Search acceleration and Learning at the Edge with Crossbar ReRAM

Sylvain Dubois

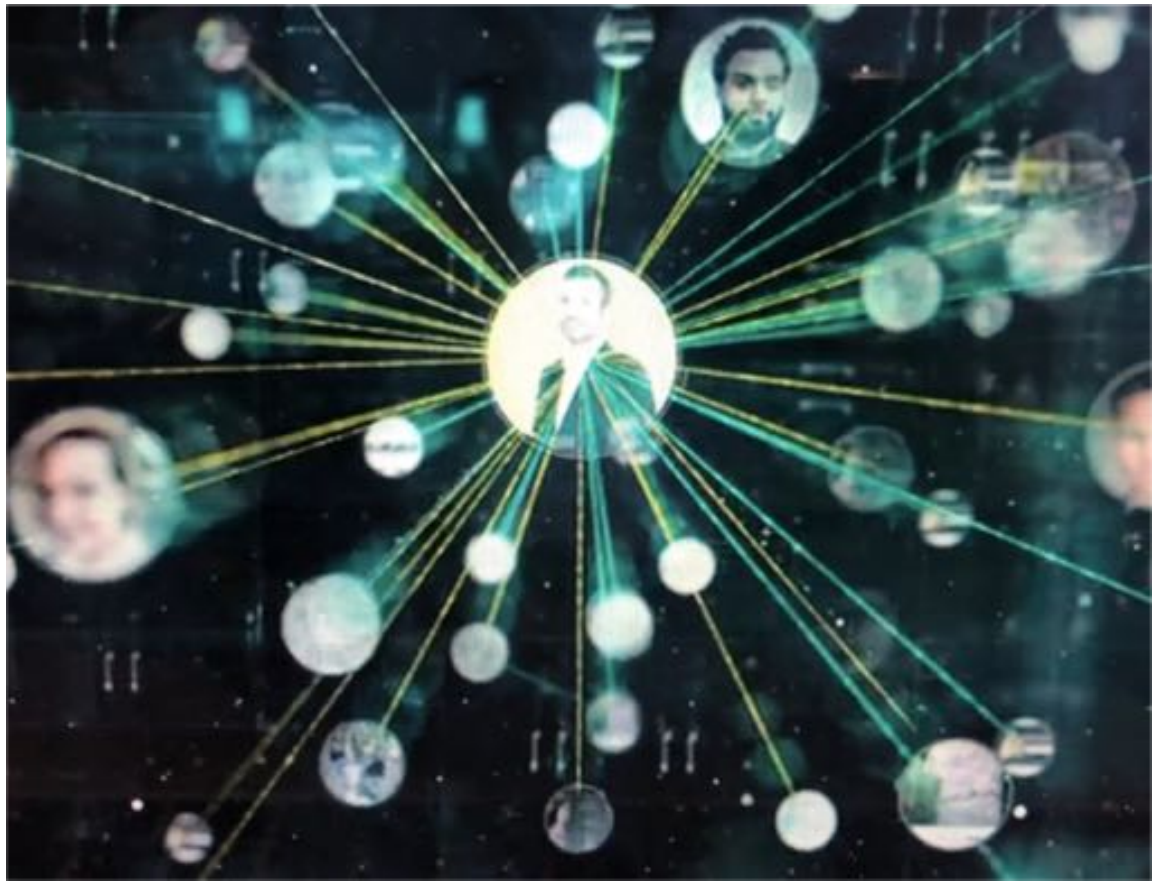
Vice President Business Development & Marketing

sylvain.dubois@crossbar-inc.com

Aug 8th, 2019

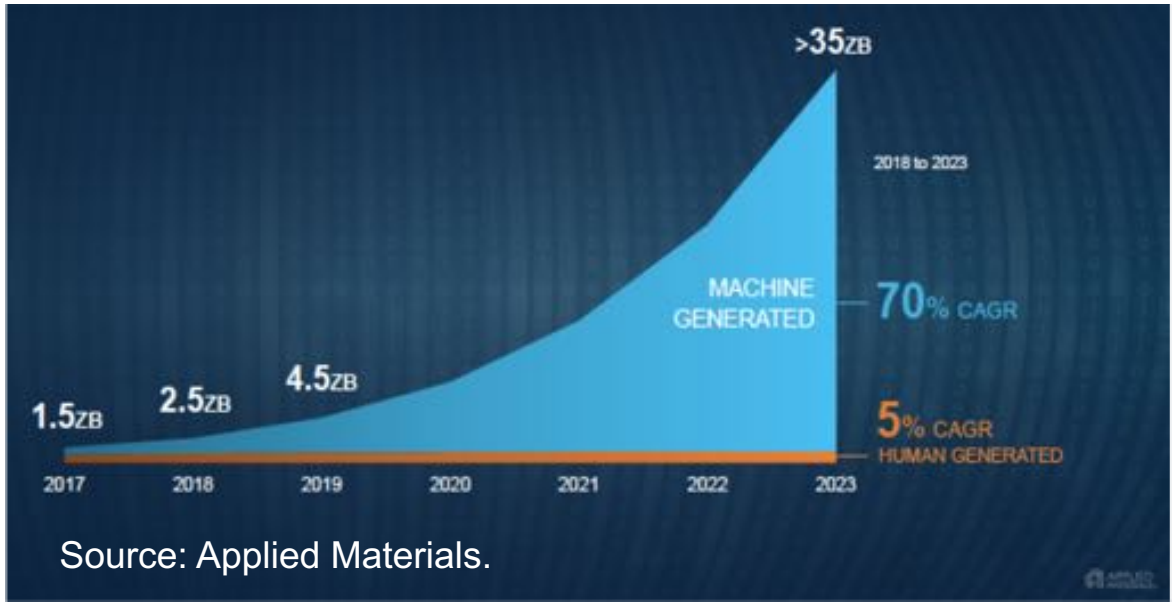
CROSSBAR

Search can be similar to finding a needle in a haystack



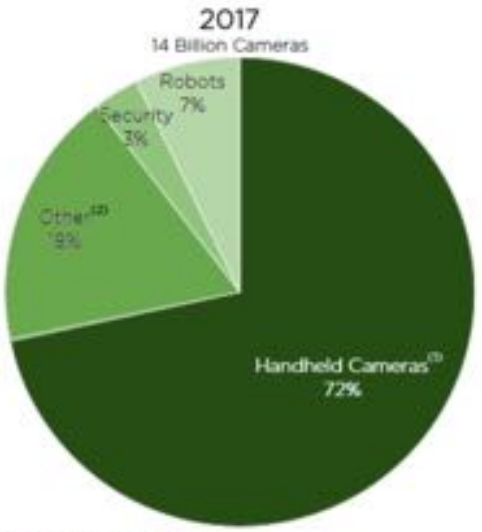
It’s getting even more difficult with machine-generated data growth

>35ZB of data generated in 2023

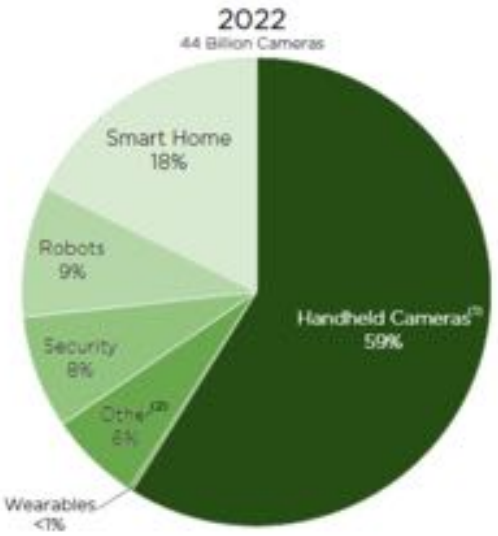


Source: Applied Materials.

14 Billion Cameras in 2017



44 Billion Cameras in 2022



LDV CAPITAL

Find the needle in a greater and greater haystack

Problem: Objects (vectors) Classification in AI

- There is a computing-intensive task required after every Neural Network

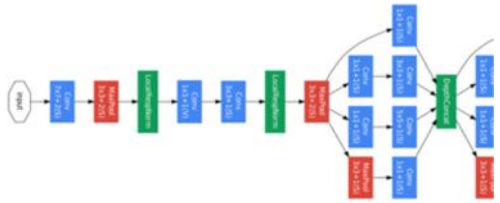
Any data sources

Unstructured datasets

Camera, microphones, sensors…

Neural Networks Accelerators

Features/Vectors Extraction



<v1,v2,……..vn>

Events

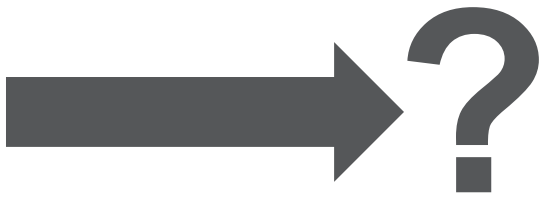
Video

Images

Speech

Keywords

Sensors



For some AI applications, the classification phase can take up to 3X the time than the features extraction with Neural Network

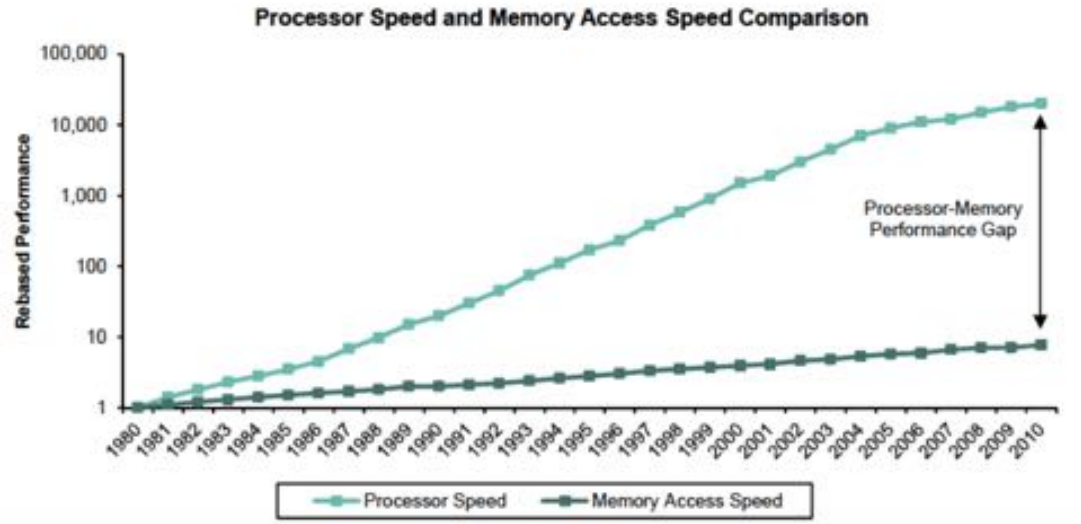
The memory bottleneck

"Memory is the key to enable true intelligence"

BERNSTEIN ARTIFICIAL INTELLIGENCE

MEMORY ACCESS SPEED LIMITATION

EXHIBIT 5: **One of the major limitations of the von Neumann architecture is the "bottleneck" created due to the divergence in performance seen between processor speeds and memory access performance**



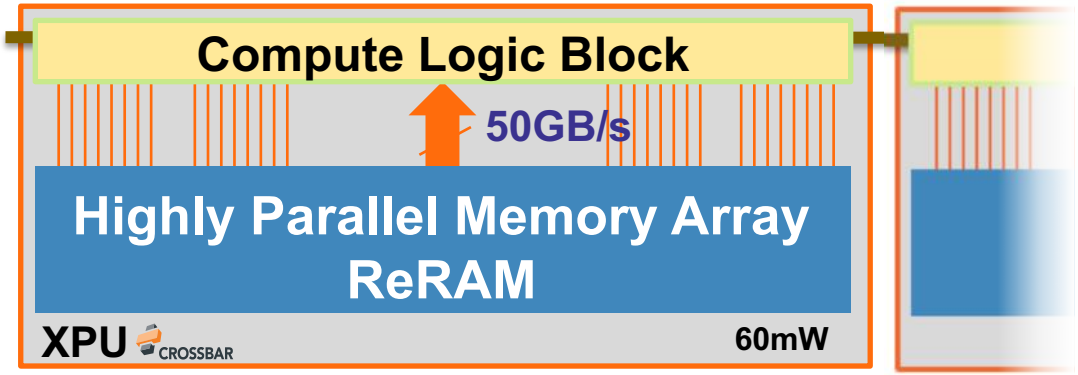
MEMORY ACCESS ENERGY CONSUMPTION

- DDR4 DIMMs: 320 pJ/Byte
- In-package HBM DRAM: 64pJ/Byte
- In-processor SRAM:
  - 6pJ/bit for 8Mbit → 47pJ/bit for 64Mbit
- In-processor Crossbar ReRAM: <0.5pJ/bit

Solution: XPU is a near-memory computing accelerator

Host interface @ 66MHz
xSPI/FIFO interface

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



➢Deterministic perf & persistent memory

- 8-bit signed integer to binary objects
- Object length of 16 to1K
- 1024 to 64K objects per macro
- Manhattan or Cosine distance
- Simultaneous processing
- 3 Billion OLUPS and 53 Billion OLU/Watt

➢Configurable

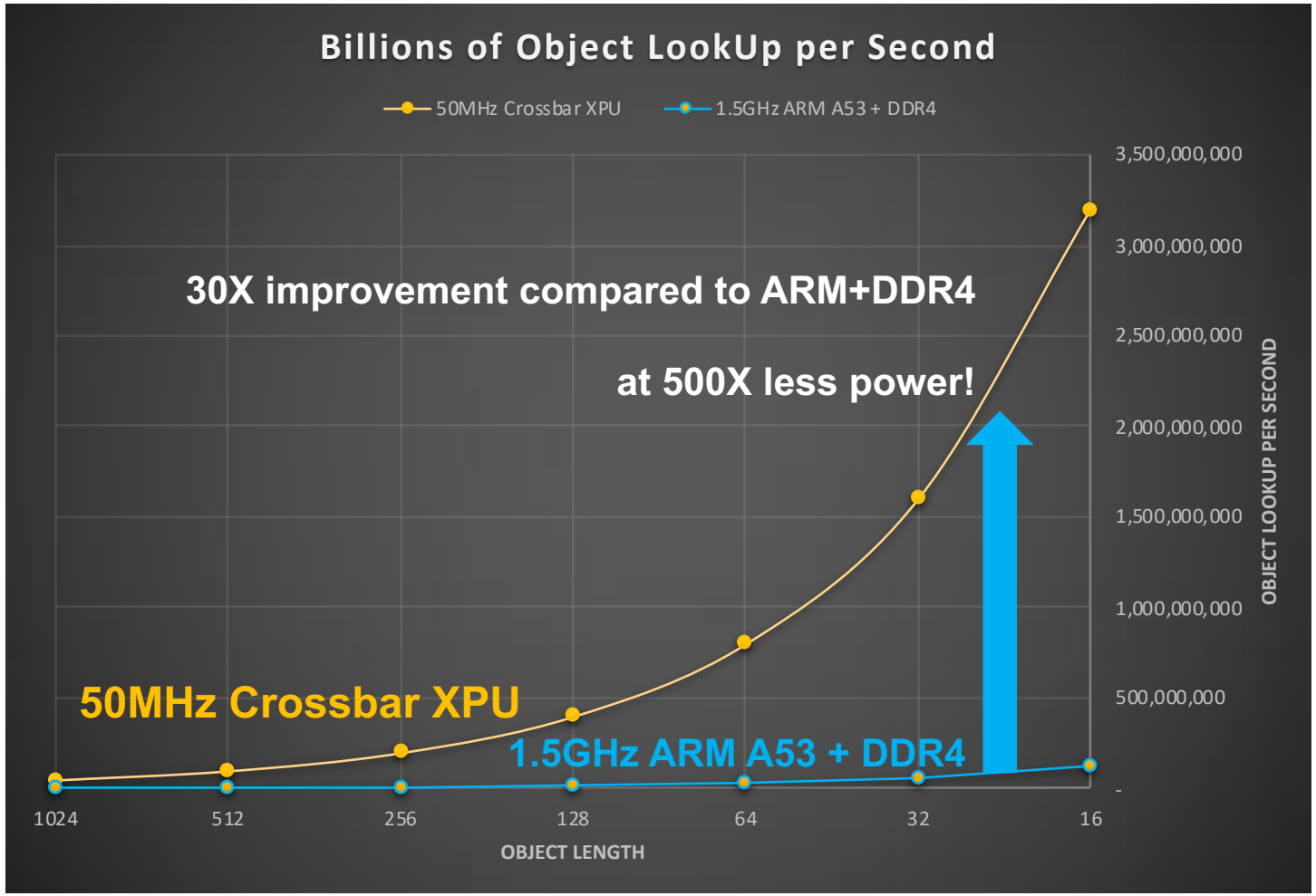
- 8-bit signed integer to binary objects
- Object length of 16 to1K
- 1024 to 64K objects per macro
- Manhattan or Cosine distance

➢Scalable

- Multiple Instances of Macros/Chips can be cascaded to increase # of Instances

Enabling Learning at the Edge

3+ Billion Objects LookUp Per Second (OLUPS)



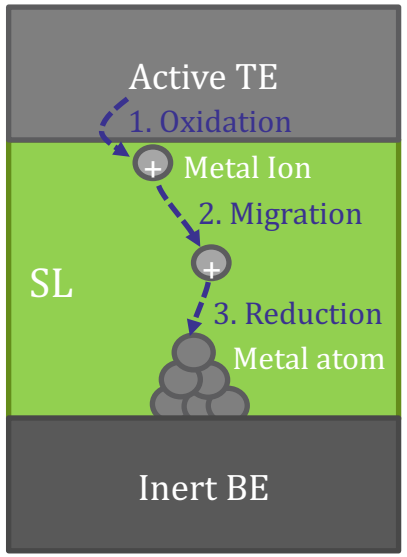
Object length	OLUPS	OLU/Watt
1024	50,000,000	833,333,333
512	100,000,000	1,666,666,667
256	200,000,000	3,333,333,333
128	400,000,000	6,666,666,667
64	800,000,000	13,333,333,333
32	1,600,000,000	26,666,666,667
16	3,200,000,000	53,333,333,333

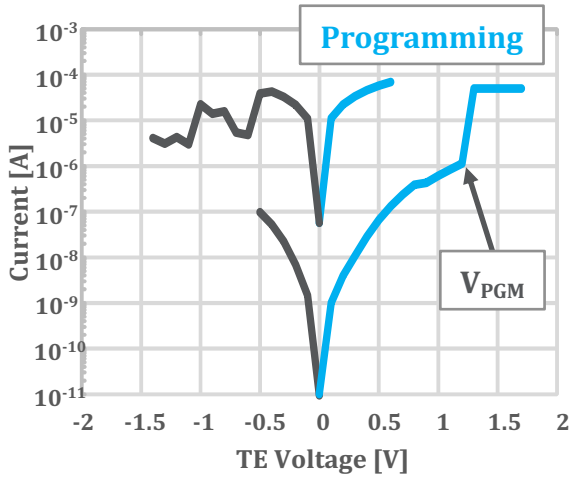
Scalable to 16 Billion OLUPS per stick

Probably more than you need !

CROSSBAR

Enabled by Crossbar ReRAM technology

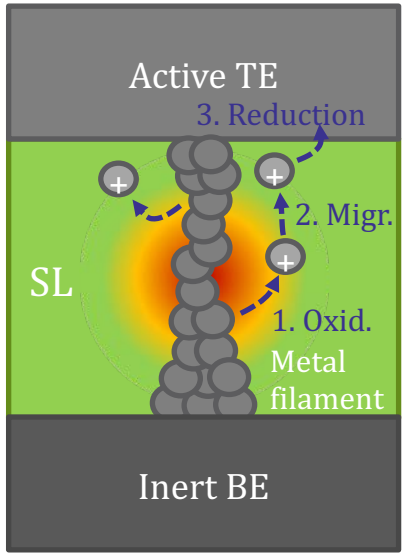


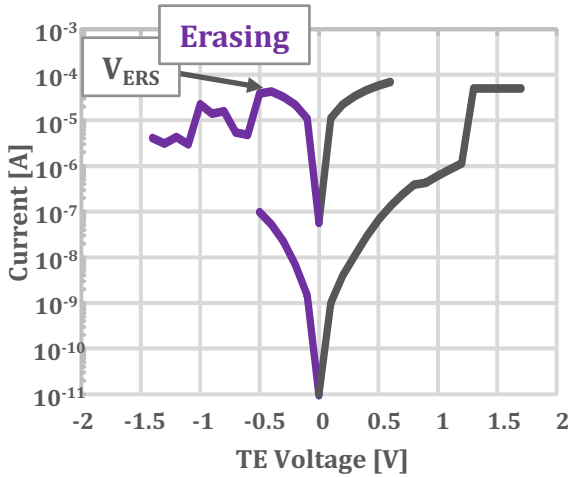


Programming: Positive Voltage on TE

1. Creation of Metal ions from TE oxidation
2. Electro-migration of the ions through the switching layer
3. Reduction of the ions and formation of the filament

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





Erasing: Positive Voltage on BE

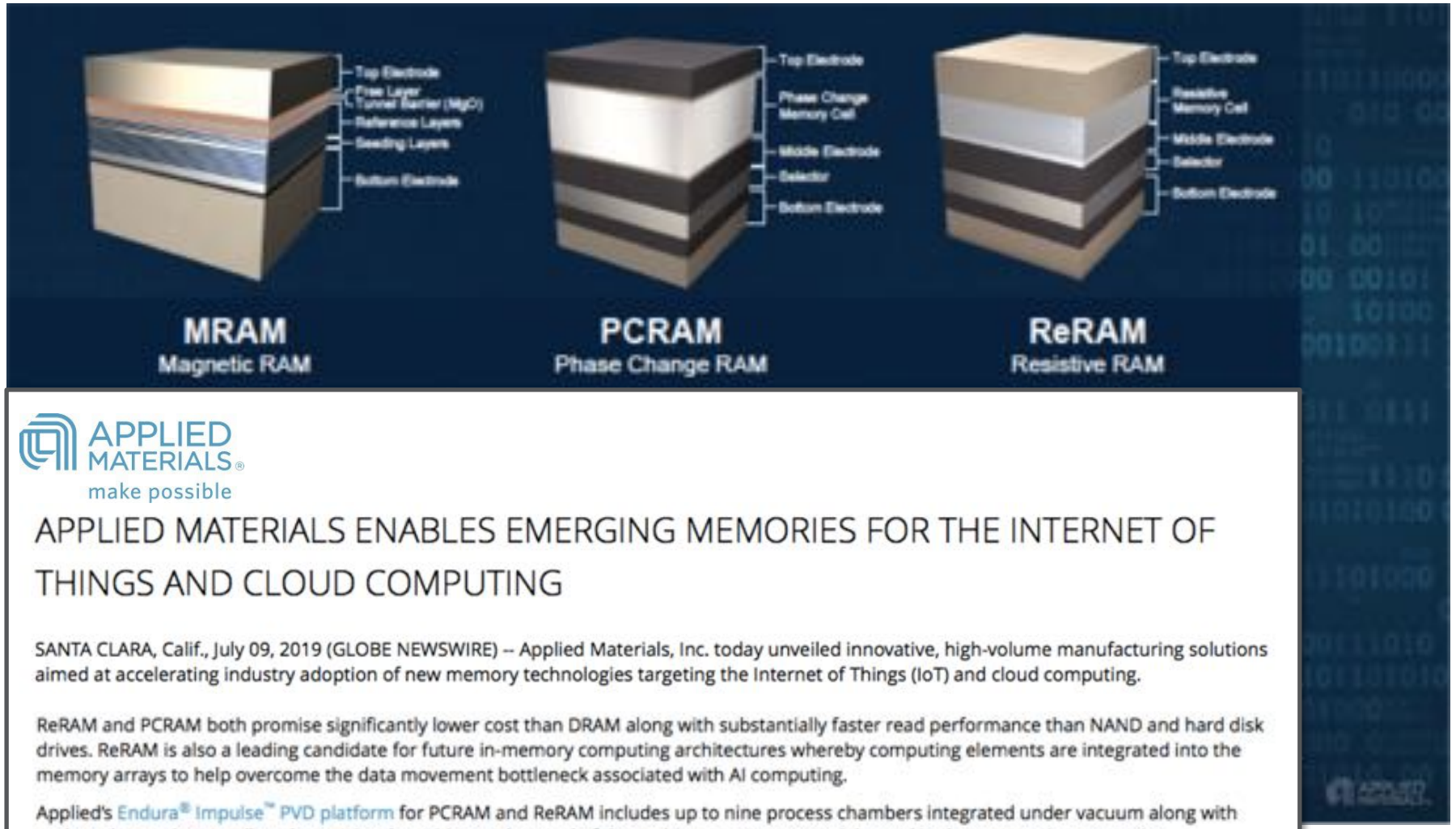
1. Oxidation of the filament atoms through electric field and temperature (Joule Heating)
2. Electro-migration of the ions through the switching layer
3. Reduction of the ions and reformation of the TE

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

Status: from lab to fab

“In a lab, you can certainly create architectures that work with certain characteristics, but then when you go from the lab to high-volume manufacturing and you want to make billions of those devices at high yield, that’s a whole different kettle of fish.”

Gary Dickerson, president and CEO
Applied Materials



APPLIED MATERIALS
make possible

APPLIED MATERIALS ENABLES EMERGING MEMORIES FOR THE INTERNET OF THINGS AND CLOUD COMPUTING

SANTA CLARA, Calif., July 09, 2019 (GLOBE NEWSWIRE) -- Applied Materials, Inc. today unveiled innovative, high-volume manufacturing solutions aimed at accelerating industry adoption of new memory technologies targeting the Internet of Things (IoT) and cloud computing.

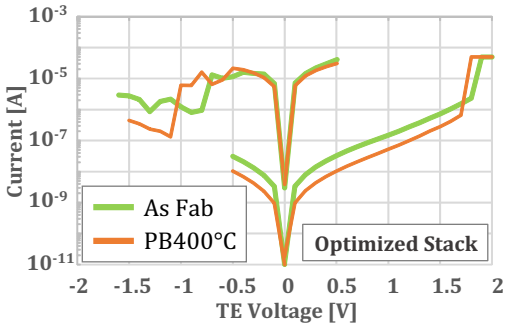
ReRAM and PCRAM both promise significantly lower cost than DRAM along with substantially faster read performance than NAND and hard disk drives. ReRAM is also a leading candidate for future in-memory computing architectures whereby computing elements are integrated into the memory arrays to help overcome the data movement bottleneck associated with AI computing.

Applied's Endura® Impulse™ PVD platform for PCRAM and ReRAM includes up to nine process chambers integrated under vacuum along with on-board metrology to allow the precise deposition and control of the multi-component materials used in these emerging memories.

"Uniform deposition of the new materials used in ReRAM memories is critical to achieving the highest possible device performance, reliability and endurance," said George Minassian, CEO and co-founder of Crossbar, Inc. "We specify the Applied Materials Endura Impulse PVD system with on-board metrology in our ReRAM technology engagements with memory and logic customers because it enables a breakthrough in these critical metrics."

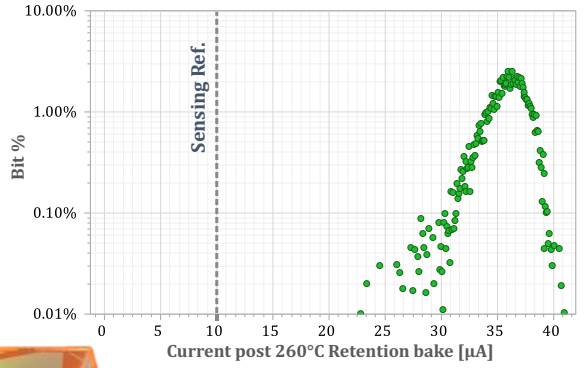
Latest silicon results

CMOS integration compatibility

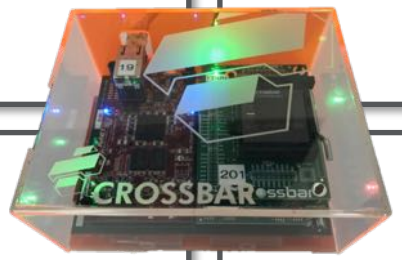


Crossbar's ReRAM is capable of withstanding standard 400C alloy annealing

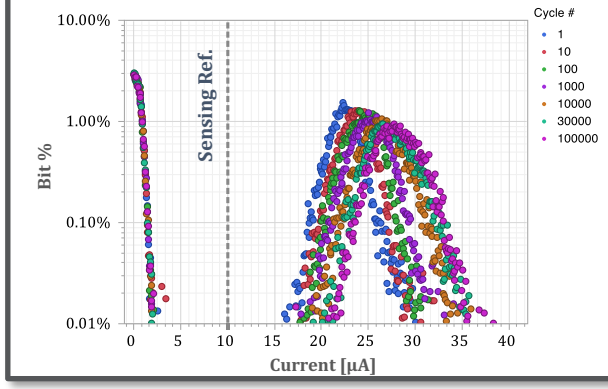
Soldering Reflow Compatibility



Crossbar's ReRAM is capable of maintaining perfect ON Retention after 260C Retention bake



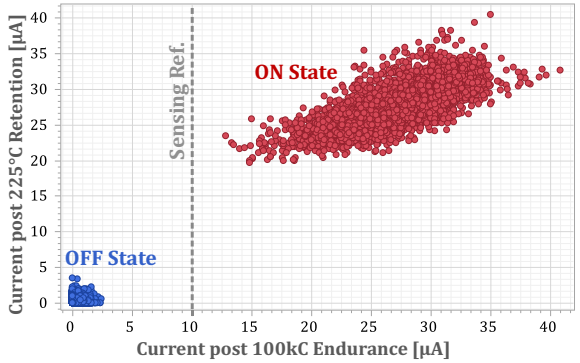
Endurance



Crossbar's ReRAM is capable of sustaining beyond 100kC

1M+ cycles demonstrated at FMS

Retention



Crossbar's ReRAM is exhibiting extremely good Retention after 225C Bake and 100kC cycles

Crossbar ReRAM Advantages

	Target Commercial Crossbar ReRAM 40/22nm	Commercial Embedded Flash 40nm	Anticipated Oxygen ions based RRAM 40nm	Anticipated Embedded MRAM 22nm
Physical Mechanism & on/off ratio	**Metal atoms storage 80~120X on/off ratio**	Electron storage 3~6X on/off ratio	Oxygen ions storage	Spin-polarized current 1.3~1.7X on/off ratio
Stack complexity	**Simple**	Complex dedicated CMOS lines	Simple	Super complex 10+ layers stack
Materials involved	**3 films Existing materials**	Existing materials	3 films Existing materials	>25 materials
Mask layer adder	**2 masks**	6+ masks	2 masks	5 masks
Speed Read	**15ns**	25ns	25ns	20ns
Speed Write	**10us**	12us	30us	300ns
Read energy	**Low 0.2 uA/MHz/bit**	Low 0.77 uA/MHz/bit	Medium 1.2 uA/MHz/bit	High 2 uA/MHz/bit
Write current	**Low ~60uA/bit**	Complex access block erase only	High > 250uA/bit	High 300uA/bit
Standby current	**Low 2 uA**	Super high > 150uA	Medium > 4uA	Super high 200 uA
Data retention	**> 10Yr**	> 10Yr	> 10Yr	> 10Yr
Endurance	**> 1M**	10K / 100K	10K	1M
Operating temp	**125C**	150C	125C	150C
Magnetic Immunity	**YES**	YES	YES	NO

Crossbar ReRAM

Scales below 2xnm

10X Simpler than MRAM

2X Fewer Masks
10X Fewer materials .vs MRAM

Faster read

3X-10X Lower energy

High reliability Magnetic immunity

Crossbar: Make an impact on Edge and Cloud computing

Intelligence & Learning at the Edge

Multi-modal event detection
People re-identification

Reduce TCO and power for hyperscale players

3X lower cost than DRAM & 8X lower energy
$1K reduction per server

Summary

- AI = Finding the needle in a haystack
- Need for more efficient memory access to absorb data explosion
- Solution is to bring data closer to computing
- Crossbar XPU delivering Billions OLUPS
- Enabled by ReRAM
  - Forming Free with DC voltage below 2V
  - Compatible with CMOS integration
  - Compatible with pre soldering reflow programing
  - Extremely good Endurance and Retention beyond 100kC
- In high-volume manufacturing

Crossbar moving the needle in Edge and Cloud Computing

CROSSBAR