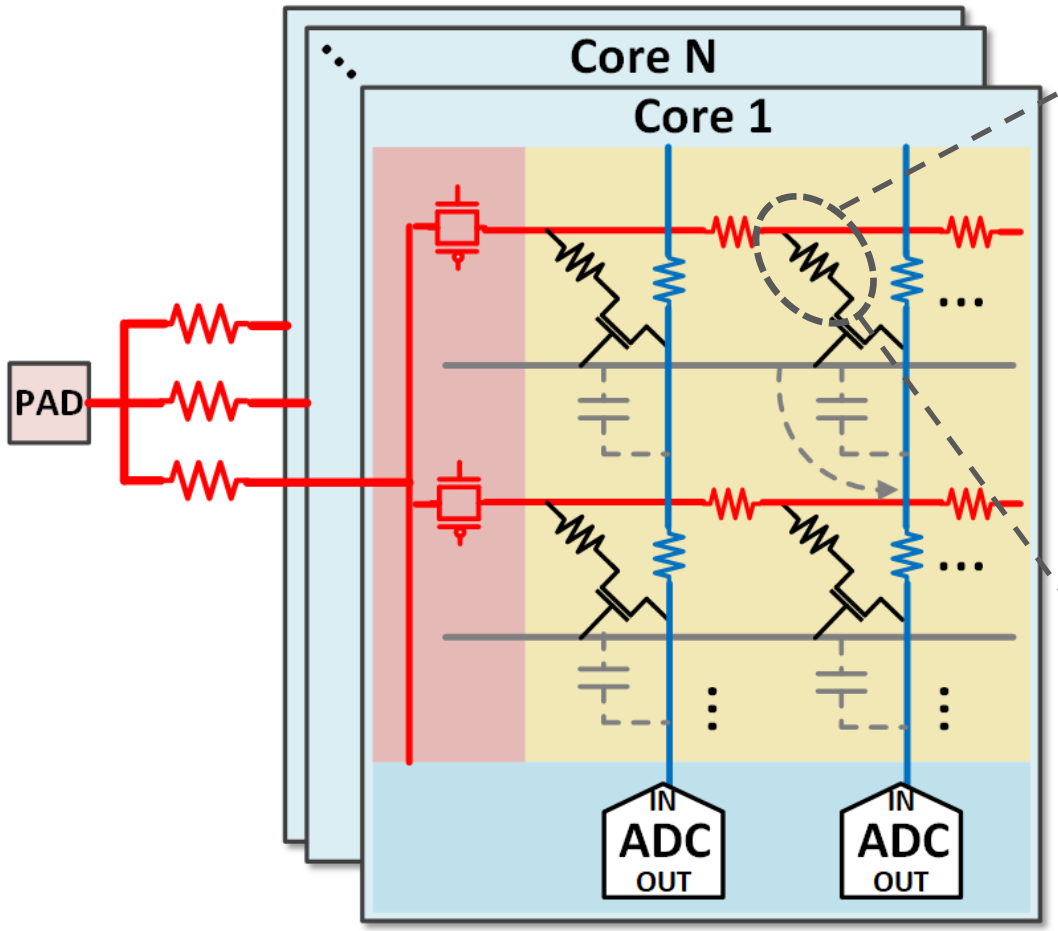
ANALOG MVM NON-IDEALITIES

IR DROPS

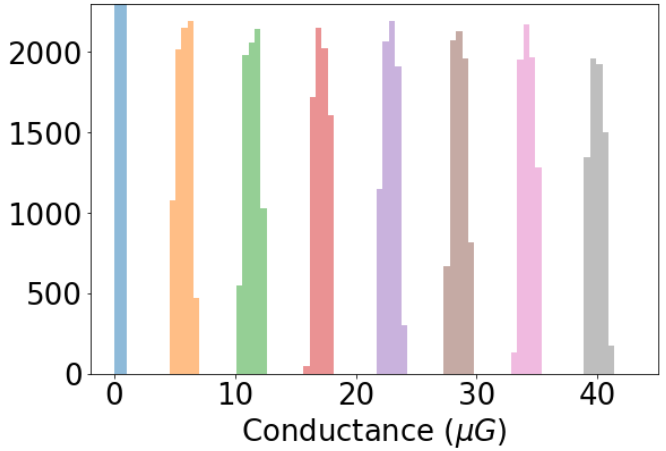


ANALOG MVM NON-IDEALITIES

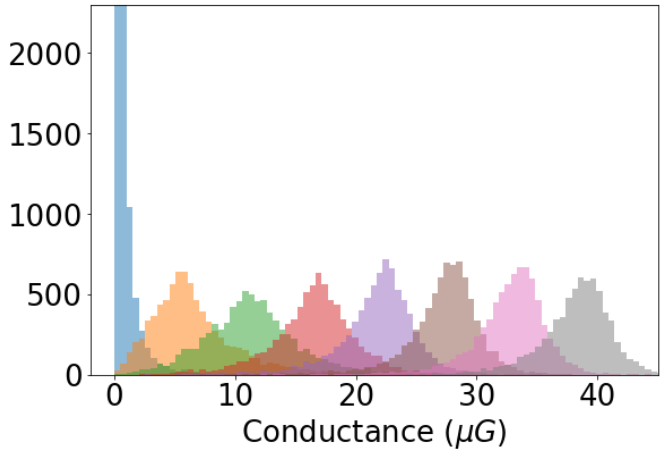
RRAM CONDUCTANCE RELAXATION



During programming

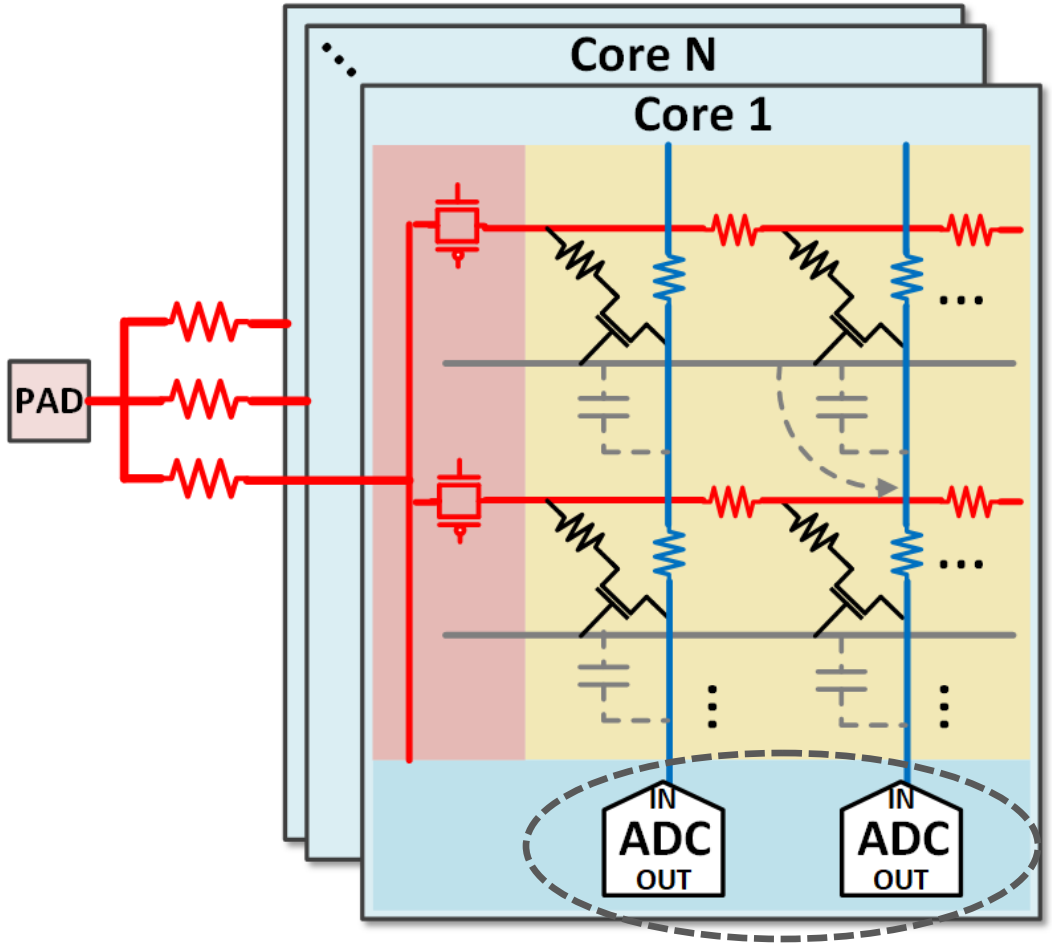


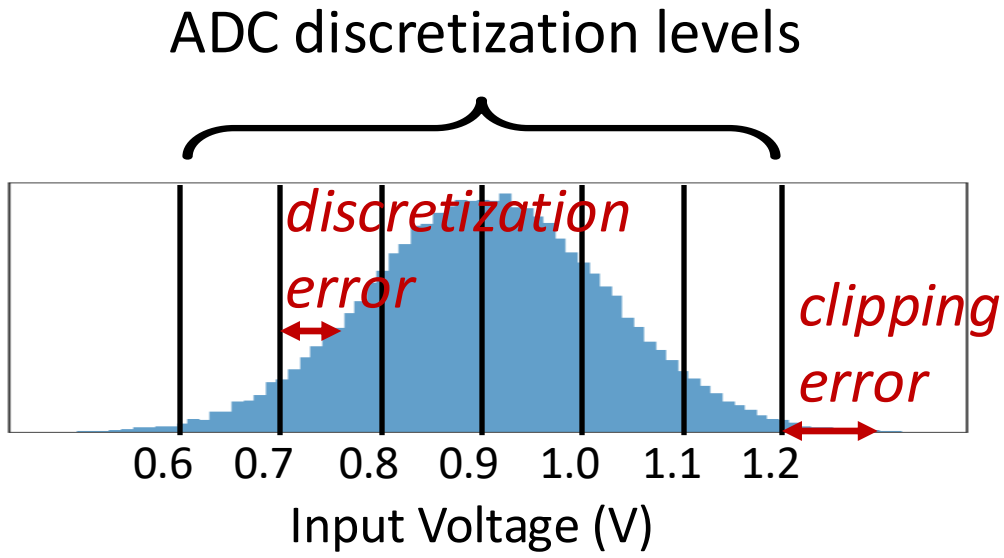
30 mins after programming



ANALOG MVM NON-IDEALITIES

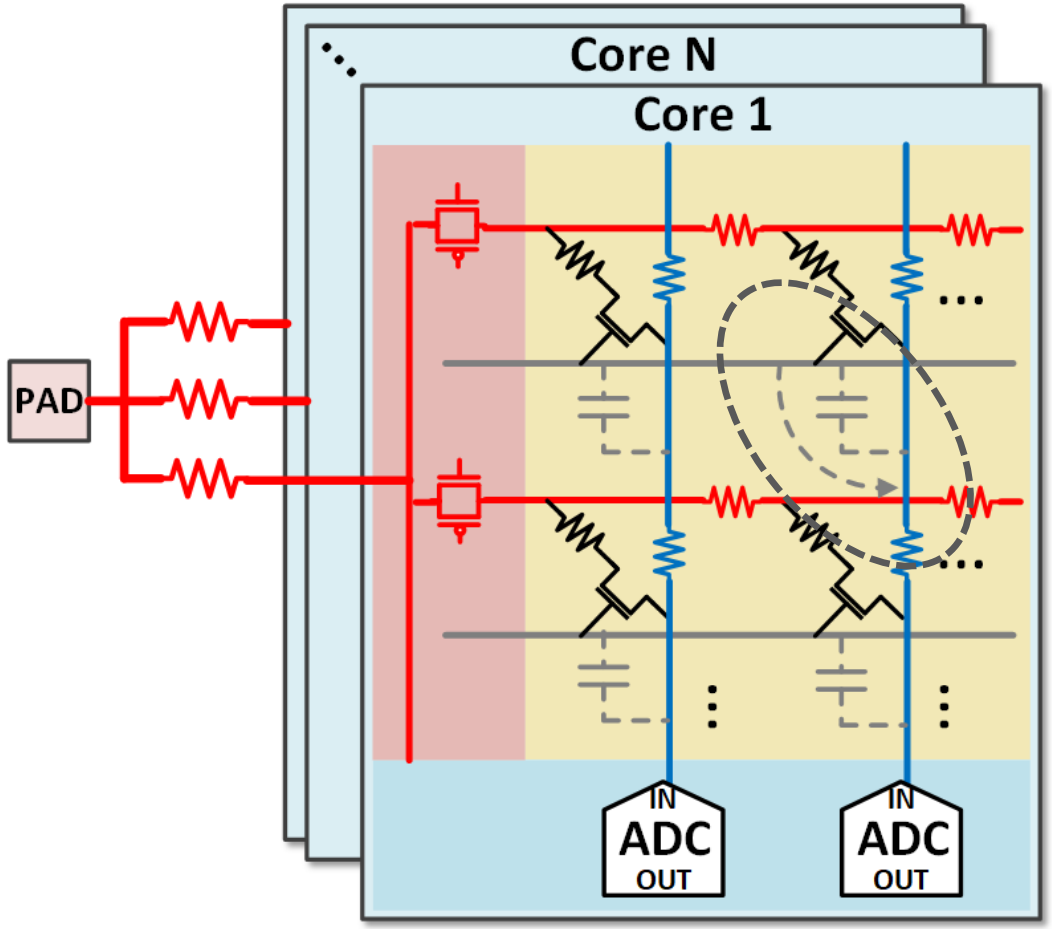
LIMITED ADC RESOLUTION

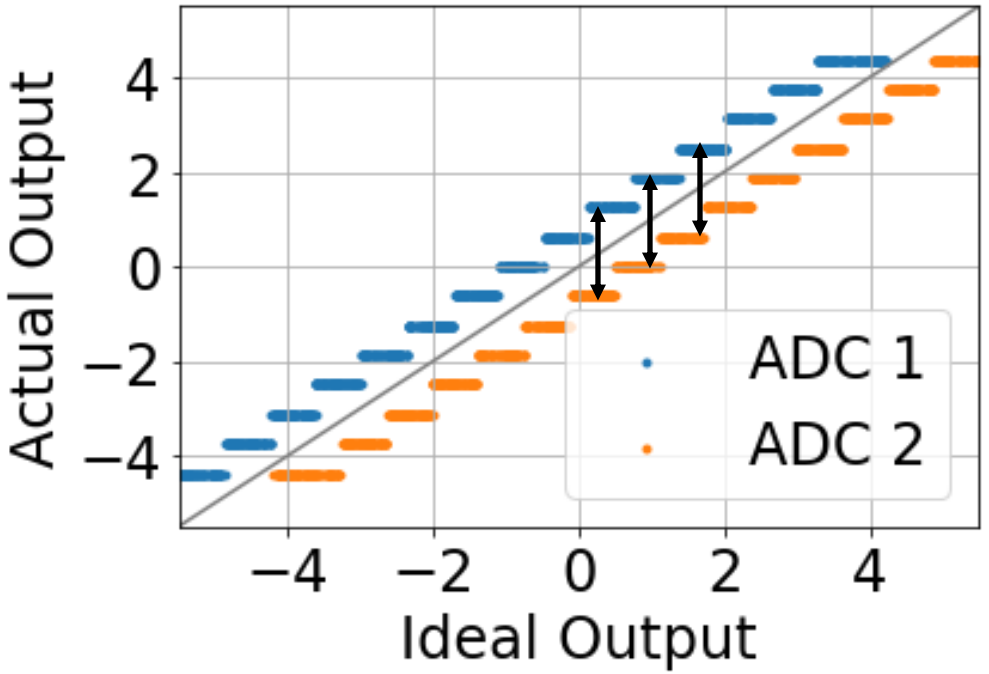




ANALOG MVM NON-IDEALITIES

CAPACITIVE COUPLING INDUCED OFFSET & VARIATION

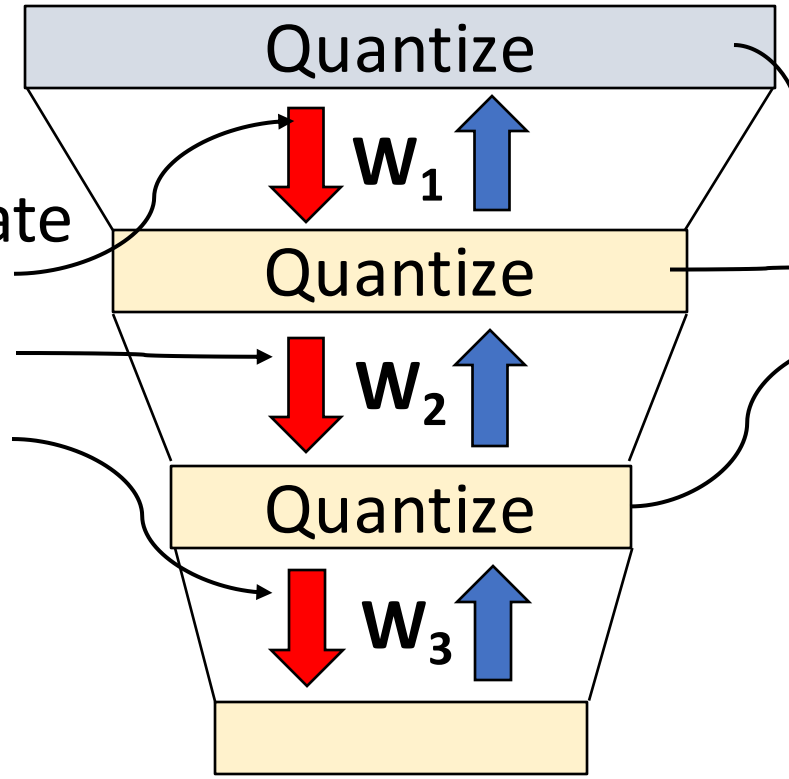




NON-IDEALITY-AWARE MODEL TRAINING

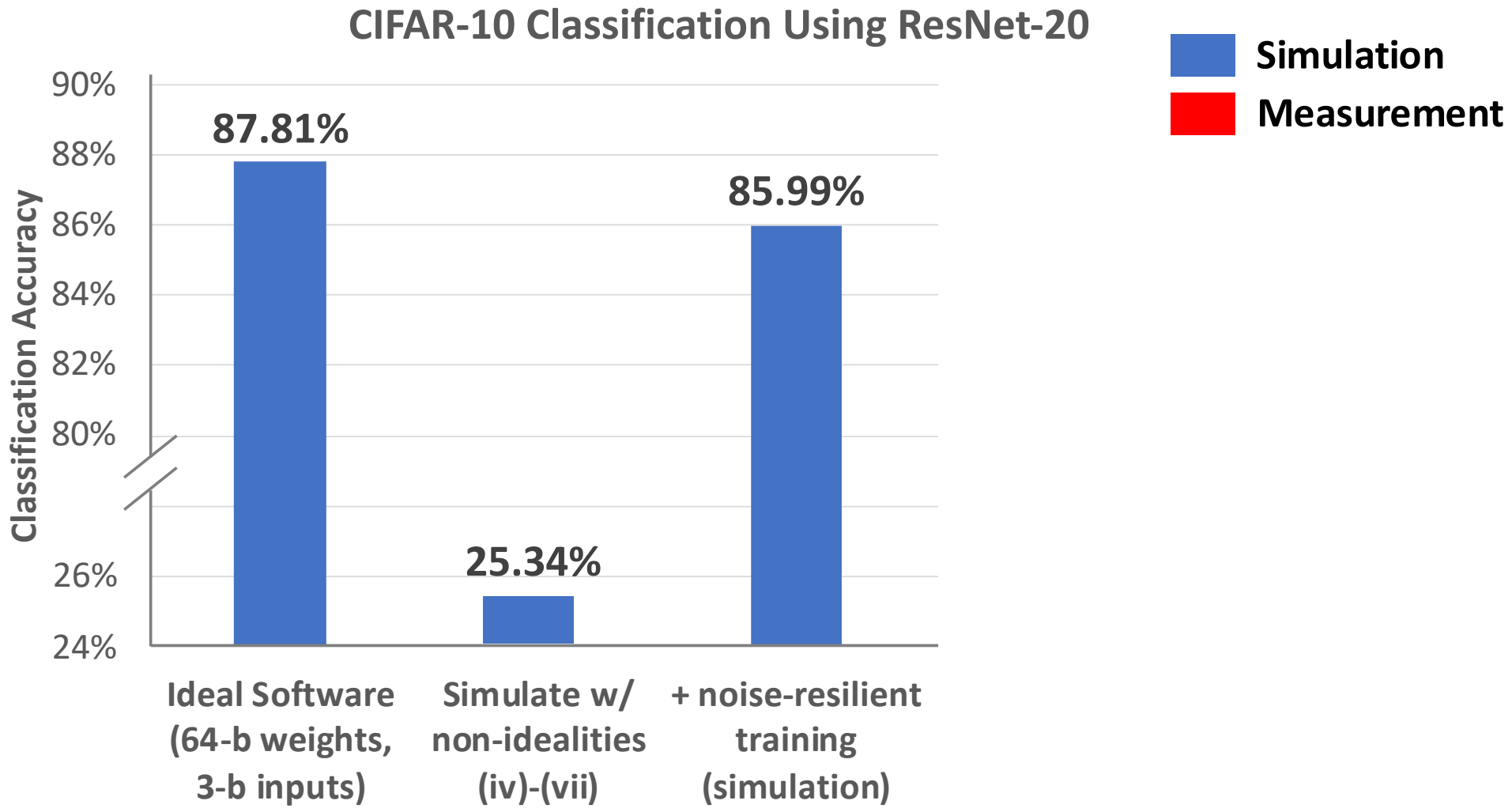
MODEL LEARNS TO ADAPT TO NON-IDEALITIES TROUGH TRAINING

Inject noises that emulate **RRAM conductance relaxation** into weights during training



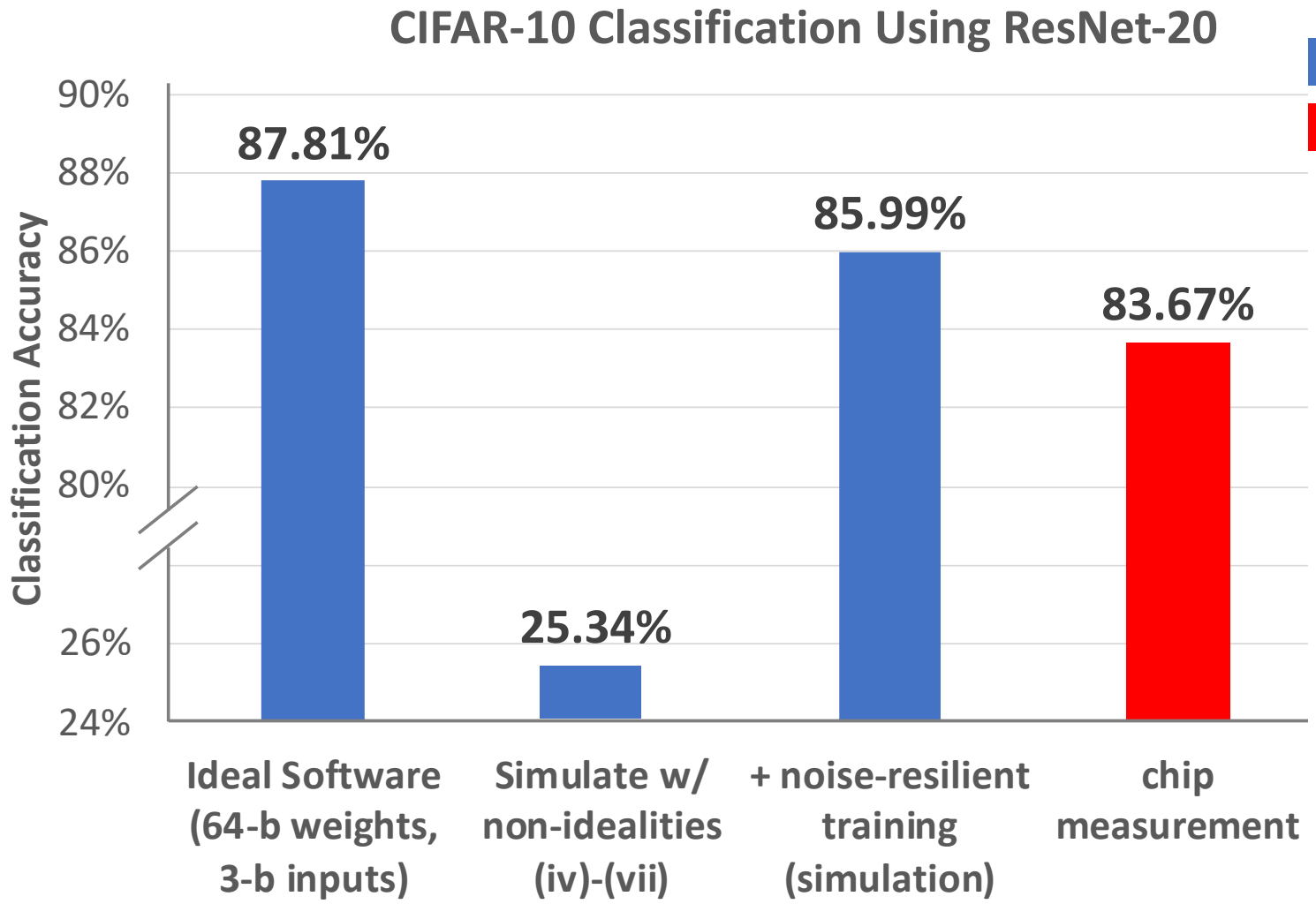
Quantize input & output of each layer to emulate **DAC & ADC discretization**

NOISE INJECTION IMPROVES MODEL'S NOISE-RESILIENCY



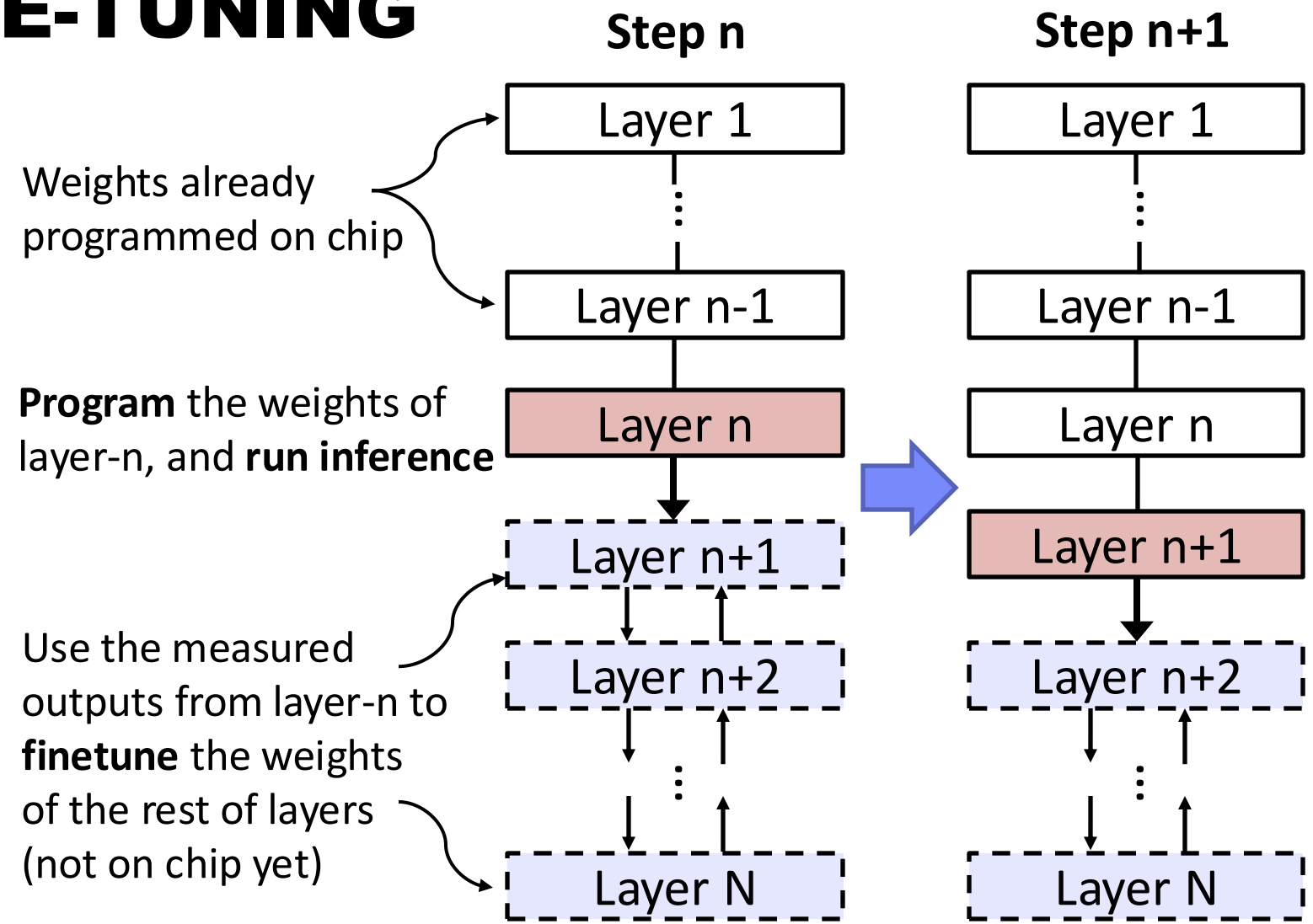
SIMULATION ≠ MEASUREMENT

SHOWS THE IMPORTANCE OF HARDWARE EXPERIMENTS…

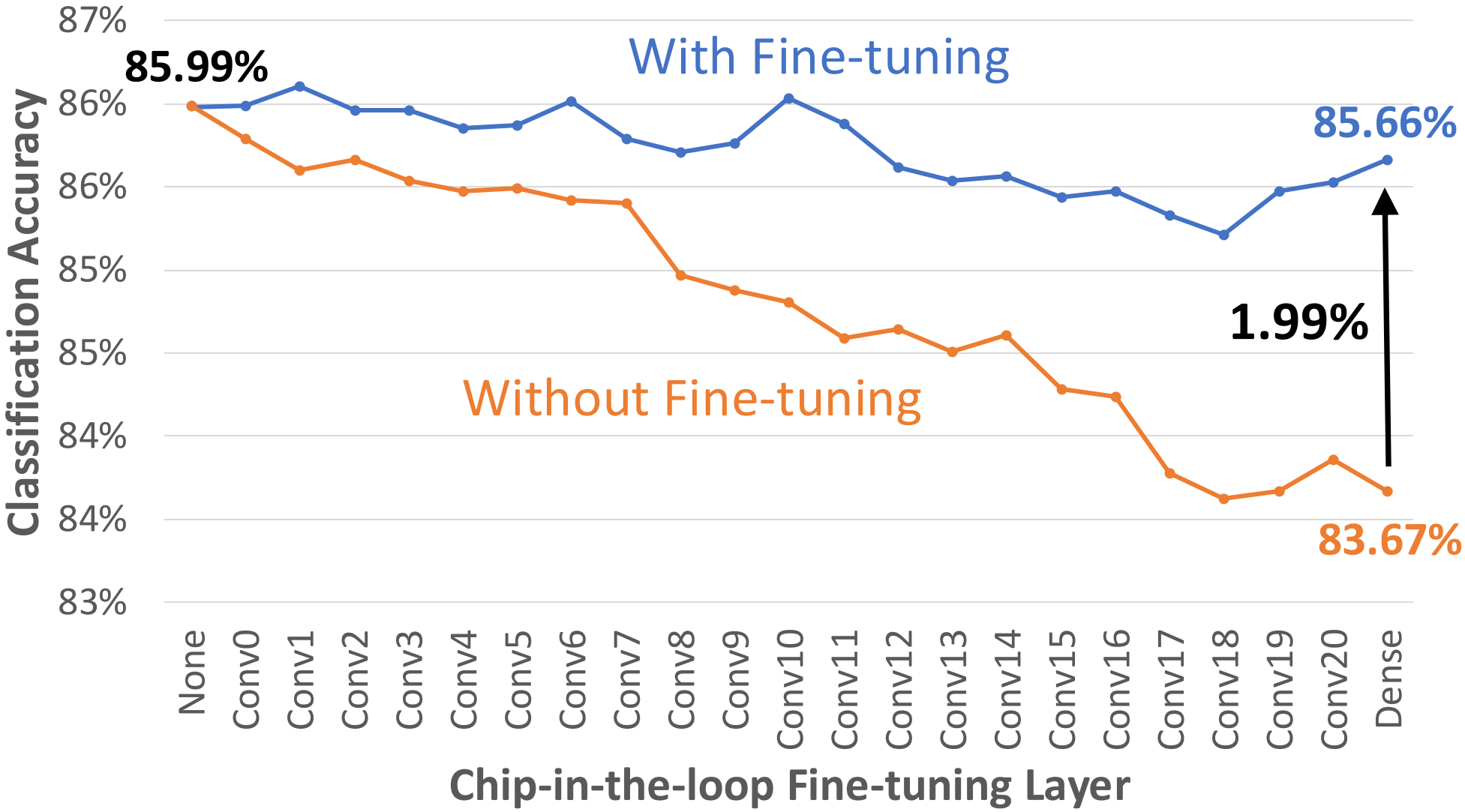


Due to non-idealities that were not modelled accurately in simulation

CHIP-IN-THE-LOOP PROGRESSIVE MODEL FINE-TUNING

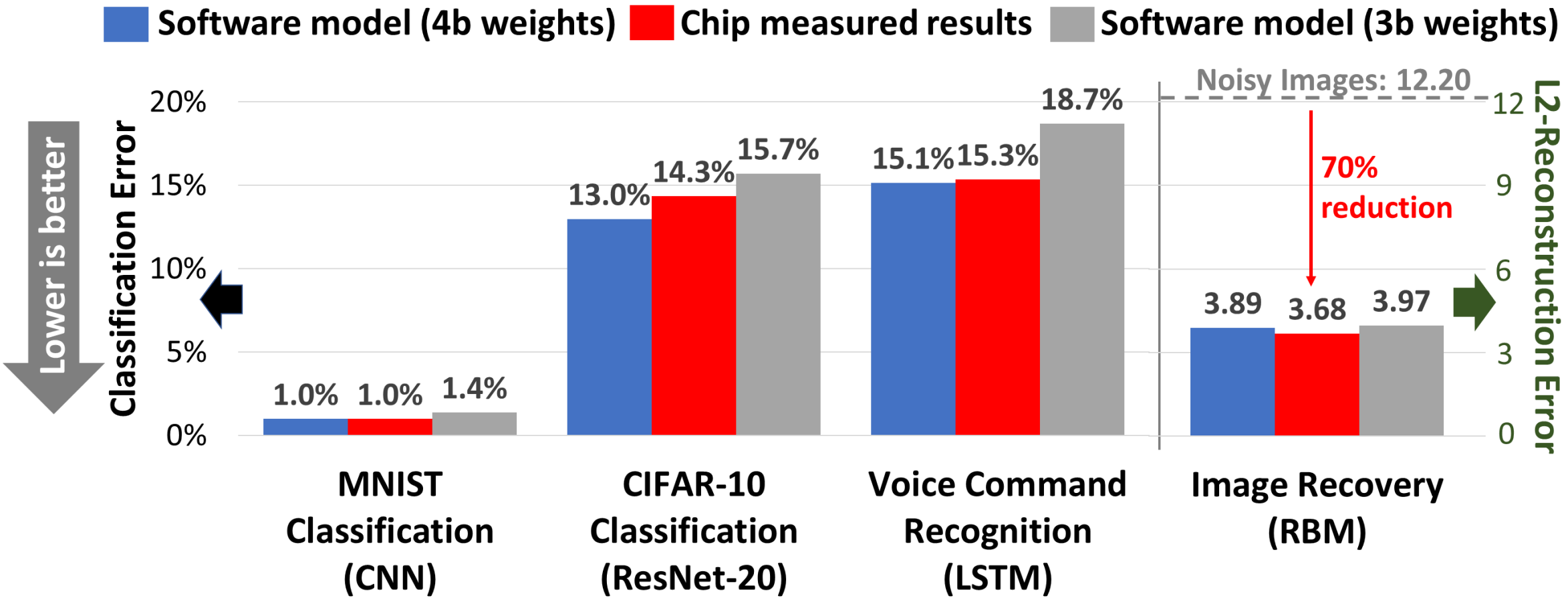


FINE-TUNING RECOVERS ACCURACY LOSS



CHIP MEASURED INFERENCE ACCURACY

COMPARABLE TO SOFTWARE MODEL W/ 4B WEIGHTS



SUMMARY

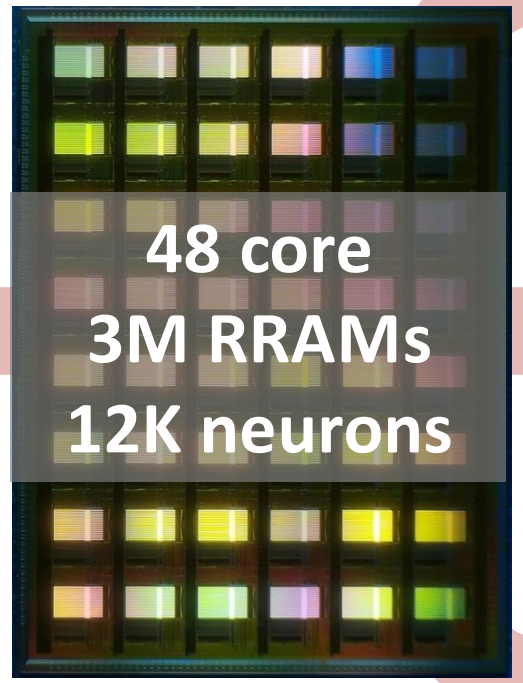
Algorithm

System

Architecture

Circuit

Technology



W. Wan *et al.*, *Nature* 2022

ENERGY-EFFICIENCY

Highest among RRAM CIM chips

RECONFIGURABILITY

CNN, MLP, LSTM, RBM, etc.

INFERENCE ACCURACY

Comparable to software model with 4-bit weights

FUTURE?

- **Device Technology**
  - Density ever reach DRAM-level?
  - Yield for MLC?
- **Analog Compute**
  - Robustness against non-idealities and PVT variation
  - Design overhead?
- **System-level Benefits**
  - Overshadowed by other components?
- **Computing Model**
  - Can we make it transparent to software developers?

SPONSORS

Stanford | SystemX Alliance

Stanford | Non-Volatile Memory Technology Research Initiative (NMTRI)

北京未来芯片技术高精尖创新中心
BEIJING INNOVATION CENTER FOR FUTURE CHIPS

UC San Diego

Questions?