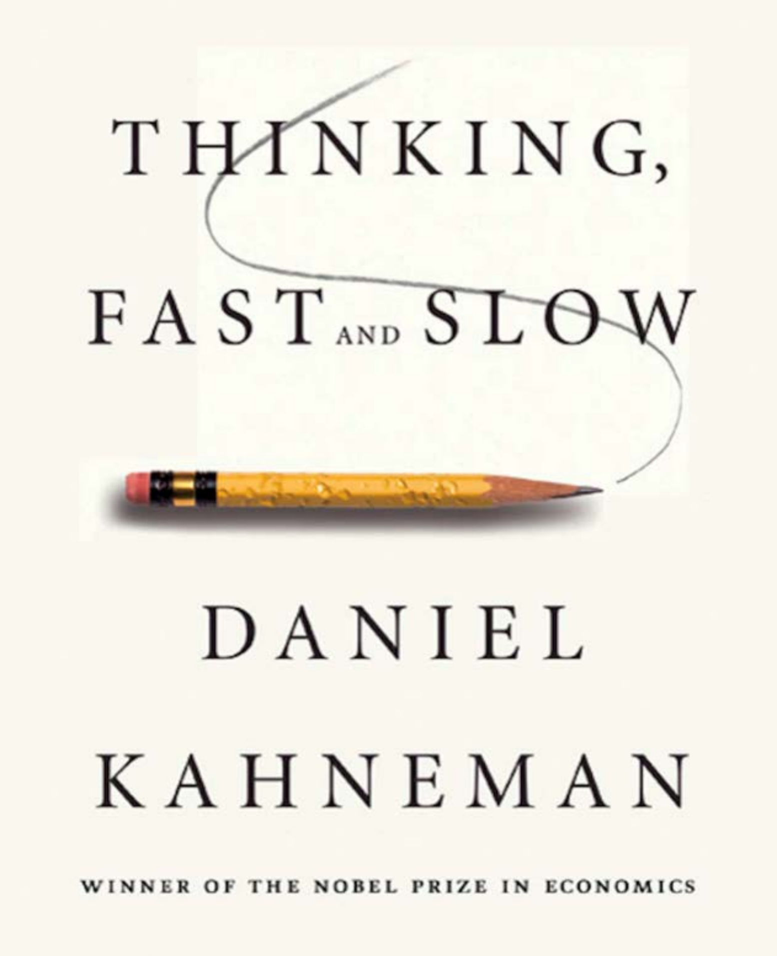
SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

Thinking Fast & Slow: Intuition, Reasoning and Emerging Memory

Dave Eggleston
Intuitive Cognition Consulting

Abstract

- Our human brain can be modeled as two distinctly different systems: a real time intuition system, and a background reasoning system.
- As we move into the AI compute era, Emerging Memory technologies play an increasingly important role in overcoming the limitations of DRAM and NAND.
- The commercialization of Emerging Memory will therefore accelerate our realization of powerful AI systems.



THINKING, FAST AND SLOW

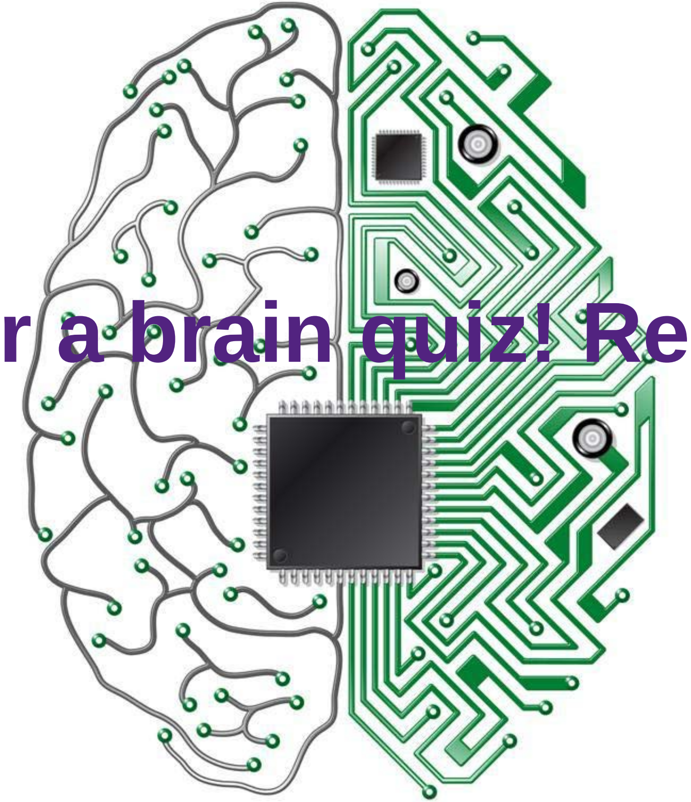
DANIEL KAHNEMAN

WINNER OF THE NOBEL PRIZE IN ECONOMICS

“A groundbreaking tour of the mind, and explains the two systems that drive the way we think.”

“System 1 is fast, intuitive, and emotional; System 2 is slower, more deliberative, and more logical.”

Daniel Kahneman is professor emeritus of psychology and public affairs at Princeton University.



Time for a brain quiz! Ready?

17 x 24 = ?

Intuition

17 x 24 = ?

Reasoning

Intuition

System 1

- ❑ Lightning fast
- ❑ Automatic
- ❑ Real time
- ❑ Effortless
- ❑ Approximate

Edge

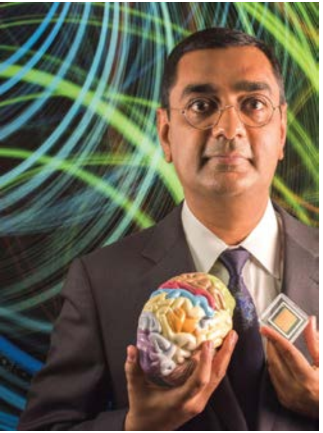
Reasoning

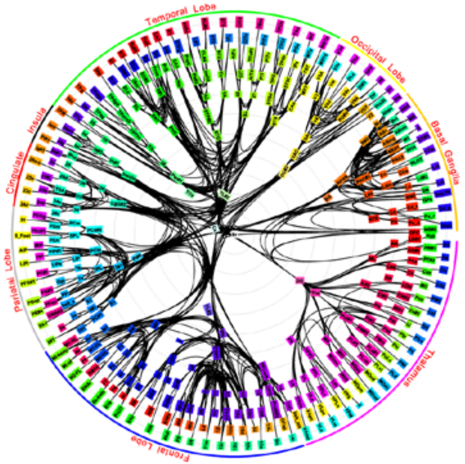
System 2

- ❑ Slow
- ❑ Interrupt driven
- ❑ Background
- ❑ Energy inefficient
- ❑ Precise

Datacenter

Edge

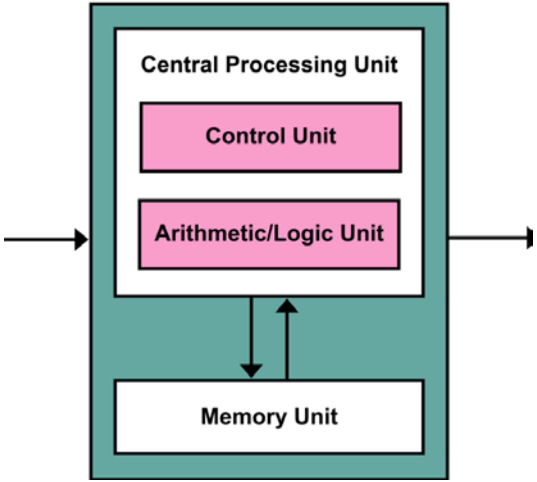




□ Non-von Neumann architecture

Intuition

Datacenter



□ von Neumann architecture

Reasoning

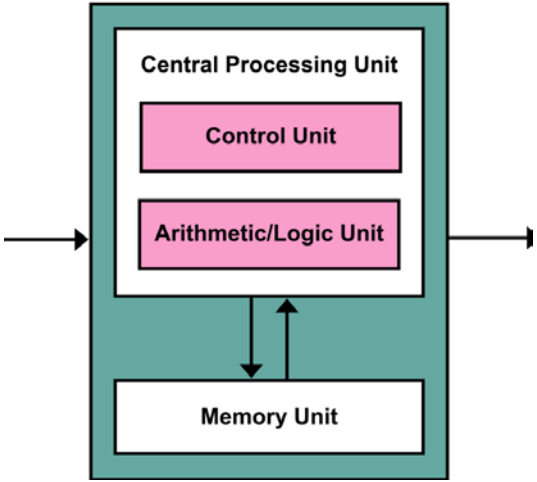
Edge

Store Dr. Moda for now; we'll come back to discuss Edge & Intuition

❑ Non-von Neumann architecture

Intuition

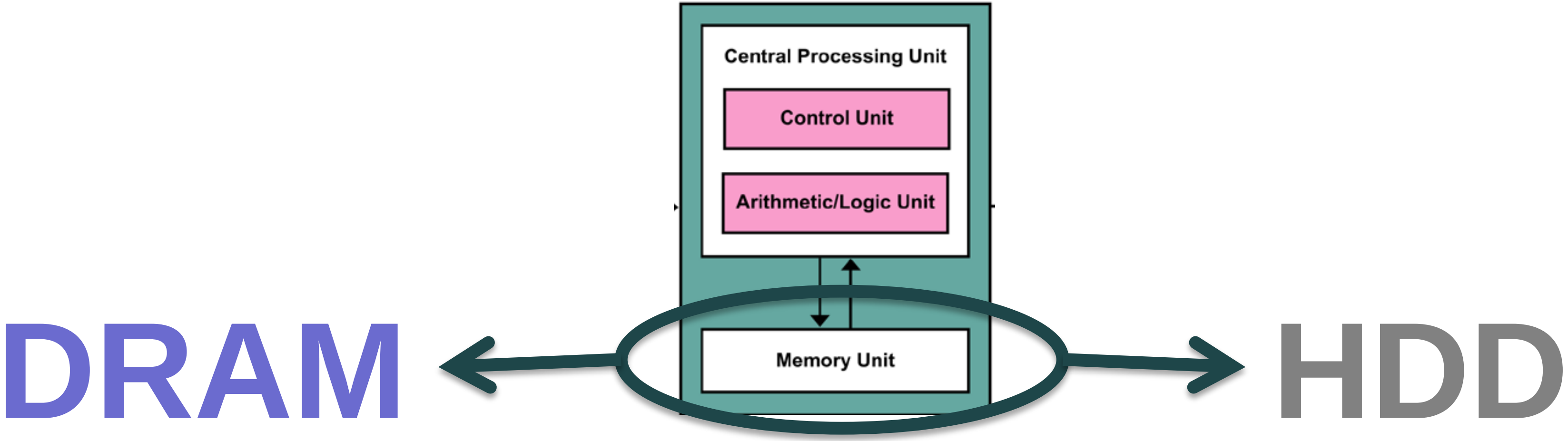
Datacenter



❑ von Neumann architecture

Reasoning

Reasoning

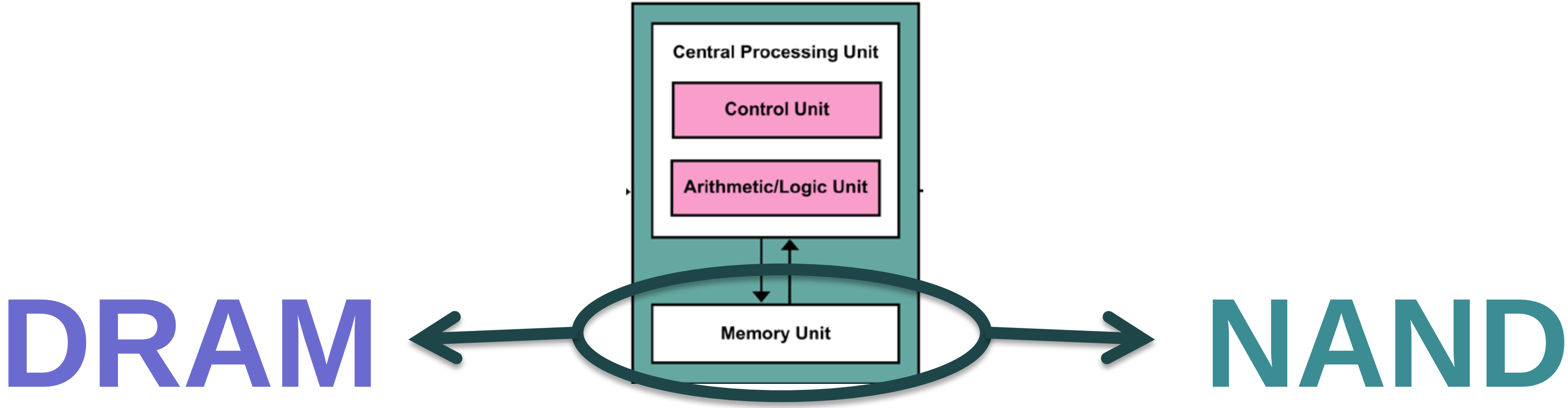


DRAM

HDD

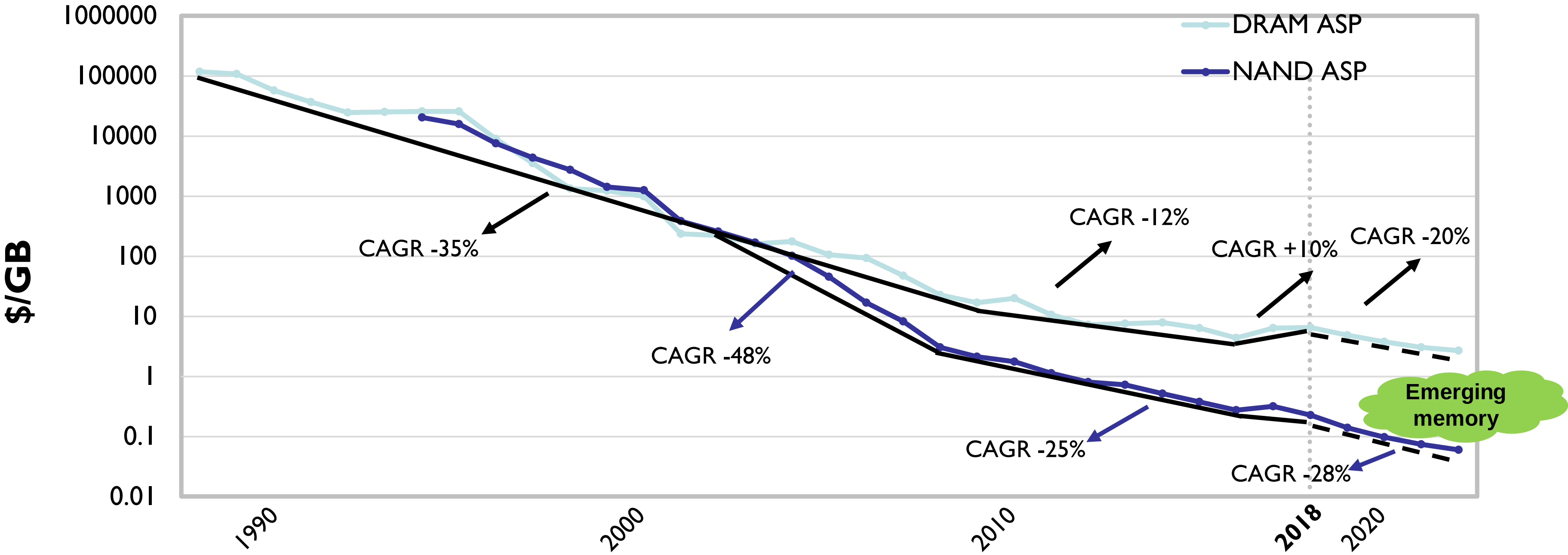
- Once upon a time, long ago…

Reasoning



- And for a while we were happy! ☺ ☺ ☺

Moore's Law is slowing – but still need low cost bits

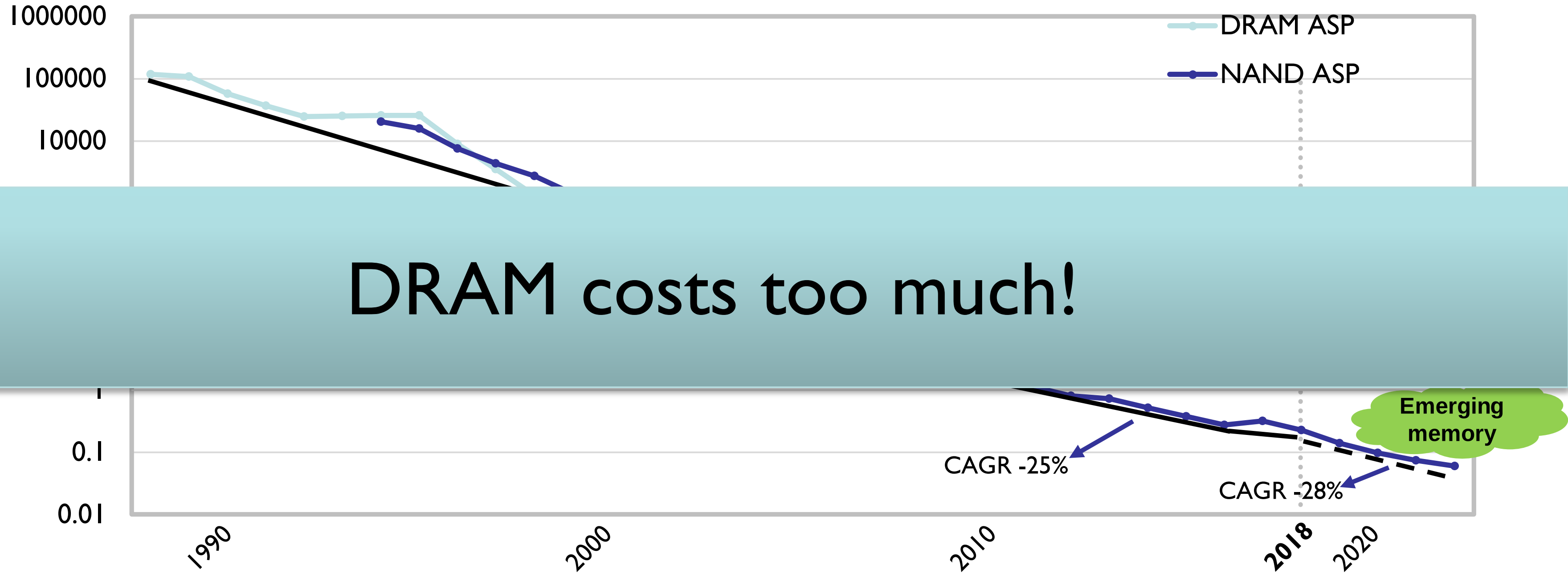


- Cost gap between DRAM and NAND continues to increase
- Need cost-effective emerging memory to fill this gap
- Trajectory for DRAM prices for the next 5 years uncertain

Rambus

Source: IDC

Moore’s Law is slowing – but still need low cost bits



DRAM costs too much!

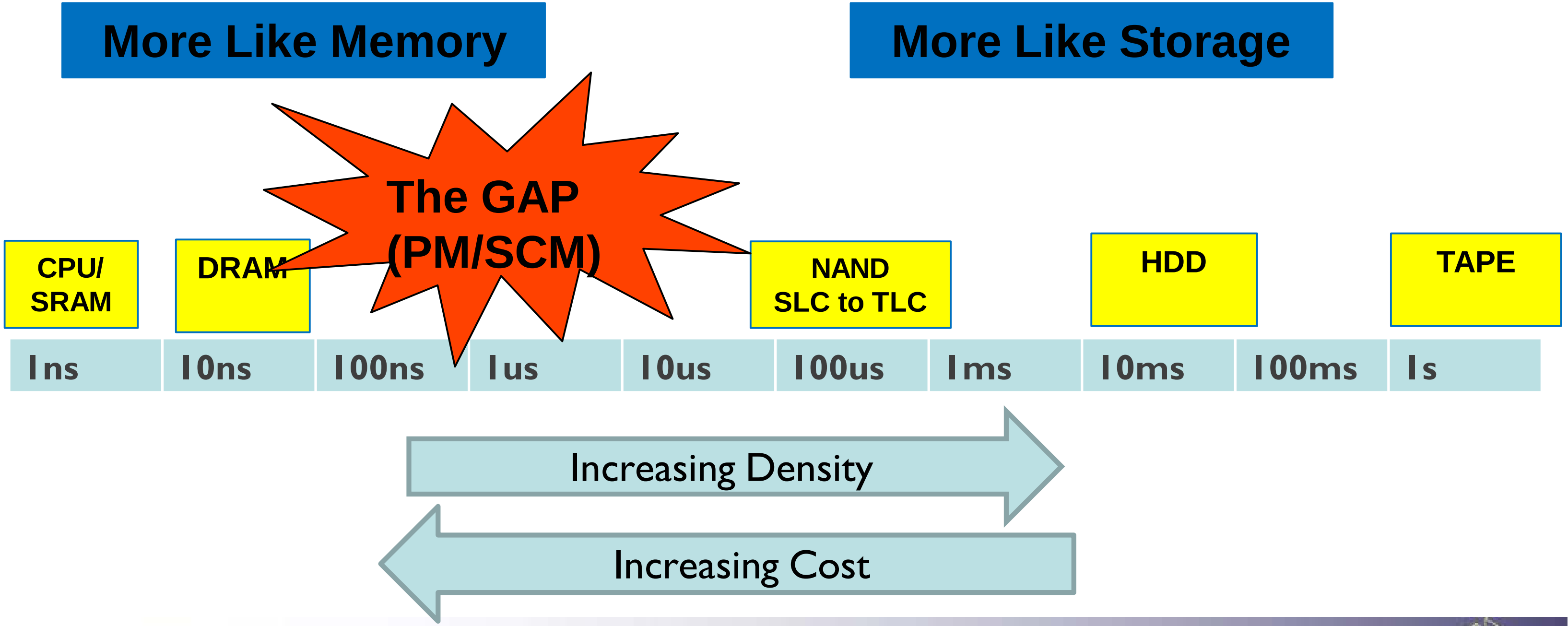
- Cost gap between DRAM and NAND continues to increase
- Need cost-effective emerging memory to fill this gap
- Trajectory for DRAM prices for the next 5 years uncertain

Source: IDC

Rambus

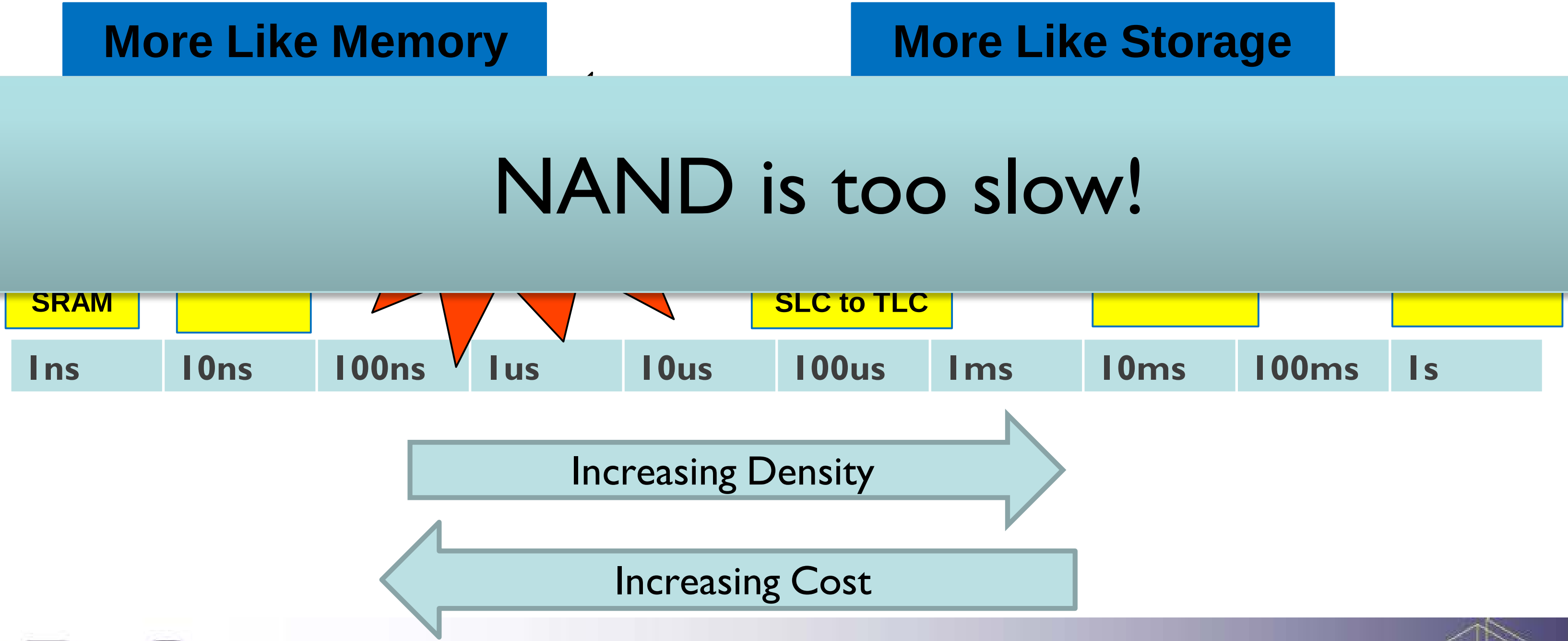
The Latency Spectrum and Gaps ~ 2015

Source: Mark Webb, MKW Ventures Consulting, FMS 2018



The Latency Spectrum and Gaps ~ 2015

Source: Mark Webb, MKW Ventures Consulting, FMS 2018



More Like Memory

More Like Storage

NAND is too slow!

SRAM

SLC to TLC

1ns | 10ns | 100ns | 1us | 10us | 100us | 1ms | 10ms | 100ms | 1s

Increasing Density

Increasing Cost

Limitations

DRAM

Cost

NAND

Latency

Emerging Memory Targets

Cost
1/3rd DRAM

Latency
Read <1µs

Emerging Memory Targets

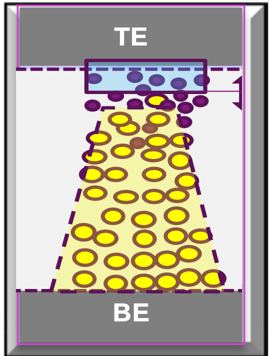
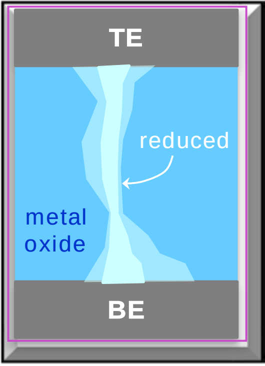
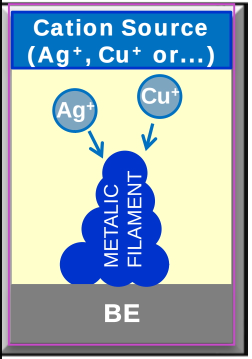
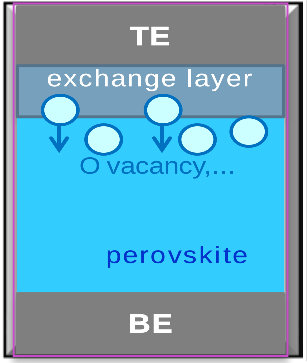
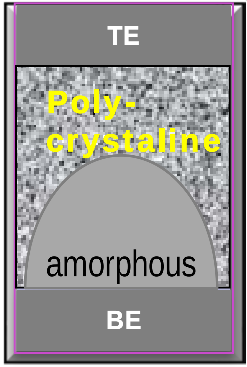
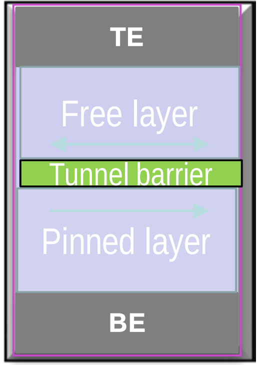
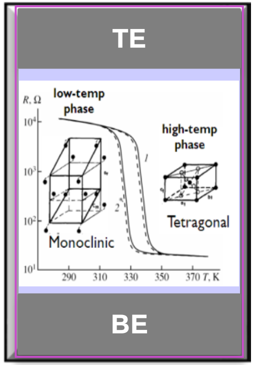
Does such an Emerging Memory even exist?

Cost
1/3rd DRAM

Latency
Read <1µs

Switching Mechanisms

Source: M. Jurczak, imec, ISSCC 2015 Memory Forum

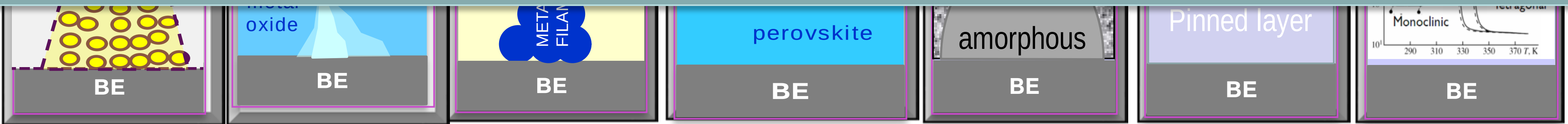
Filamentary ReRAM			Interfacial ReRAM	Bulk Transition		
Oxygen vacancy migration	Thermo-Chemical Fuse / antifuse	Electro-Chemical ECM	Schottky or Tunnel Barrier	Phase Change PCM	Tunnel Magneto resistance	Electronic MIT (Mott)
 	 	 Cation Source (Ag⁺, Cu⁺ or...) Ag⁺ Cu⁺ METALIC FILAMENT BE	 	 	 	
BIPOLAR	UNIPOLAR	BIPOLAR	BIPOLAR	UNIPOLAR	BIPOLAR	UNIPOLAR
TMO: HfO2, TaO2	TMO: NiO2	CBRAM: Cu, Ag based	Memristor VMCO, PCMO, TiO2	Chalcogenide alloys: GST	STT RAM (CoFeB, MgO)	VO2, NbO2

Switching Mechanisms

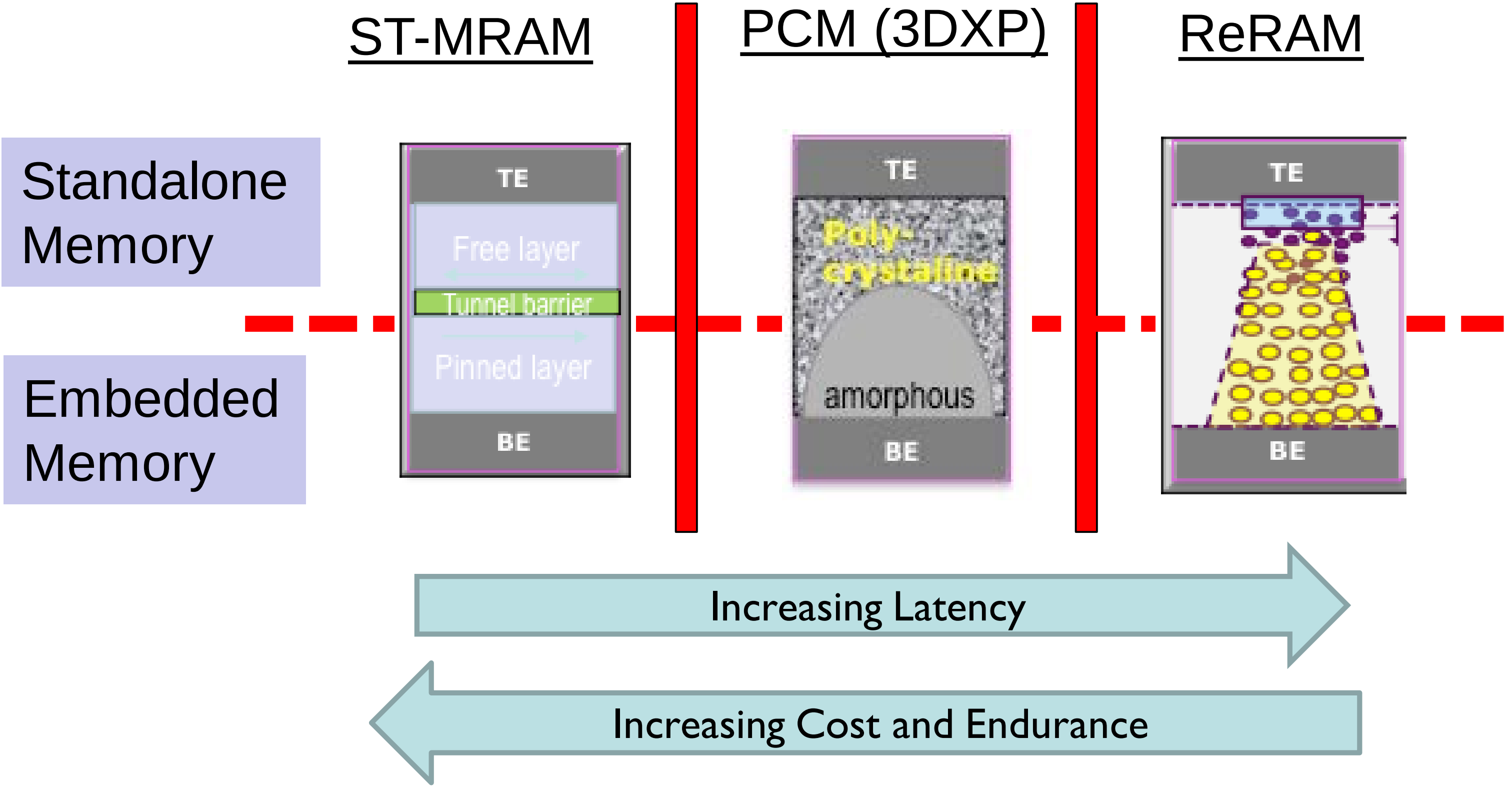
Source: M. Jurczak, imec, ISSCC 2015 Memory Forum

Filamentary ReRAM			Interfacial ReRAM	Bulk Transition		
Oxygen vacancy	Thermo-Chemical	Electro-Chemical	Schottky or Tunnel	Phase Change	Tunnel Magneto	Electronic MIT
BIPOLAR	UNIPOLAR	BIPOLAR	BIPOLAR	UNIPOLAR	BIPOLAR	UNIPOLAR
TMO: HfO2, TaO2	TMO: NiO2	CBRAM: Cu, Ag based	Memristor VMCO, PCMO, TiO2	Chalcogenide alloys: GST	STT RAM (CoFeB, MgO)	VO2, NbO2

(Too) many switching mechanisms!



Classifying the What



ST-MRAM

PCM (3DXP)

ReRAM

Standalone Memory

Embedded Memory

Increasing Latency

Increasing Cost and Endurance

WHO is doing what?

	ST-MRAM	PCM (3DXP)	ReRAM
Standalone	EVERSPIN® TECHNOLOGIES	intel® Micron®	SONY adesto™ TECHNOLOGIES
Embedded	GLOBALFOUNDRIES® SAMSUNG tsmc UMC	ST	CROSSBAR Panasonic

Logos are the property of their respective trademark owners

Let’s focus on the two shipping technologies!

	ST-MRAM	PCM (3DXP)	ReRAM
Standalone	EVERSPIN TECHNOLOGIES	intel Micron	SONY adesto TECHNOLOGIES
Embedded	GLOBALFOUND SAMSUNG tsmc		CROSSBAR Panasonic

WHO is doing what?

ST-M

ReRAM

Standalone

Embedded

Logos are the property of their respective trademark owners

Does such an Emerging Memory even exist?

	Cost 1/3rd of DRAM	**Latency** Read <1us
PCM (3DXP)	👍	👎
ST-MRAM	👎	👍

😢

Uh oh…

That’s not good.

A Word to the Wise: ~~Replacements~~

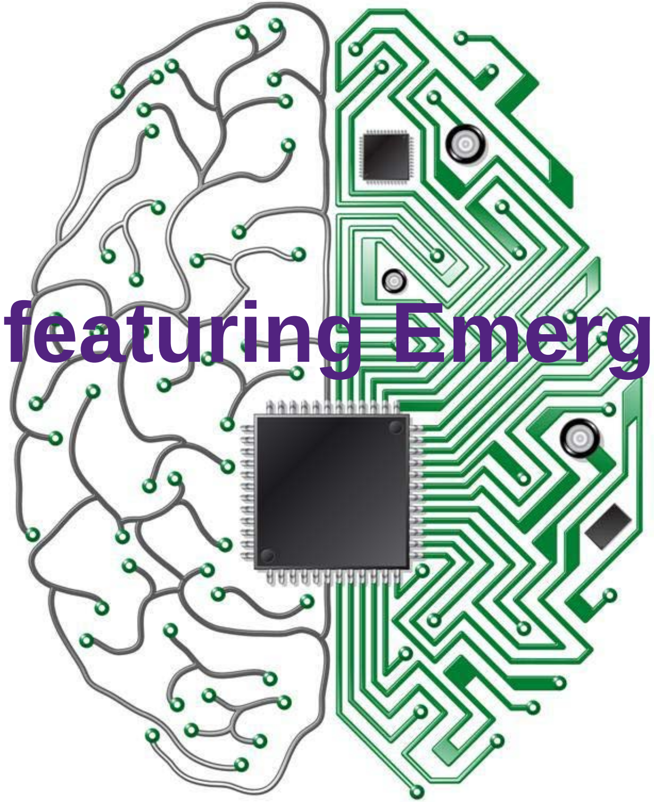
- There are no 1:1 memory replacements.
- Stop. looking. for. Them.
- Moving from DRAM+NAND world into one with Combinations of memory.

Just one word.
Combinations.

Let’s use memory combinations!

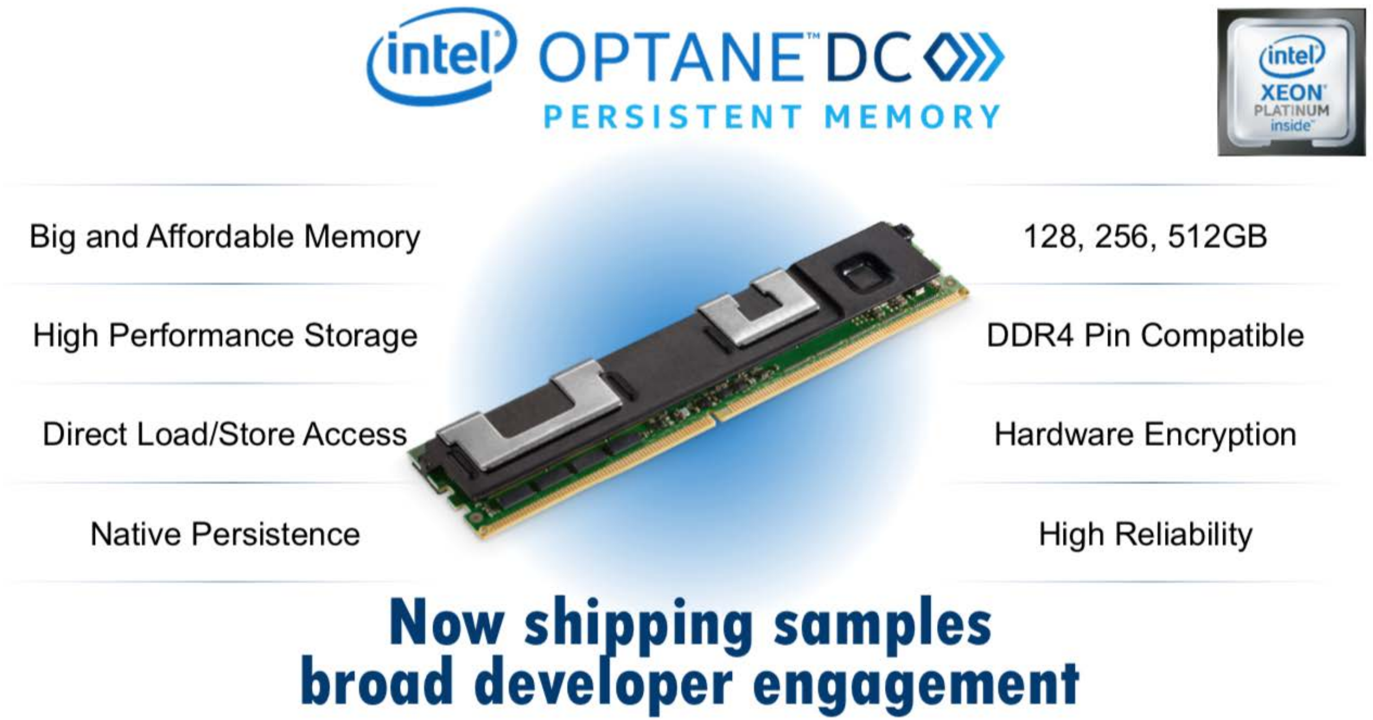
	Cost 1/3rd of DRAM	**Latency** Read <1us	Winning Combination
PCM (3DXP)	🟡	🟢	+DRAM DIMMs for reduced latency!
ST-MRAM	🟢	🟢	+NAND for reduced SSD cost!

2018 products featuring Emerging Memory



FMS18: Intel Optane DIMM (3DXP on DRAM bus)

Source: Mark W. Henderson, Intel, FMS 2018



intel OPTANE™ DC

PERSISTENT MEMORY

Big and Affordable Memory

High Performance Storage

Direct Load/Store Access

Native Persistence

128, 256, 512GB

DDR4 Pin Compatible

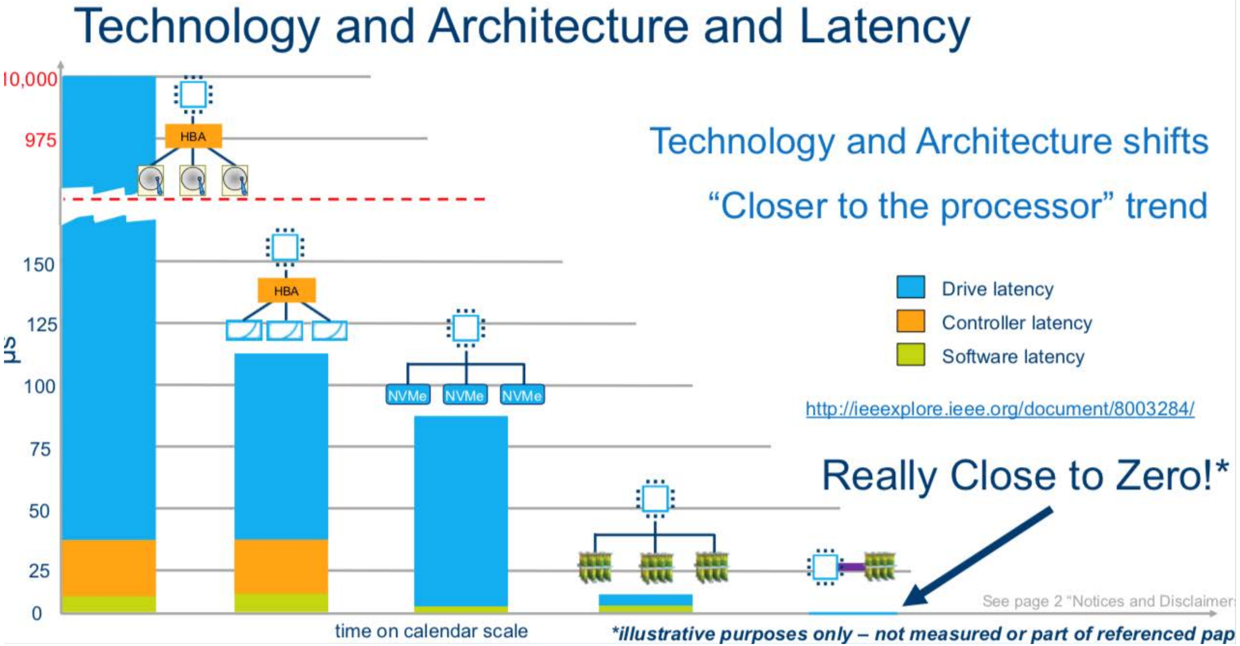
Hardware Encryption

High Reliability

Now shipping samples broad developer engagement

FMS18: Intel Optane DIMM (3DXP on DRAM bus)

Source: Mark W. Henderson, Intel, FMS 2018



- Arch, software, hardware total effort
- Reduces Optane read latency to a few μs
- DRAM operates as "near memory", Optane operates as "far memory"
- Intel controlled

FMS18: IBM Flash Core Module (MRAM+NAND)

Source: Brent Yardley, IBM, FMS 2018

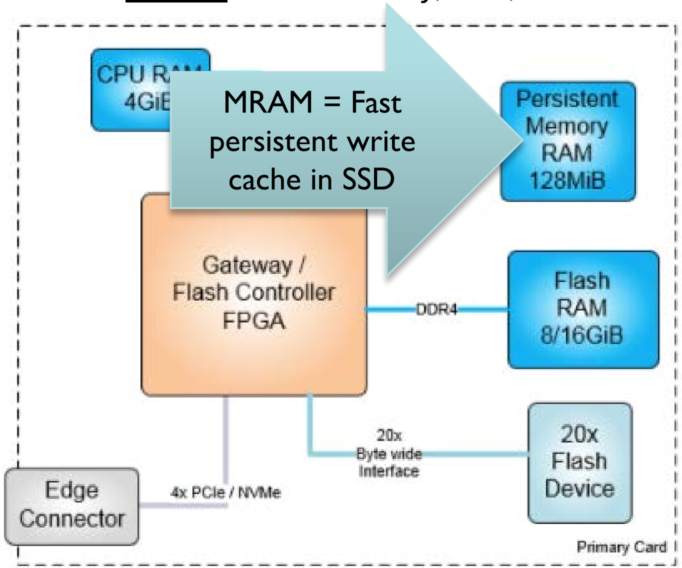
Flash Memory Summit

Introducing The IBM FlashCore Module

IBM **FlashCore™** technology delivers key differentiators

- Built in, performance neutral **hardware compression** and **encryption**
- Using **64 layer 3DTLC NAND**
- Enterprise data **reliability**
- Cognitive Algorithms for **Wear Levelling, Health binning, Heat segregation** and media management
- Intelligent media management that **keeps settings ideal** to keep performance consistent.
- **Endurance** without latency penalty
- **FIPS 140** certification

4.8TBu, 9.6TBu, 19.2TBu capacity options with up to 3:1 compression



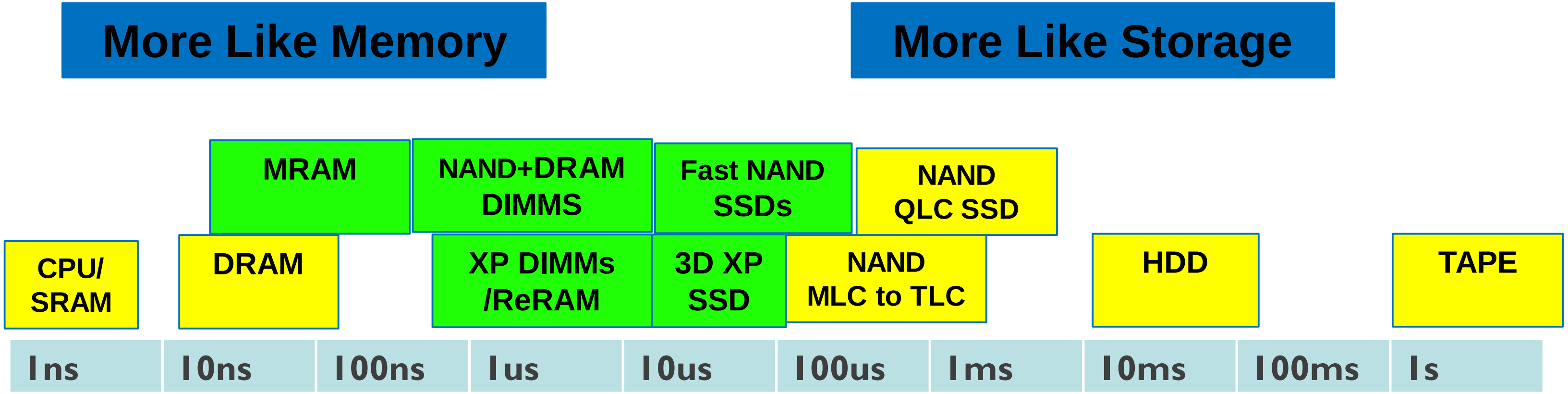
MRAM = Fast persistent write cache in SSD

IBM®

EVERSPIN® TECHNOLOGIES

The Latency Spectrum and Gaps ~ Now

Source: Mark Webb, MKW Ventures Consulting, FMS 2018



More Like Memory

More Like Storage

MRAM

NAND+DRAM DIMMS

Fast NAND SSDs

NAND QLC SSD

CPU/ SRAM

DRAM

XP DIMMs /ReRAM

3D XP SSD

NAND MLC to TLC

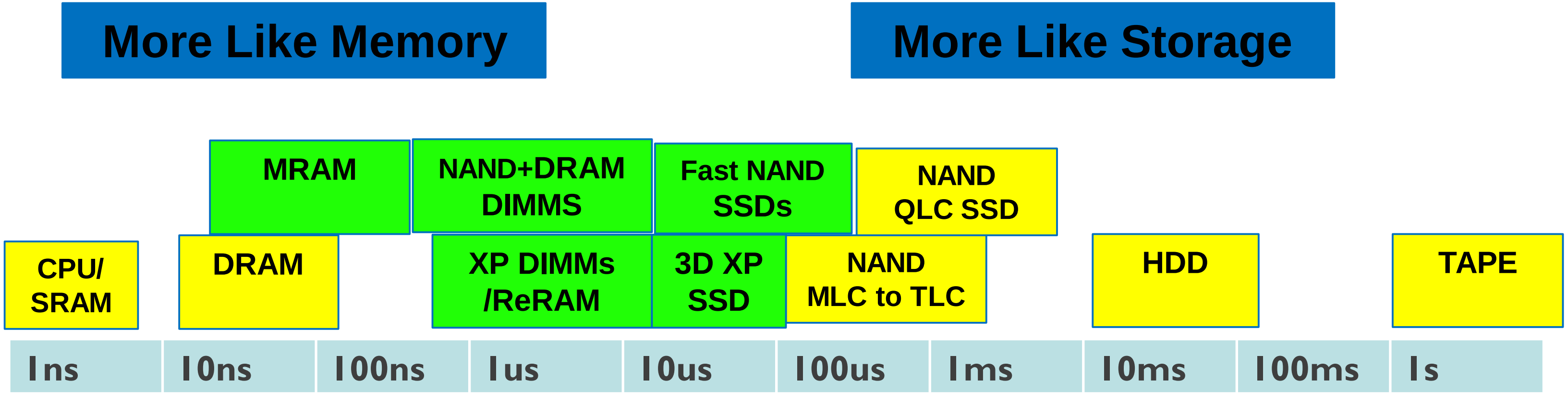
HDD

TAPE

1ns | 10ns | 100ns | 1us | 10us | 100us | 1ms | 10ms | 100ms | 1s

The Latency Spectrum and Gaps ~ Now

Source: Mark Webb, MKW Ventures Consulting, FMS 2018



More Like Memory

More Like Storage

MRAM

NAND+DRAM DIMMS

Fast NAND SSDs

NAND QLC SSD

CPU/ SRAM

DRAM

XP DIMMs /ReRAM

3D XP SSD

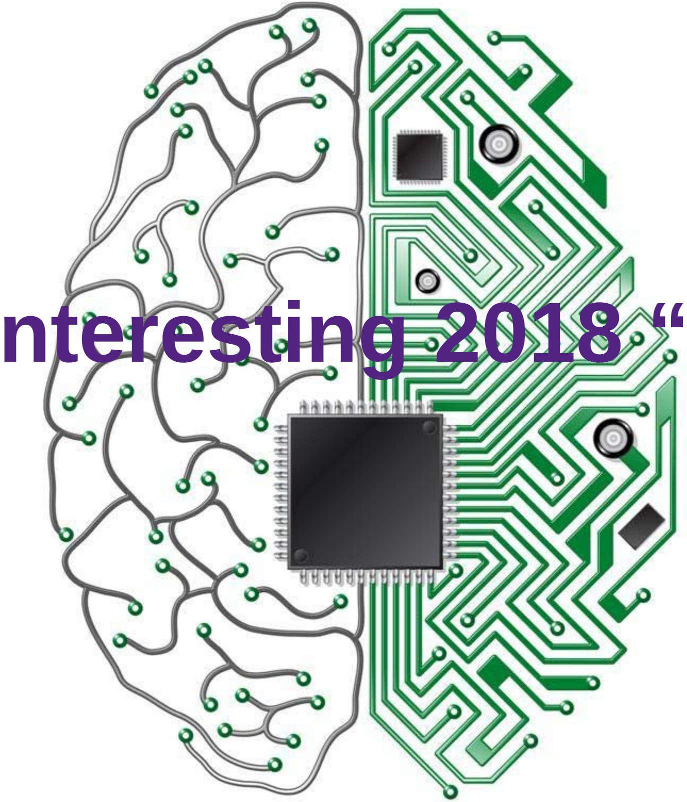
NAND MLC to TLC

HDD

TAPE

1ns 10ns 100ns 1us 10us 100us 1ms 10ms 100ms 1s

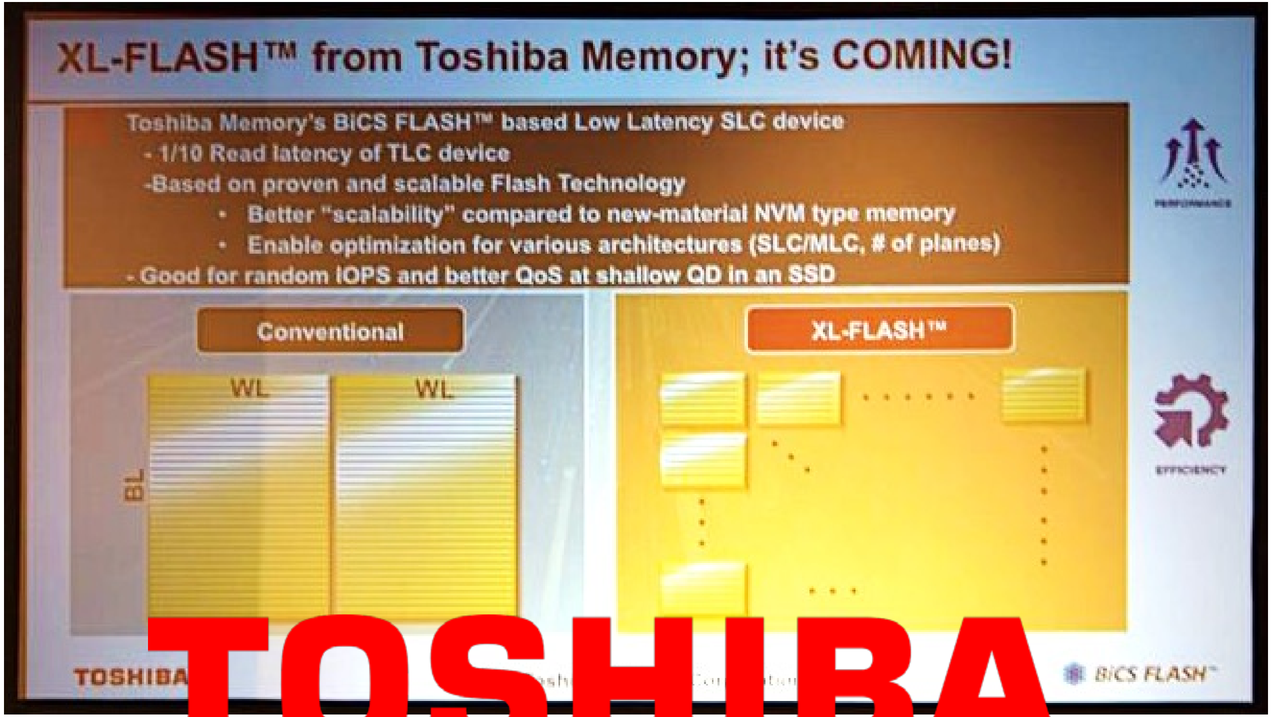
Several memory combinations reduce latency!



Other interesting 2018 “stuff”

FMS18: Toshiba XL-Flash (Low Latency SSD)

Source: Jeff Oshima, Toshiba, FMS 2018



XL-FLASH™ from Toshiba Memory; it's COMING!

Toshiba Memory's BiCS FLASH™ based Low Latency SLC device

- 1/10 Read latency of TLC device
- Based on proven and scalable Flash Technology
  - Better "scalability" compared to new-material NVM type memory
  - Enable optimization for various architectures (SLC/MLC, # of planes)
- Good for random IOPS and better QoS at shallow QD in an SSD

Conventional

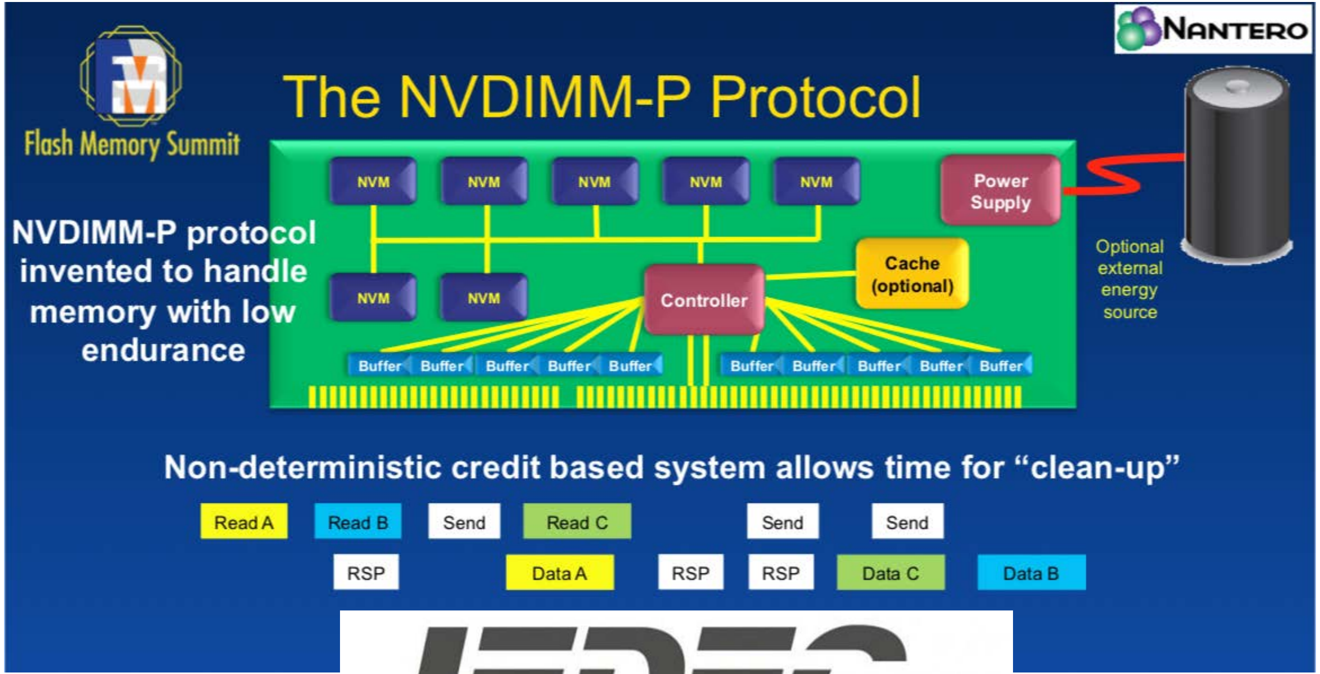
XL-FLASH™

TOSHIBA

- ❐ 10x reduced latency vs. TLC
- ❐ Still not 1us ☹
- ❐ For low latency SSDs, attached to compute nodes
- ❐ Samsung Z-NAND, Intel Optane SSDs are competitors

FMS18: JEDEC NVDIMM-P

Source: Bill Gervasi, Nantero, FMS 2018



The NVDIMM-P Protocol

NVDIMM-P protocol invented to handle memory with low endurance

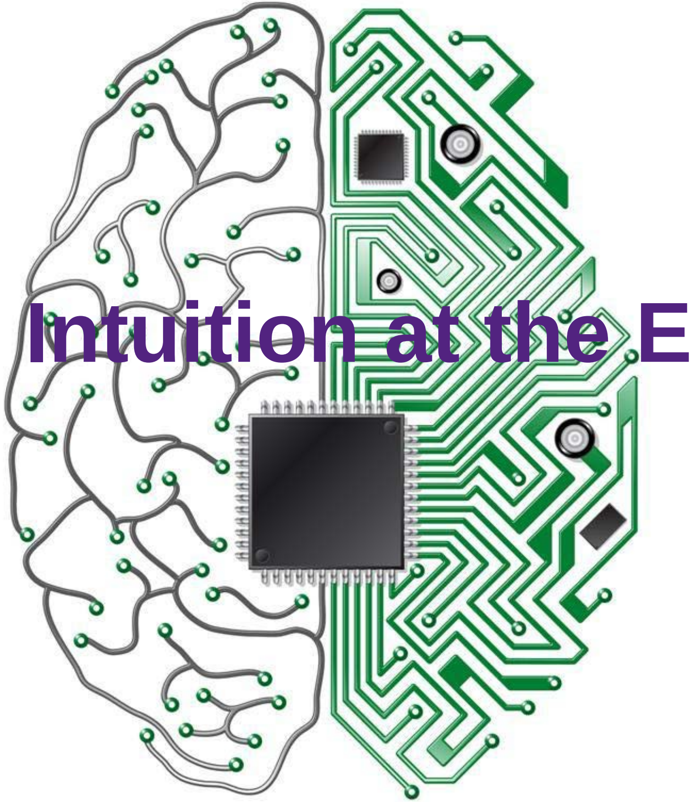
Non-deterministic credit based system allows time for “clean-up”

JEDEC®

- ❒ Emerging memory and DRAM on the same DDR bus
- ❒ Open standard
- ❒ Non-deterministic behavior allowed
- ❒ Will compete with Intel Optane DIMMs
- ❒ Backed by all major memory companies

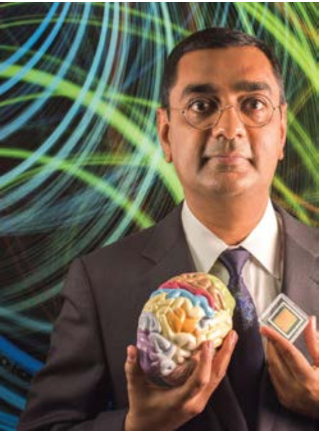
Summary: Your 2018 SSD/DIMM Watch List

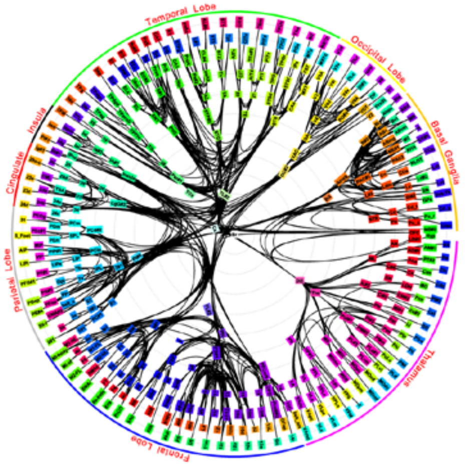
- Everspin MRAM in low latency SSDs
- Intel 3DXP in lower cost server Optane DIMMs
- Toshiba XL NAND in lower latency SSDs
- JEDEC NVDIMM-P in open standard DIMMs



Back to Intuition at the Edge!

Edge





- □ Non-von Neumann architecture

Intuition

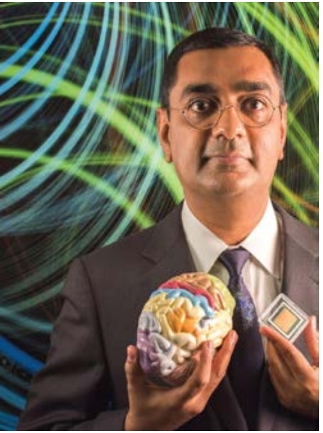
Datacenter

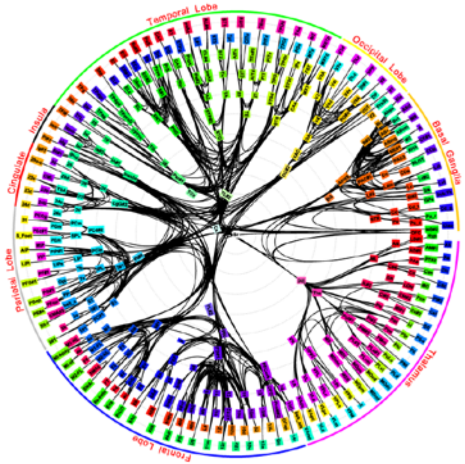
Let's talk about Intuition at the Edge now!

- □ von Neumann architecture

Reasoning

Edge





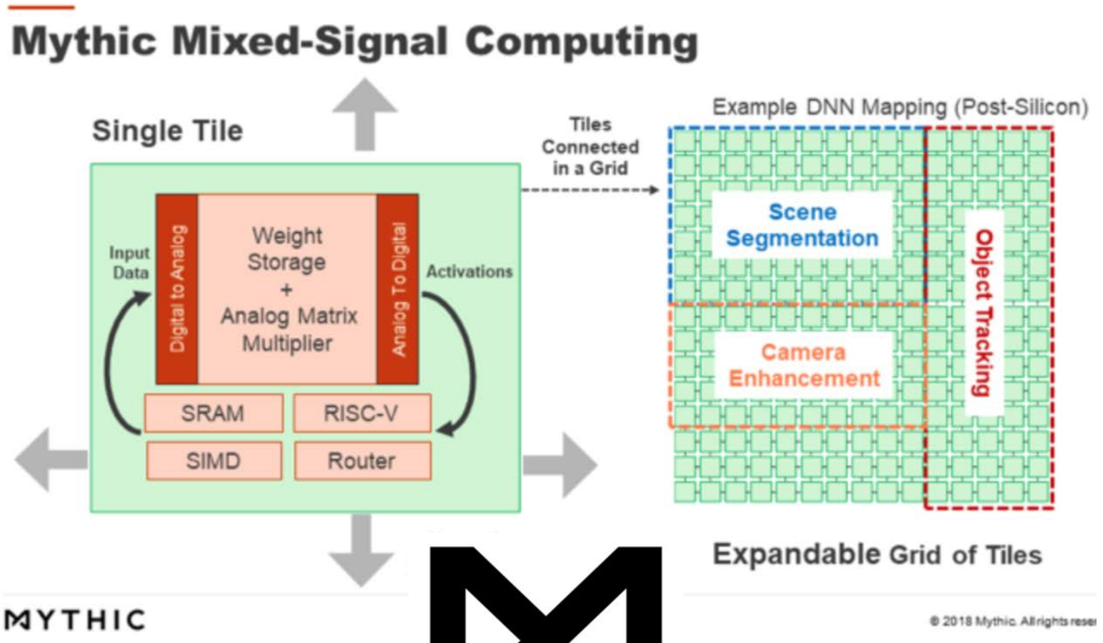
- Non-von Neumann architecture

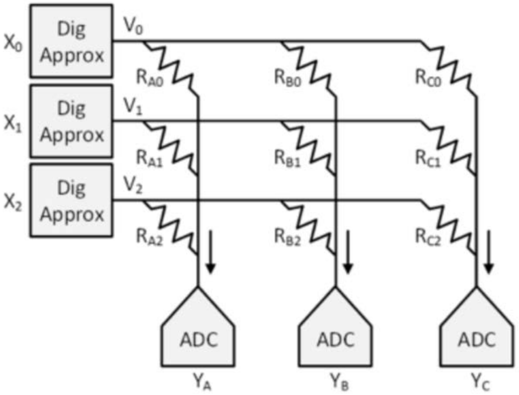
Intuition

- Intuition system in the human brains use <20 watts of power
- Using von-Neumann architecture for Intuition >20 Gigawatts!
- Highly networked, local compute nodes
- Trained neural nets (NN) perform lightning fast intuition
- Embedded Emerging Memory used to hold weights inside NN
- Sum the weights using analog combine
- Most efficient implementation is analog memory (6+ levels per cell)
- Doesn’t require 7nm/5nm process!

Hot Chips 2018: Edge Intuition SoC

Source: Dave Fick, Mythic, Hot Chips 2018

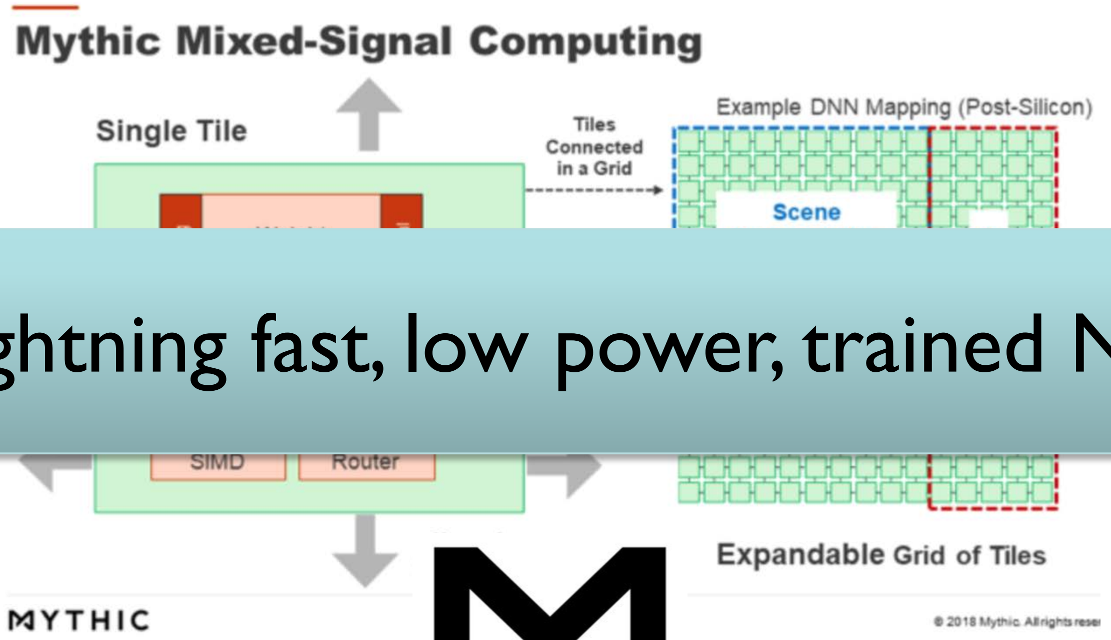


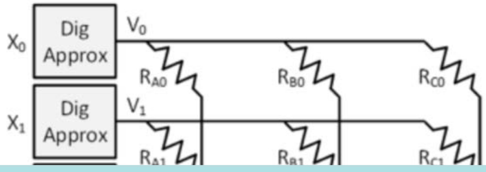


- Mythic currently uses embedded NOR Flash in analog mode to hold weights

Hot Chips 2018: Edge Intuition SoC

Source: Dave Fick, Mythic, Hot Chips 2018


Mythic Mixed-Signal Computing
Single Tile
Tiles Connected in a Grid
Example DNN Mapping (Post-Silicon)
Scene
SIMD
Router
Expandable Grid of Tiles
MYTHIC
© 2018 Mythic. All rights reser

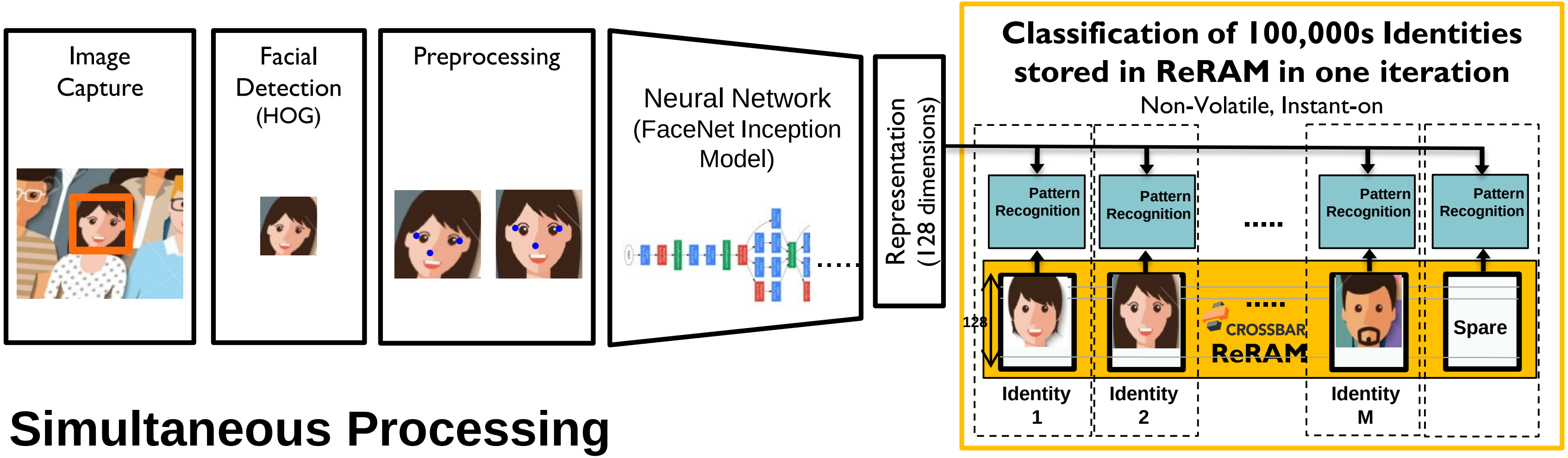


Lightning fast, low power, trained NN, analog memory

- Mythic currently uses embedded NOR Flash in analog mode to hold weights

FMS18: Future Edge Intuition SoCs use ReRAM?

Source: Hagop Nazarian, Crossbar, FMS 2018

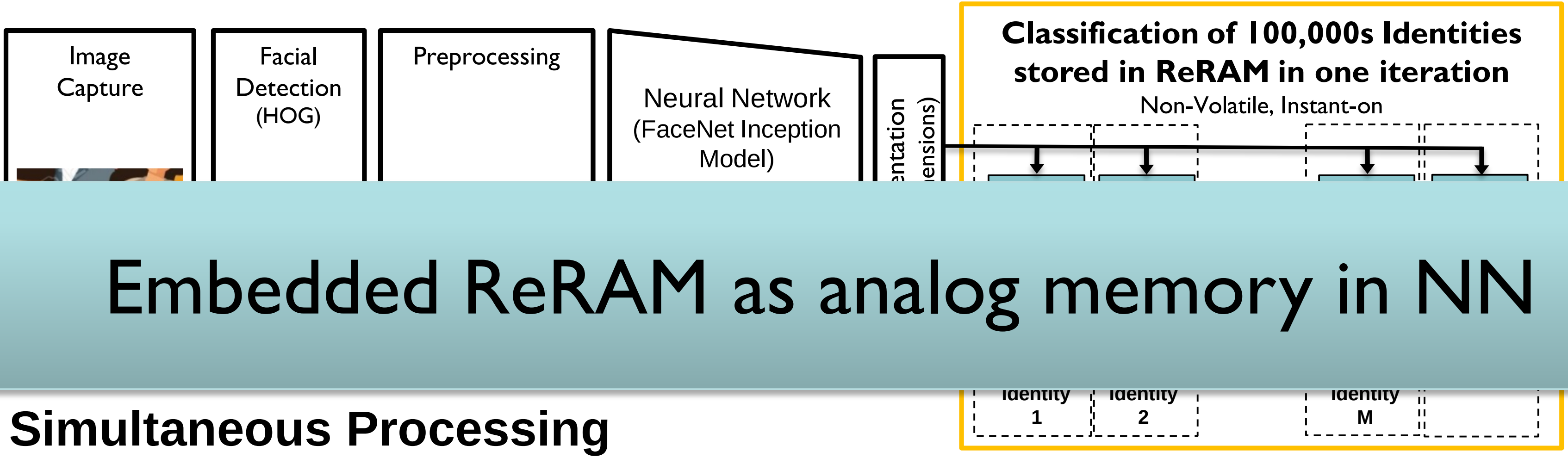


Simultaneous Processing with Deterministic Performance

- Parallel comparison against all identities
- If no match, new identity created (learning)
- Classification performed in one cycle independent of number of identities

FMS18: Future Edge Intuition SoCs use ReRAM?

Source: Hagop Nazarian, Crossbar, FMS 2018



Embedded ReRAM as analog memory in NN

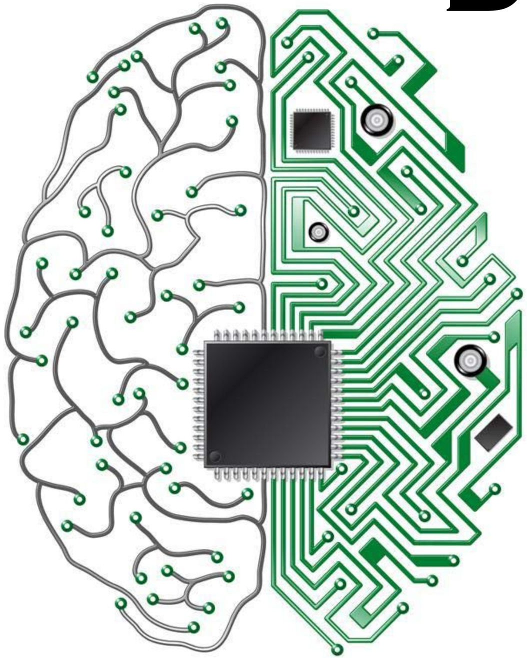
Simultaneous Processing with Deterministic Performance

- Parallel comparison against all identities
- If no match, new identity created (learning)
- Classification performed in one cycle independent of number of identities

Why do I care? What is Emerging Memory enabling for AI?

Edge

- ❐ Analog memory in NN
- ❐ Lightning fast intuition at low power



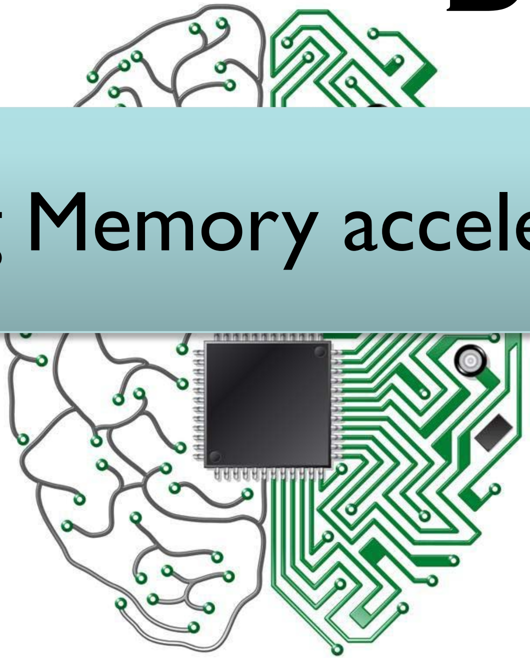
DATACENTER

- ❐ TBs of memory
- ❐ Better reasoning, based on more data

Why do I care? What is Emerging Memory enabling for AI?

Edge

Datacenter



Emerging Memory accelerates AI!

- ❒ Lightning fast intuition at low power

- ❒ Faster reasoning, based on more data

Take-away points:

- Your brain has two distinct systems
- Reasoning system needs reduced latency and cost
- There are NO 1:1 memory replacements
- Emerging memory combos with NAND and DRAM
- Intuition system needs new architecture and low power
- Emerging memory utilized as analog weight in neural nets
- Emerging memory accelerates AI systems

Register NOW for SNIA 2019 PM Summit!

SNIA

PERSISTENT MEMORY

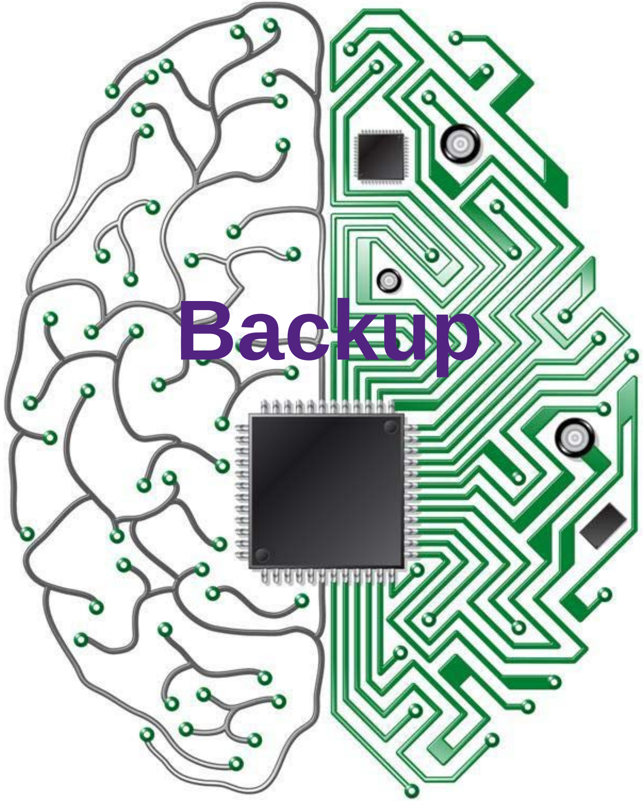
PM SUMMIT

JANUARY 24, 2019 | SANTA CLARA, CA

https://www.snia.org/events/persistent-memory-summit/persistent-memory-summit-2019-registration

Dave Eggleston
Intuitive Cognition Consulting
Technology & Business Strategy

Email: dave@in-cog.com
Twitter: @NVM_DaveE
LinkedIn:
linkedin.com/in/deggleston/



Backup

Talk Outline

1. Demonstrate the human brain has two distinct systems
2. Discuss the best fit compute architecture to model each AI system
3. Articulate how DRAM and NAND are applied to the compute architectures
4. Identify the key limitations of DRAM and NAND
5. Present and classify some Emerging Memory alternatives
6. Discuss the Emerging Memory system enhancements
7. Identify who is doing what (by when) in the Emerging Memory landscape
8. Articulate the unique challenges in realizing an AI intuition system
9. Propose how Emerging Memory may solve some intuition problems
10. Point to the future of AI systems based on Emerging Memory