# Memory is Key to Intelligence

## From near-memory computing to in-memory computing

Sylvain Dubois
VP Business Development & Marketing

Sep 18th, 2019

# Human Learning
# is Based on Experience

# Machine Learning is Based on Massive Amount of Data

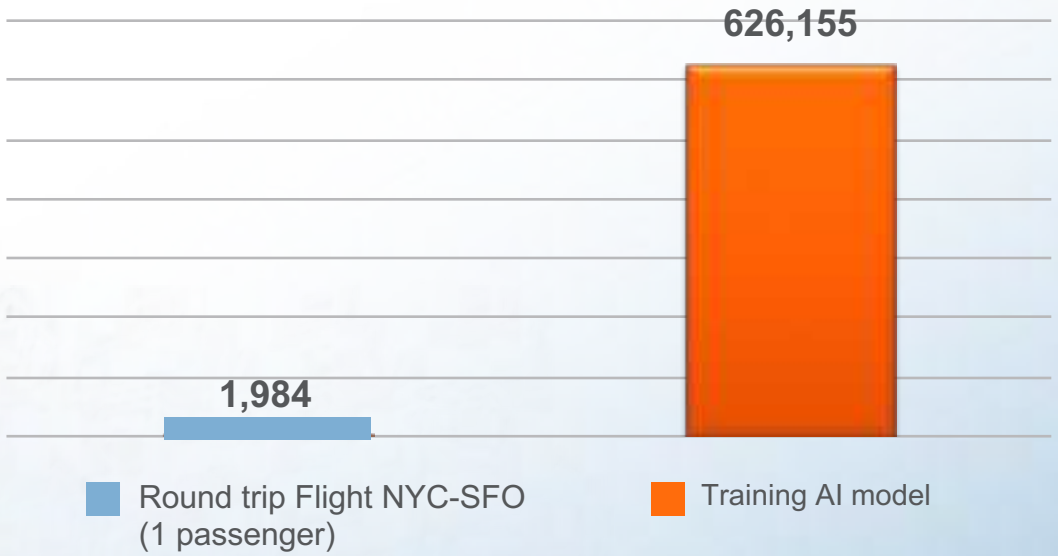
**A Lot of Data…**

# Training a single AI model consumes a lot of energy

*"Training a single AI model can emit as much as 300X a roundtrip flight from NYC to SFO"*

MIT TECHNOLOGY REVIEW, JUNE 2019

**Common carbon footprint benchmarks**
in lbs of $CO_2$ equivalent




Source: MIT Technology Review Source: Strubell et al.
Transformer (213M parameters) w/ neural architecture search

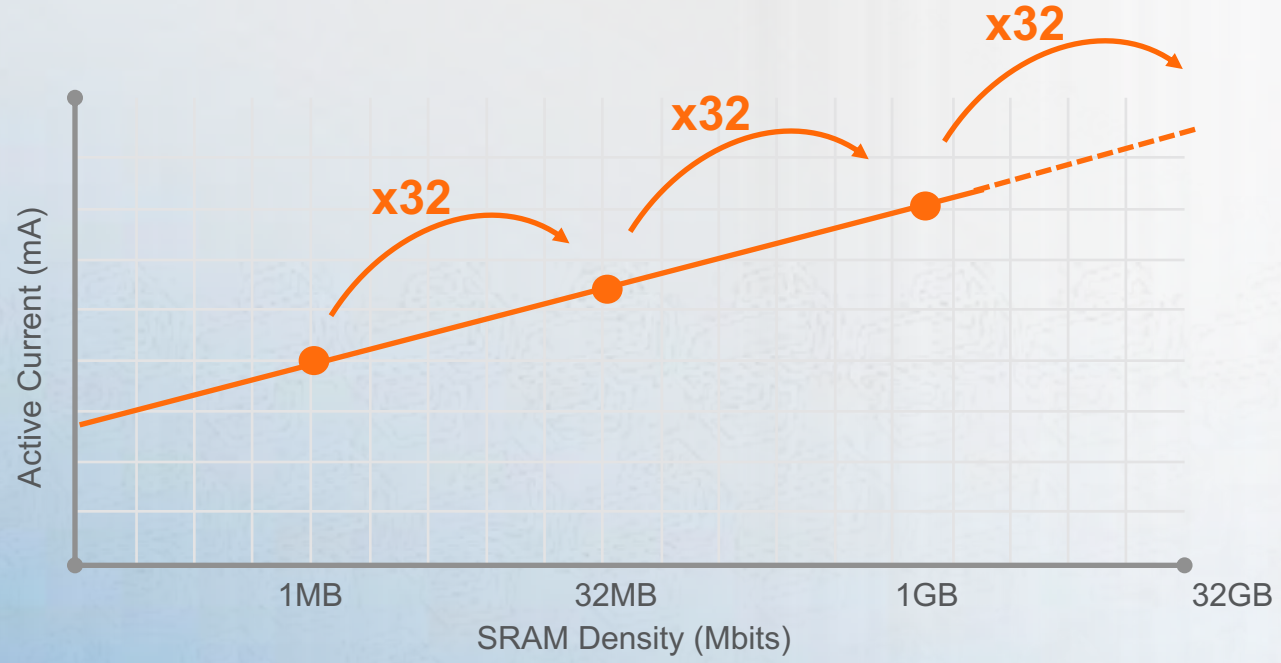
# Because Moving Bits Over a DRAM Bus Consumes a Lot

Reads from DDR4 DIMMs: 320 pJ/Byte
Reads from HBM DRAM: 64 pJ/Byte

In AI, the data movement off-chip causes a lot of penalty in energy and latency. **That is a bottleneck.**

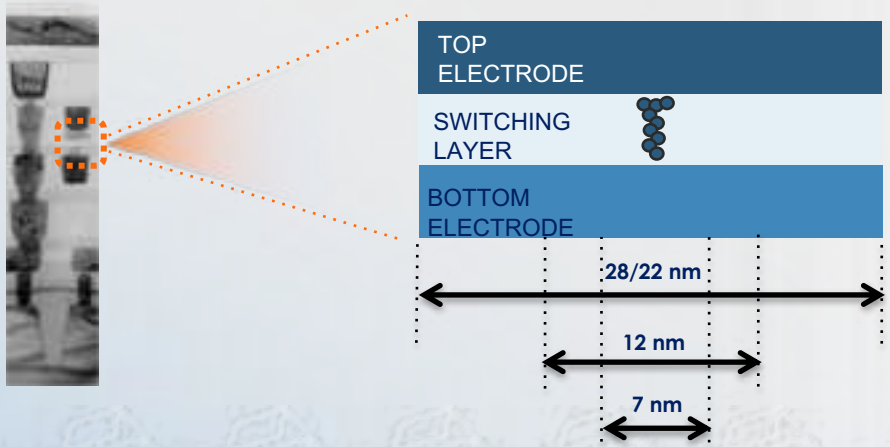
# And Using Embedded SRAM
# Has a Leakage Problem

on-chip SRAM  has shorter wires to processor
but has a fundamental leakage problem



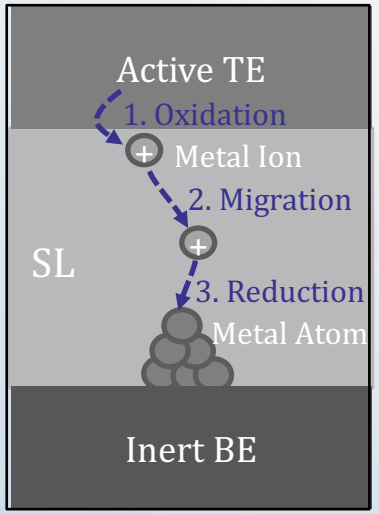

Reads from SRAM at 6 pJ/bit for 8 Mbit but getting worse for higher densities

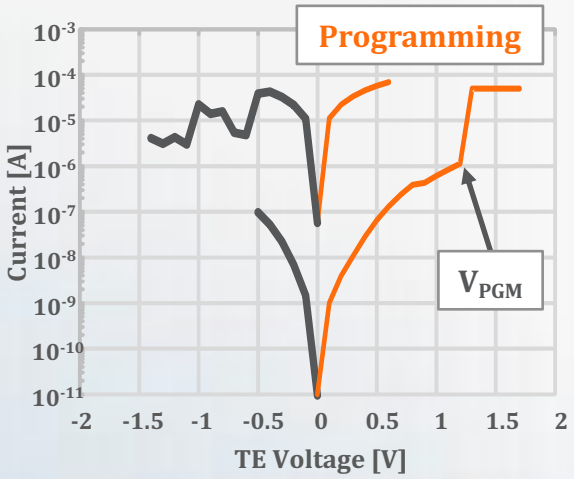
# One Alternative is to Use Embedded ReRAM




1. Non-volatile memory cell – **no leakage**
2. Monolithic CMOS integration: **located between any metal layers**
3. Scalable: **5nm metal filament**
4. Low energy: **15ns reads - <0.5pJ/bit**
5. 1M write cycles – **10 years retention**

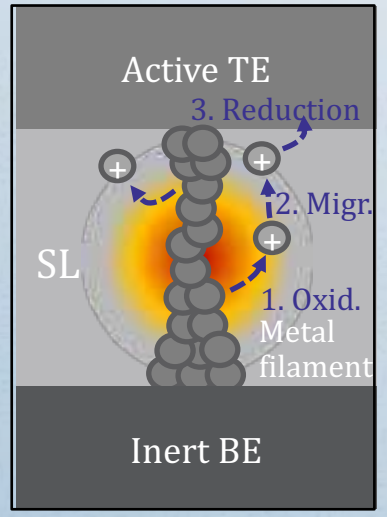
# ReRAM Fundamentals

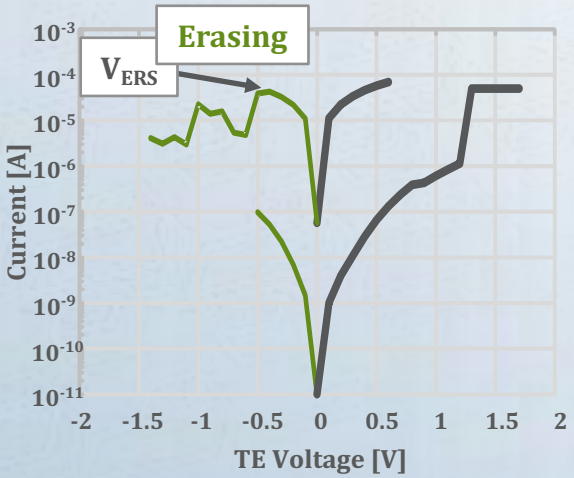



**Programming:** Positive Voltage on Top Electrode

→ **ON state** is reached when a complete filament is created between both electrodes




**Erasing:** Positive Voltage on Bottom Electrode

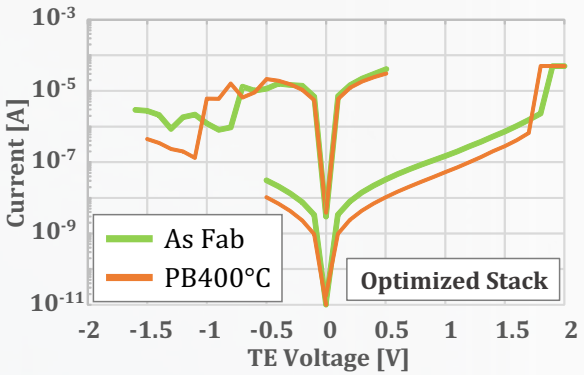
→ **OFF state** is reached when the conductive path is broken

# How Does ReRAM Compare With Other eNVM

| | **Target Commercial Crossbar ReRAM** | Target Commercial Embedded MRAM | Commercial Embedded Flash |
|---|---|---|---|
| Scalability | **28 / 22 nm** | 22 nm | 40 nm |
| Cost | **3 layers**<br>**0 new elements**<br>**2 additional masks**<br><br>**0.03 um$^2$ cell size** | 20+ layers<br>9+ new elements<br>5 additional masks<br><br>0.05 um$^2$ cell size | dedicated CMOS front-end<br>12 additional masks<br><br>0.07 um$^2$ cell size |
| Energy | **0.4 uA/MHz/bit read current**<br>**4 uA standby current** | 2 uA/MHz/bit read current<br>200 uA standby current | 1.6 uA/MHz/bit read current<br>150 uA standby current |
| Other Characteristics | **15ns read latency**<br>**1M write cycles**<br>**10 years retention**<br>**Magnetic immunity** | 15ns read latency<br>1M write cycles<br>10 years retention<br>No magnetic immunity | 25ns read latency<br>100K write cycles<br>10 years retention<br>Magnetic immunity |

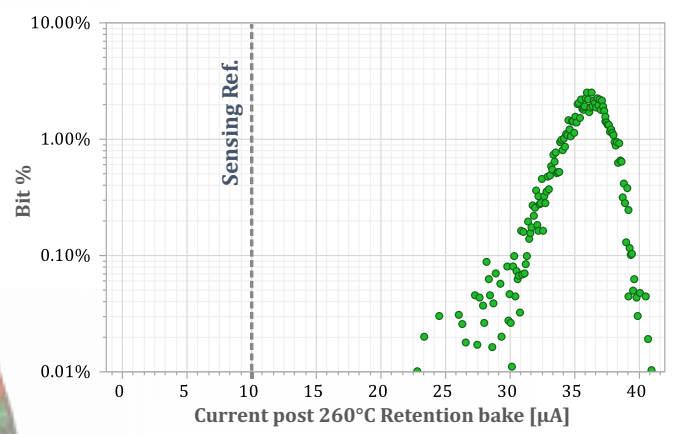
# Ready to engage. Here's Our Latest Results

## CMOS Integration Compatibility

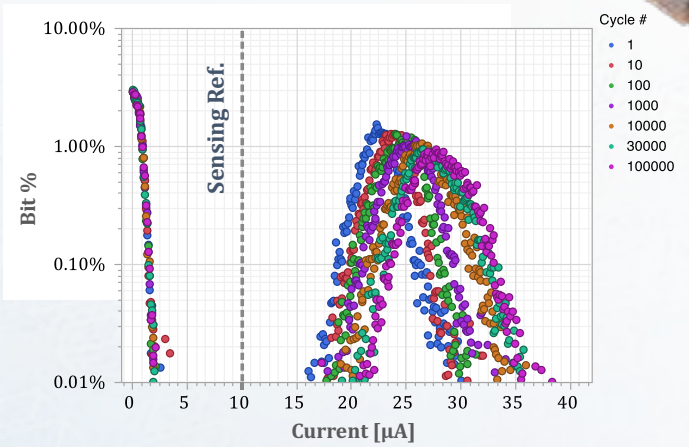




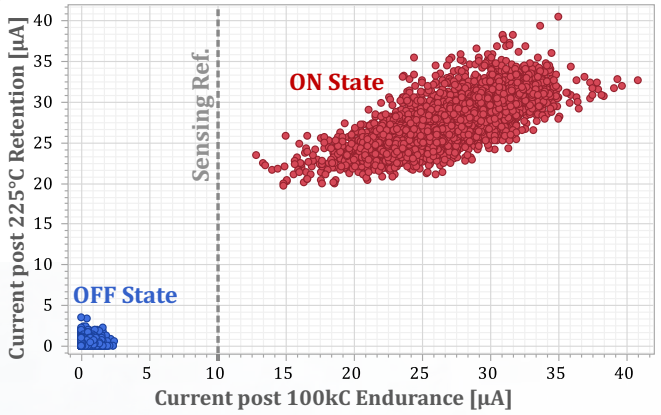
## Soldering Reflow Compatibility




## Endurance




## Retention



# Today: Near-Memory Computing

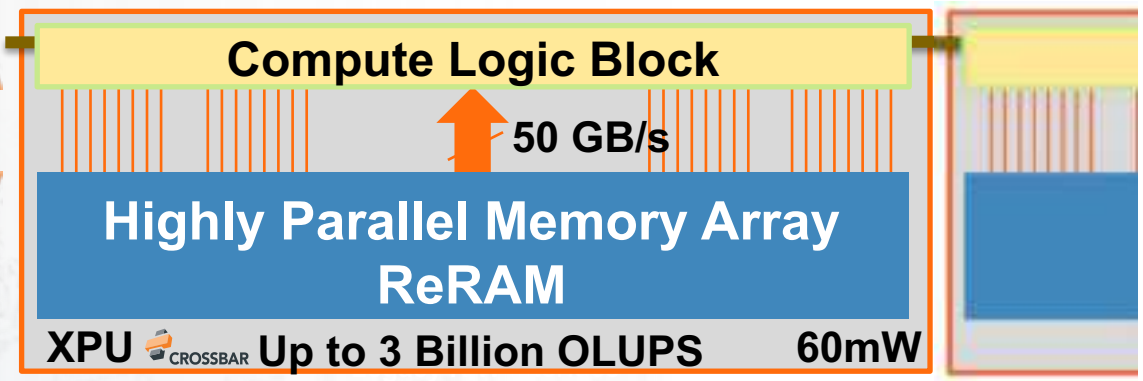
# Today: Near-Memory Computing

## Compute Logic Block

- Host interface @ 66MHz
- 1000s of computation engines
- Up to 3 Billion Object LookUp Per Sec (OLUPS)
- Decision module
- Wide internal memory bus 50GB/s

## Memory Array

- Non-volatile ReRAM
- 50GB/s memory bus
- 8K bit lines

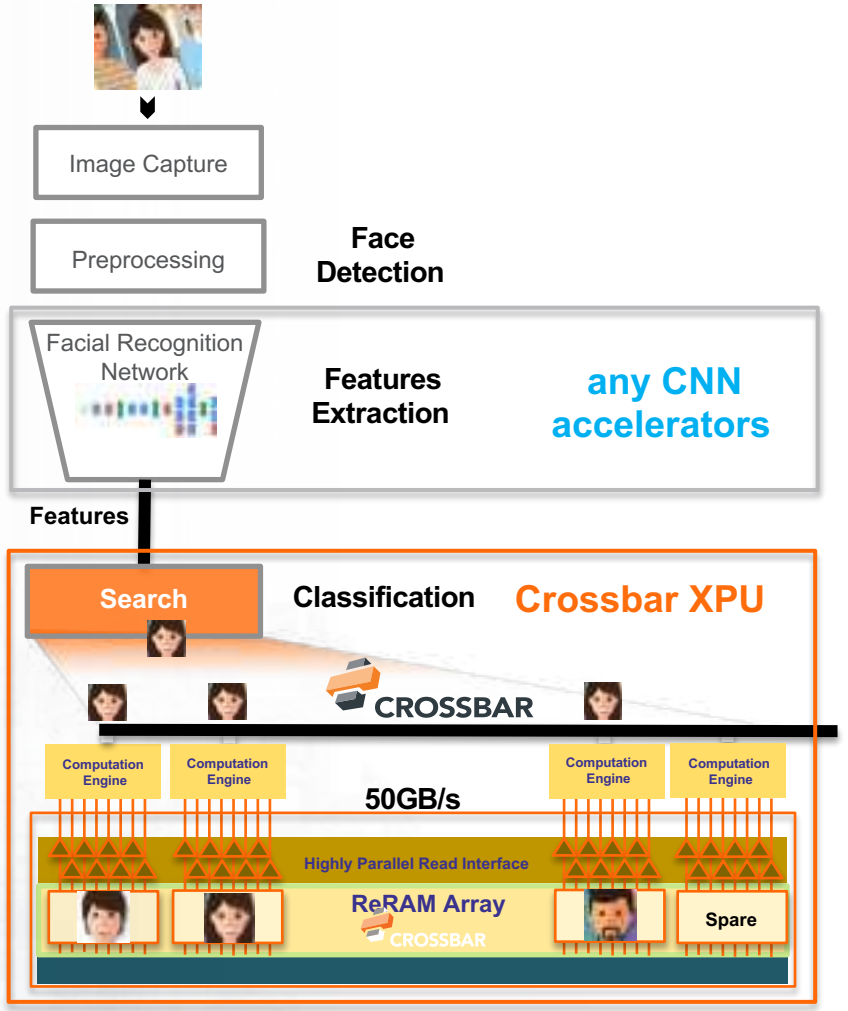



**Targeted for massive search/lookups, kNN, RBF, CBIR, Softmax**

# Example: Face Recognition

Classification is currently performed on host CPU

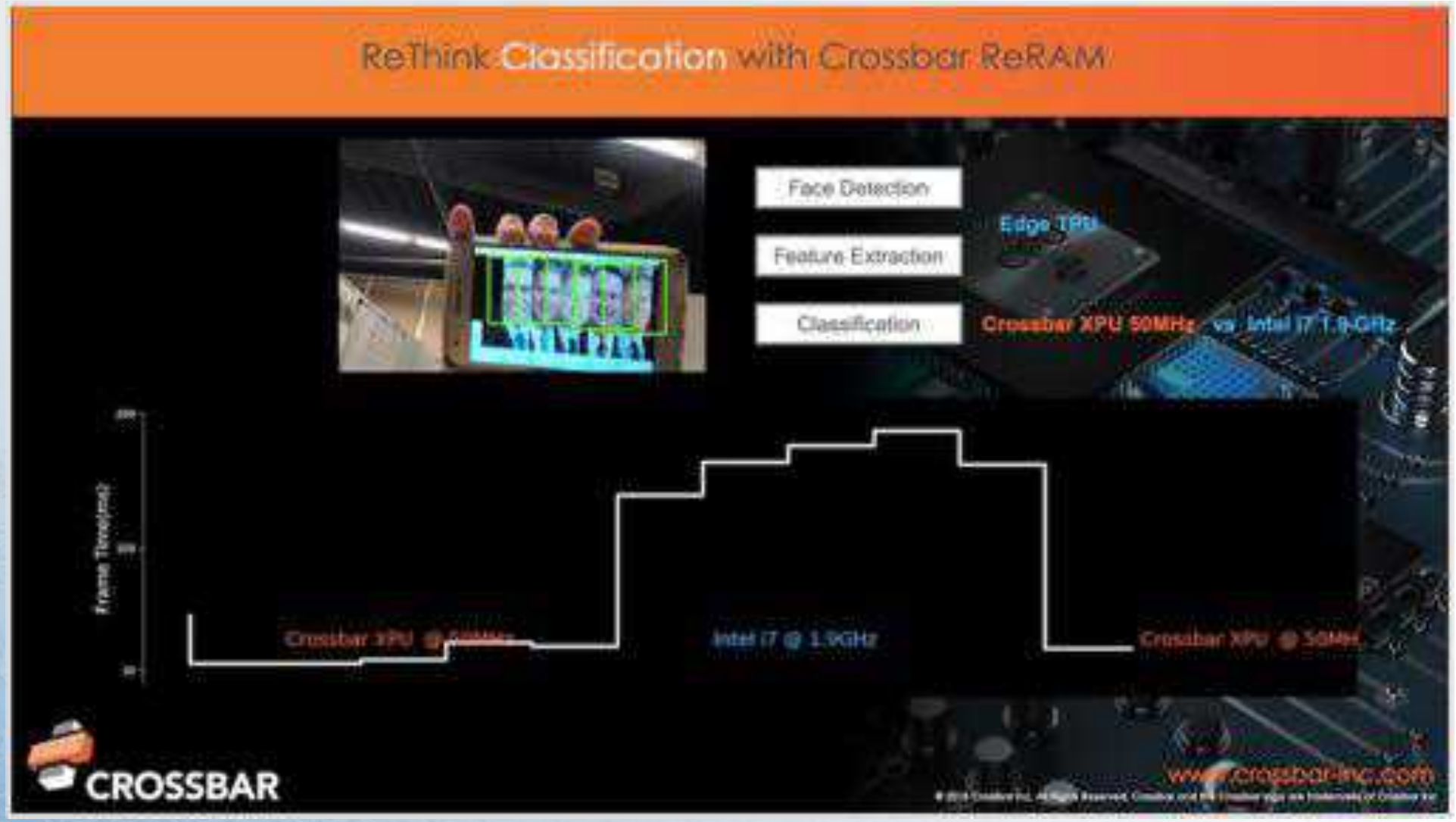
→ Use Crossbar XPU to compare feature representations against many




**Deterministic Performance**
**Persistent Memory**
**Configurable**
**Scalable**
**Low Power**

***Enabling Learning at the Edge***

# Ready to Engage




Visit Booth #15

# What's Next: Persistent Memory

# High-density Persistent Memory for Data Centers


## 128GB to 1TB NV-DIMM

**Read performance** - 25.6GB/s - 64 IOs – 250ns random reads

**Low power -** persistent memory and <1W active reads

**Endurance|Retention** – 10M | 6 months

## 8GB ReRAM chip

**DRAM read performance** - 3.2GB/s interface – 8 IOs

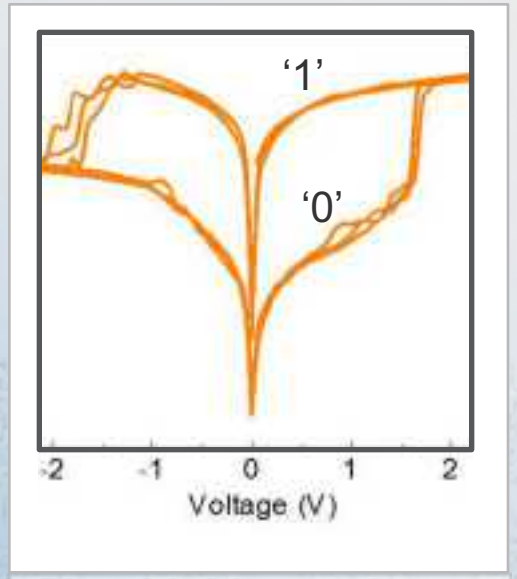
**Non-volatile** - No current drawn during inactivity

**Low energy** - 68mA active read current – 3.6 pJ/bit

**Lower cost than DRAM** 1~2$/GB

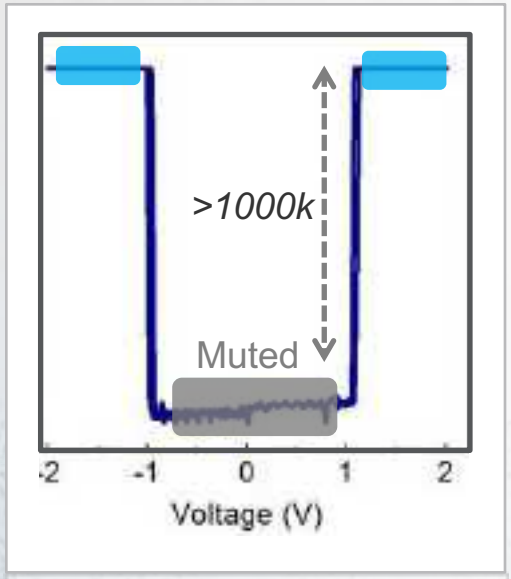
# How is it Possible: Crossbar Selector

**Crossbar ReRAM**



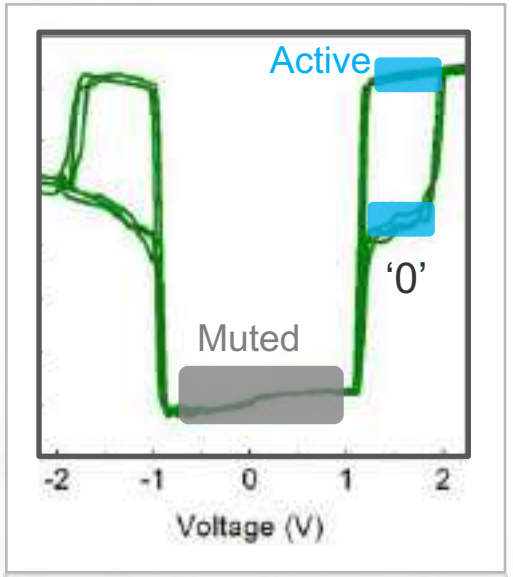

+

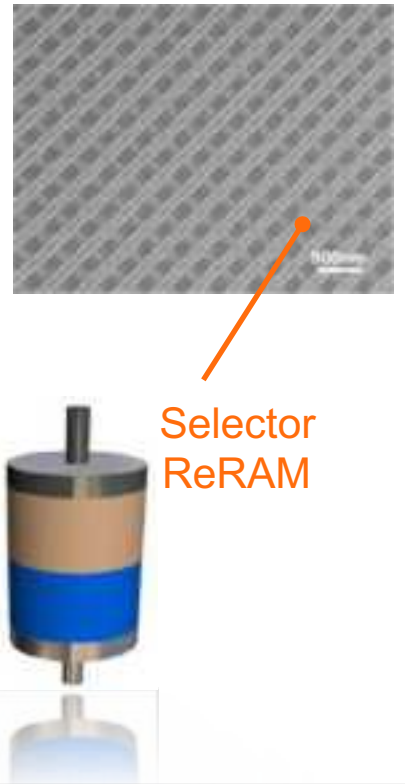
**Crossbar built-in Selector**




=

**Integrated ReRAM + Selector**

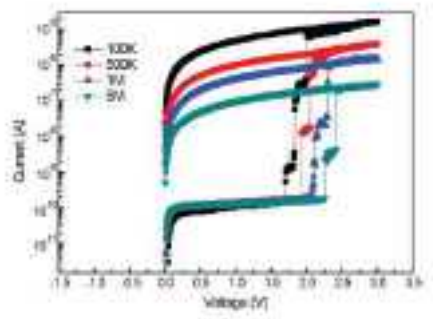


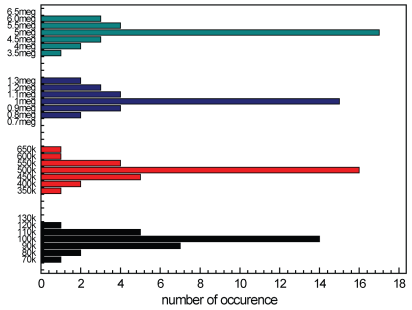



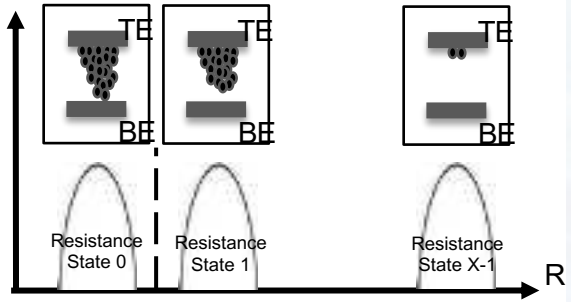
# Future is In-Memory Computing

# In-Memory Computing

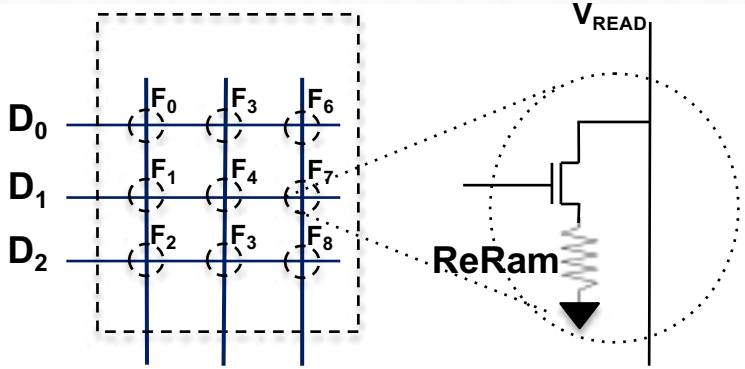
## Device & Circuits






## In-Memory Logic

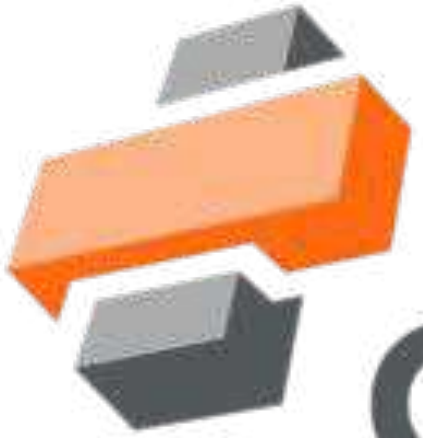



**Example: 1x3 to 3x3 Matrix Multiplication**

$$\begin{bmatrix} D_0 & D_1 & D_2 \end{bmatrix} \times \begin{bmatrix} F_0 & F_3 & F_6 \\ F_1 & F_4 & F_7 \\ F_2 & F_5 & F_8 \end{bmatrix}$$

# Crossbar: From Near-Memory to In-Memory Computing

- Based in Santa Clara, CA, U.S.A.
- $120M in raised capital to date
- Leader in **Resistive RAM technology**
- New class of non volatile memory: **Metal Filament Resistor**
- Back-end of line Non Volatile Memory: **2x nm, 1x nm**
- Patented **Technology:** 310 filed / 160 issued
- Applications in **Persistent Memory, AI, FPGAs, eNVM**
- Efficient search and computing with **Highly Parallel Memory**



# CROSSBAR