Search acceleration and Learning at the Edge with Crossbar ReRAM

Sylvain Dubois

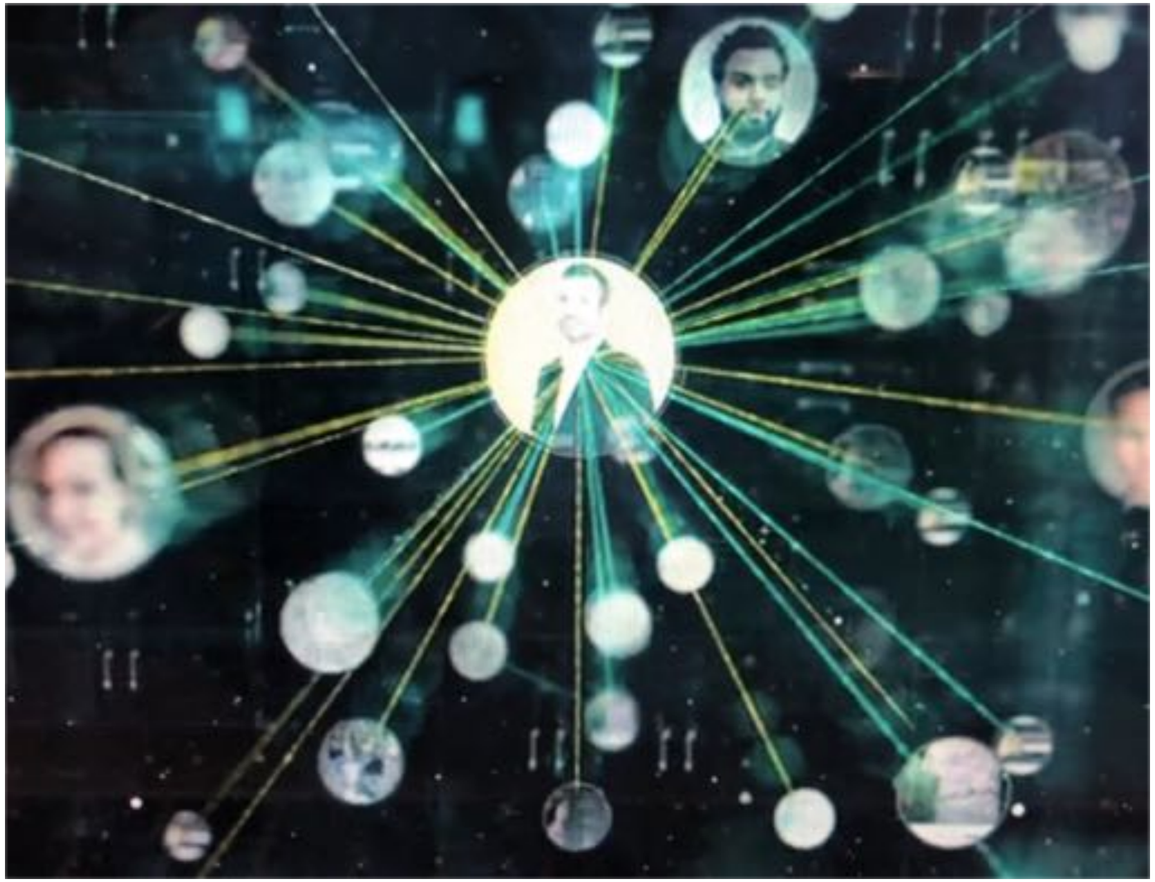
Vice President Business Development & Marketing

sylvain.dubois@crossbar-inc.com

Aug 29th, 2019

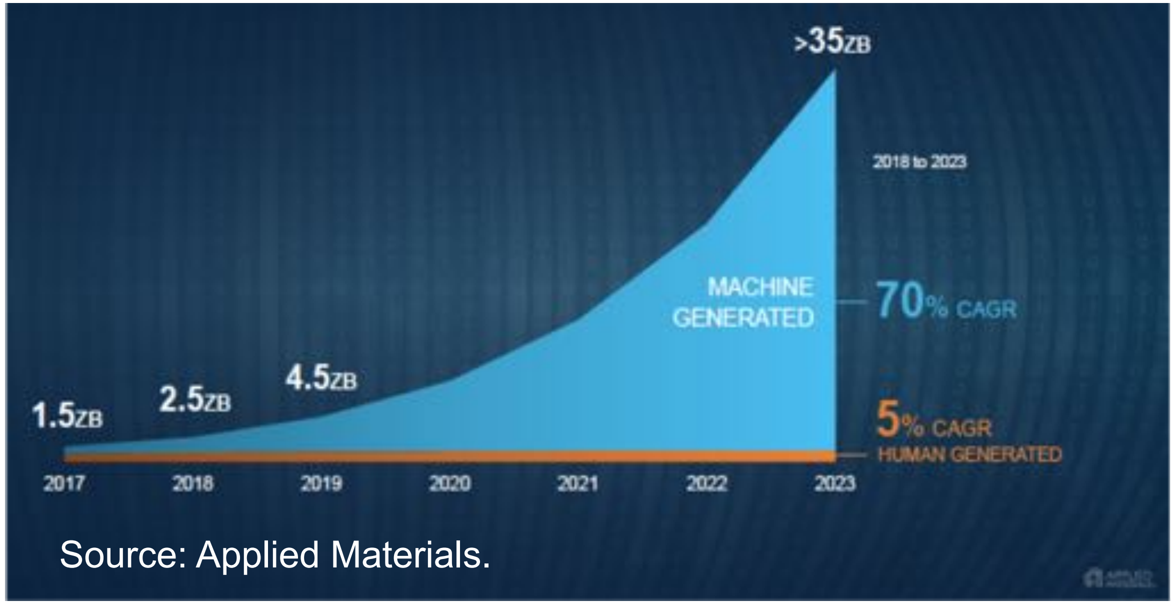
CROSSBAR

Search can be similar to finding a needle in a haystack



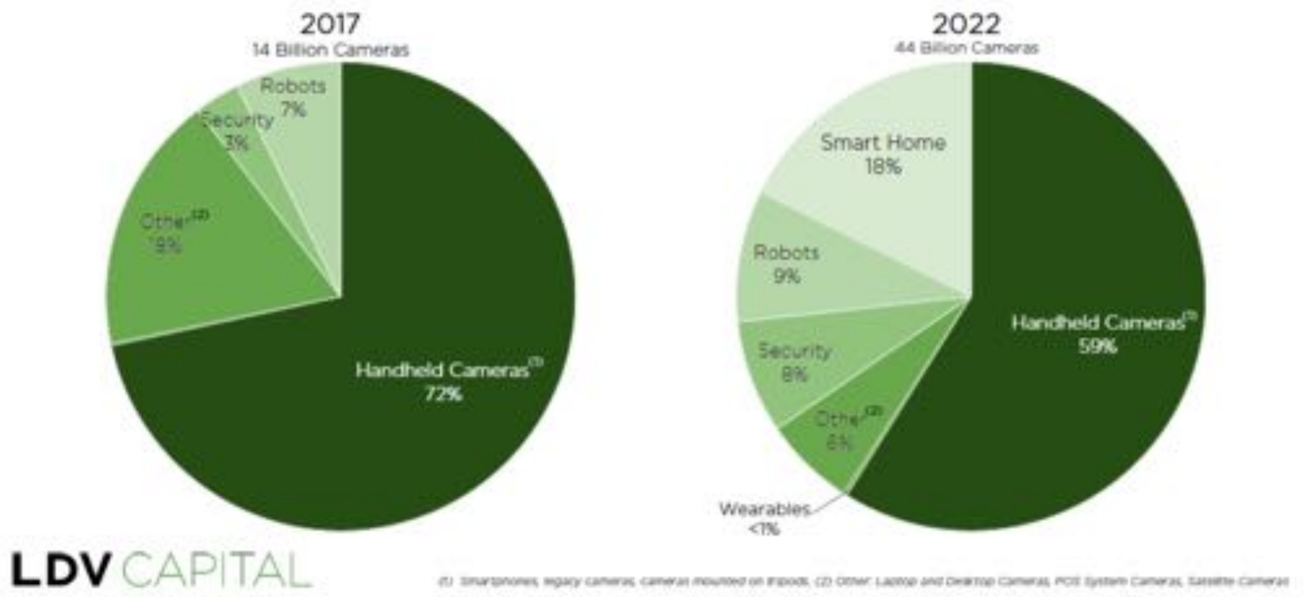
It’s getting even more difficult with machine-generated data growth

>35ZB of data generated in 2023



Source: Applied Materials.

14 Billion Cameras in 2017 44 Billion Cameras in 2022



(1) Smartphones, legacy cameras, cameras mounted on tripods. (2) Other: Laptop and Desktop Cameras, POS System Cameras, Satellite Cameras

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

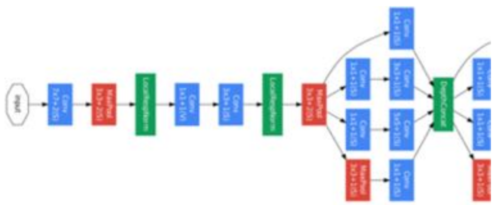
Any data sources

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

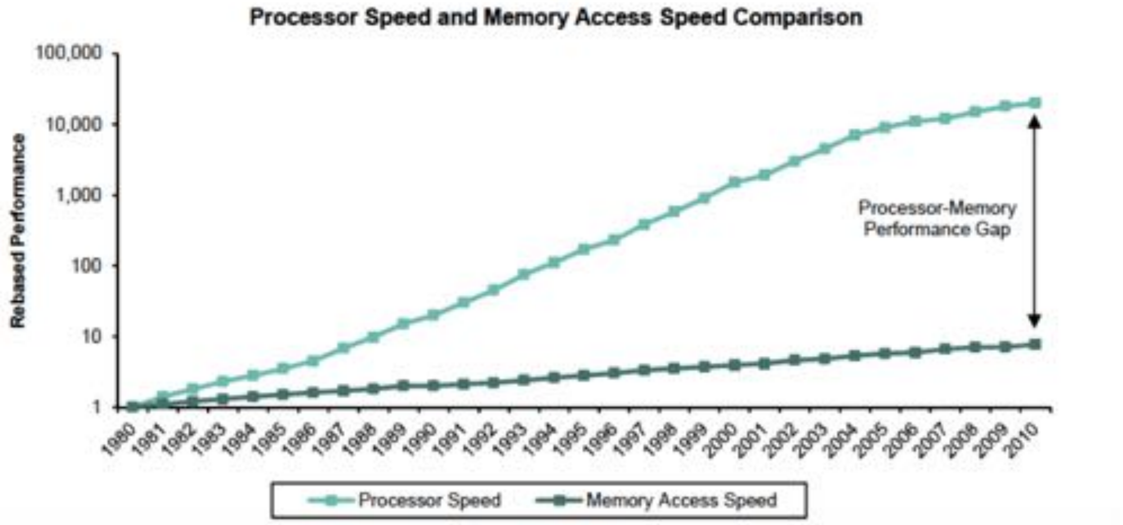
The memory bottleneck

> “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

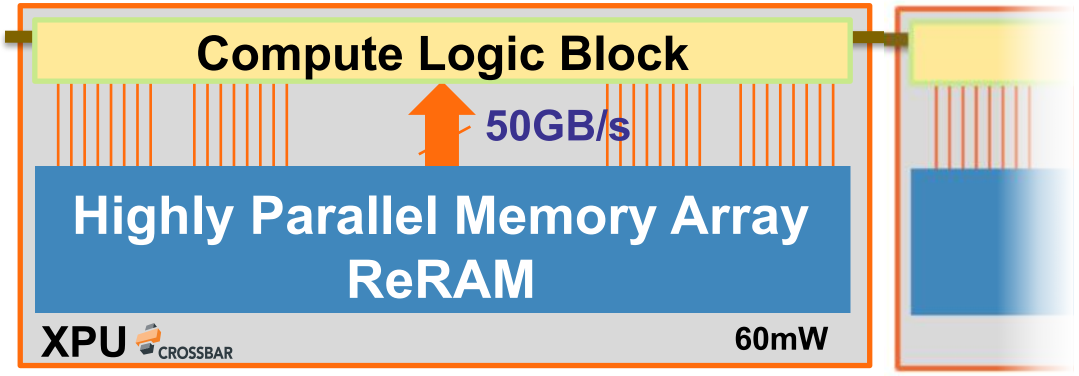
- DDR4 DIMMs: 320 pJ/Byte
- In-package HBM DRAM: 64pJ/Byte
- In-processor SRAM:
  - 6pJ/bit for 8Mbit → 47pJ/bit for 64Mbit
- In-processor Crossbar ReRAM: <0.5pJ/bit

Solution: XPU is a near-memory computing accelerator

Host interface @ 66MHz

xSPI/FIFO interface

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



Compute Logic Block

50GB/s

Highly Parallel Memory Array ReRAM

XPU CROSSBAR

60mW

➢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

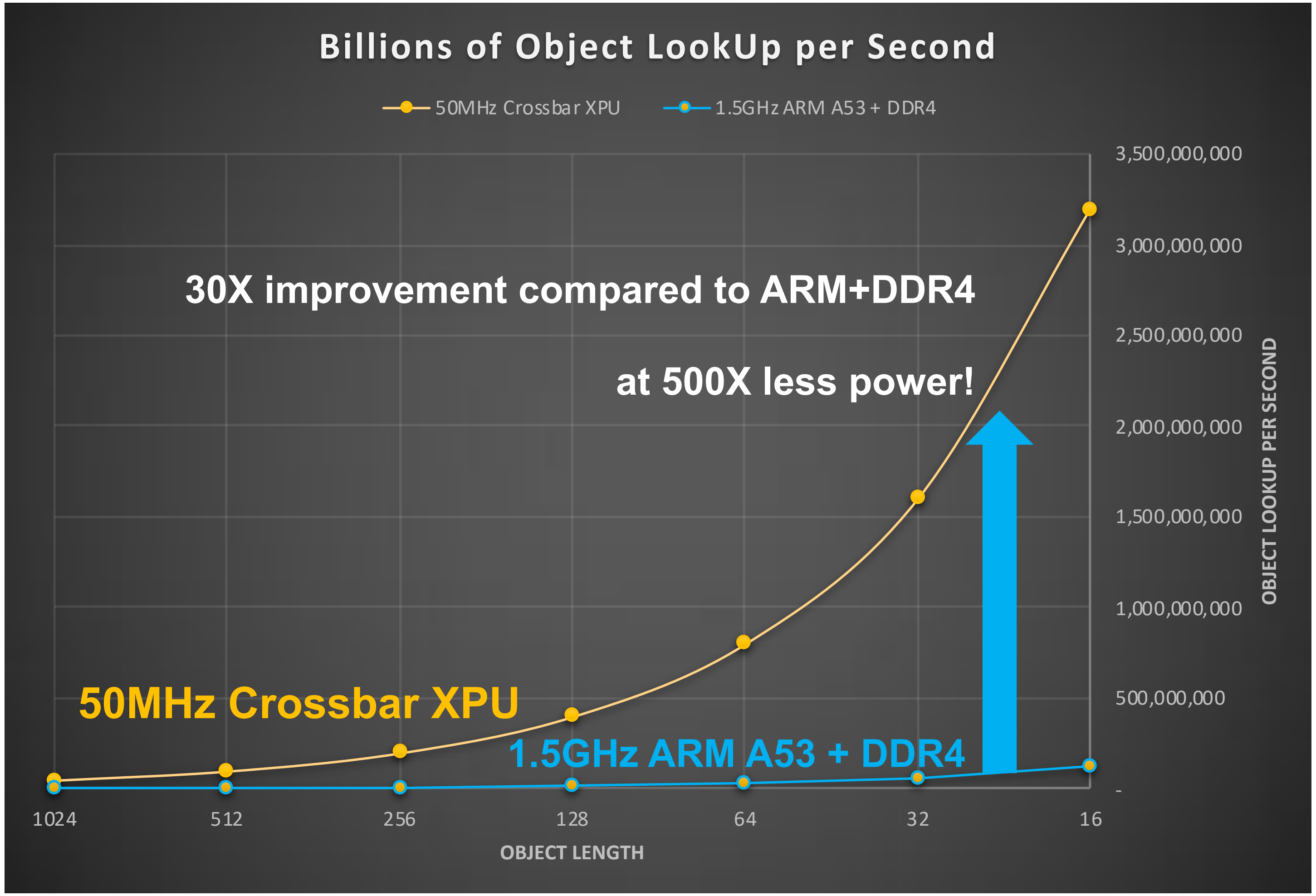
- ○ 8-bit signed integer to binary objects
- ○ Object length of 16 to1K
- ○ 1024 to 64K objects per macro
- ○ Manhattan or Cosine distance

➢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)



Billions of Object LookUp per Second

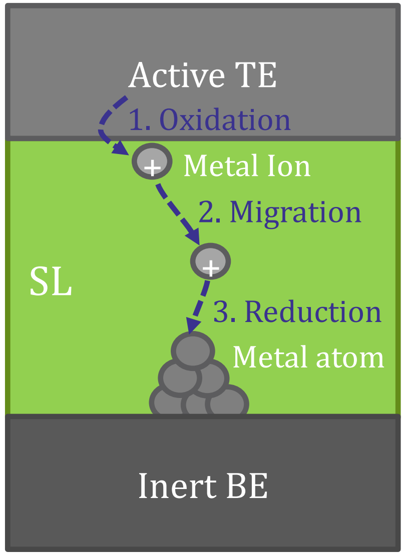
30X improvement compared to ARM+DDR4

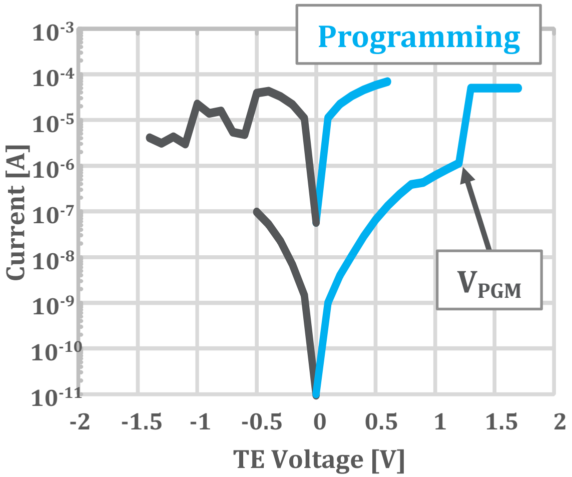
at 500X less power!

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

Enabled by Crossbar ReRAM technology

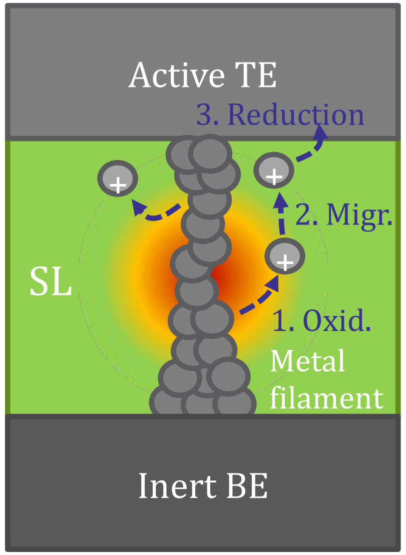


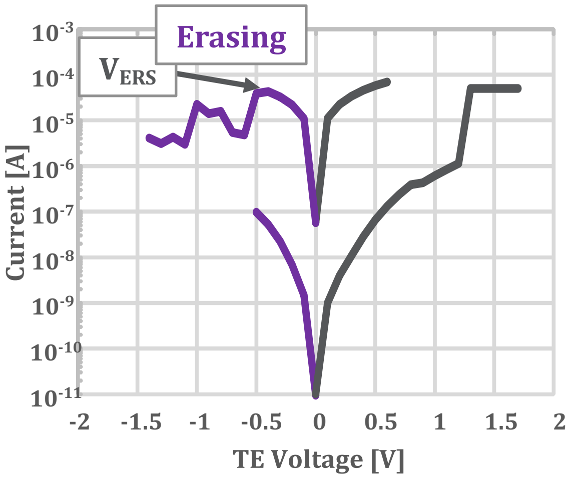


Programming: Positive Voltage on TE

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**



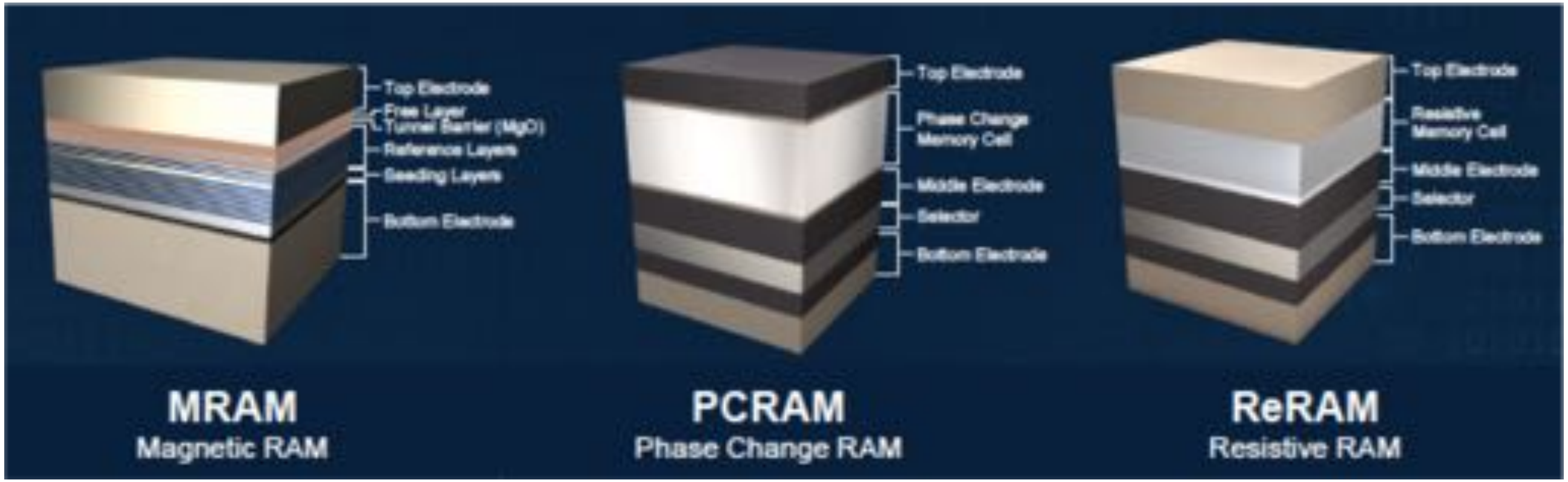


Erasing: Positive Voltage on BE

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

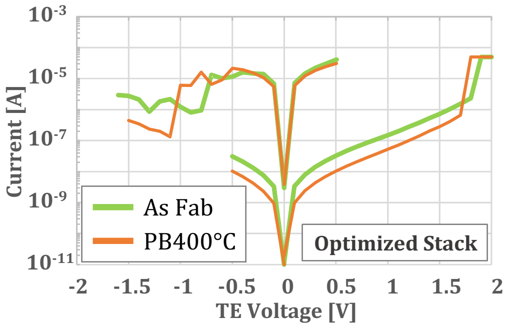


“In a lab, you can certainly create architectures that work with certain characteristics, but then when you go from the lab to high-volume manufacturing and you want to make billions of those devices at high yield, that’s a whole different kettle of fish.”

Gary Dickerson, president and CEO
Applied Materials

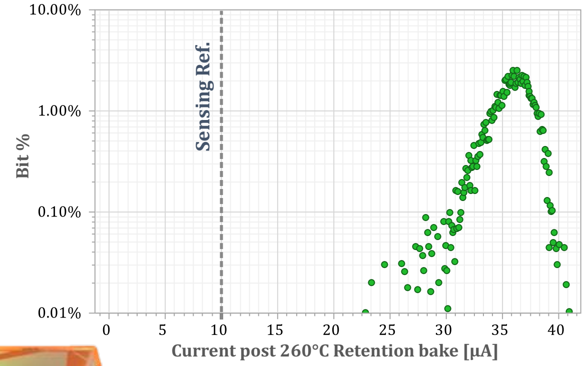
Latest silicon results

CMOS integration compatibility



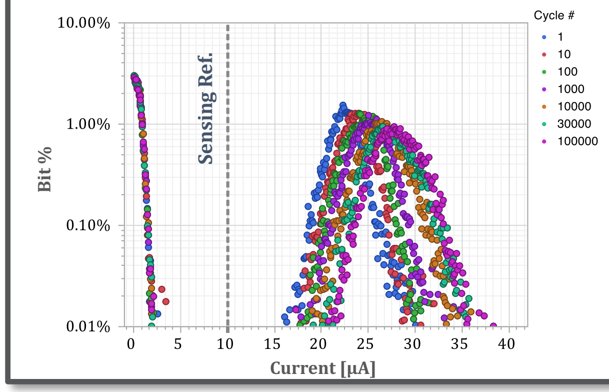
Crossbar’s ReRAM is capable of withstanding standard 400C alloy annealing

Soldering Reflow Compatibility



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

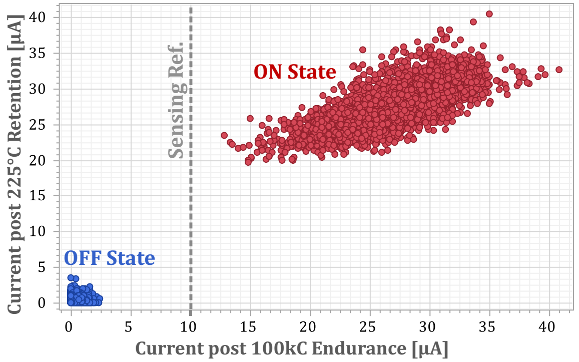
Endurance



Crossbar’s ReRAM is capable of sustaining beyond 100kC

1M+ cycles demonstrated at Flash Memory Summit

Retention



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

Crossbar ReRAM Advantages

	Target Commercial Crossbar ReRAM 28/22nm	Commercial Embedded Flash 40nm	Anticipated Oxygen ions based RRAM 40nm	Anticipated Embedded MRAM 22nm	Crossbar ReRAM
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	**Scales below 2xnm**
Stack complexity	**Simple**	Complex dedicated CMOS lines	Simple	Super complex 10+ layers stack	**10X Simpler than MRAM**
Materials involved	**3 films Existing materials**	Existing materials	3 films Existing materials	>25 materials	**2X Fewer Masks 10X Fewer materials .vs MRAM**
Mask layer adder	**2 masks**	6+ masks	2 masks	5 masks	
Speed Read	**15ns**	25ns	25ns	20ns	**Faster read**
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	**3X-10X Lower energy**
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	**High reliability Magnetic immunity**
Endurance	**> 1M**	10K / 100K	10K	1M	
Operating temp	**125C**	150C	125C	150C	
Magnetic Immunity	**YES**	YES	YES	NO	

CROSSBAR

Crossbar: Make an impact on Edge and Cloud computing

Intelligence & Learning at the Edge

Multi-modal event detection
People re-identification

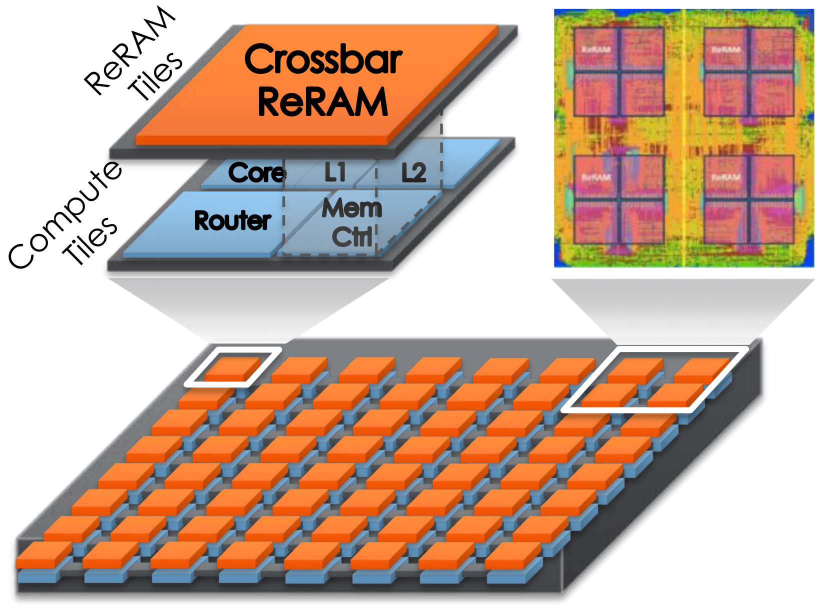
Reduce TCO and power for hyperscale players

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

EDGE COMPUTING

CLOUD COMPUTING

Research programs for ReRAM-based monolithic computers



ReRAM Tiles integrated with 256-bit RISC-V

Monolithic ReRAM + CPUs die

- 1TB/s access to ReRAM tiles
- 500X denser than SRAM
- 10X energy efficiency over stacked DRAM

University of Maryland Technical Report UMIACS-TR-2019-01, July 2019.

Design and Evaluation of Monolithic Computers Implemented Using Crossbar ReRAM

Meenatchi Jagasivamani, Candace Walden, Devesh Singh, Shang Li, Luyi Kang, Mehdi Asnaashari, Sylvain Dubois, Bruce Jacob, and Donald Yeung

University of Maryland and Crossbar Incorporated

ABSTRACT

A *monolithic computer* is an emerging architecture in which a multicore CPU and a high-capacity main memory system are all integrated in a single die. We believe such architectures will be possible in the near future due to non-volatile memory technology, such as the resistive random access memory, or *ReRAM*, from Crossbar Incorporated. Crossbar's ReRAM can be fabricated in a standard CMOS logic process, allowing it to be integrated into a CPU's die. The ReRAM cells are manufactured in between metal wires and do not employ per-cell access transistors, leaving the bulk of the base silicon area vacant. This means that a CPU can be monolithically integrated directly underneath the ReRAM memory, allowing the cores to have massively parallel access to the main memory.

This paper presents the characteristics of Crossbar's ReRAM technology, informing architects on how ReRAM can enable monolithic computers. Then, it develops a CPU and memory system architecture around those characteristics, especially to exploit the unprecedented memory-level parallelism. The architecture employs a tiled CPU, and incorporates memory controllers into every compute tile that support a variable access granularity to enable high scalability. Lastly, the paper conducts an experimental evaluation of monolithic computers on graph kernels and streaming computations. Our results show that compared to a DRAM-based tiled CPU, a monolithic computer achieves 4.7x higher performance on the graph kernels, and achieves roughly parity on the streaming computations. Given a future 7nm technology node, a monolithic computer could outperform the conventional system by 66% for the streaming computations.

1. INTRODUCTION

In the post-Moore era, computer architects will no longer be able to rely on technology scaling. Instead, they will need to look for new technologies to continue fueling architectural innovation. This paper explores one such possibility: *monolithic computers*. A monolithic computer relies on new logic-memory integration technology to fabricate a CPU and a high-capacity main memory system all on a single die. Compared to conventional package-level integration involving multiple dies—*e.g.*, stacking DRAM dies on top of a logic layer or integrating DRAM and CPU dies over a silicon interposer—a monolithic computer achieves much higher integration of the CPU and memory system.

Whereas package-level integration can support thousands of wires between the CPU and main memory, monolithic integration will be able to support *millions of wires*, providing a much wider main memory interface than is currently possible. This will enable architects to deliver greater memory parallelism and bandwidth to data-intensive computations, and achieve superior bandwidth-per-watt. In addition, monolithic computers will reduce the physical distance that memory requests will need to travel. Because all memory requests can stay on the CPU die, there won't be a need to cross the silicon interposer, or worse, to traverse the system motherboard. This locality benefit can provide significant additional improvements in power efficiency.

Monolithic computers do not exist yet, but some researchers believe emerging non-volatile memories will change that in the future [1, 2]. Aly *et al* [1] argue that spin-transfer torque magnetic RAM (STT-MRAM) or resistive RAM (ReRAM) are such enabling memory technologies. Unlike conventional DRAM which requires special memory fabrication processes, *STT-MRAM and ReRAM can be fabricated in a standard CMOS logic process. Hence, they have the potential to be integrated into a CPU's die.*

The main hurdle, though, is identifying a suitable integration technology. Fine-grained 3D monolithic integration [1, 2] has been proposed as a possible solution, one that assumes a process technology in which multiple planar layers of silicon can be fabricated monolithically in 3D. This would allow compute logic and non-volatile memory to be integrated in alternating planar layers. While this results in extremely high logic and memory densities, unfortunately, it requires advanced process technology that is still at the developmental stages in research labs.

In our work, we assume a much simpler integration approach that exploits non-volatile memories with 3D crosspoint architectures. In crosspoint memories-examples include Intel's Optane [3] as well as the ReRAM technology from Crossbar Incorporated [4]-the memory cells are fabricated in between metal wires of a CMOS logic process, *i.e.* at the intersection of wires laid out perpendicularly in adjacent metal layers. Rather than isolate individual cells using access transistors, crosspoint arrays provide inter-cell isolation via selector devices integrated with the memory cells. So, there are no transistors within the core of the crosspoint arrays. Instead, *the bulk of the silicon area underneath the memory are free for implementing non-memory circuits.*

It is well known that such crosspoint memories can be layered on top of compute logic during back-end of line (BEOL) processing, for example in the top metal layers of a CPU's die. Although there are no per-cell transistors, some peripheral logic is still needed at each crosspoint array for

http://maggini.eng.umd.edu/pub/UMIACS-TR-2019-01.pdf
http://maggini.eng.umd.edu/pub/monolithic-memsys18.pdf

Summary

- Huge demand for more efficient memory access in AI
- 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

Crossbar moving the needle in Edge and Cloud Computing

CROSSBAR