In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks

M. Bocquet[*†], T. Hirtzlin[*‡], J.-O. Klein[‡], E. Nowak[§], E. Vianello[§], J.-M. Portal[†] and D. Querlioz[‡]

[†] *Aix Marseille Univ, Université de Toulon, CNRS, IM2NP, Marseille, France*

[‡] *C2N, Univ Paris-Sud, CNRS, Orsay, France*

[§] *CEA, LETI, Grenoble, France*

[*]*These authors contributed equally to the work*

In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks

M. Bocquet[*†], T. Hirtzlin[*‡], J.-O. Klein[‡], E. Nowak[§], E. Vianello[§], J.-M. Portal[†] and D. Querlioz[‡]

[†] Aix Marseille Univ, Université de Toulon, CNRS, IM2NP, Marseille, France

[‡] C2N, Univ Paris-Sud, CNRS, Orsay, France

[§] CEA, LETI, Grenoble, France

[*]These authors contributed equally to the work

The Energy Challenge of Artificial Intelligence

- Deep learning

Mask R-CNN

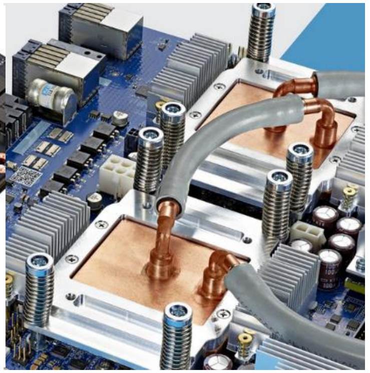
- Energy consumption

Amazon Data-center

- CPU / GPU / TPU > 100W

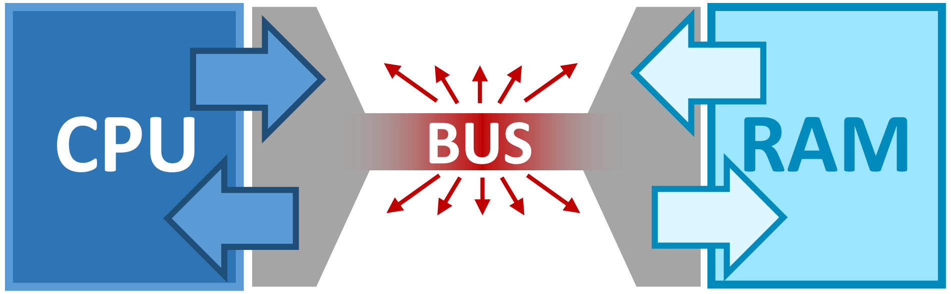
Intel Xeon CPU

NVIDIA GPU
Titan V100



TPU 3 Google

- von Neumann bottleneck

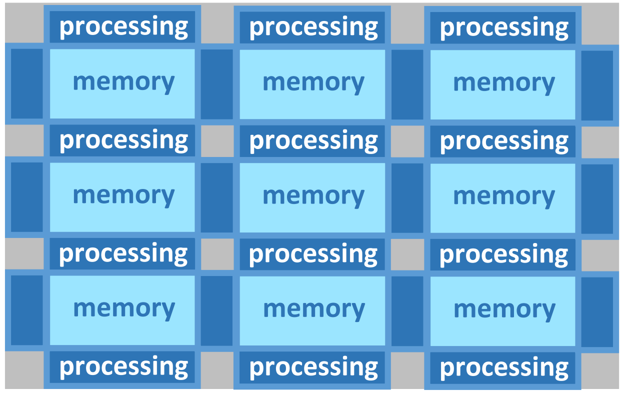


Beyond von Neumann

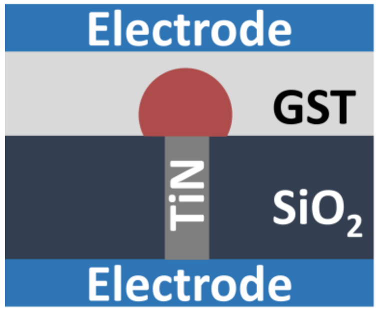
100W → Incompatible with IoT

Microsoft IoT

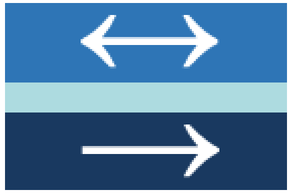
- In memory computing


processing processing processing
memory memory memory
processing processing processing
memory memory memory
processing processing processing
memory memory memory
processing processing processing

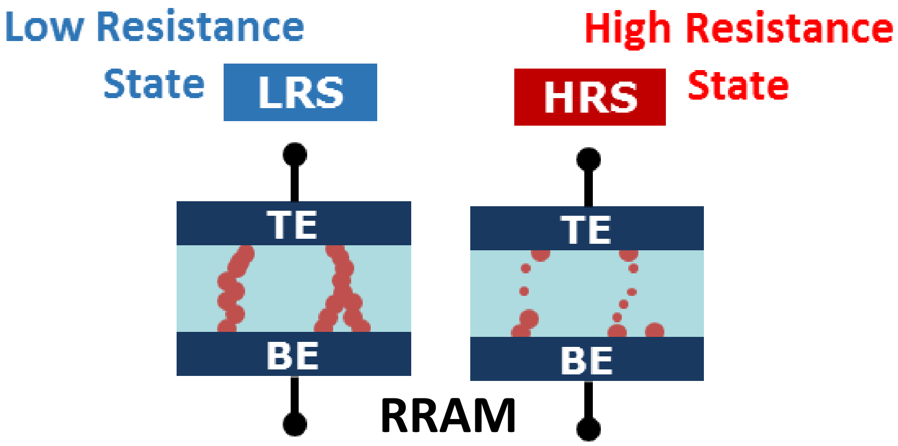
- New **Non-Volatile** memory technology



PCM



MRAM



RRAM

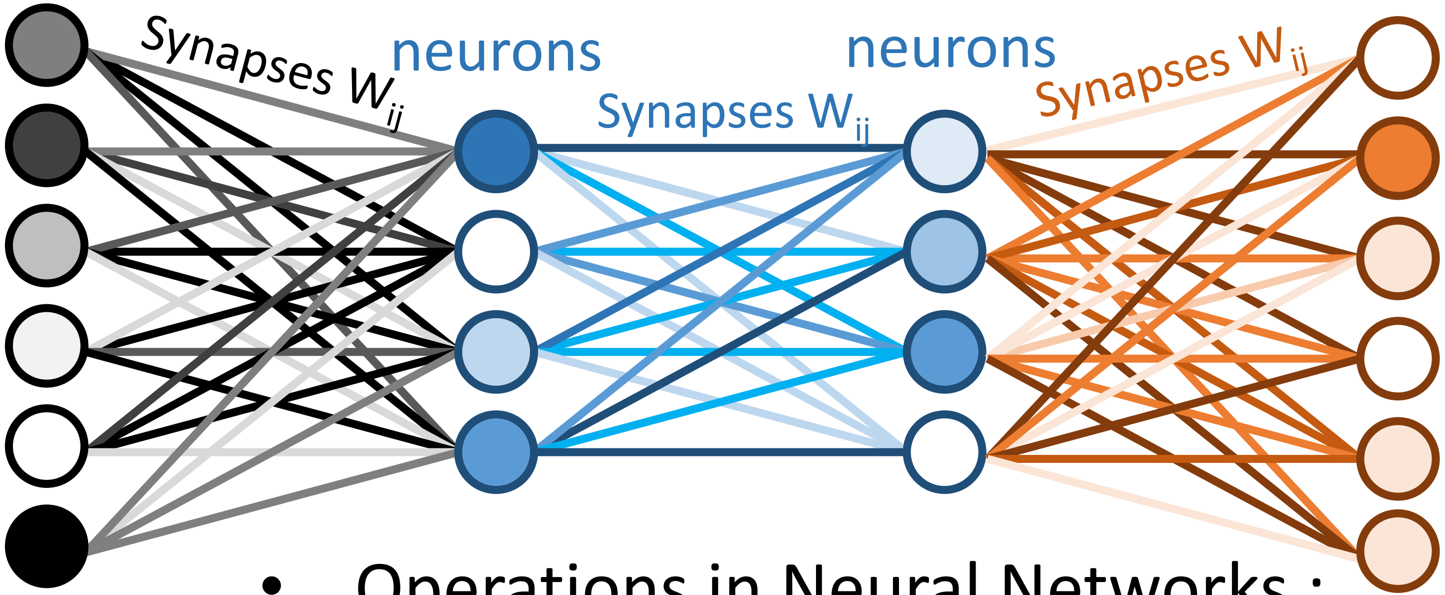
NEURAL NETWORKS seems Especially Adapted for In Memory Computing

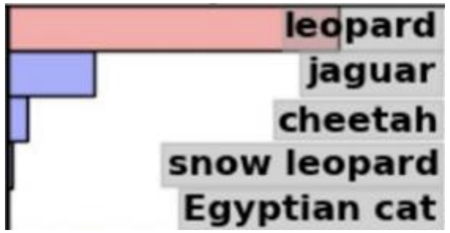
input neurons

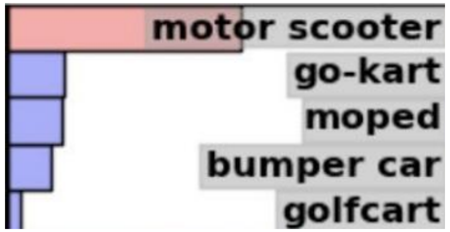
hidden neurons

hidden neurons

output neurons

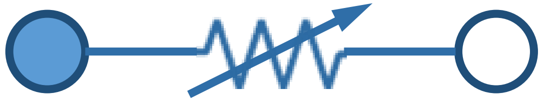






- Operations in Neural Networks :

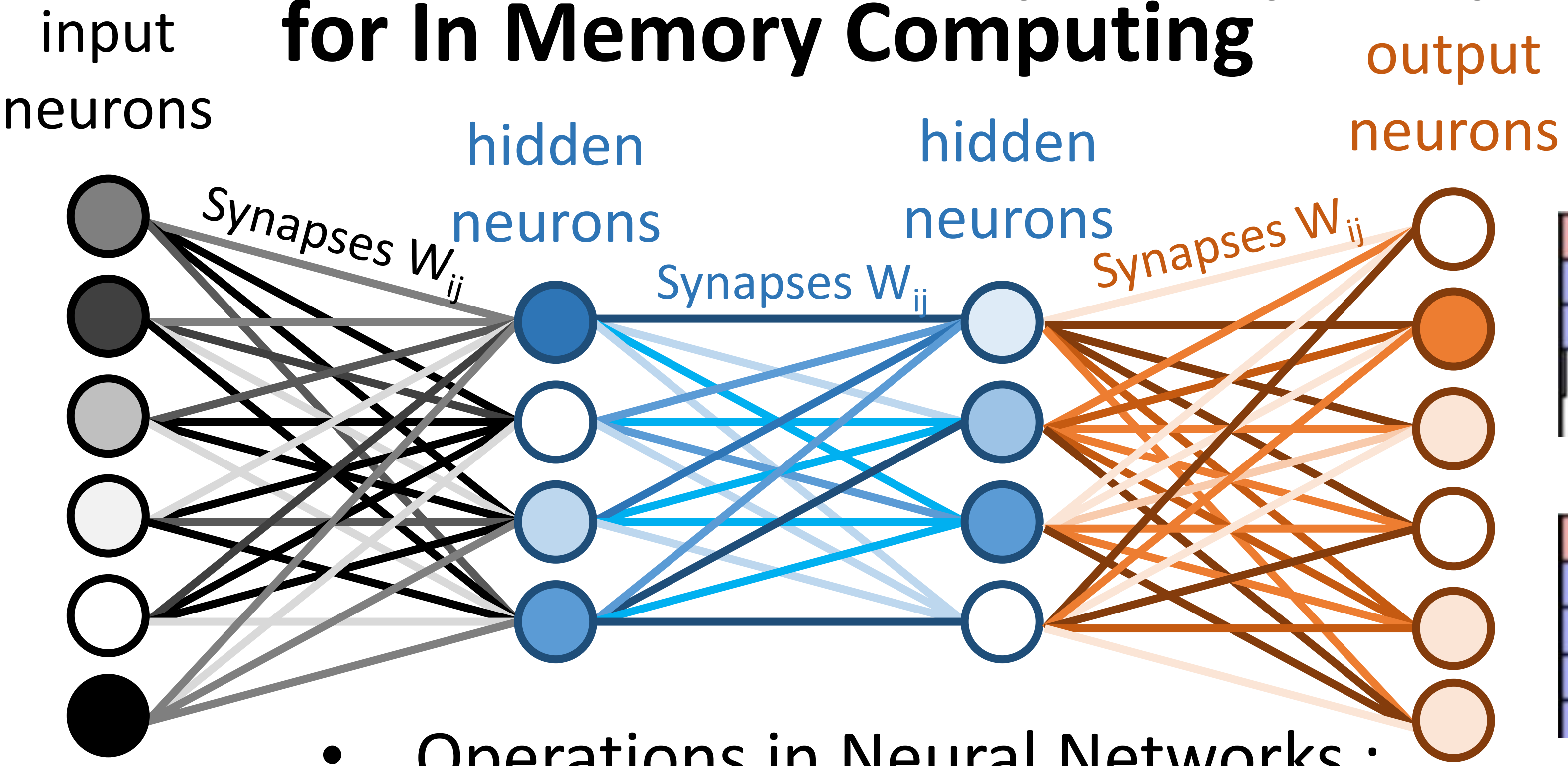
Multiplication → Accumulation → Non-linear function



Lot of research on RRAM as analog synapse

$$a_i = f\left(\sum_j W_{ij} \,.\, a_j\right)$$

NEURAL NETWORKS seems Especially Adapted for In Memory Computing



leopard
jaguar
cheetah
snow leopard
Egyptian cat

motor scooter
go-kart
moped
bumper car
golfcart

- Operations in Neural Networks :

Multiplication → Accumulation → Non-linear function

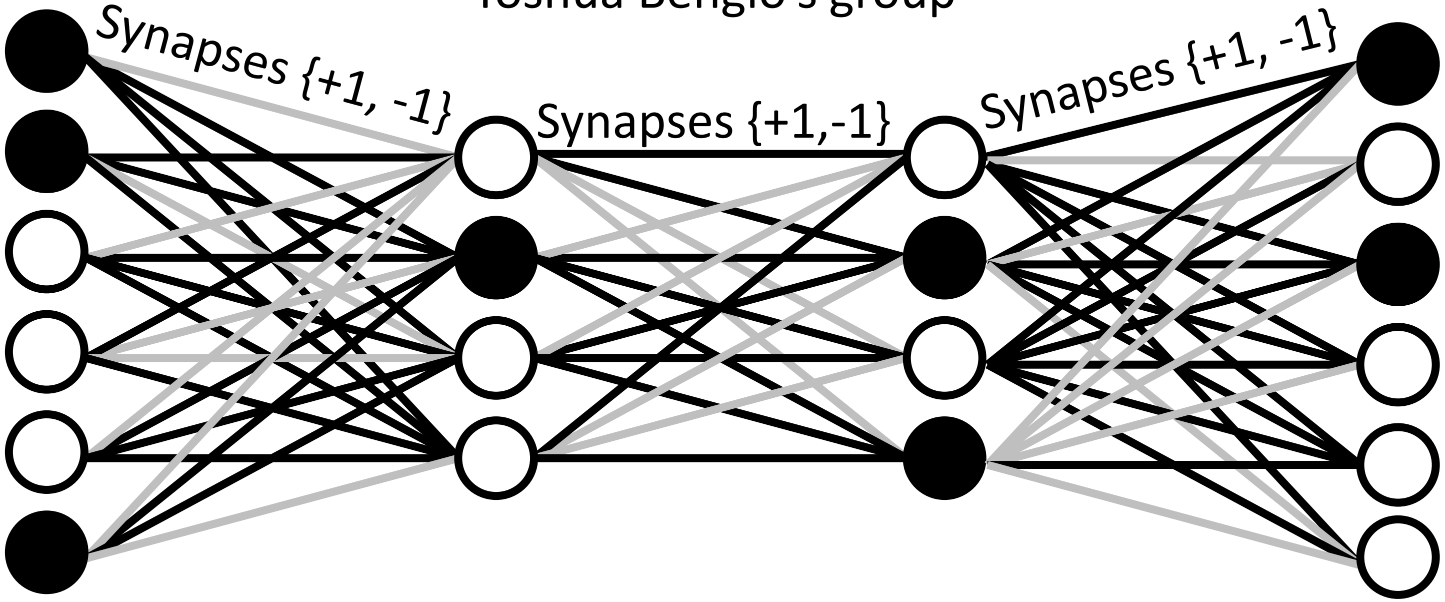
~~Lot of research on RRAM~~
as analog synapse

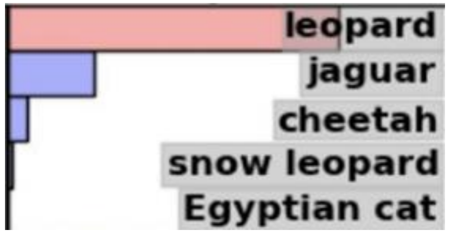
Our work :
RRAM as binary synapse

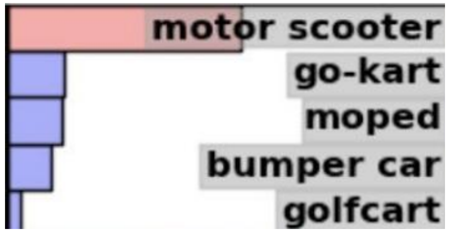
Recent Breakthrough: Binarized Neural network

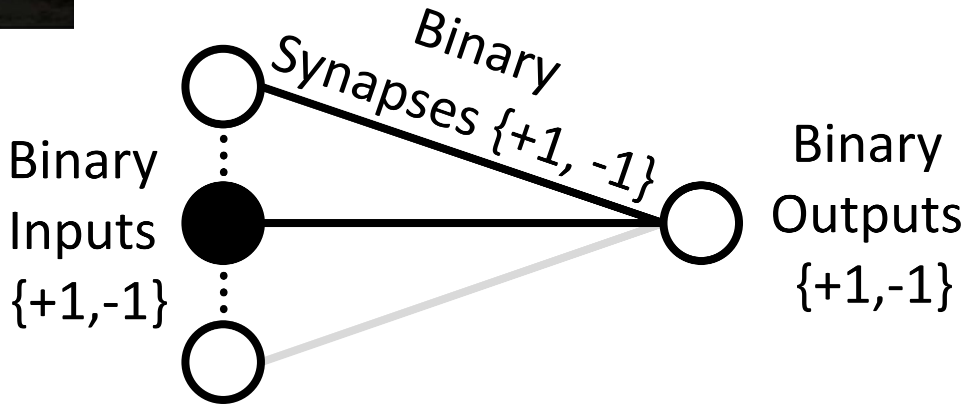
Hubara, Courbariaux et al. NIPS 2016

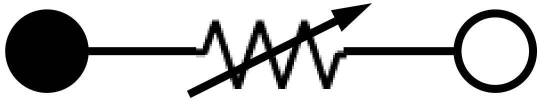
Yoshua Bengio's group











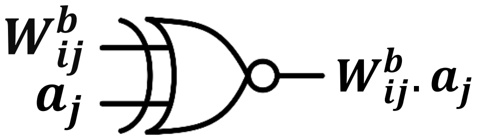
Our work :
RRAM as binary synapse

Binarized Neural network: Very Simple Logical Operations

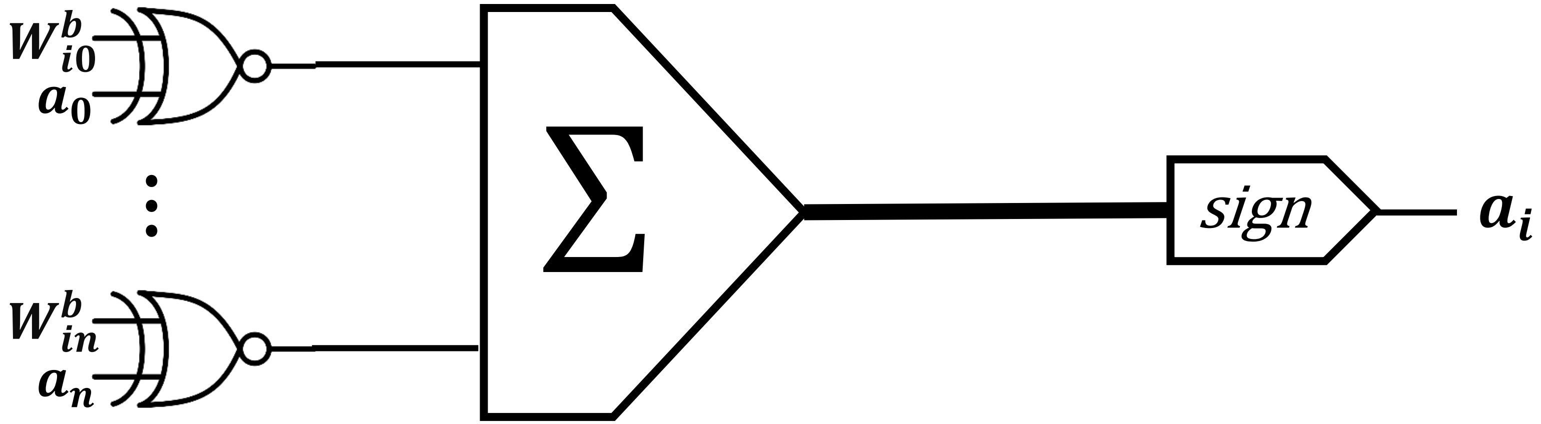
Multiplication → Accumulation → Non-linear function

$$\boldsymbol{W_{ij}\,.\,a_j} \qquad \boldsymbol{\sum_j W_{ij}\,.\,a_j} \qquad \boldsymbol{a_i = f\left(\sum_j W_{ij}\,.\,a_j\right)}$$

XNOR → Bitcount → Sign

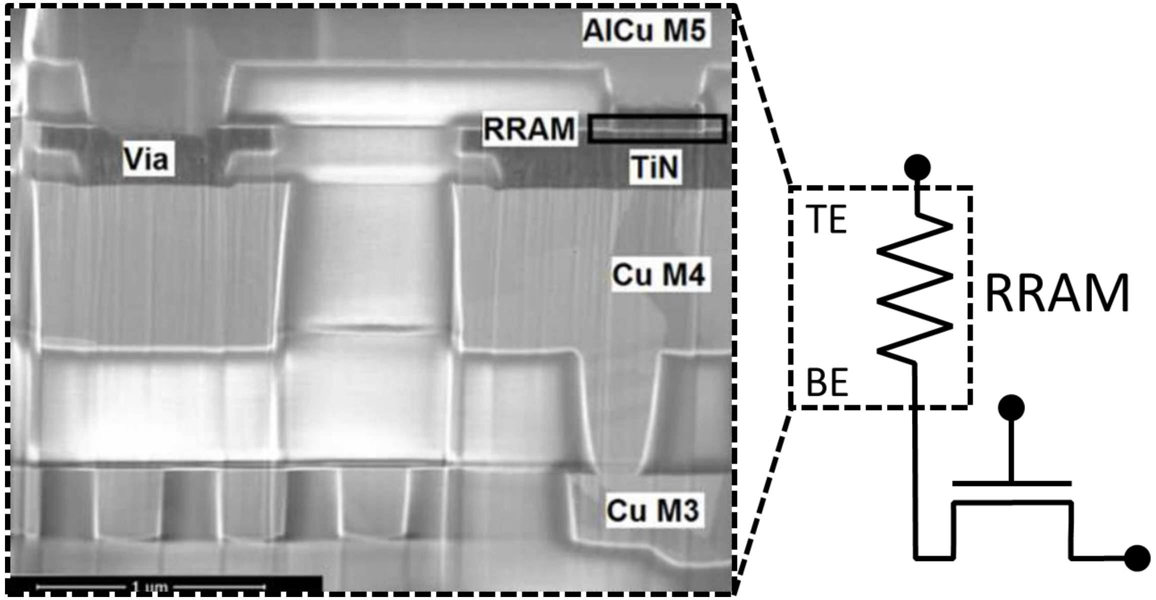


a_j	W^b_{ij}	$W^b_{ij}.\,a_j$
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1



RRAM Technology Involved

- **HfO_2-based OxRAM** integrated in a 130 nm CMOS logic process
- Fast / High retention time / High endurance

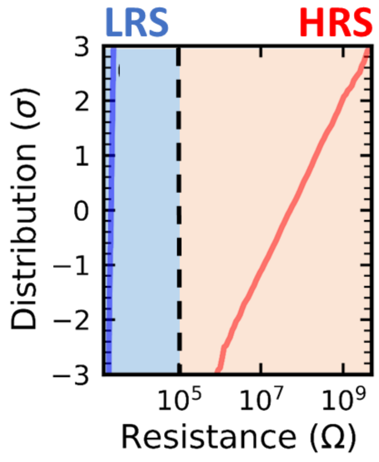
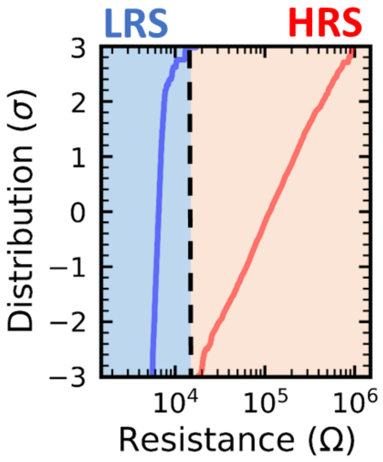
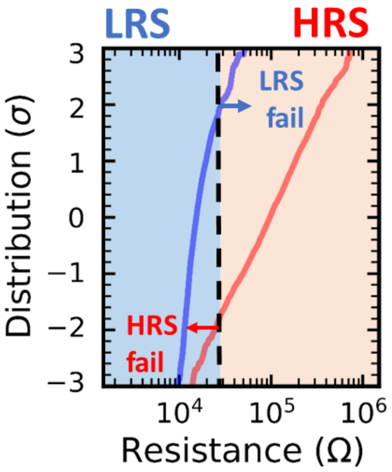


SEM cross-section – $TiN/HfO_2/Ti/TiN$

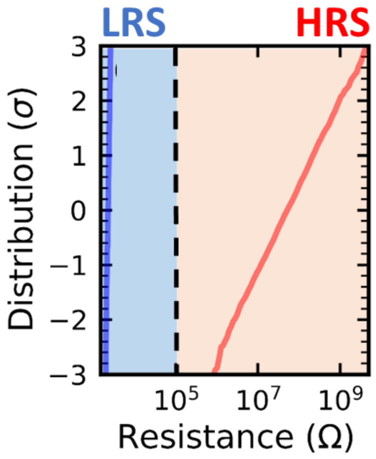
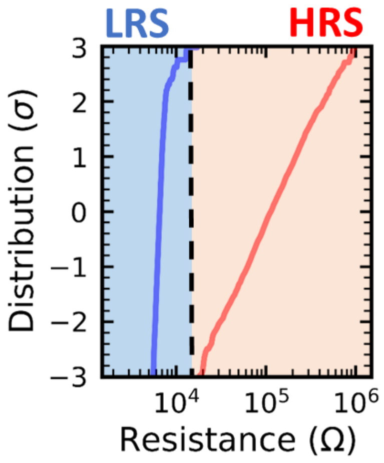
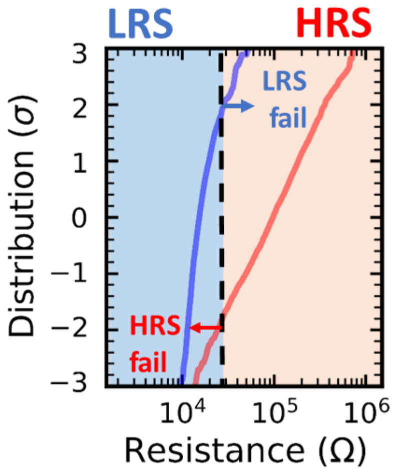
LETI MPW Shuttle managed through CMP (today in 200mm; twice a year; next start may 2019)

RRAM Challenge ⟶ **high variability**

RRAM Variability Depends Extensively on the Programming Regime

Programming condition	Very strong	Strong	Weak
SET compliance current	600μA	55μA	20μA
RESET voltage	2.5V	2.5V	1.5V
Resistance Distribution			
Bit error rate (1T1R)	$< 10^{-6}$	9.7×10^{-5}	3.3×10^{-2}
Prog. Energy (SET/RESET)	120/150*pJ*	11/14*pJ*	4/5*pJ*
Cyclability	100	> 10000	$> 10^6$

RRAM Variability Depends Extensively on the Programming Regime

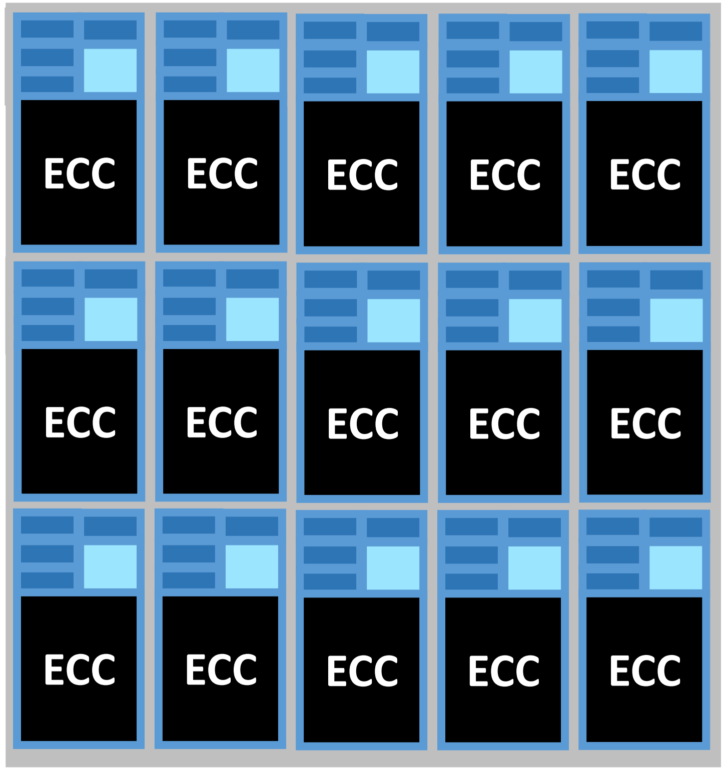
Programming condition	Very strong	Strong	Weak
SET compliance current	600μA	55μA	20μA
RESET voltage	2.5V	2.5V	1.5V
Resistance Distribution			
Bit error rate (1T1R)	$< 10^{-6}$	9.7×10^{-5}	3.3×10^{-2}
Prog. Energy (SET/RESET)	120/150*pJ*	11/14*pJ*	4/5*pJ*
Cyclability	100	> 10000	$> 10^{6}$

Exploiting weak programming conditions for BNN

How to deal with bit errors?

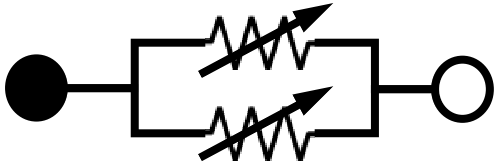
- Classical Approach : Error Correction Code

→ **Incompatible with in memory computing**



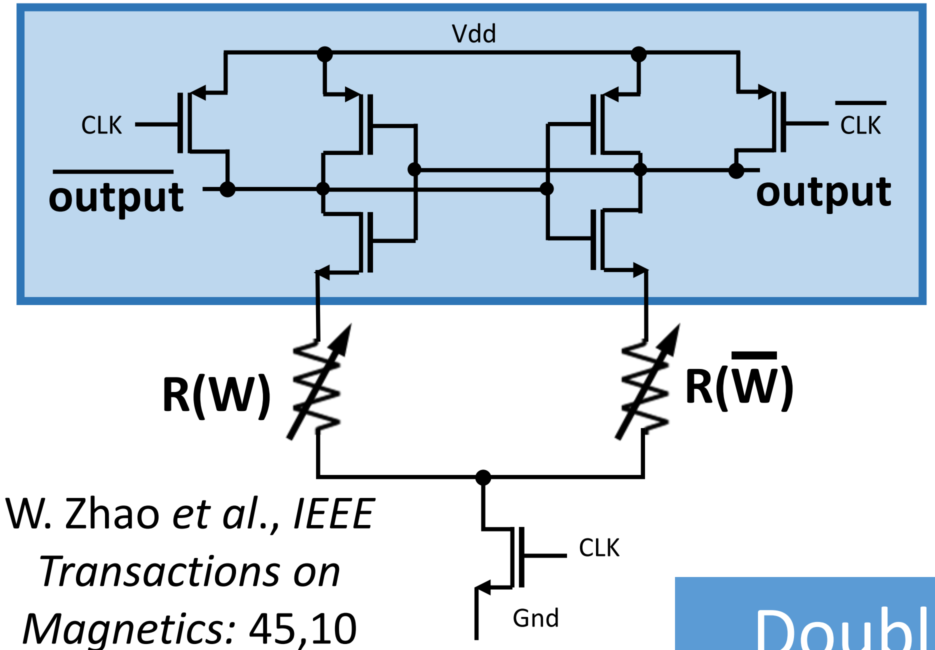
ECC leads to a big overhead

Our approach : Two RRAM Devices as One Binary Synapse to reduce bit error rate



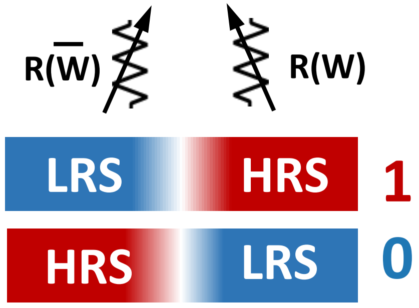
- Classical circuit to differentiate resistance state

Pre-Charge Sense Amplifier (PCSA)

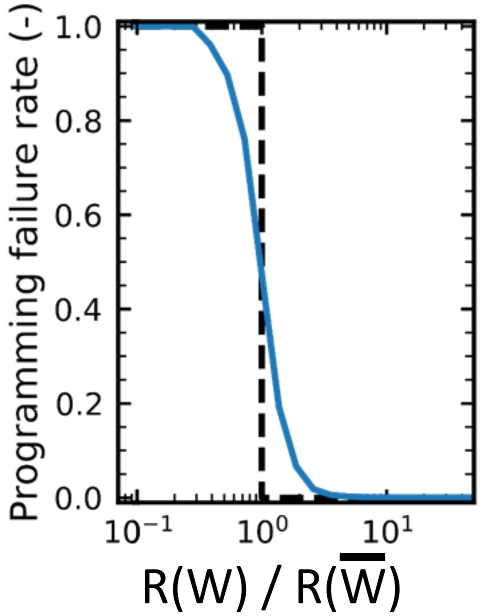


W. Zhao *et al.*, *IEEE Transactions on Magnetics:* 45,10 (2009)

- Devices programmed in a complementary fashion :



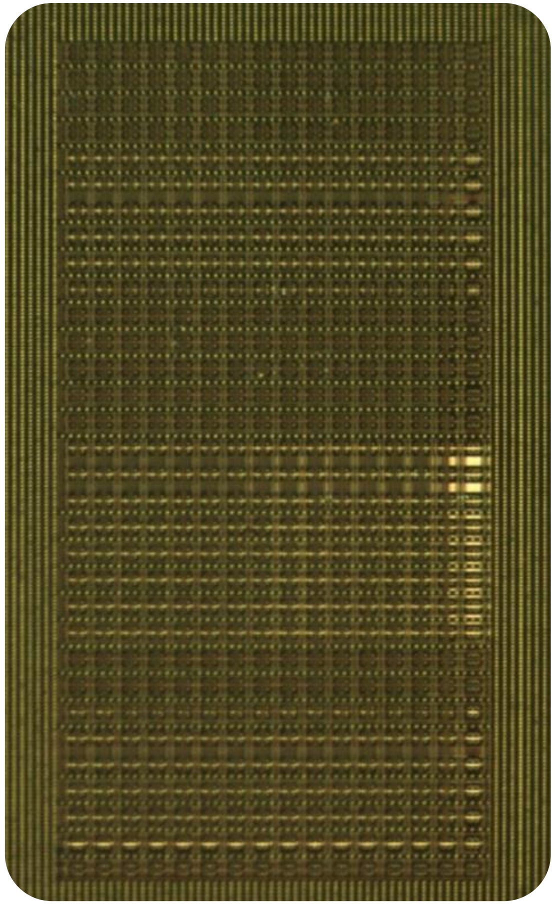
- Reading circuit behavior



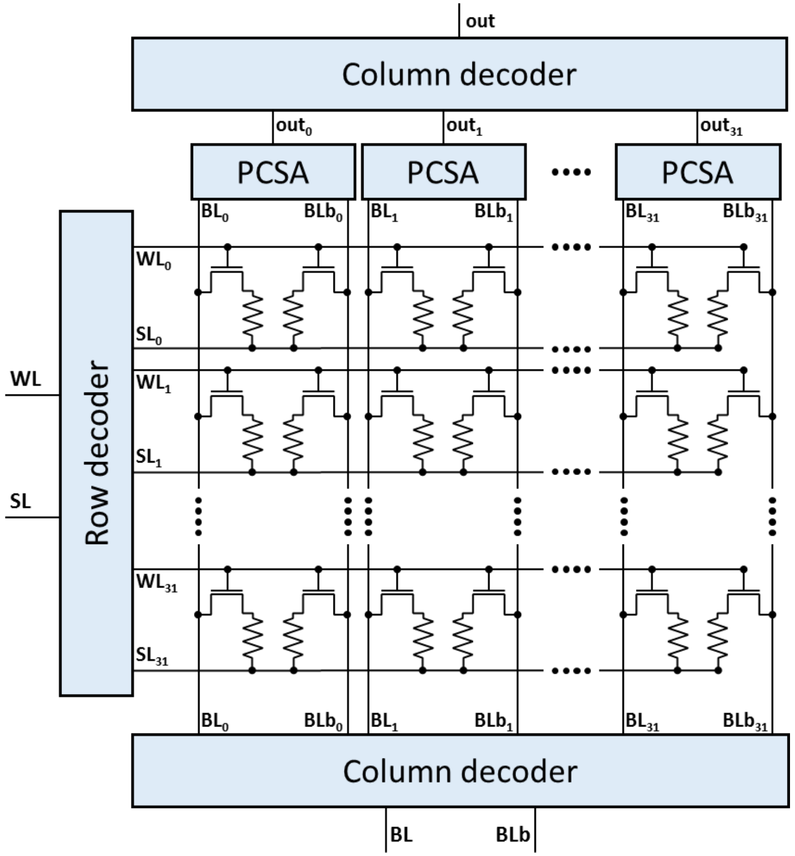
Double the amount of memory

Array structure: 2kBits devices

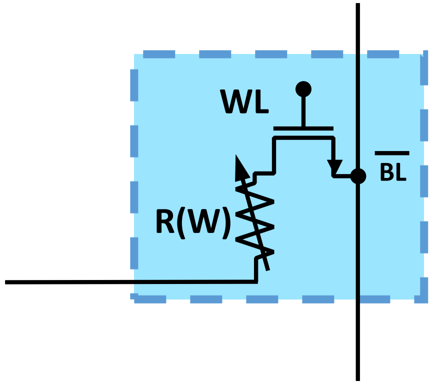
- Photograph of our circuit



- Schematic of the array

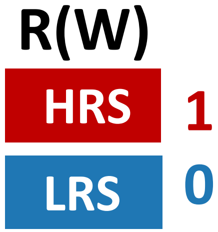


What if we used
1 Transistor 1 Resistor (1T1R) array structure ?



RRAM matrix (1T1R)

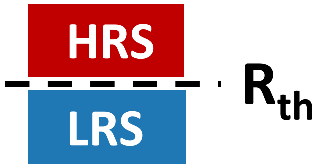
Devices resistance states :


R(W)

HRS 1

LRS 0

Resistance threshold between the two resistance states :

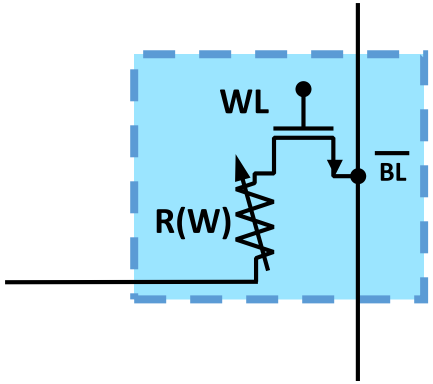

HRS

LRS

R_{th}

Big memory window

1 Transistor 1 Resistor (1T1R) is Prone to Errors



RRAM matrix (1T1R)

Devices resistance states :

R(W)

HRS	1
LRS	0

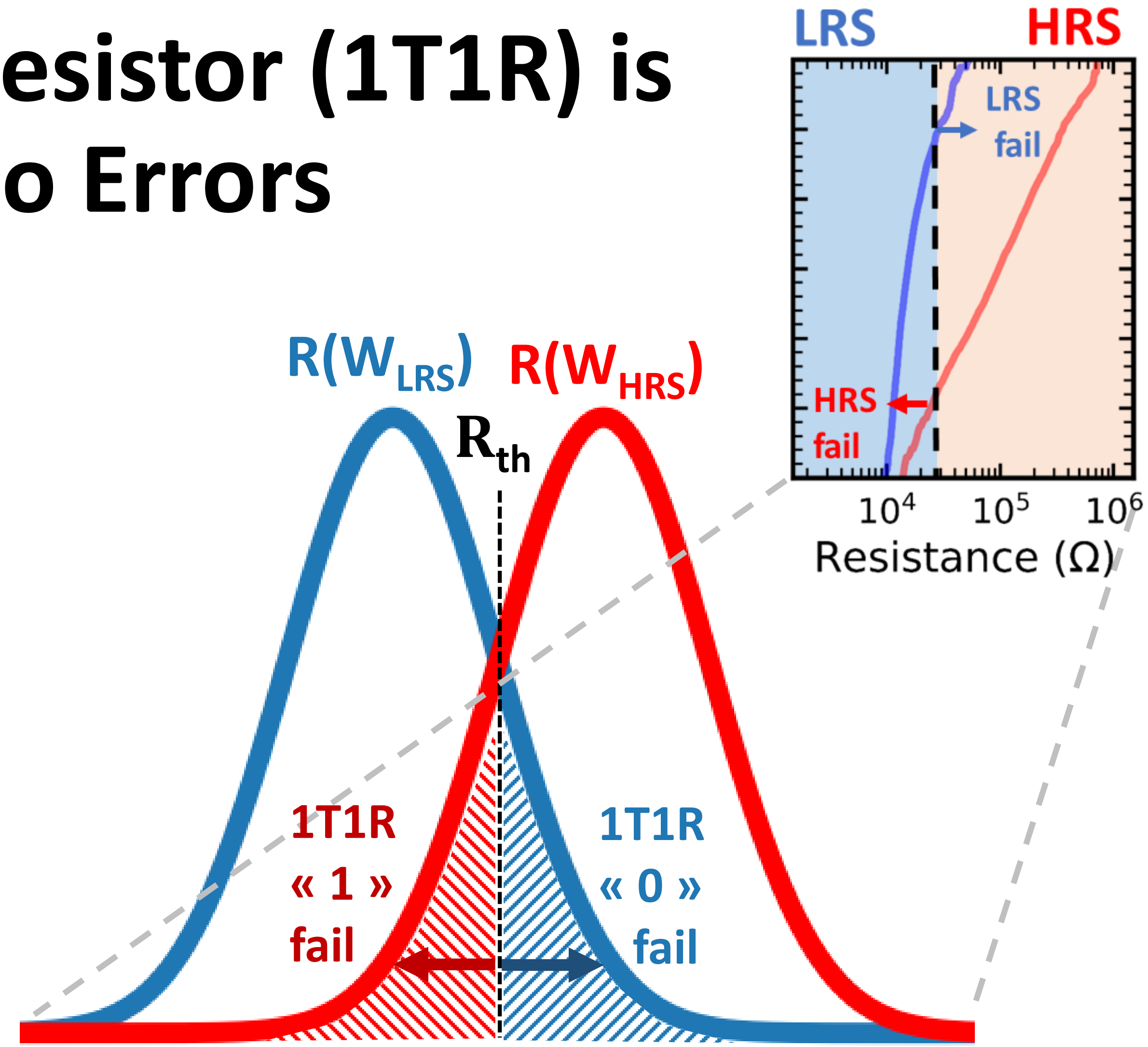
Resistance threshold between the two resistance states :

HRS

- - - - - R_{th}

LRS

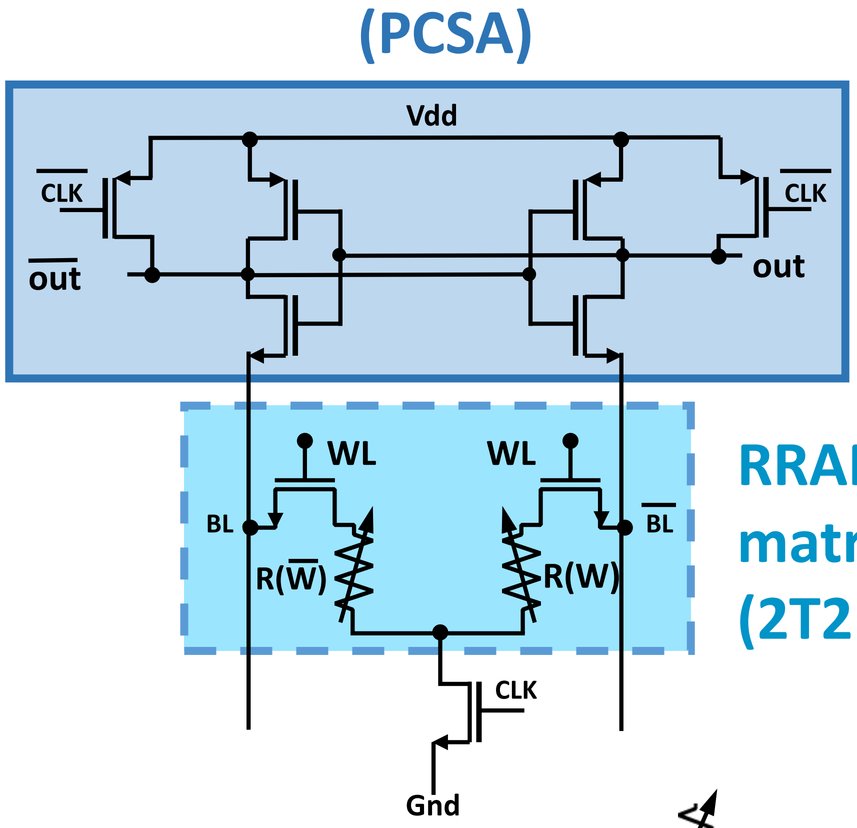
Resistances States overlap



- Errors when :

$R(W_{LRS}) > Rth$ or $R(W_{HRS}) < Rth$

2 Transistors 2 Resistors (2T2R) configuration :

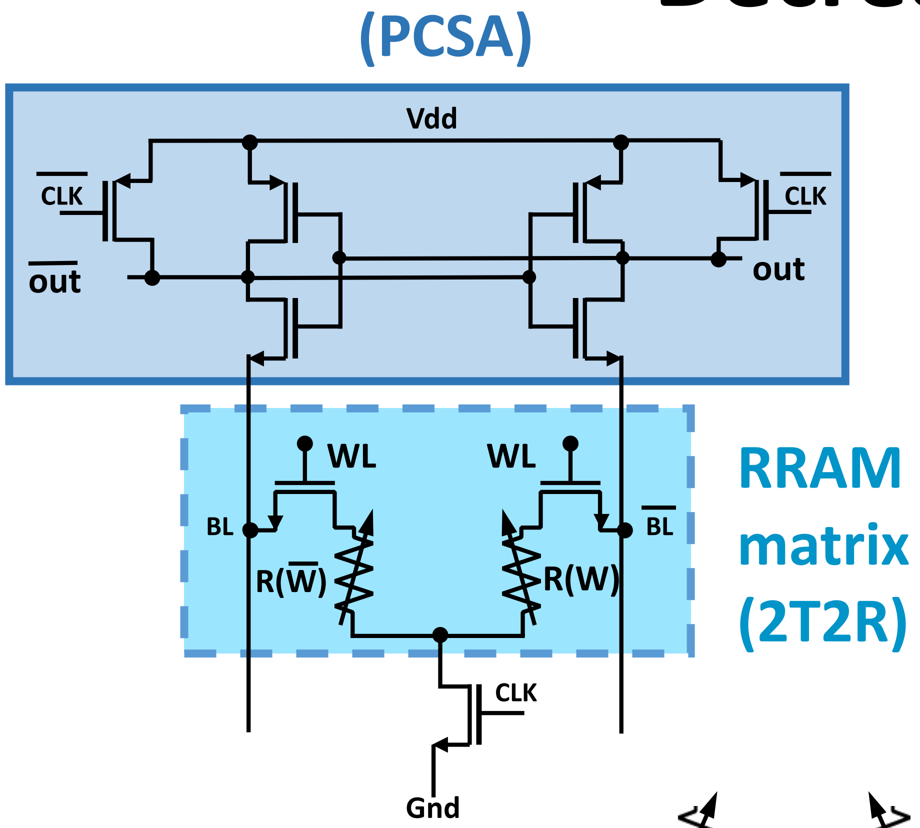


RRAM matrix (2T2R)

Devices programmed in a complementary fashion :

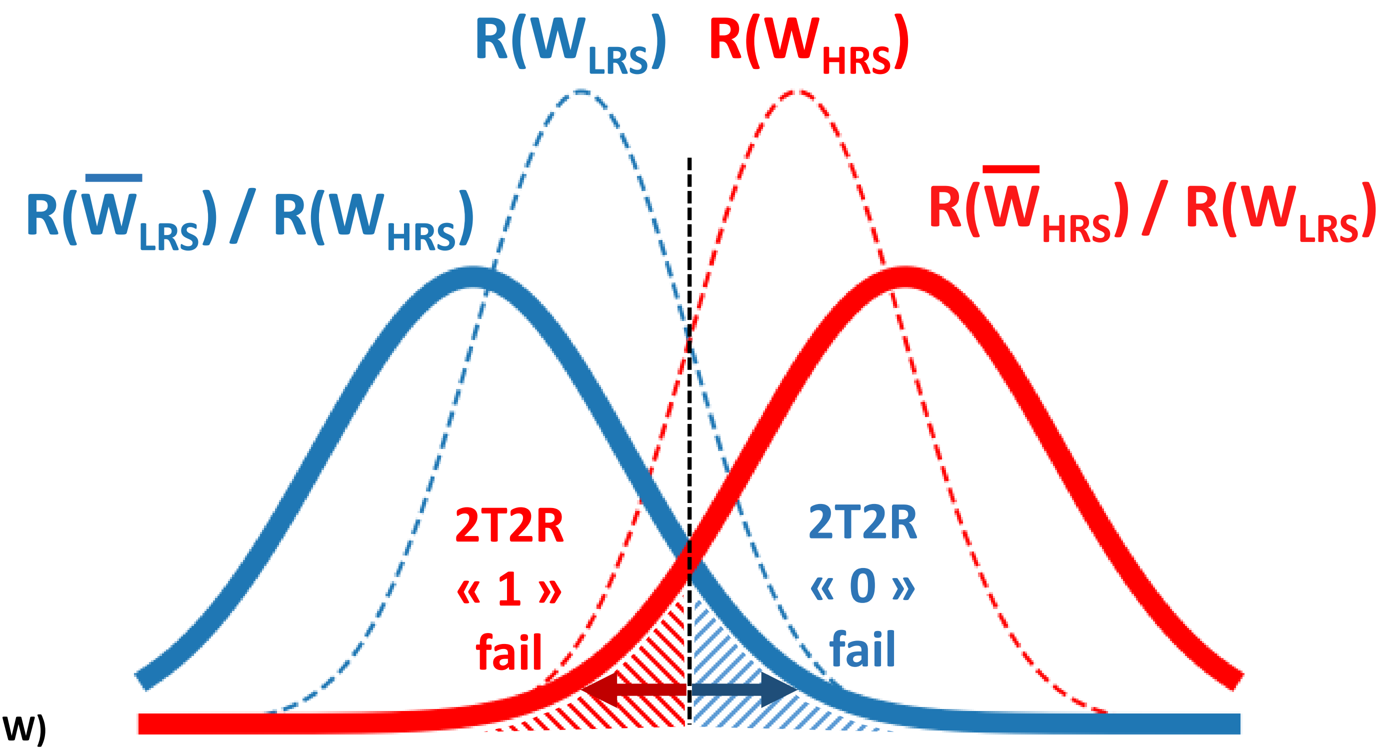
$R(\overline{W})$	R(W)	
LRS	HRS	1
HRS	LRS	0

2 Transistors 2 Resistors (2T2R): Decreasing Error Rate



Devices programmed in a complementary fashion :

$R(\overline{W})$	$R(W)$	
LRS	HRS	1
HRS	LRS	0

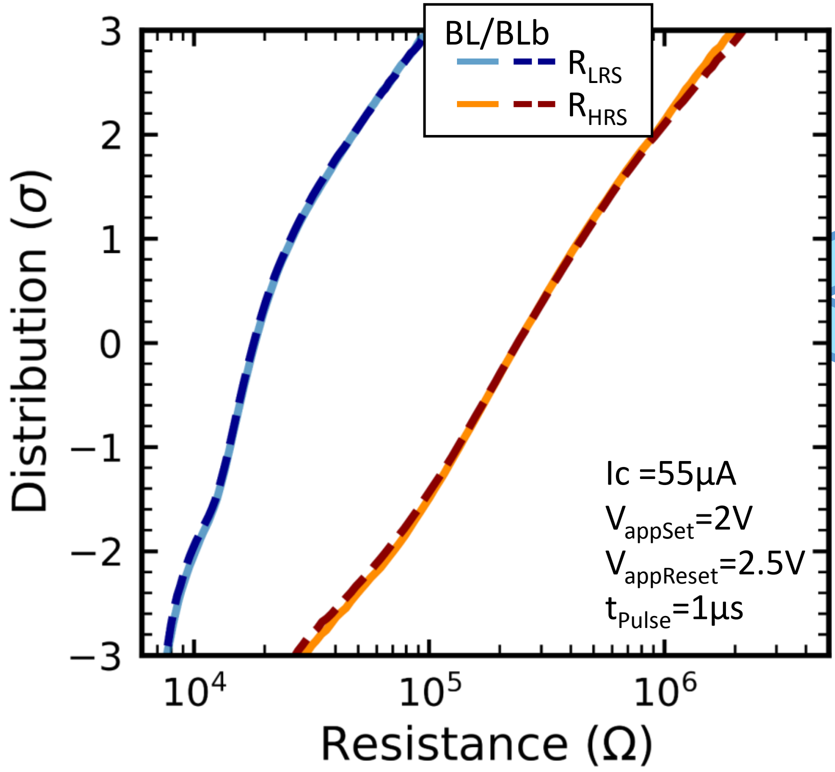


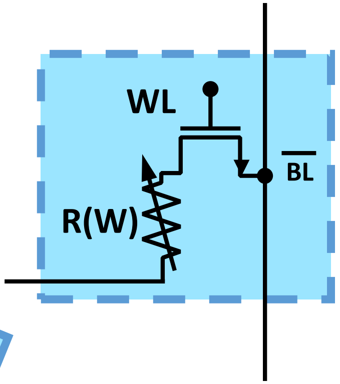
- Errors when :

$R(\overline{W}_{LRS}) > R(W_{HRS})$ or $R(\overline{W}_{HRS}) < R(W_{LRS})$

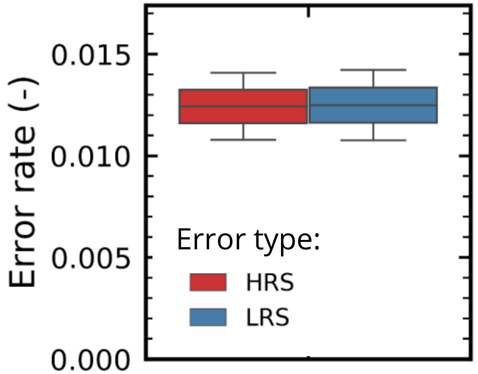
Comparison between 1T1R & 2T2R

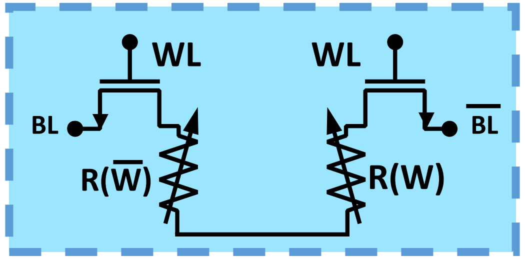
Bit error rates extraction



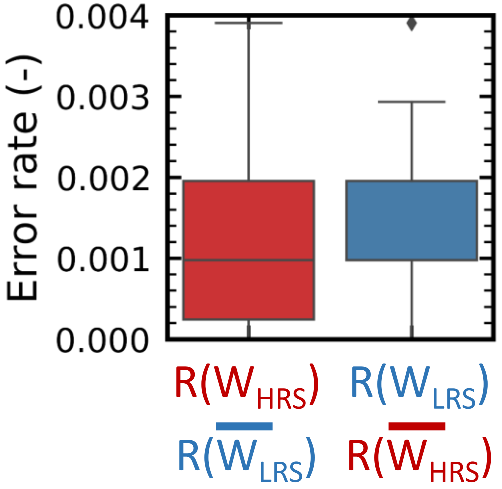


RRAM matrix (1T1R)



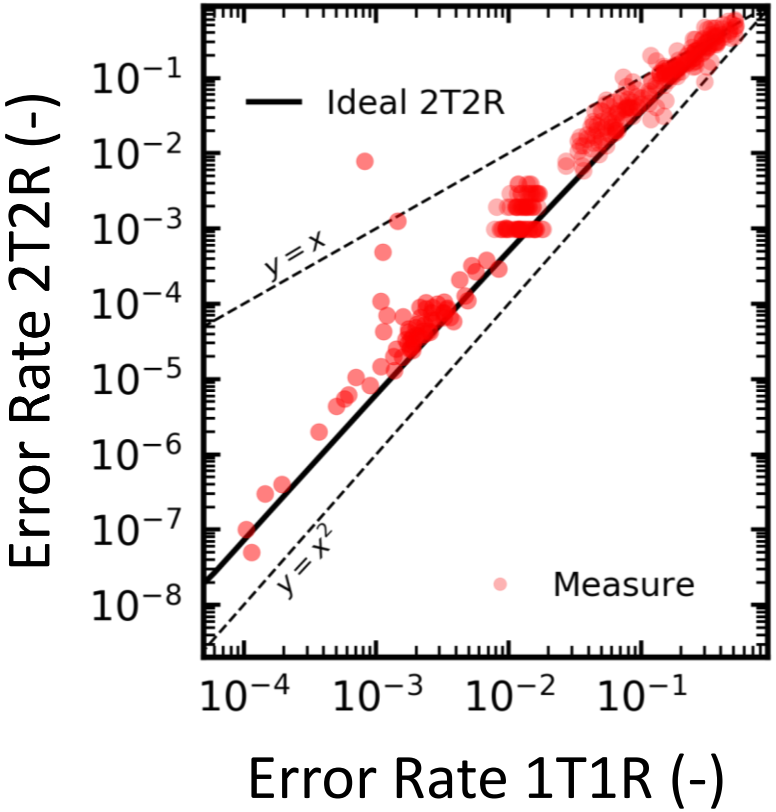


RRAM matrix (2T2R)



Comparison between 1T1R & 2T2R

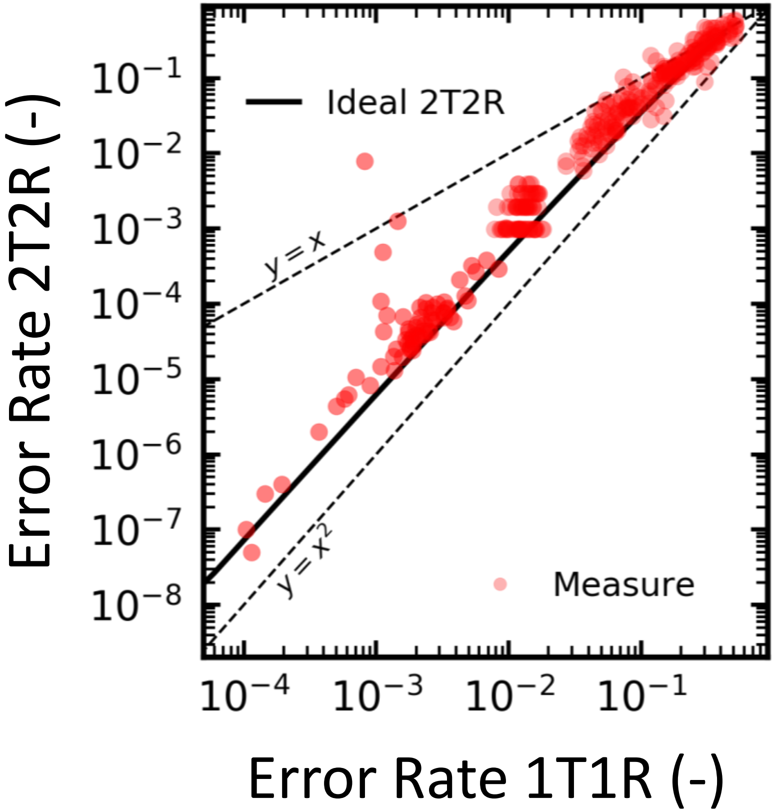
- Experimental bit error rate



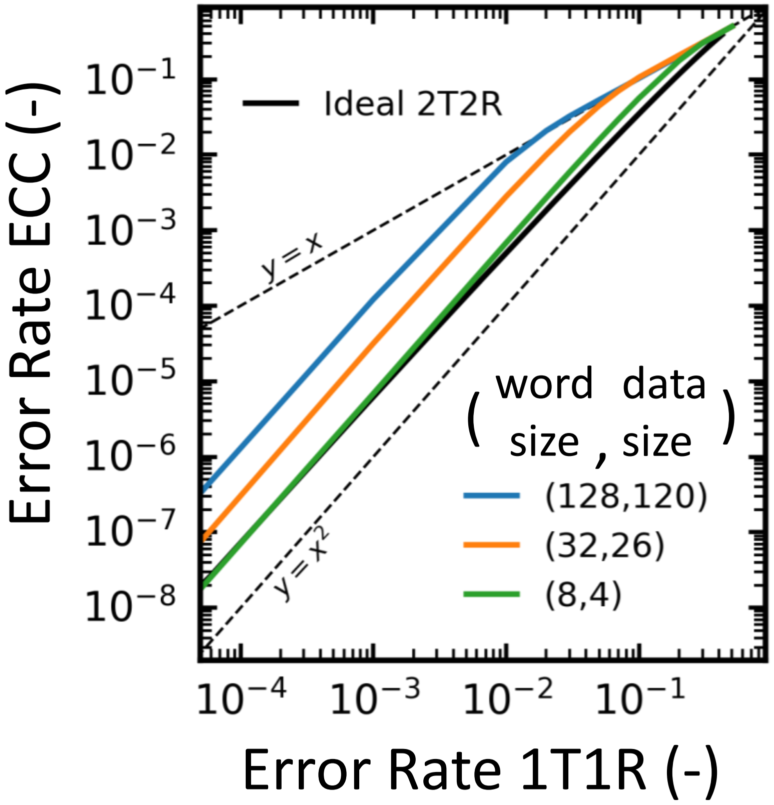
2T2R reduces error rate

Comparison with Error Correction Code

- Experimental bit error rate



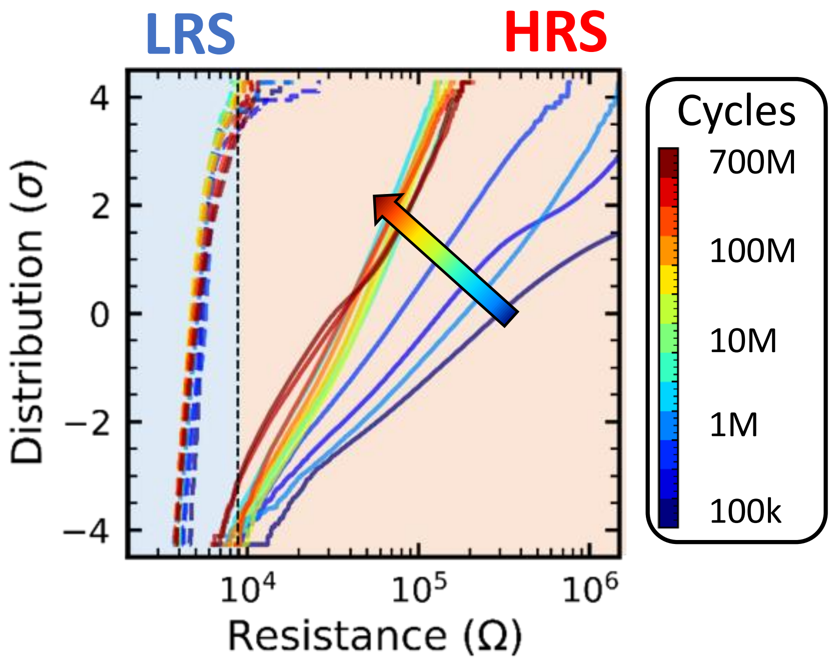
- Error Correction Code SECDED



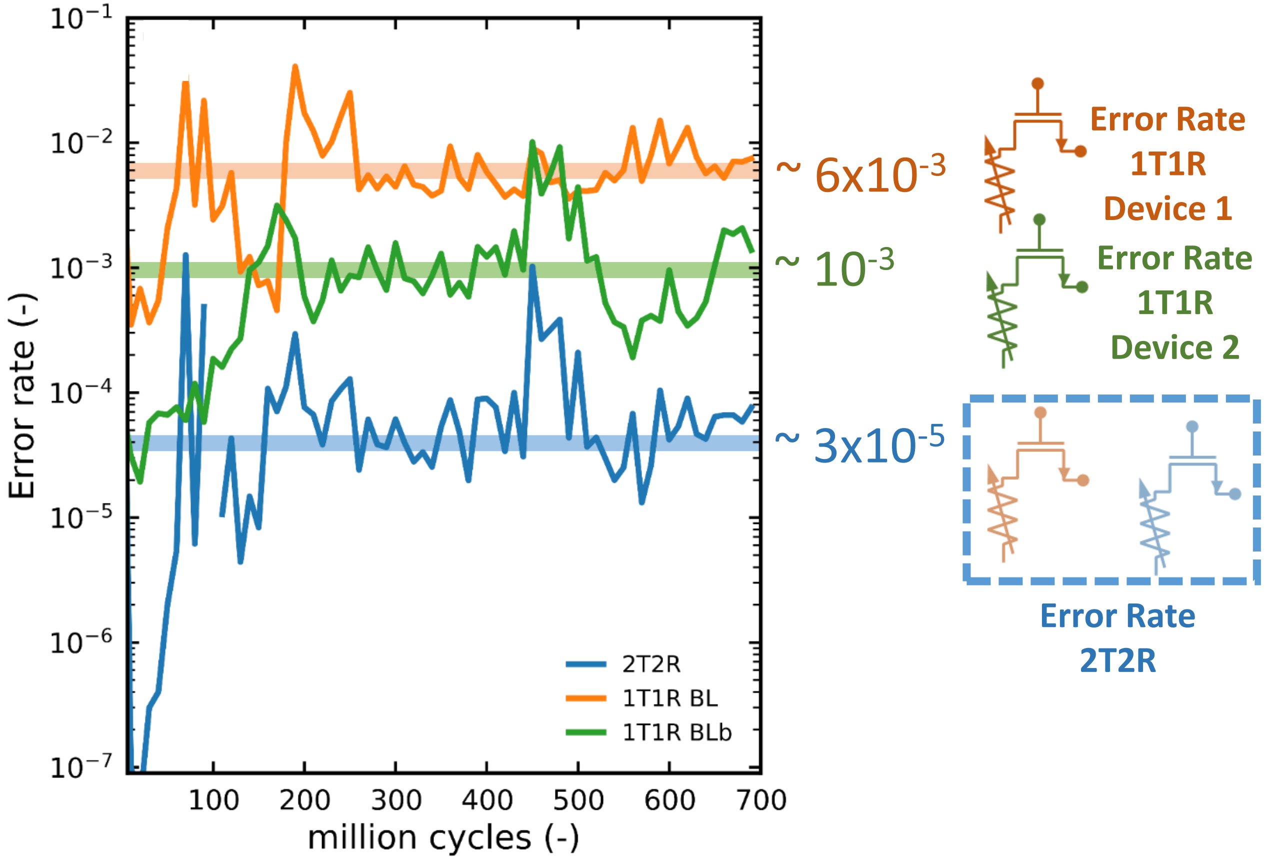
No need of ECC overhead

Characteristics over aging devices with weak programming conditions

- Distribution



- Error rate for 2 devices

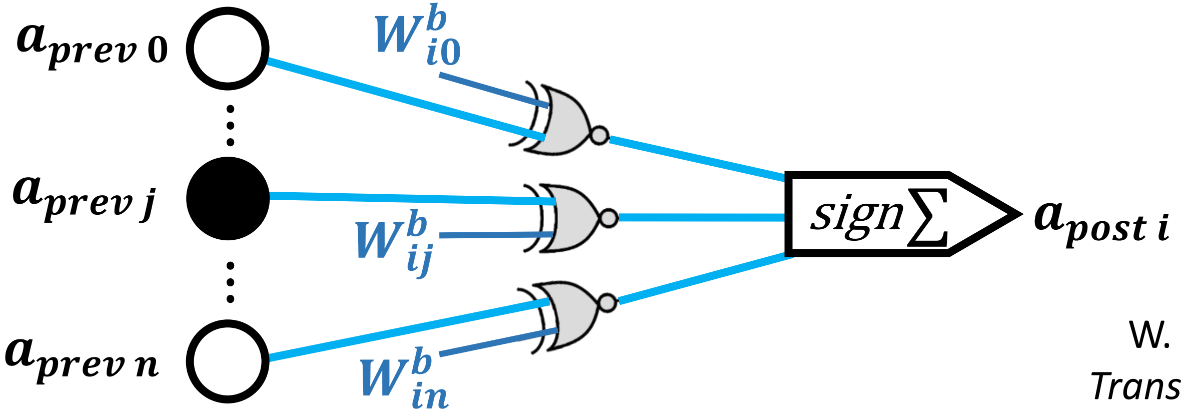


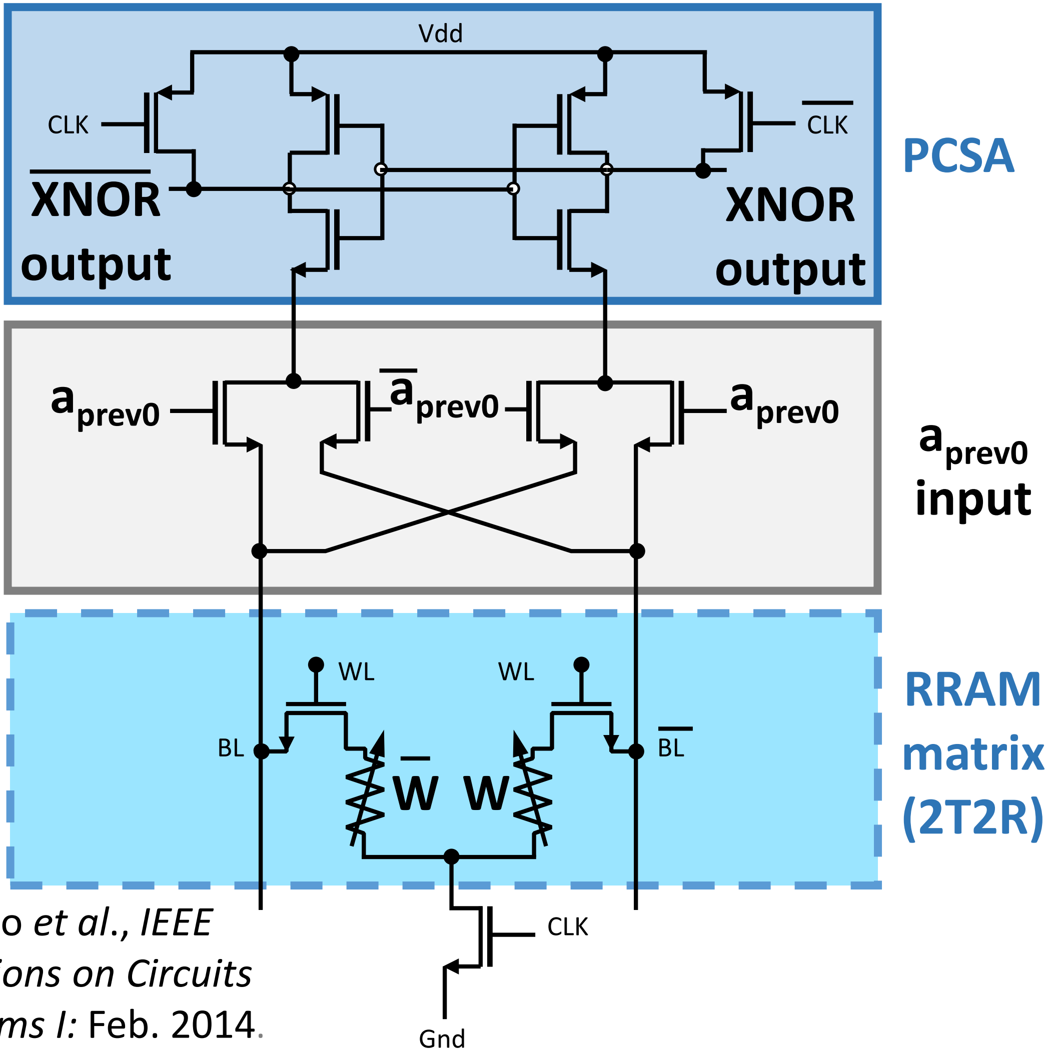
No need of ECC overhead

Logic in Memory Reading Circuit

- No ECC offers opportunity for in memory operation
- XNOR operation directly in PCSA circuit

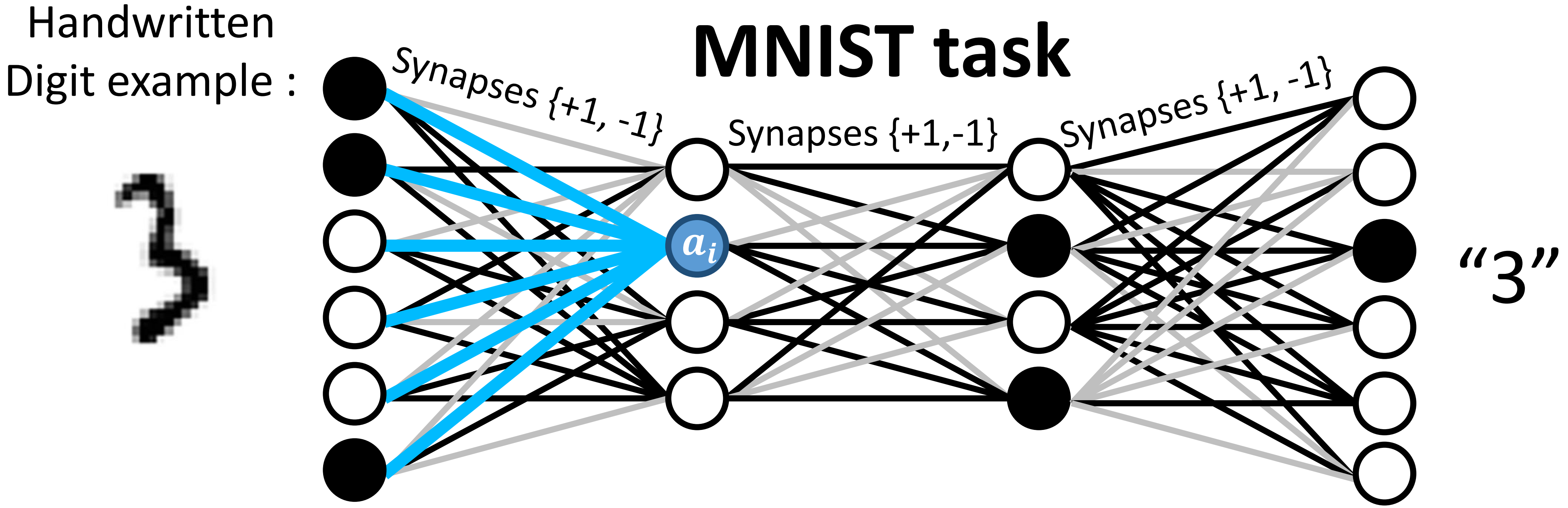
XNOR → Bitcount → Sign

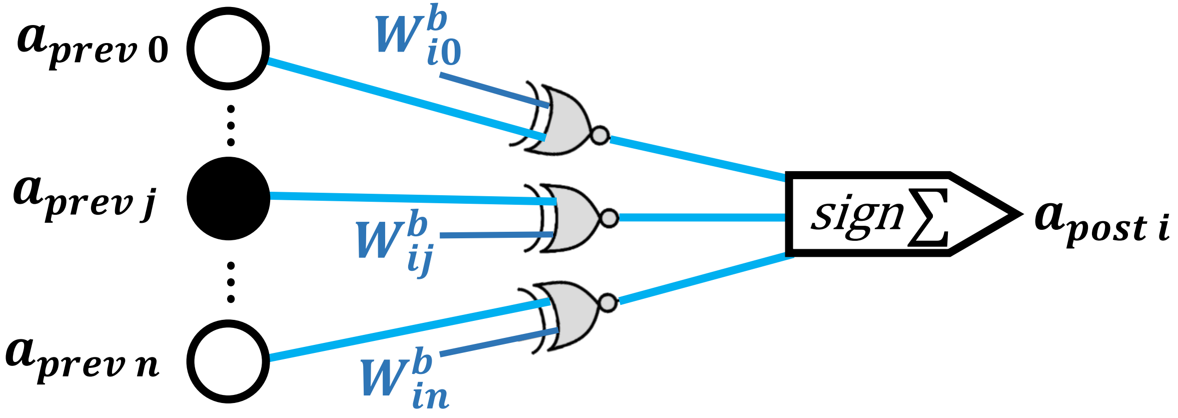


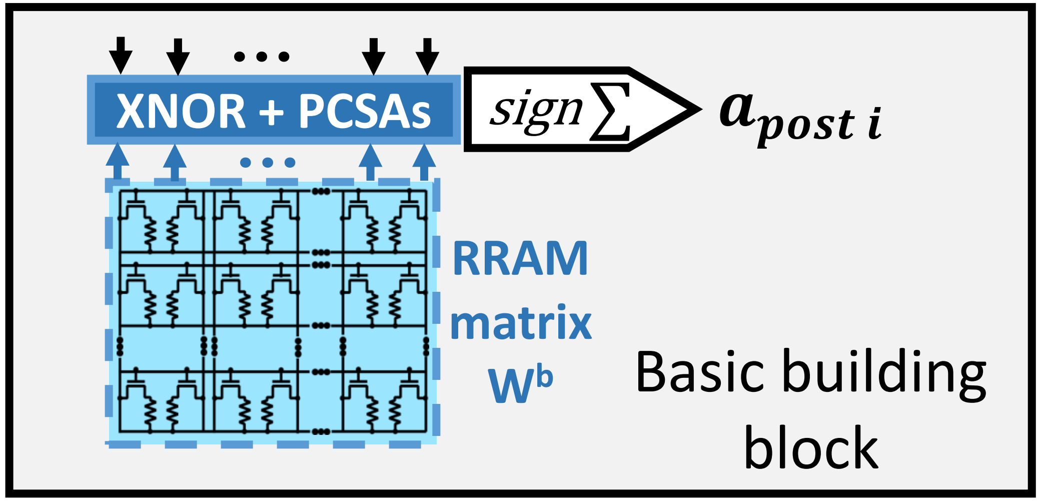


W. Zhao *et al.*, *IEEE Transactions on Circuits and Systems I:* Feb. 2014.

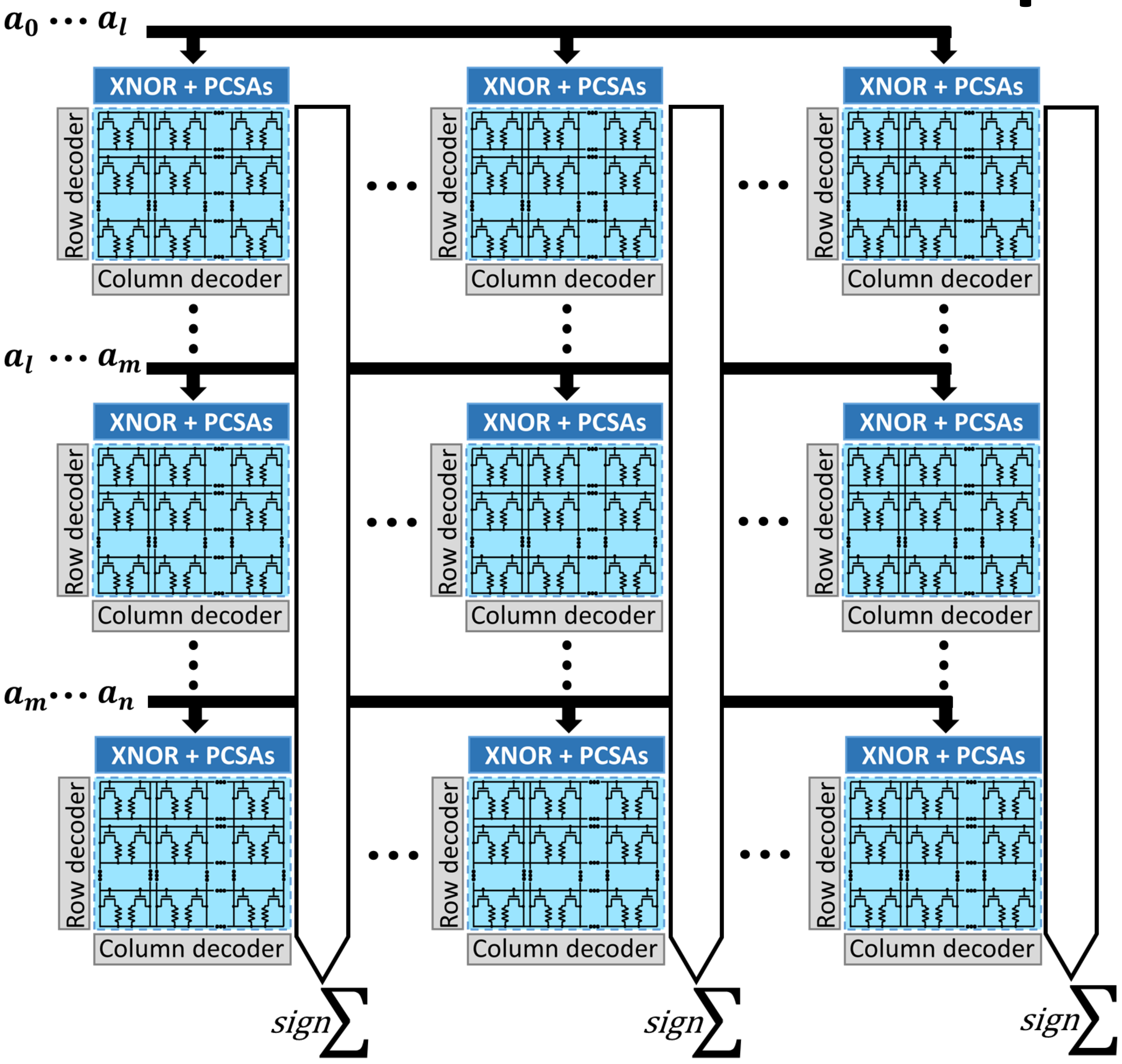
Binarized Neural Network: MNIST task







Hardware implementation



- Fully-connected BNN:
  2 layers with 1024 neurons each

RRAM In-memory BNN (this work)	**25nJ**
RRAM In-memory 8-bit fixed point	80nJ
Analog Phase Change Memory*	~56nJ
GPU (Tesla V100)	~μJ
CPU (Xeon E5)	~mJ

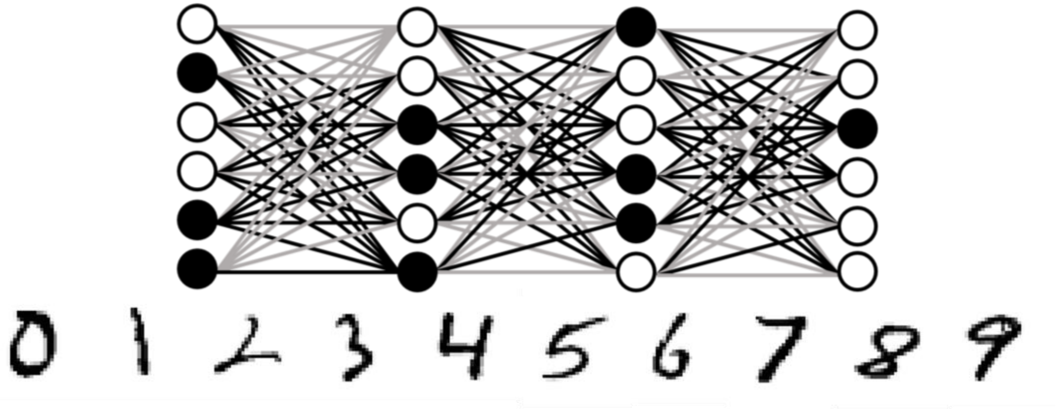
- Projection to 28 nm technology

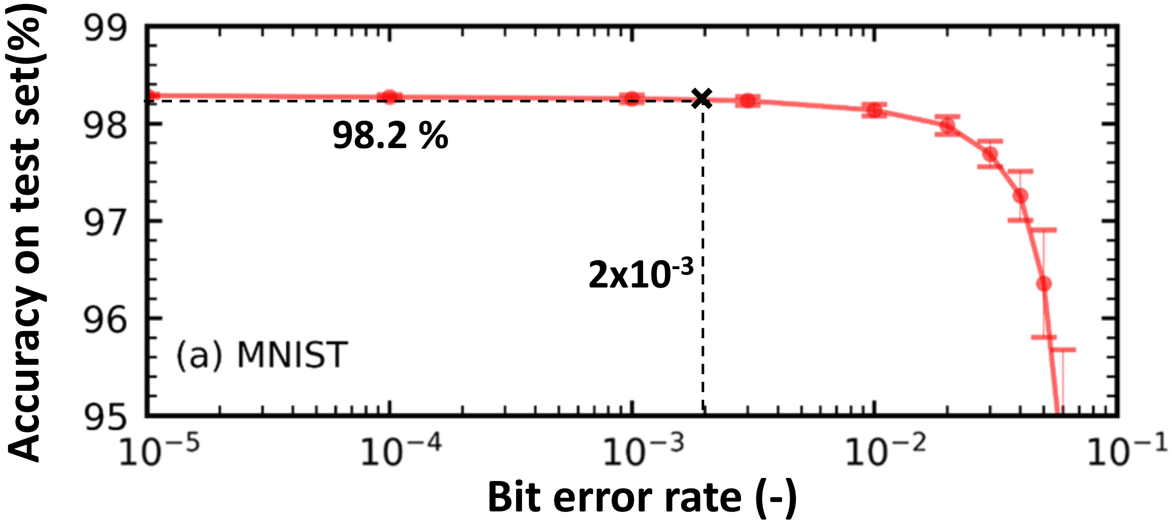
A technology ready
↳ low energy and low area

*Ambrogio, Stefano, et al. "Equivalent-accuracy accelerated neural-network training using analogue memory." *Nature* (2018)

Error evaluation on two different tasks

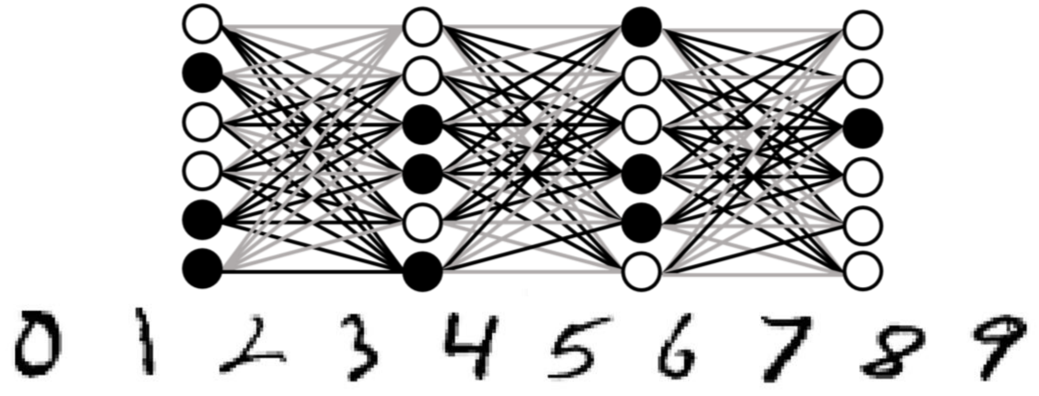
- MNIST with fully-connected NN

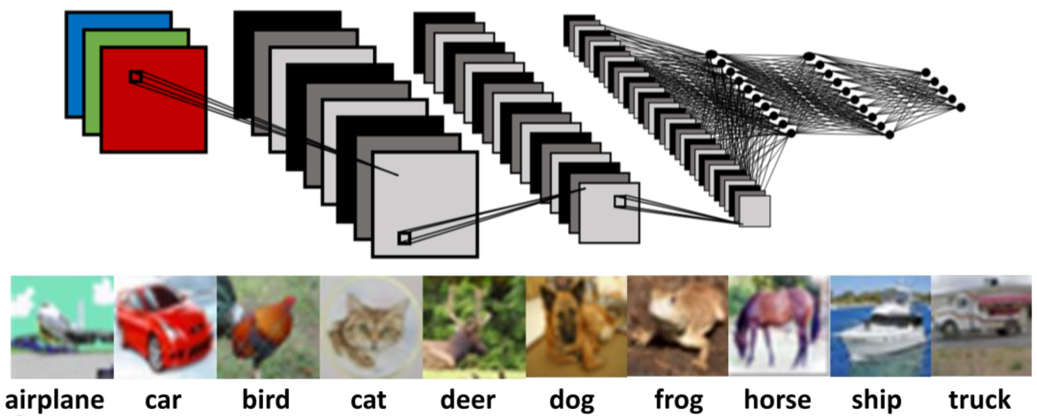


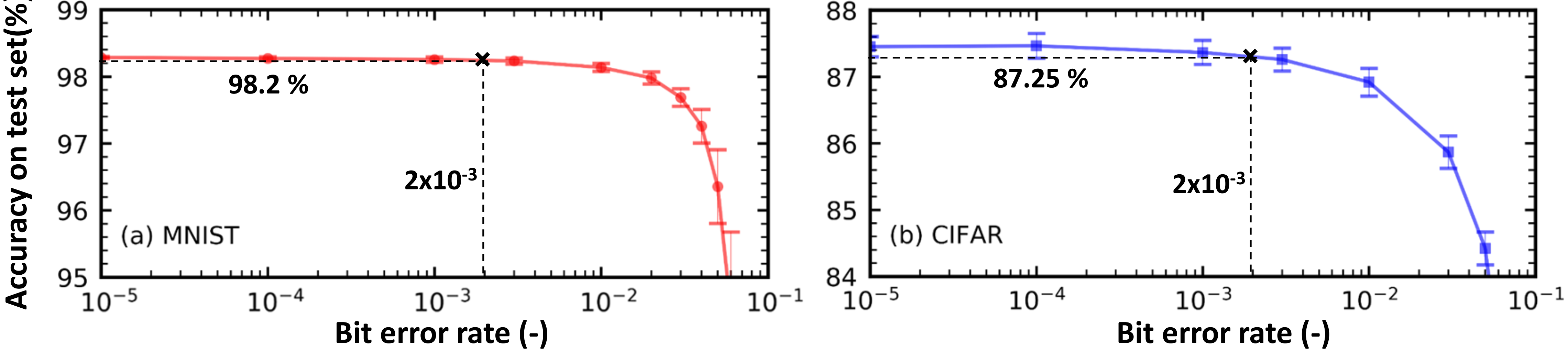


Error evaluation on two different tasks

- MNIST with fully-connected NN
- CIFAR 10 with Convolutional NN



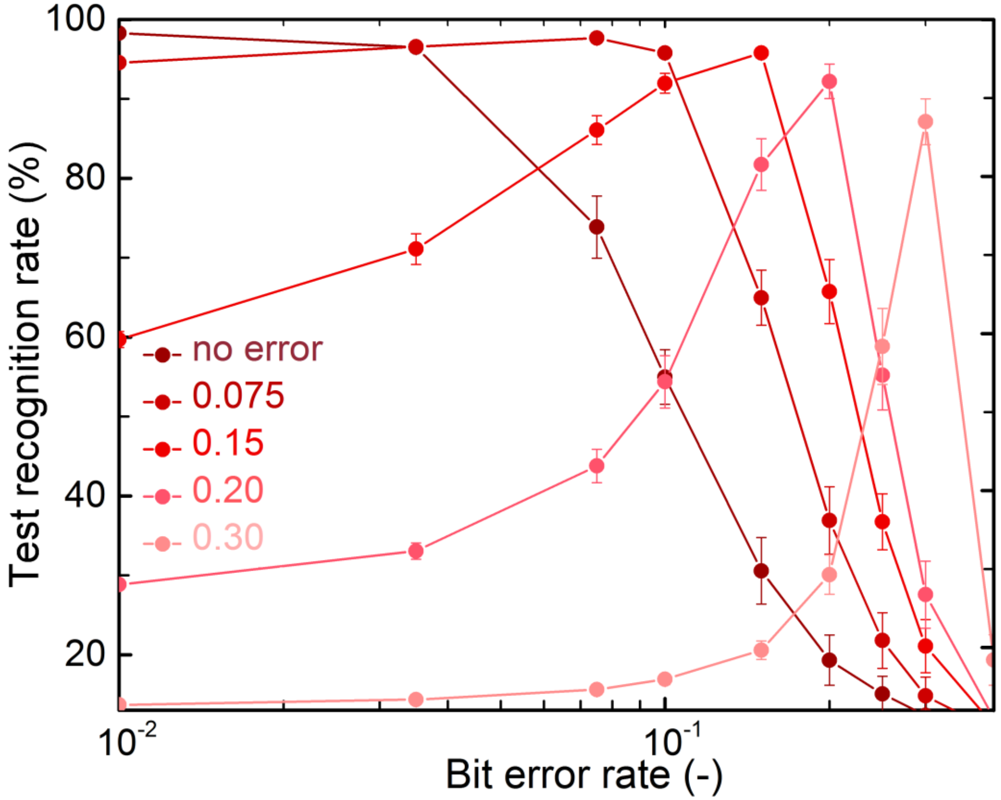




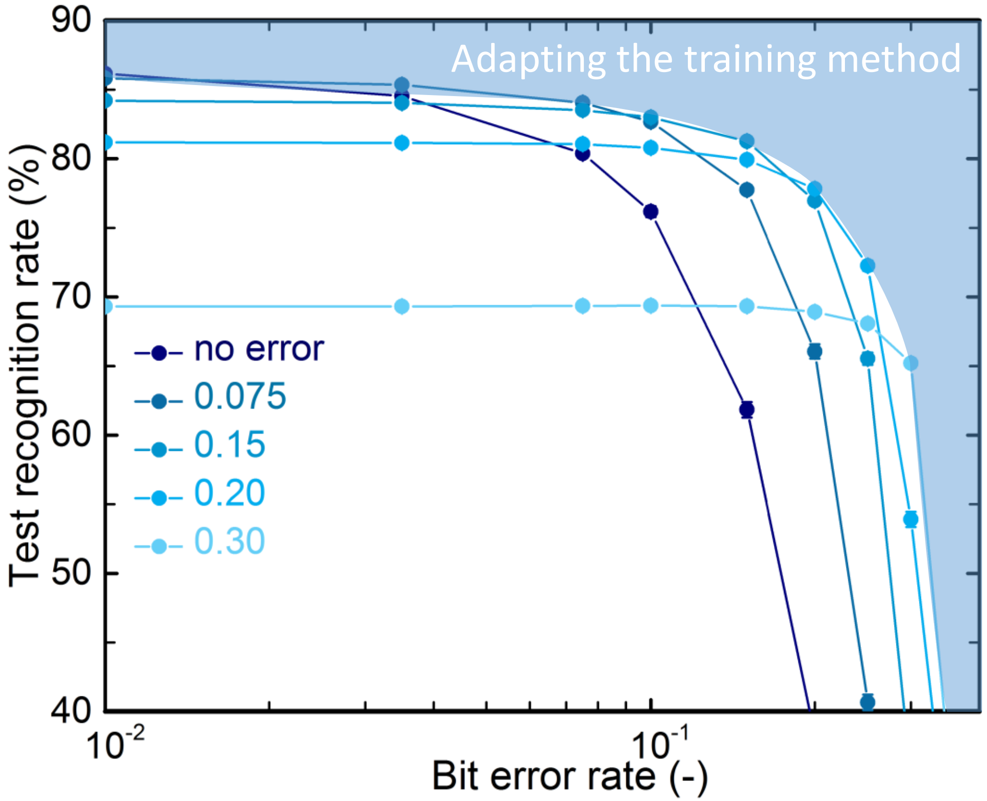
Binarized Neural Networks are resilient to errors

Including bit errors during the training process

- MNIST with fully-connected NN

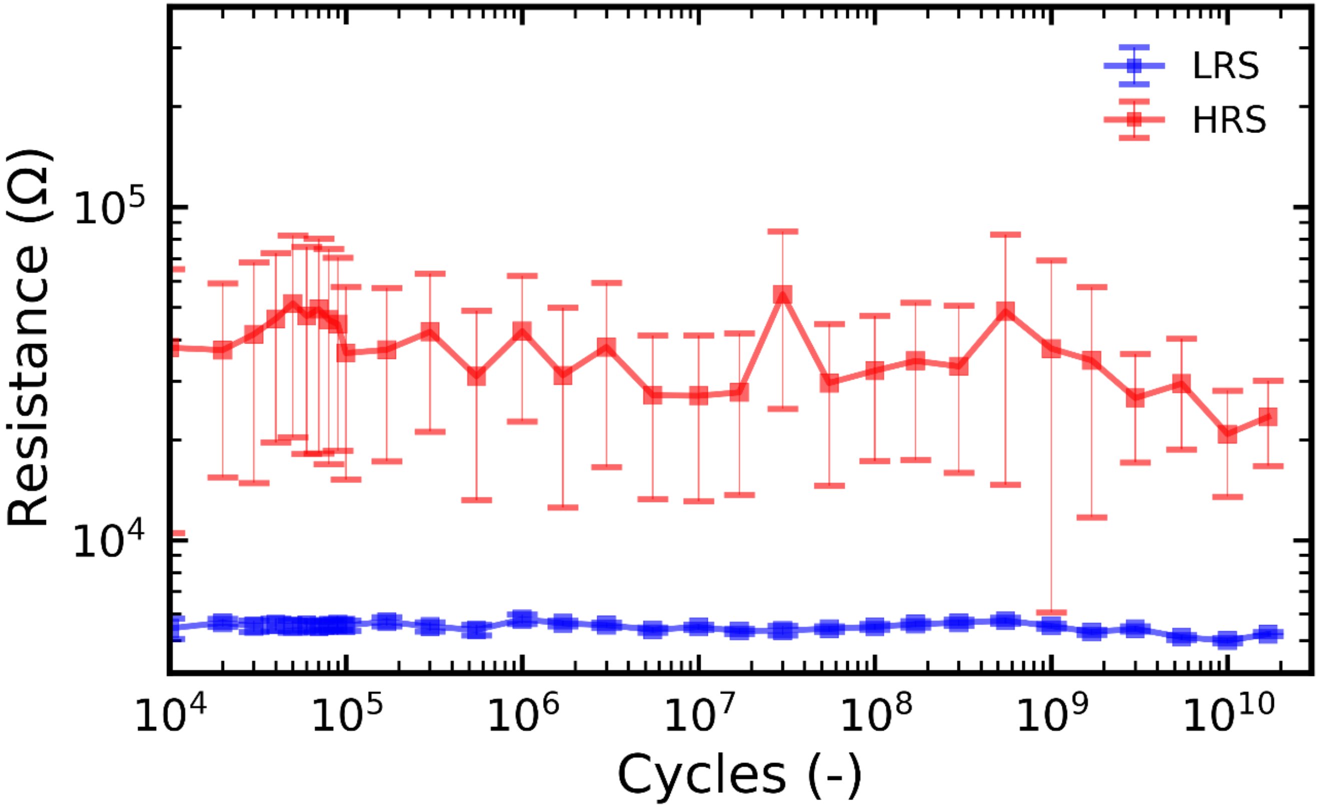


- CIFAR 10 with Convolutional NN



Adapting the training method can extend the bit error tolerance

Reliance on 2T2R + Tolerance to Errors Allows Using Weak Programming Condition



- V_{reset} = 1.5V
- Compliance current = 200 µA
- Error rate 1T1R = **~10^{-2}**
- Error rate 2T2R = **$2\text{x}10^{-3}$**

Endurance > **10^{10}** cycles

To sum up

- Binarized Neural Network ideal for in memory computing
- 2T2R decreases bit error in comparison to 1T1R and ECC
  - Avoid the use of ECC
  - Logic operation integrated in reading circuit
- Binarized Neural Network are resilient to errors
  - Low voltage/current for reading/programming

  ⮑ leads to : Low Energy & high endurance
- Technology ready today

Hardware & learning algorithm co-development

Thank you for your attention !