AICAS 2020

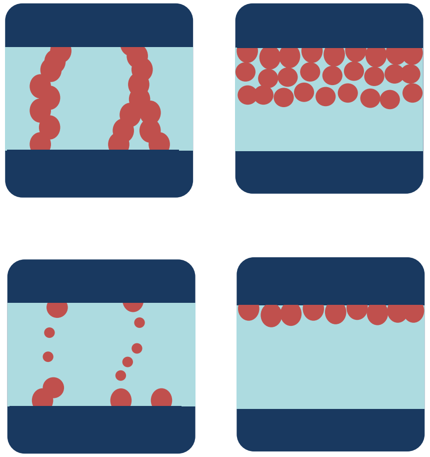
Special Session: AI accelerators with Memristive Technology

Low Power In-Memory Implementation of Ternary Neural Networks with Resistive RAM-Based Synapse

Axel Laborieux[1], Marc Bocquet[2], Tifenn Hirtzlin[1], Jacques-Olivier Klein[1], Liza Herrera-Diez[1], Etienne Nowak[3], Elisa Vianello[3], Jean-Michel Portal[2] and Damien Querlioz[1]

[1]C2N, Univ. Paris-Saclay, CNRS, Palaiseau, France
[2]IM2NP, Univ. Aix-Marseille et Toulon, CNRS, France
[3]CEA, LETI, Grenoble, France

Memristive technology promising for Neuromorphic Computing



RRAM:
Filamentary/Interfacial

Resistance state

Low: « 1 »

High: « 0 »

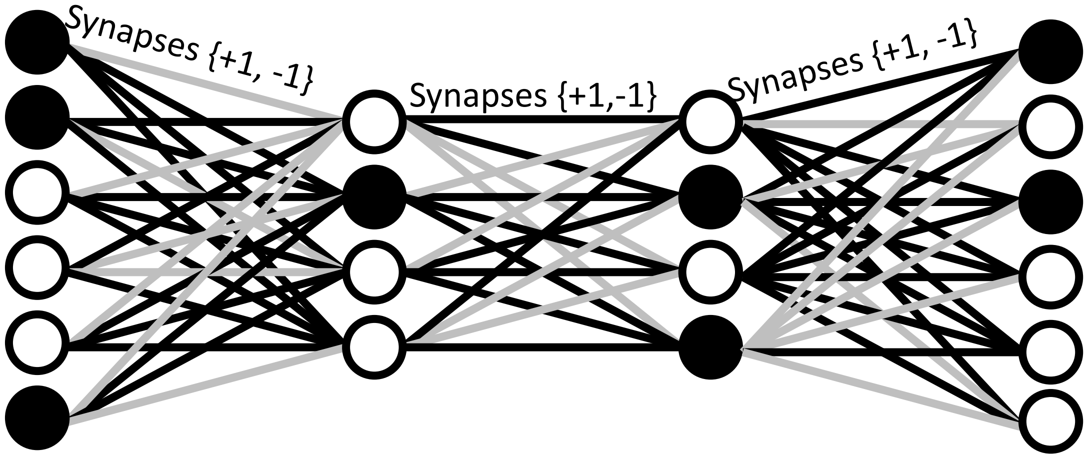
- **Fast, non-volatile memory that can be embedded at the core of CMOS**
- Memory state is the electrical resistance of the device (high or low)
- Many variations (oxide, phase change, magnetoresistive)
- In industry test production (Samsung, TSMC, Intel, ST Microelectronics...)

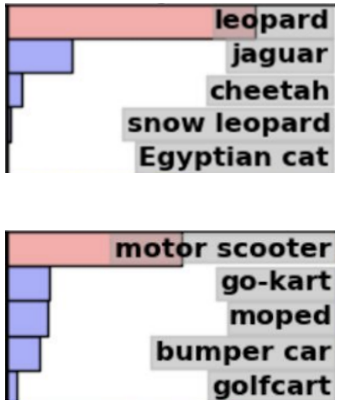
However, challenge of device imperfections/variations

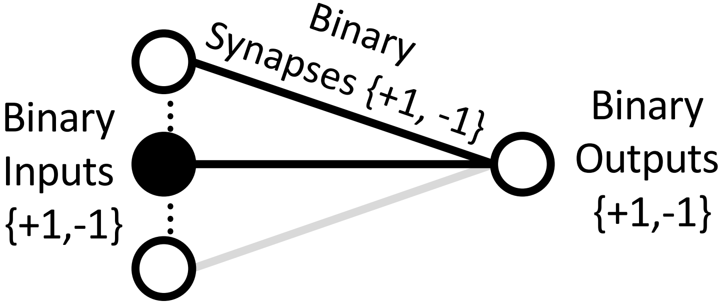
Low precision Neural Networks

Hubara, Courbariaux et al. NIPS 2016

Yoshua Bengio's group





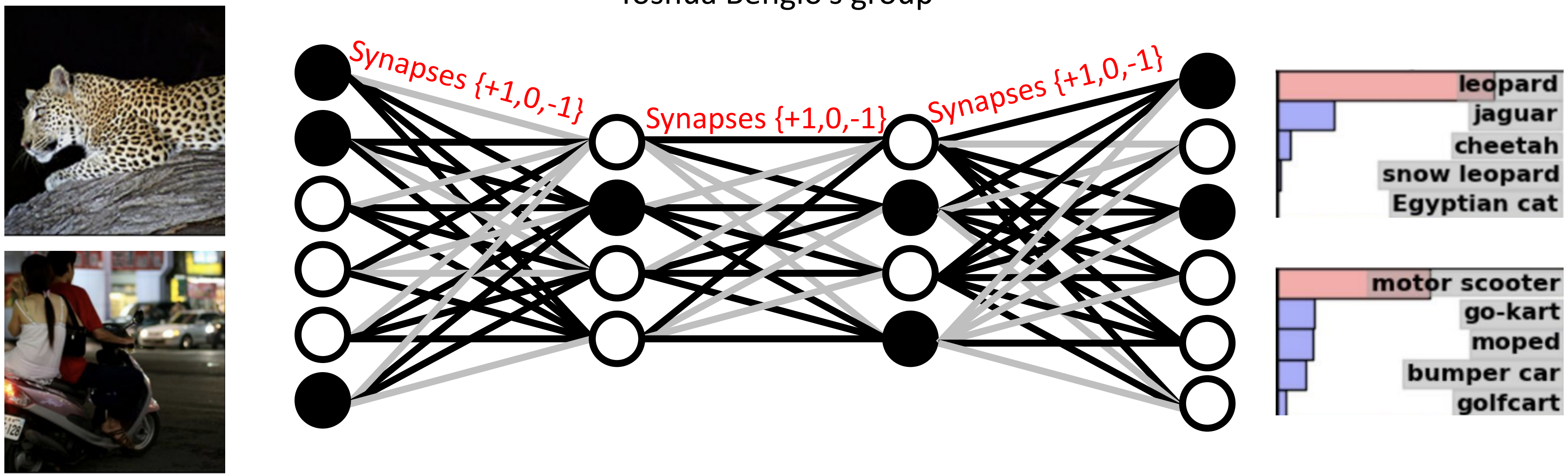


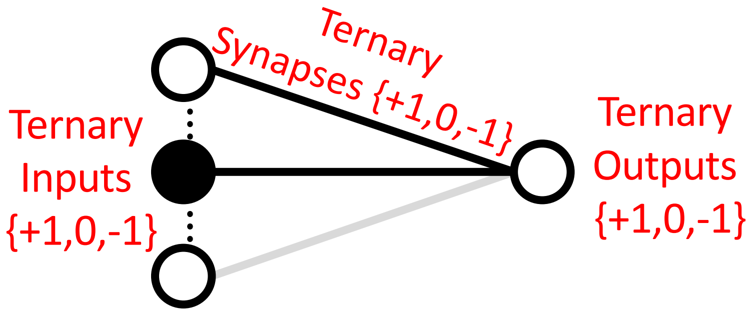
Previous work: implementation of Binarized weights

Low precision Neural Networks

Hubara, Courbariaux et al. NIPS 2016

Yoshua Bengio's group





In this work: implementation of Ternarized weights

Outline of the talk

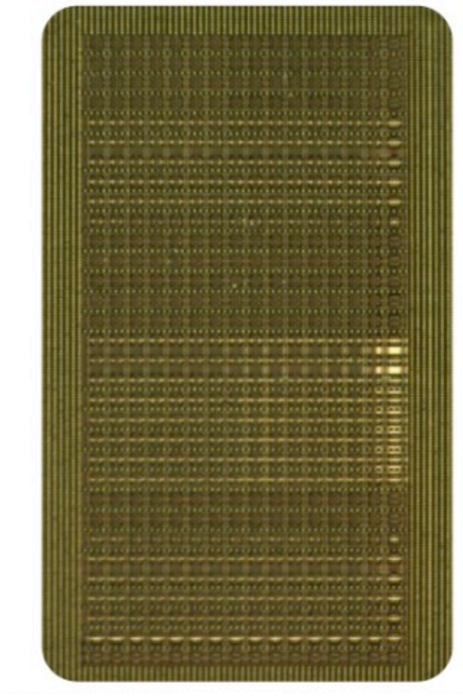
1. **Hybrid CMOS/Resistive RAM experimental implementation of ternarized weight** using a precharge sense amplifier in the **low supply voltage regime**

2. **PyTorch simulations** demonstrating that Ternarized Neural Networks **consistently outperform** Binarized Neural Networks

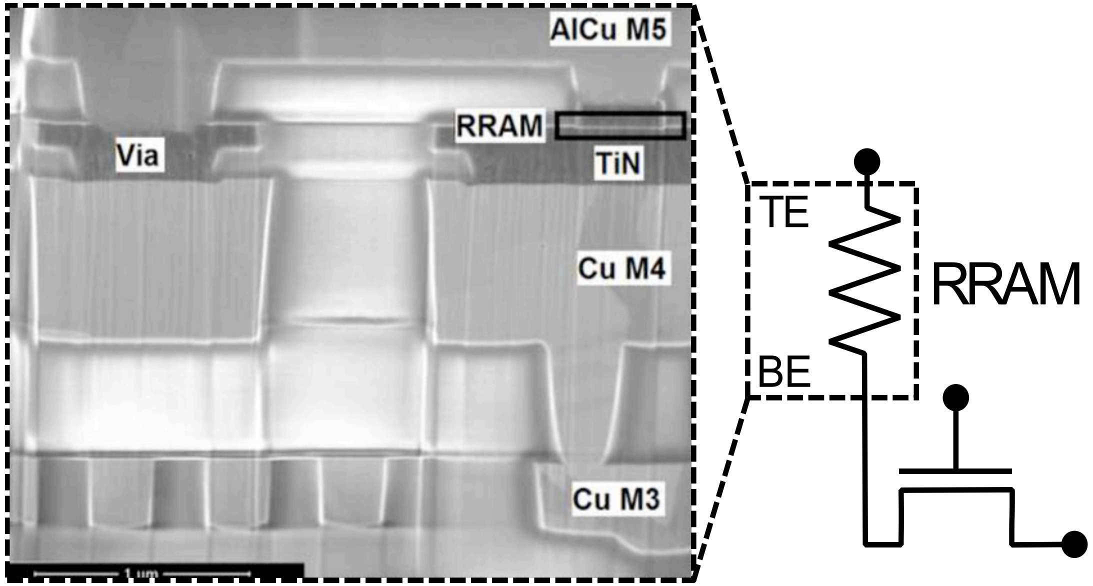
3. Demonstration of the network **robustness to the device imperfections** in the system

Our kilobit array

Our die



fabricated 130 nm
RRAM/CMOS hybrid chip

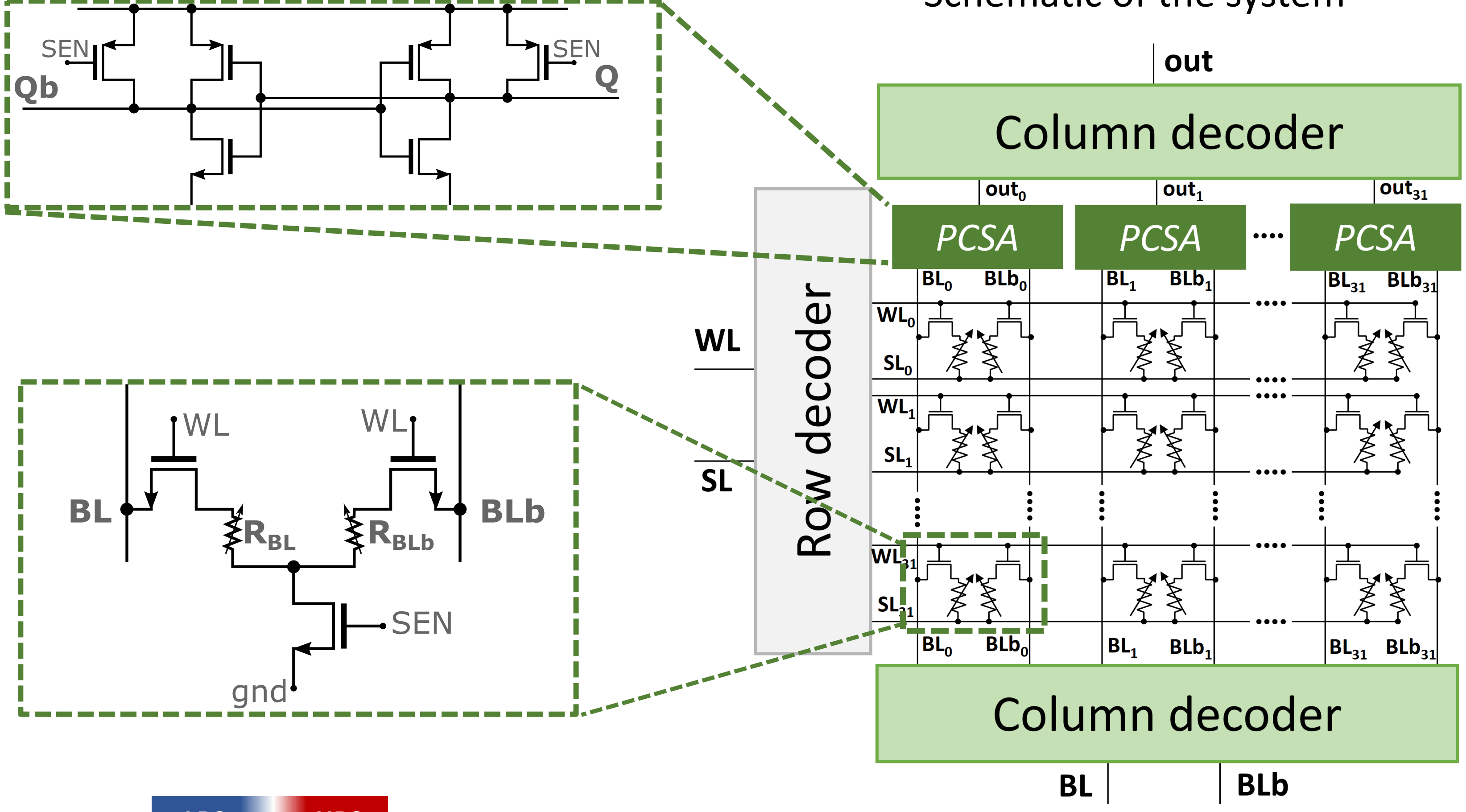


HfO_2 RRAM

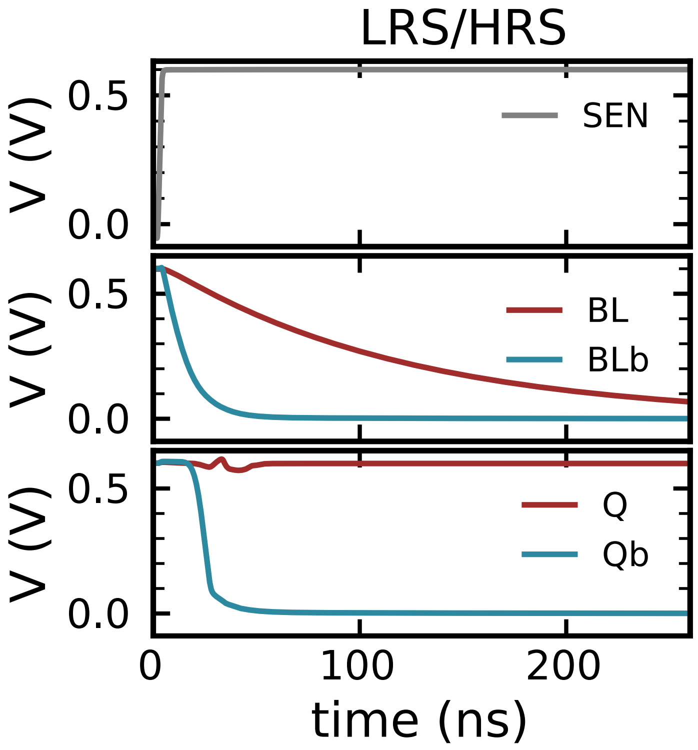
Background: two Resistive RAM devices as one synapse

Peripheric circuit to differentiate resistance states

Schematic of the system



Reading binary weight

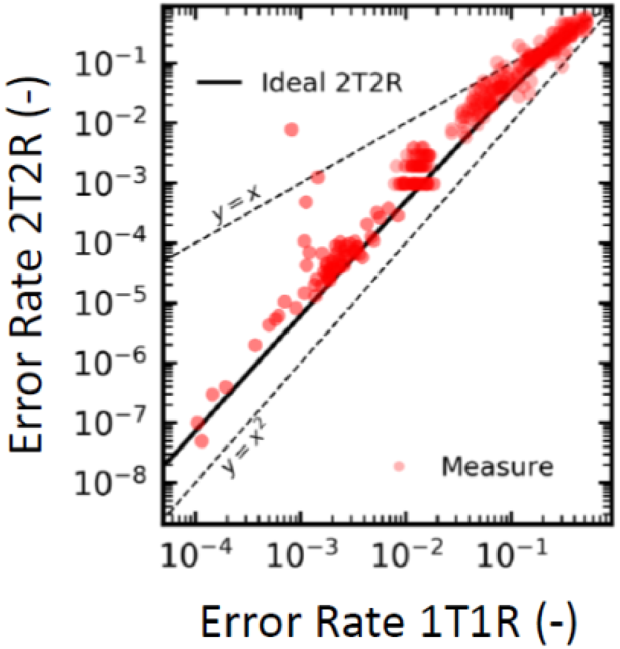


LRS HRS **+1**

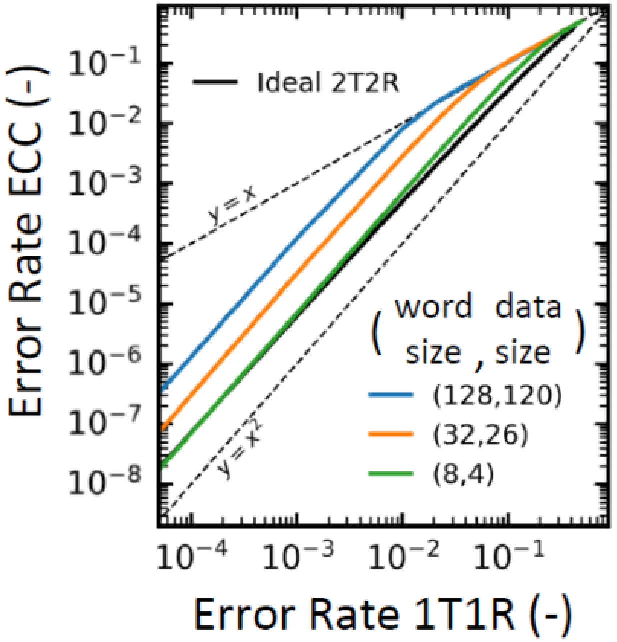
HRS LRS **-1**

Background: two Resistive RAM devices as one synapse

- Experimental bit error rate



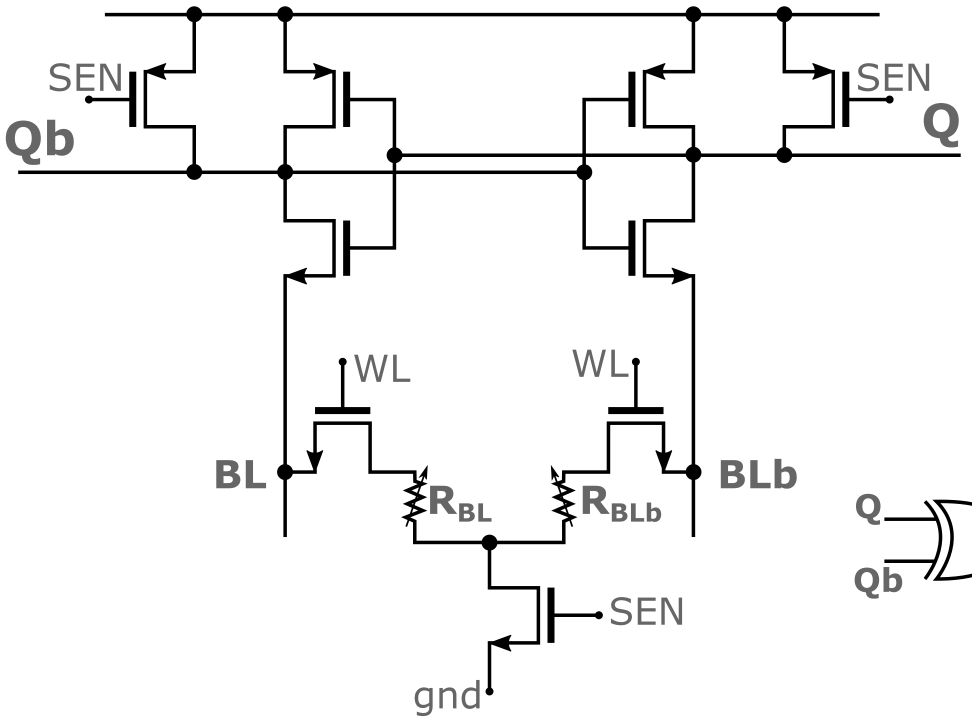
- Error Correction Code SECDED

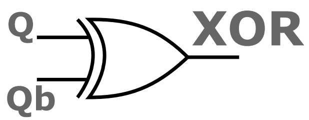


Hirtzlin et al., *Frontiers in Neuroscience*, 2020

Device pairs programmed in a complementary fashion **reduce error rate without computation overhead**

In this work: encode a third state with HRS/HRS



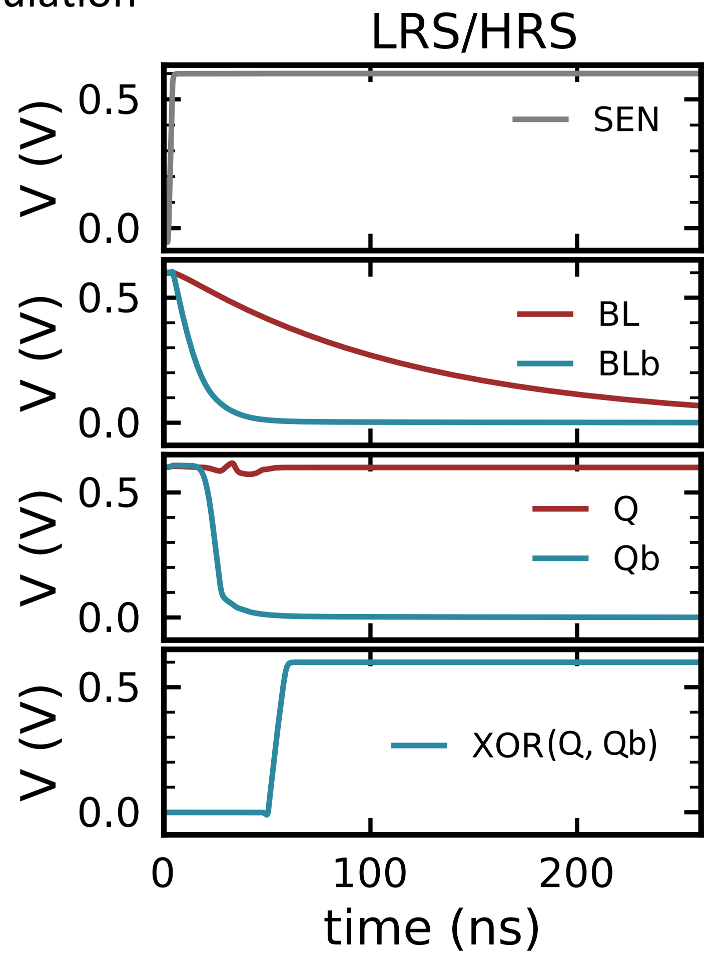


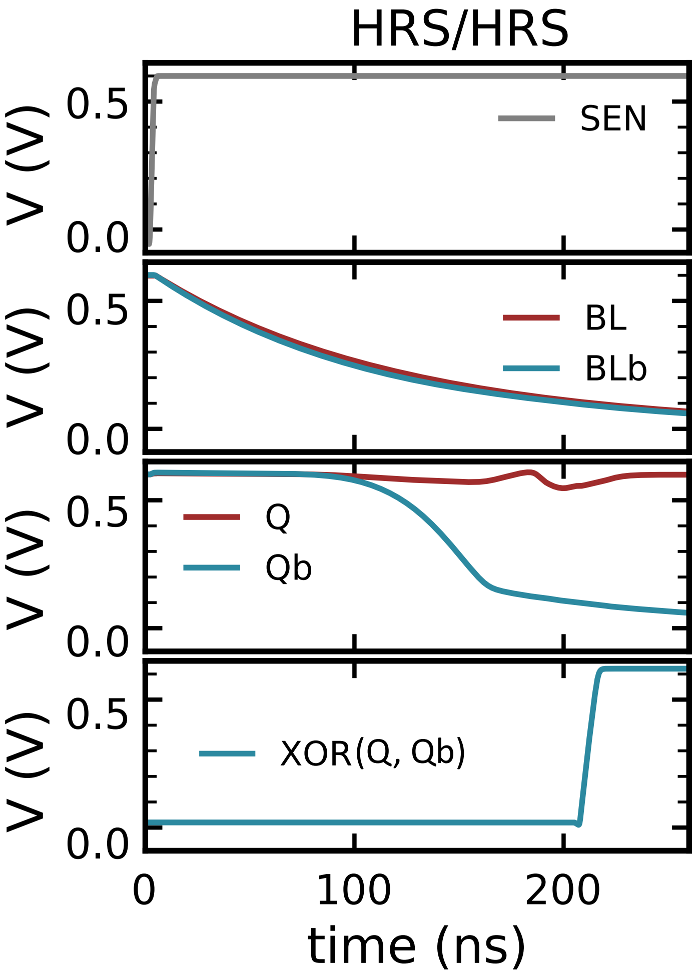
LRS	HRS	+1
HRS	HRS	0
HRS	LRS	-1

No memory overhead

The Sense converges slowly when operated with low supply voltage

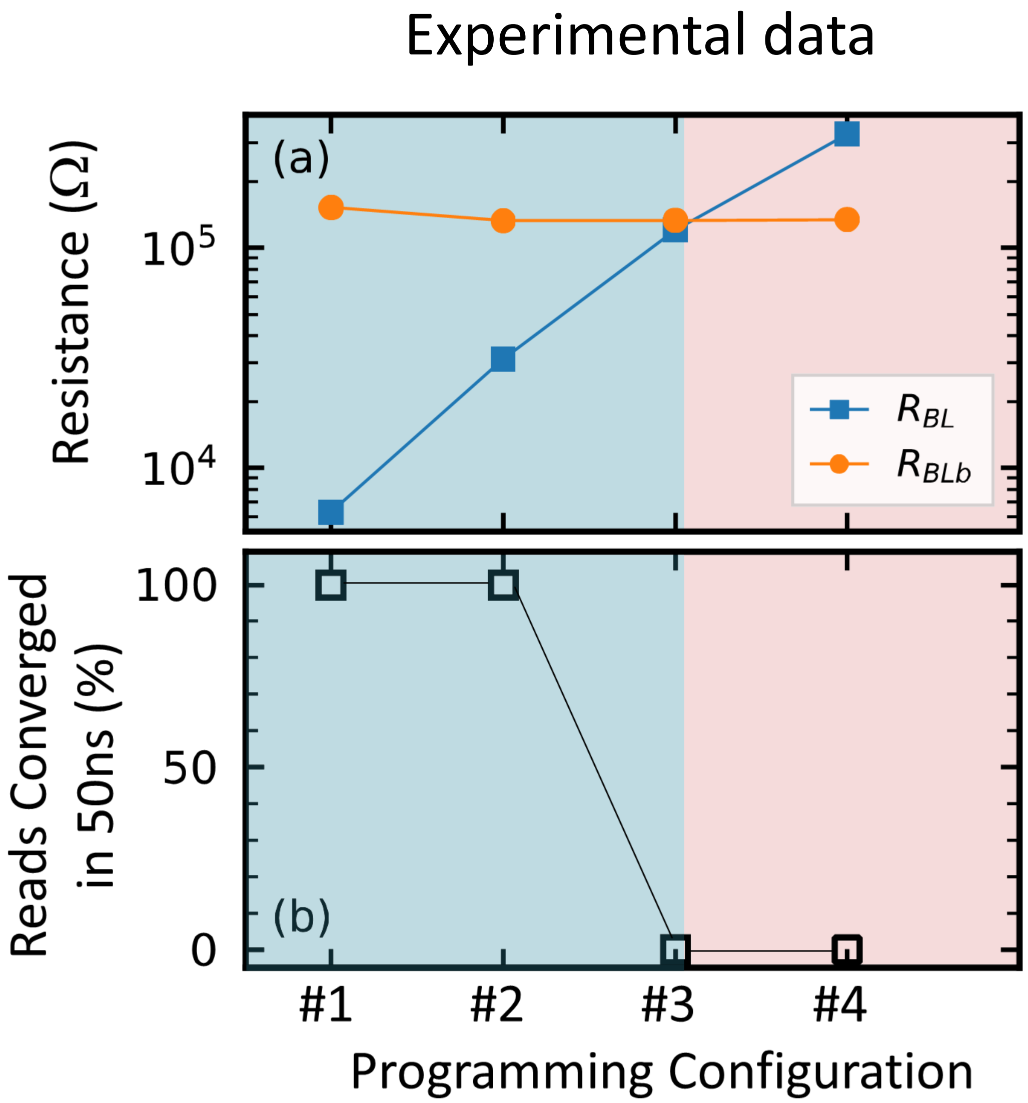
Spice simulation



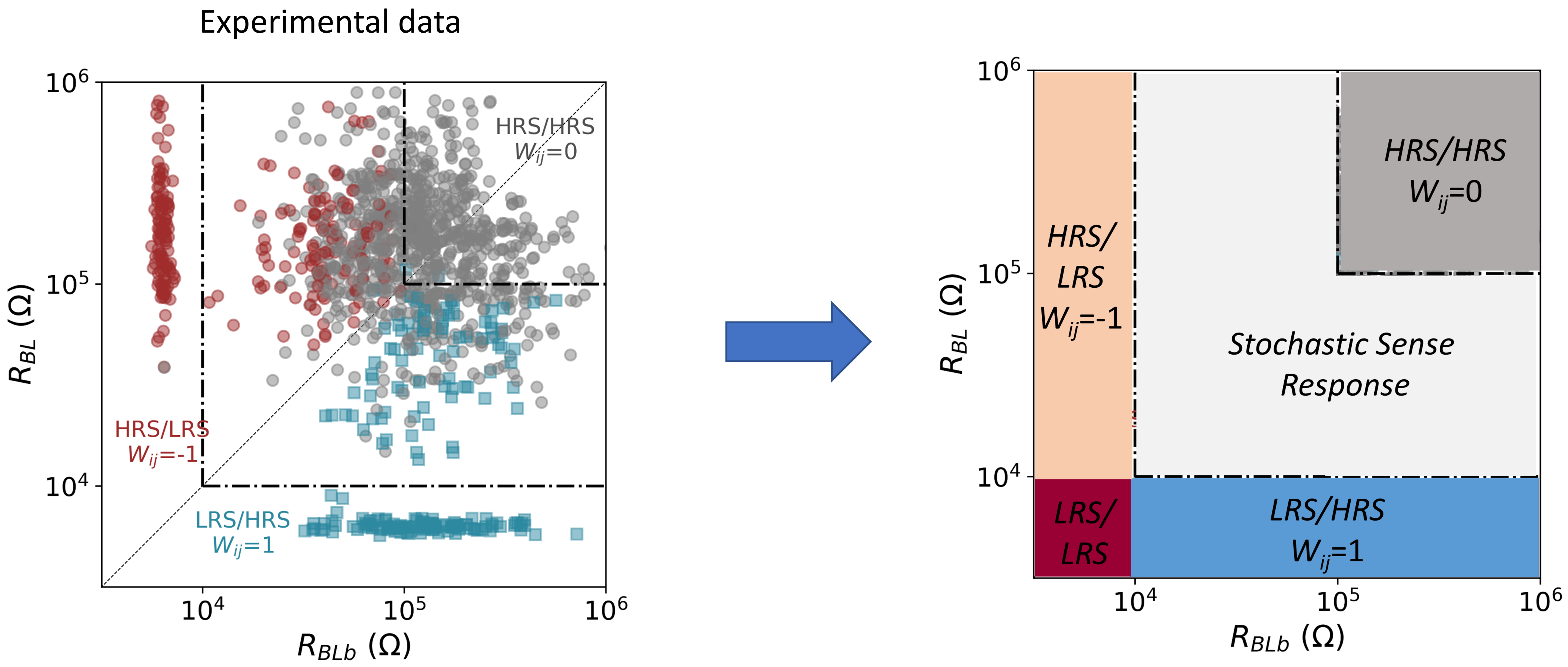


This work: leverage the speed of the Sense to store a new value

The Sense converges slowly when operated with low supply voltage

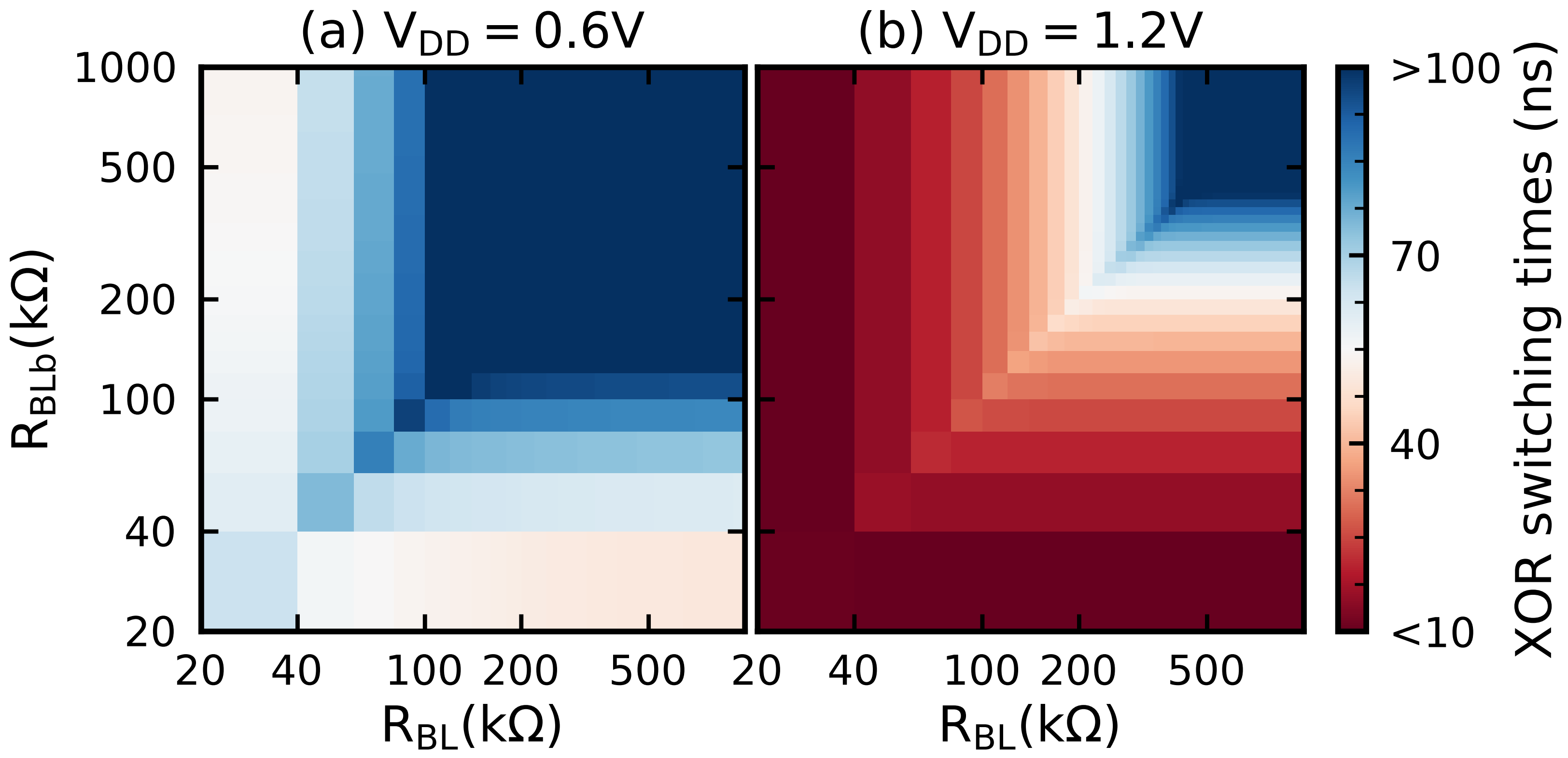


Implementation of ternary weights



No memory overhead and in a single Sense operation

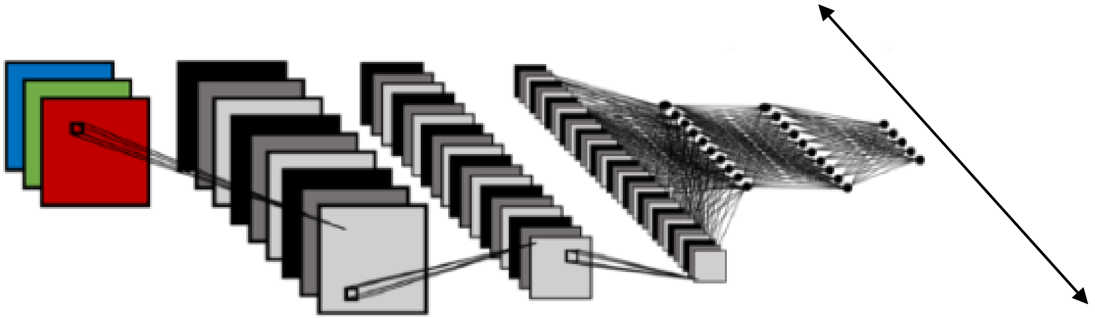
A behavior magnified in the low supply voltage regime



A behavior magnified in low supply voltage regime of the Sense
-> Better for energy efficiency

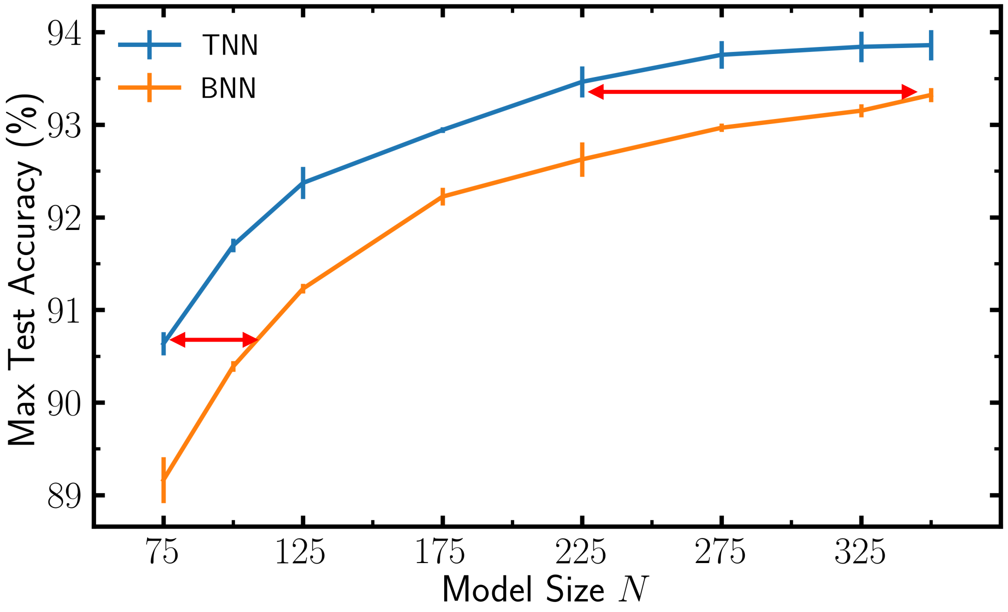
Ternarized Neural Networks (TNNs) outperform Binarized Neural Networks

CIFAR-10 Vision task (*Krizhevsky et al., 2009*)

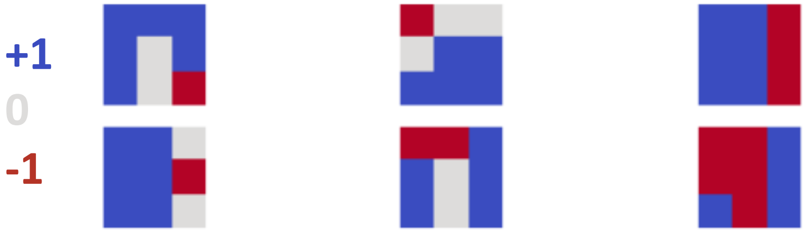


Model Size = Number of filters

PyTorch simulations



TNNs outperform BNNs consistently

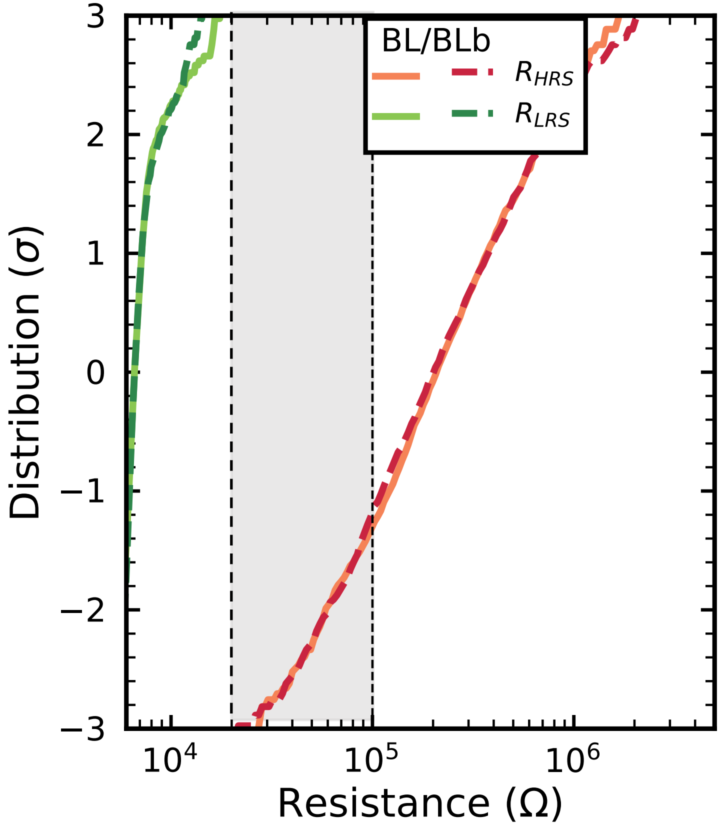


3^9 = 19,683 ternarized kernels

2^9 = 512 binarized kernels

Device pairs programmed in the stochastic area create new errors

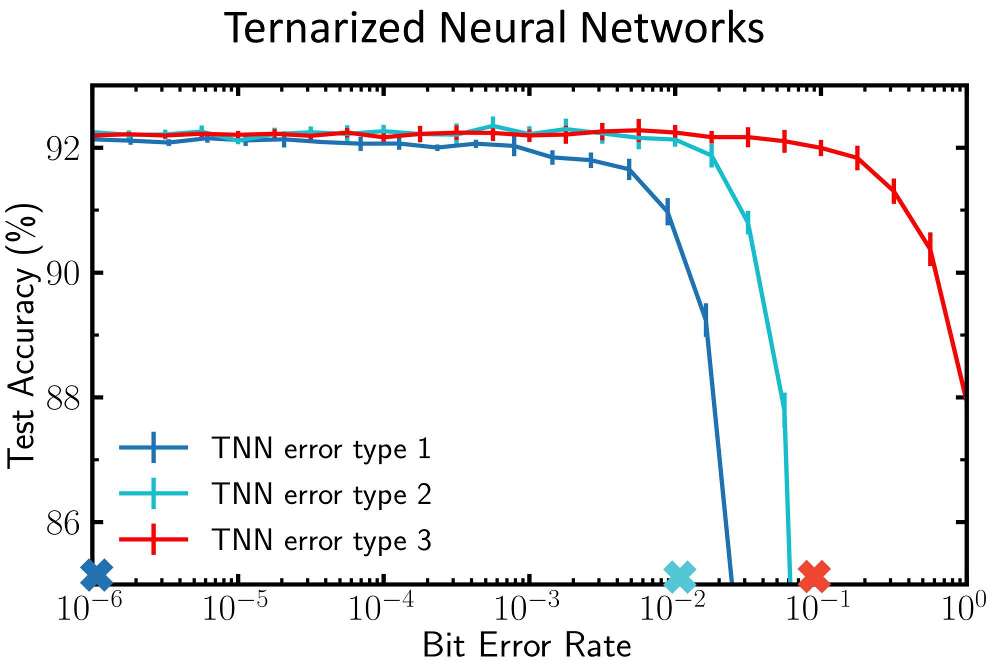
Experimental Distribution of the LRS and HRS

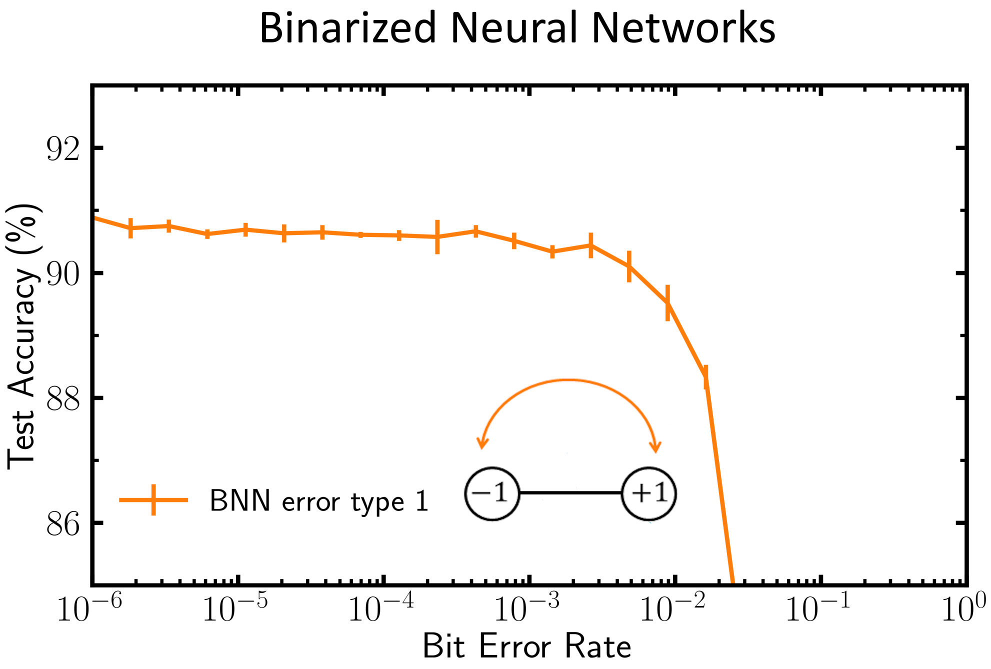


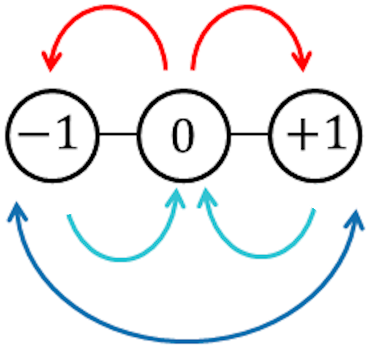
- SET compliance: 200µA
- RESET voltage: 2.5V
- Programming pulses: 100µs

New type of errors: 0 can be read as ±1

TNNs are resilient to errors due to device variability







TNNs still outperform BNNs when devices errors are taken into account

Concluding remarks

- Impact of this work: best envisionned for **low-power**, **high-performing** dedicated hardware for **edge intelligence** (wireless sensors, medical applications…)
- Low supply voltage regimes can give room for new functionalities
- Device imperfection should be embraced rather than fought against

Thank you for your attention!

Fundings:

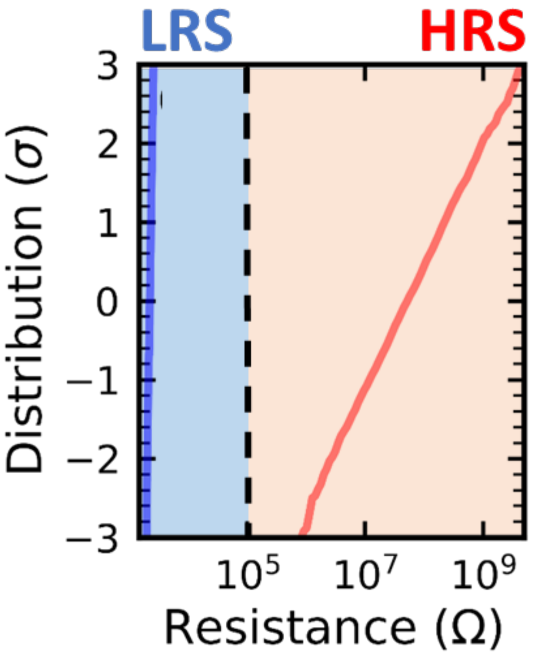
Agence Nationale de la Recherche
ANR

erc
European Research Council
Established by the European Commission

However, device to device variation is a challenge

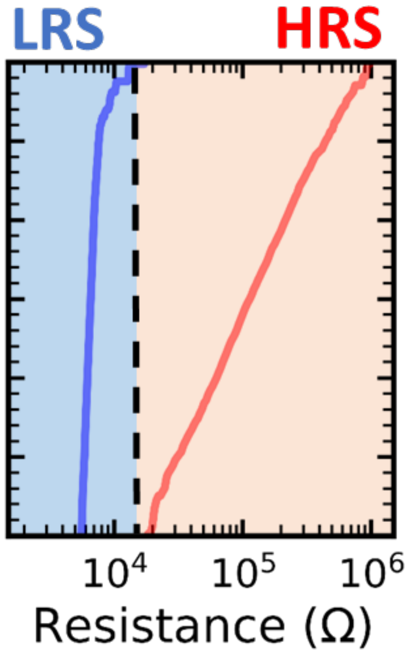
Programmation conditions

Very Strong



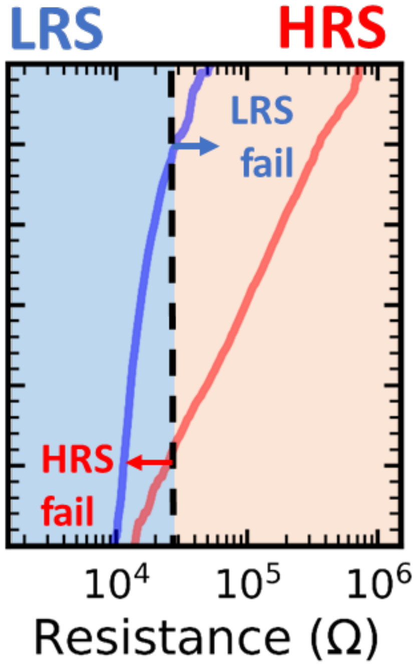
High Programmation energy
Low cyclability

Strong



Trade-off

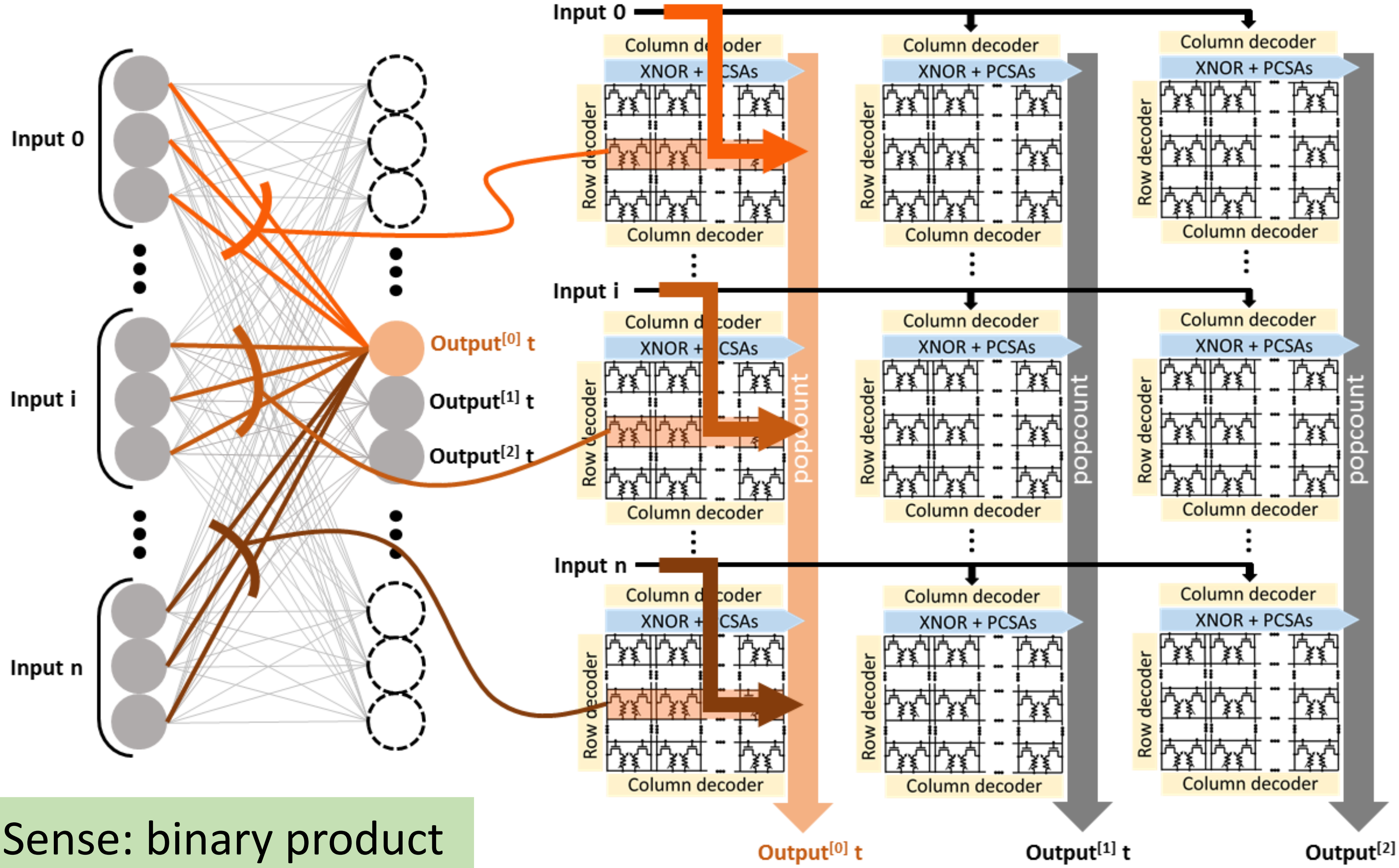
Weak



High 1T1R error rate

Appealing for low precision neural networks

Example of a BNN implemention

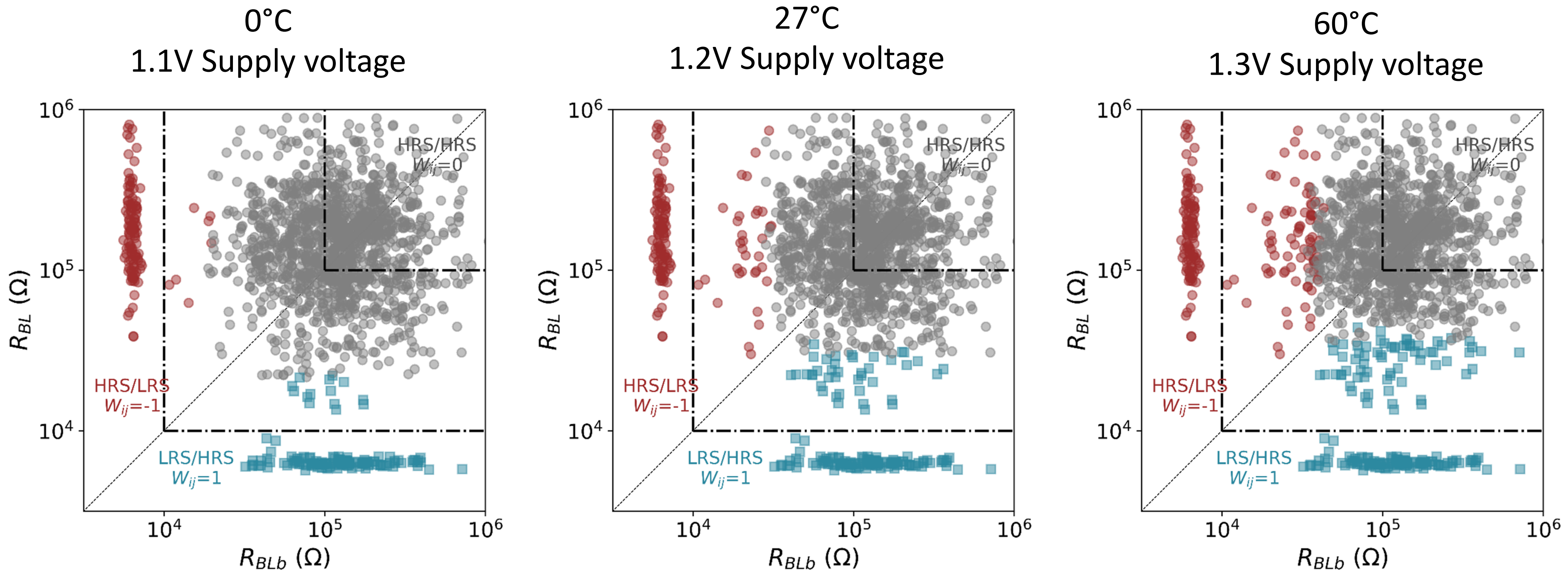


XNOR in the Sense: binary product
« In-memory computing »

Hirtzlin et al., 2019

Where do errors come from ?

Process Voltage Temperature variation analysis:


0°C
1.1V Supply voltage

27°C
1.2V Supply voltage

60°C
1.3V Supply voltage