SNIA PERSISTENT MEMORY + SUMMIT 2022 COMPUTATIONAL STORAGE

VIRTUAL EVENT • MAY 24-25, 2022

Inventing Our Way Around the Memory Wall

Presented by: Jim Handy, Objective Analysis

Thomas Coughlin, Coughlin Associates

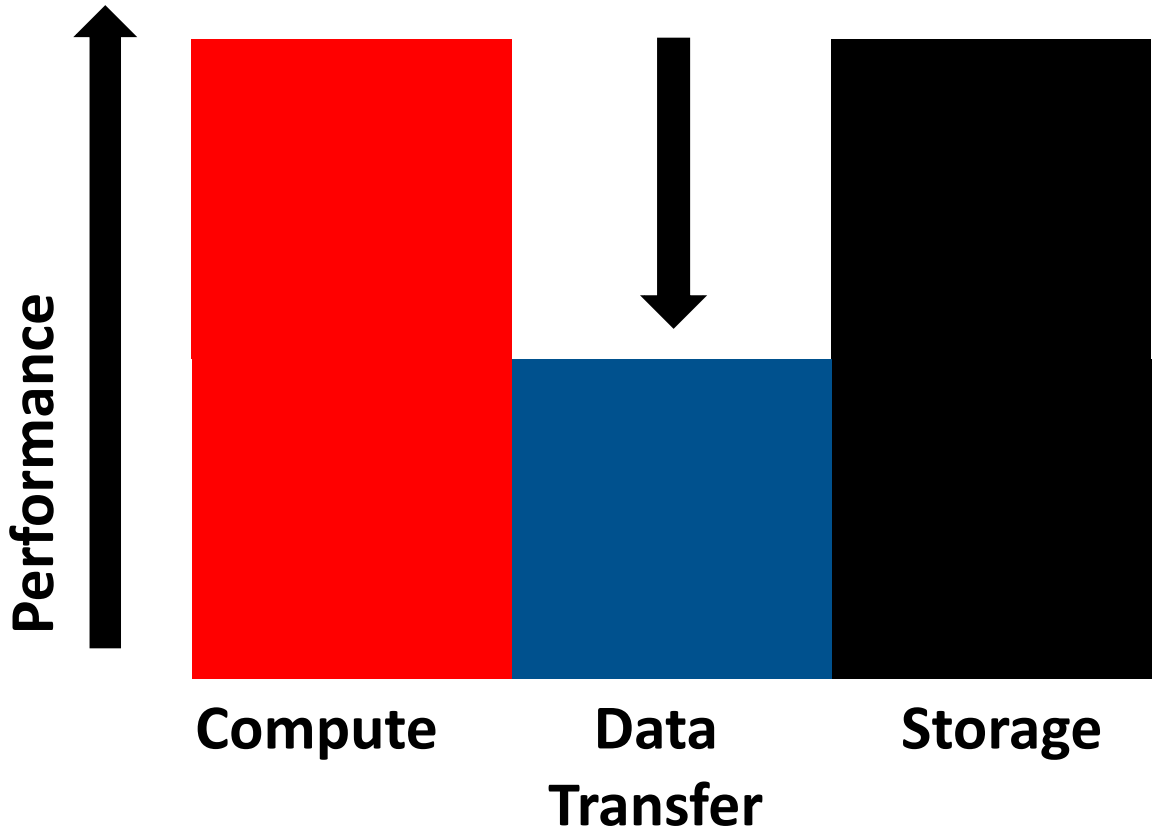
Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

Outline

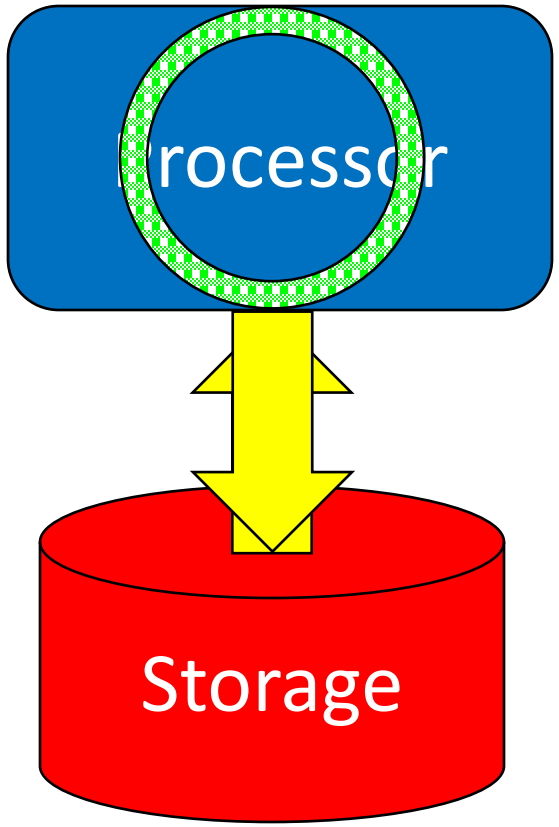
- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

Data Transfer Has Become The Bottleneck



How Work Gets Done

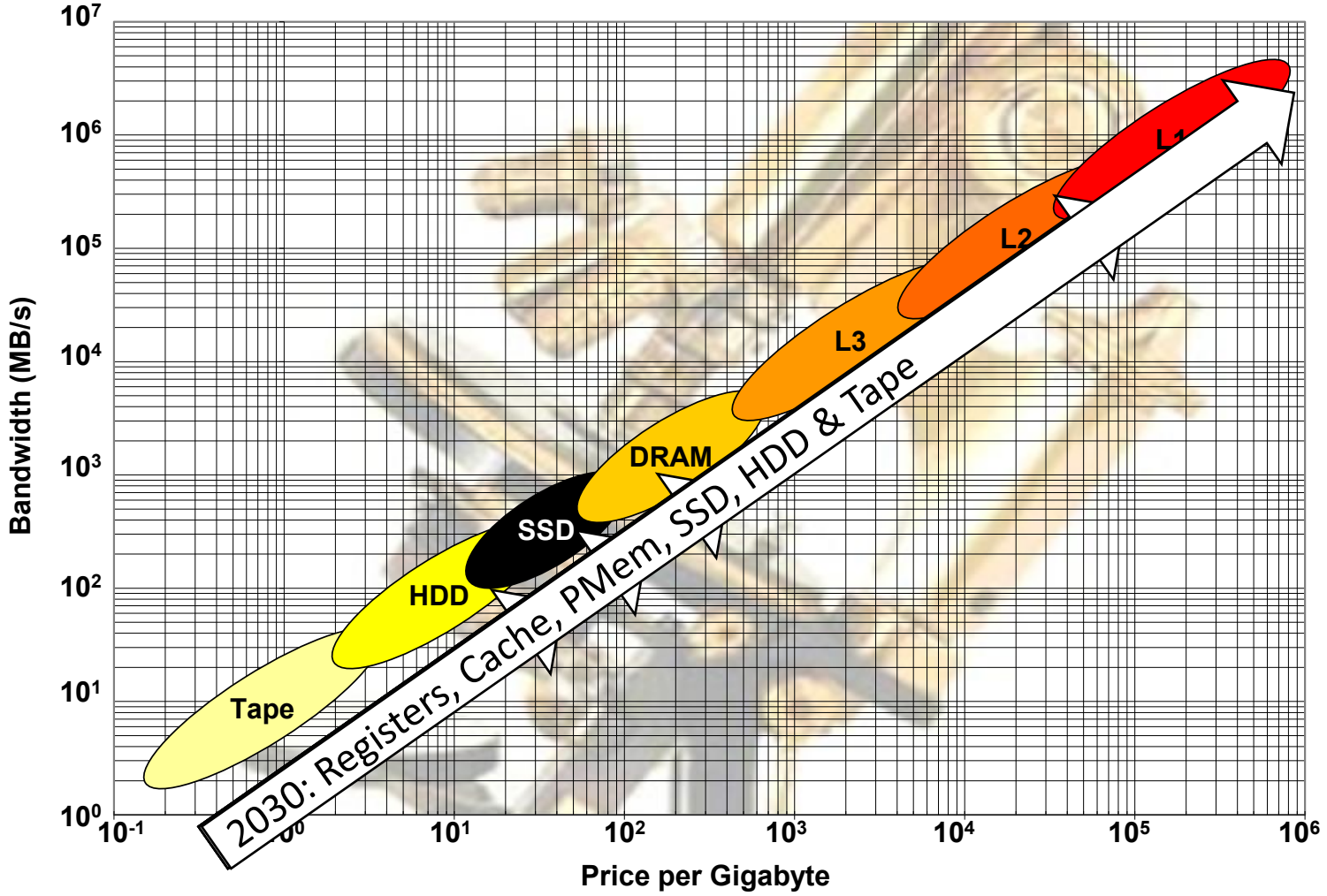
1. Request Data
2. Receive Data
3. Process Data
4. Write Data



Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

Move Persistence Up the Memory/Storage Hierarchy

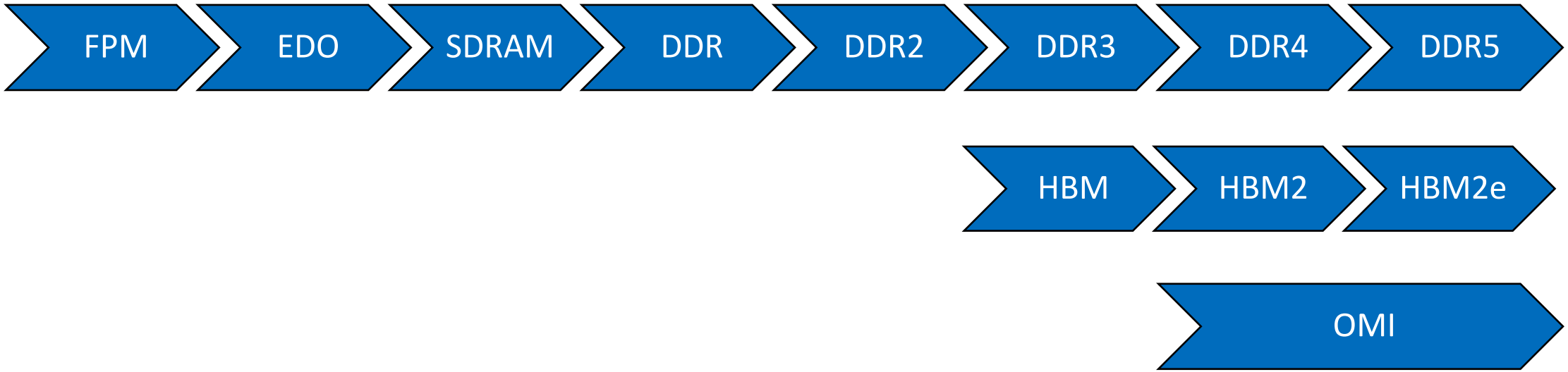


From Report: Emerging Memories Take Off

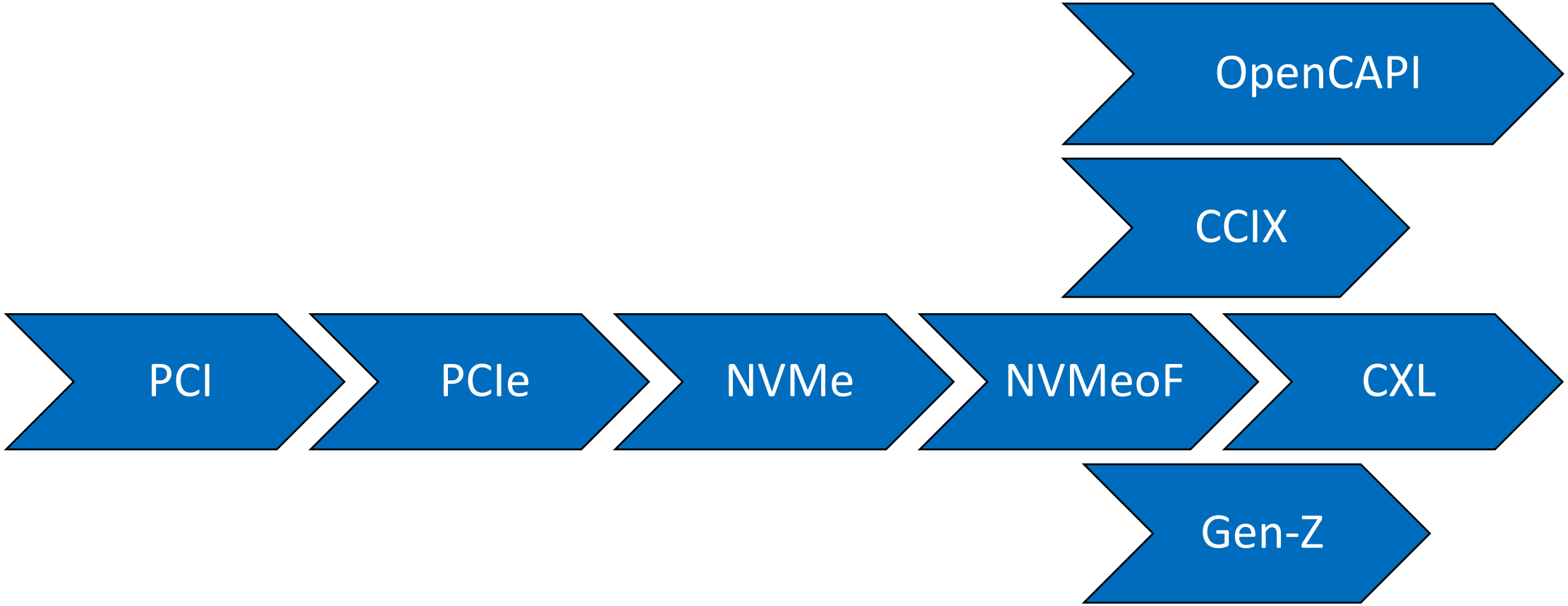
Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

DRAM: Faster Interfaces and More of 'em



System-Level Interfaces

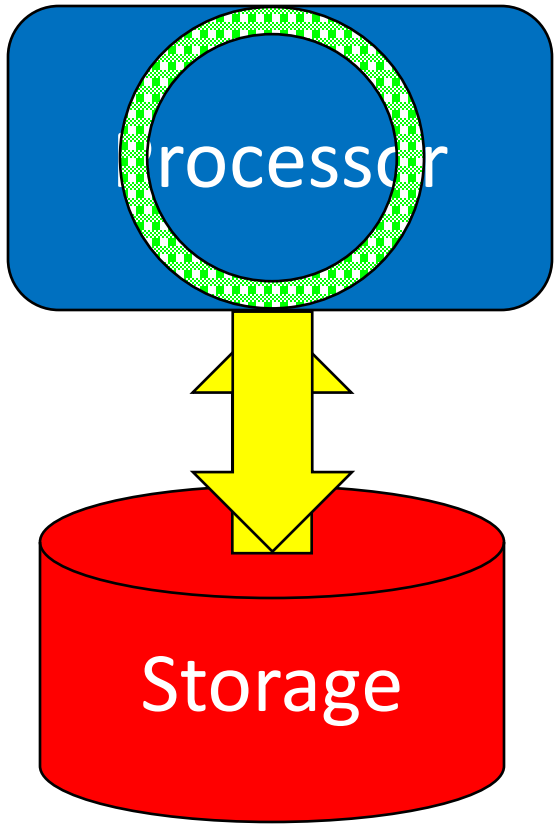


Outline

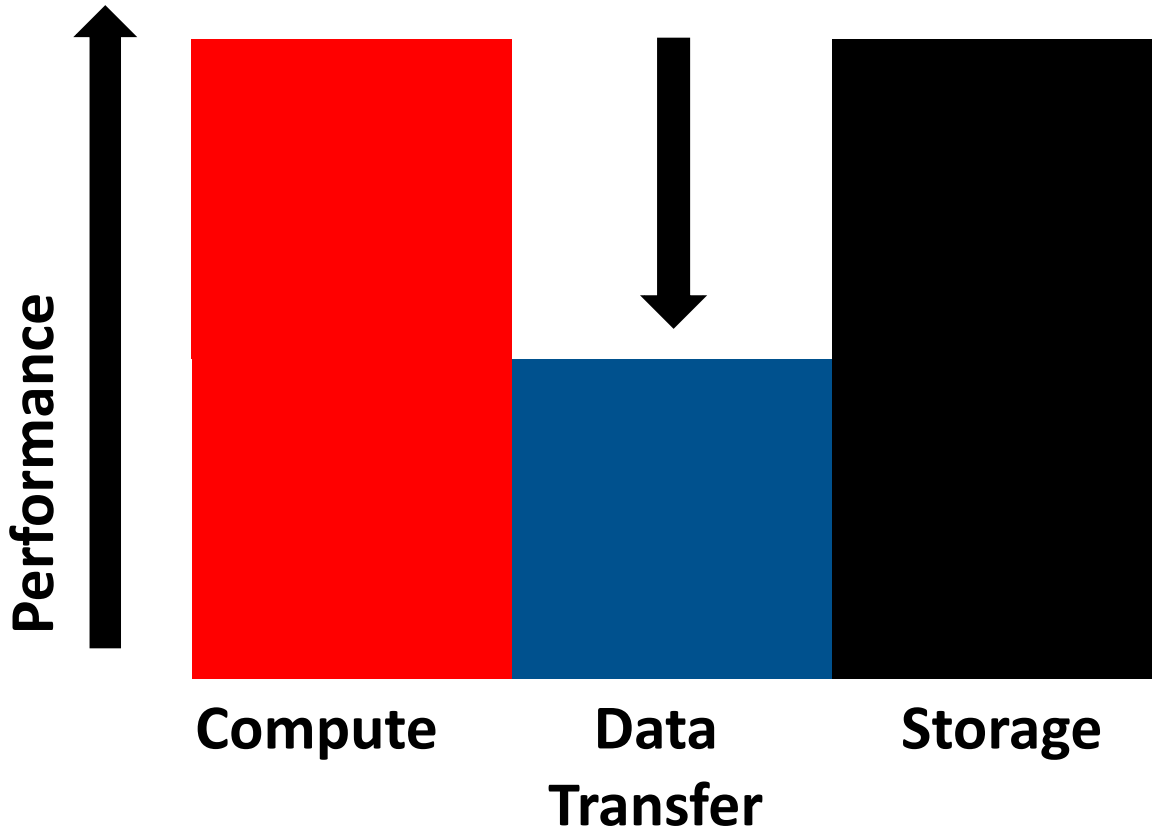
- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

How Work Gets Done

1. Request Data
2. Receive Data
3. Process Data
4. Write Data

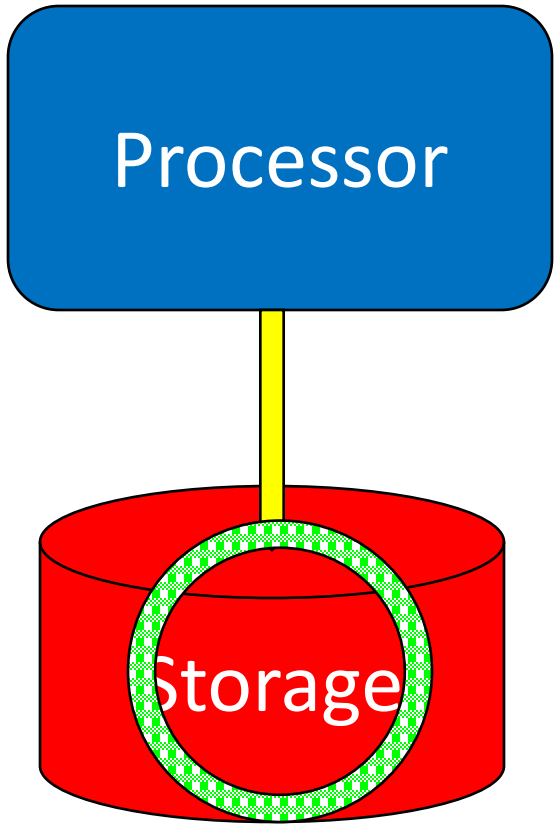


The Network Bottleneck



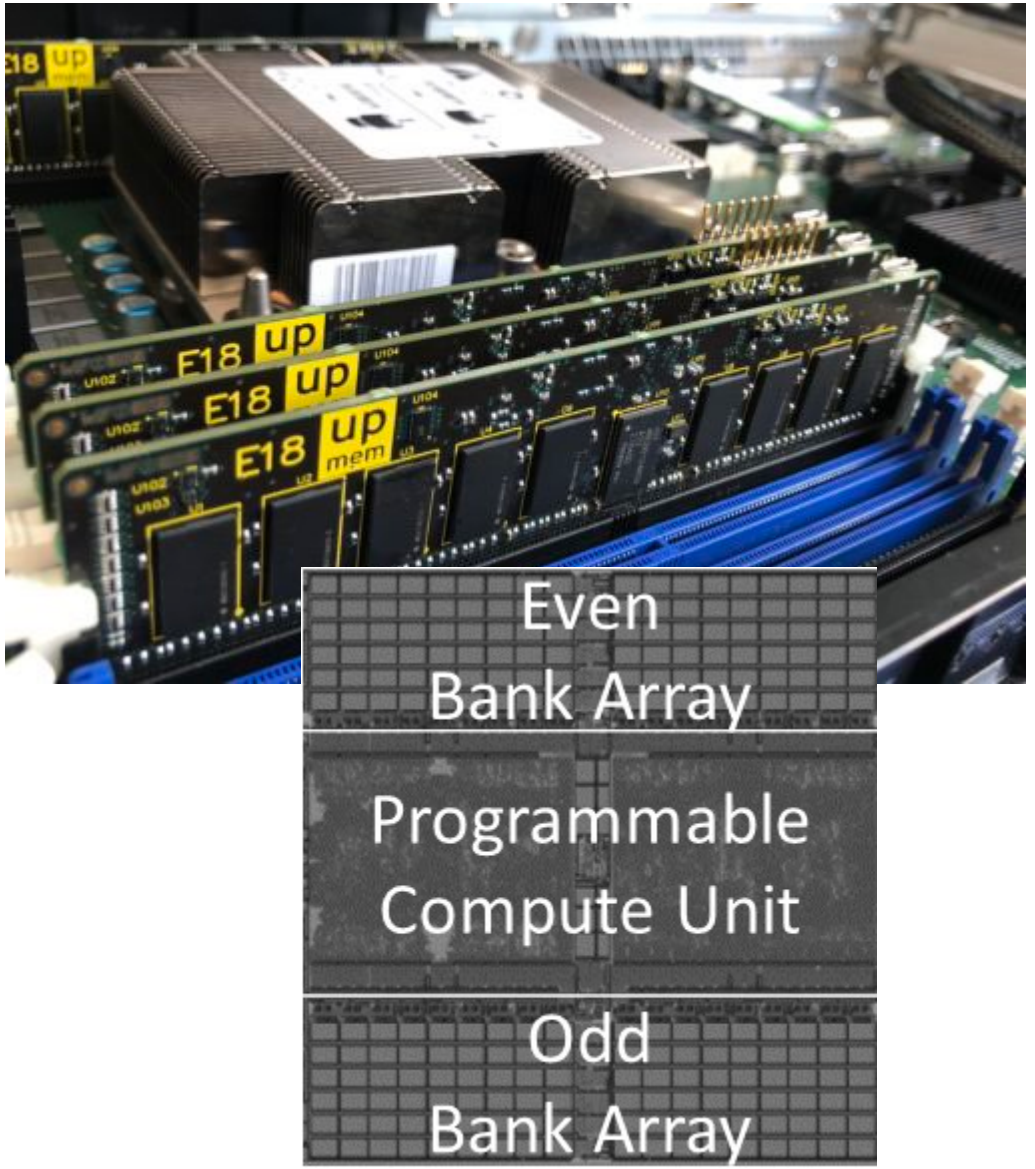
Improved Approach

1. Initiate Process
2. Process Data in Place



Compute In Memory/ Processing in Memory (PIM)

- **Automata: Micron, Natural Intelligence**
- **TOMI: Ven-Ray**
- **PIM DPU: UPmem**
- **Gemini APU: GSI**
- **Aquabolt-XL: Samsung**
- **SAPEON: SK hynix**
- **Various Neural Networks**



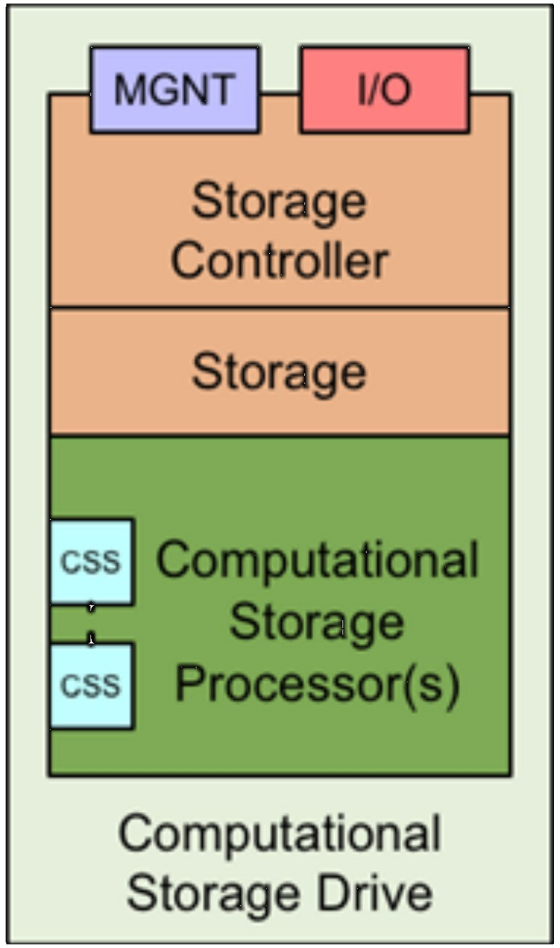
Goal is to reduce data movement

Computational Storage

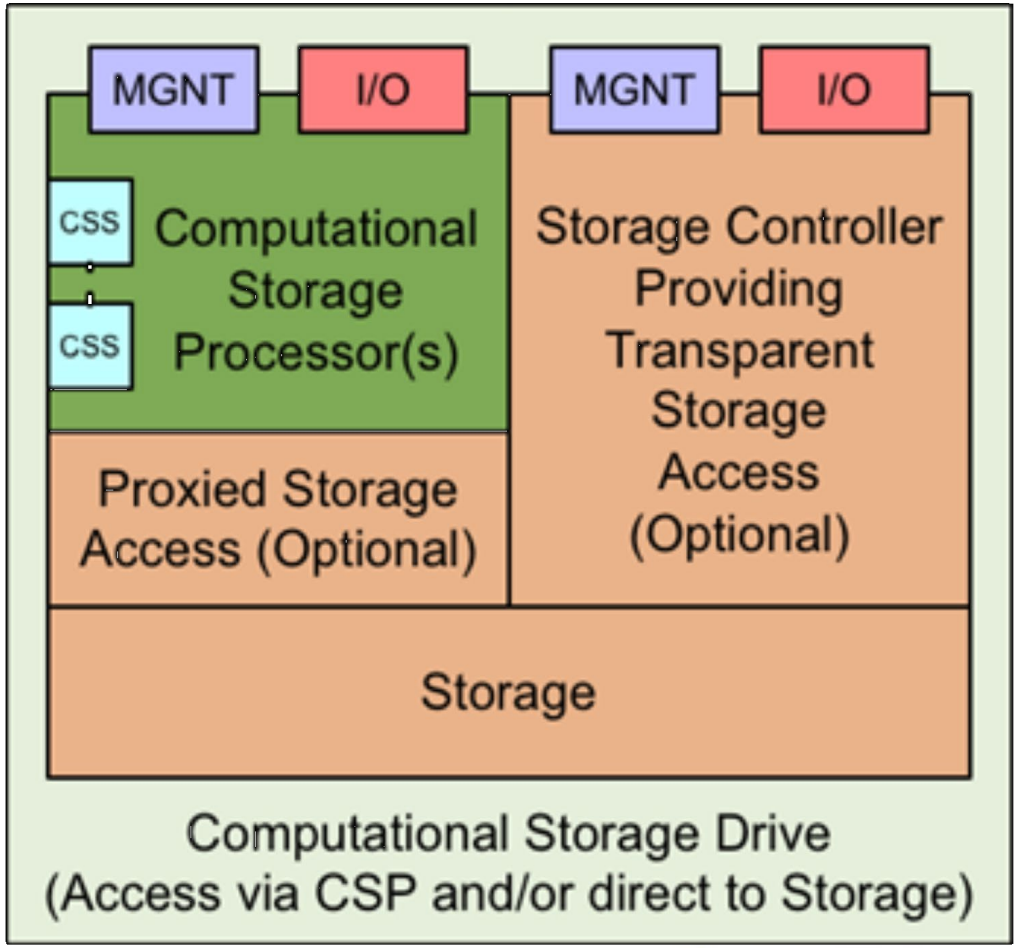
- NGD
- ScaleFlux
- Eideticom
- NVXL
- Samsung
- InSpur
- Cohesity
- IBM

Goal is to reduce data movement

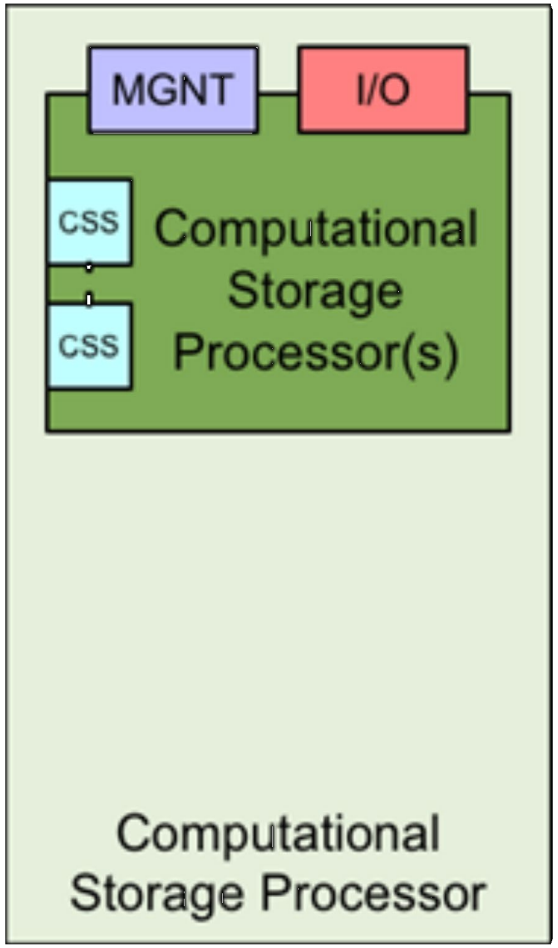
Computational Storage Drive (CSD)



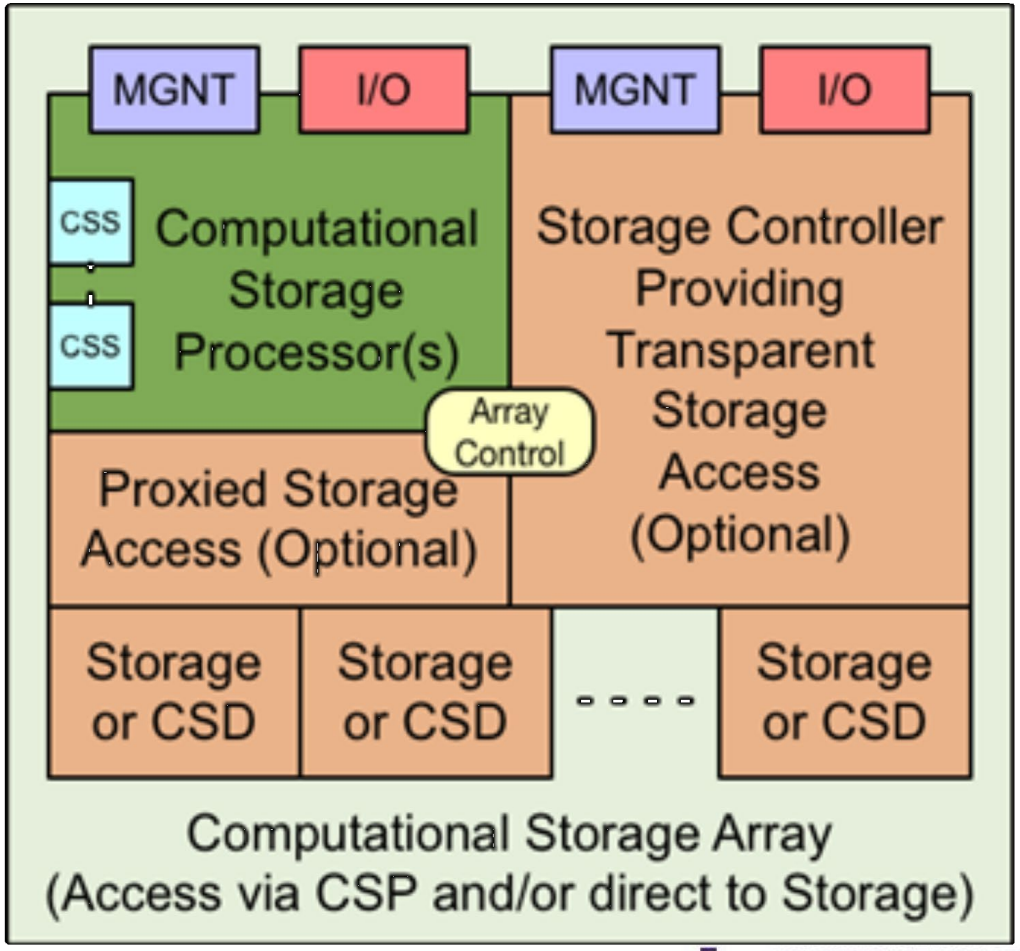
CSD with Two Access Paths



Computational Storage Processor



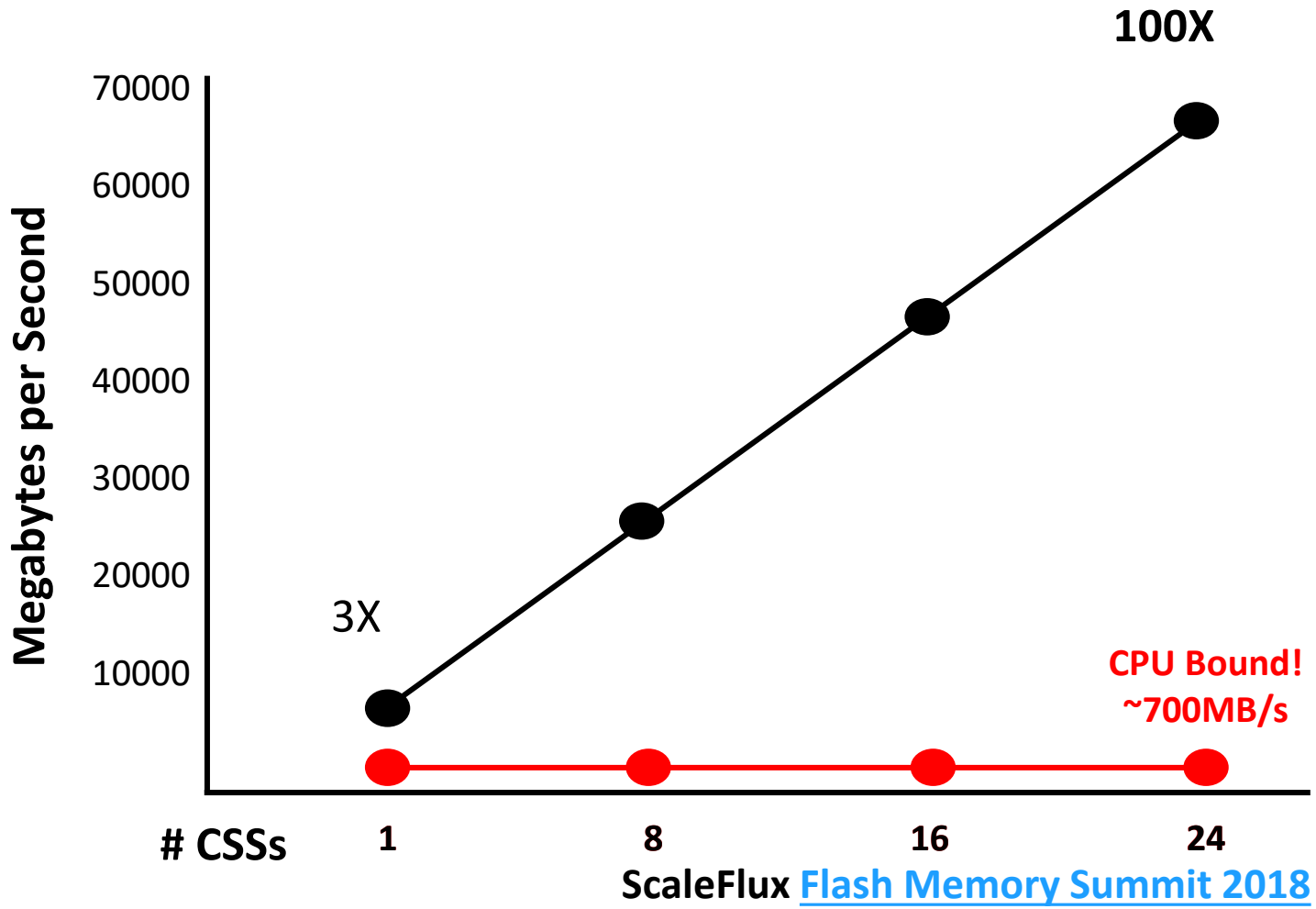
Computational Storage Array



Performance Scales with CSS Count

Fuzzy Search

(POC Unindexed Text Data, Edit Distance = 8, E5-2637v3)



ScaleFlux Flash Memory Summit 2018

Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

Tuning Algorithms for Computational Storage & PIM

- **Step 1: Standard application programs, but broken apart**
  - This part's for the server, that part's for computational storage
- **Step 2: Optimized routines to improve benefits**
  - Lightly-restructured programs to keep both sides busy
- **Step 3: Altogether new algorithms**
  - Wow! Can we really do that?

- **It's all baby steps**

Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

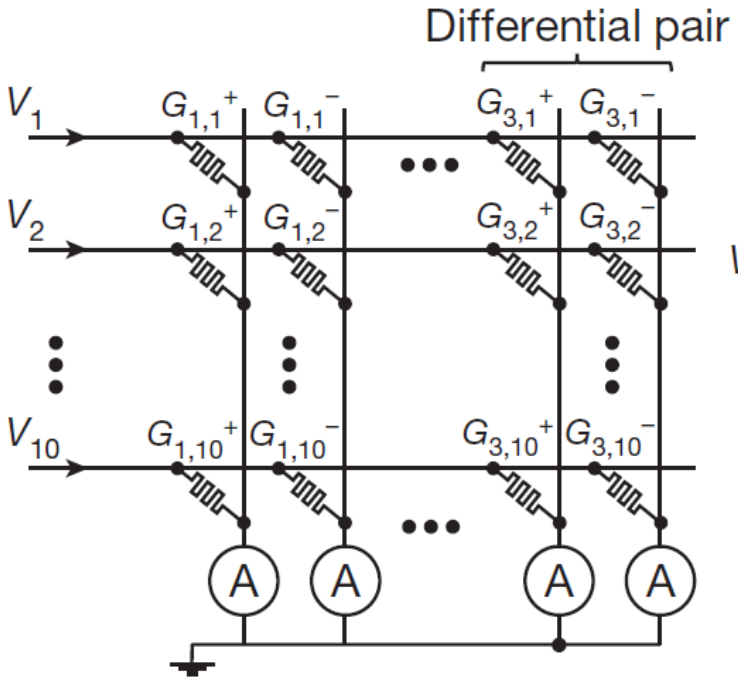
Harnessing DRAM's Internal Peculiarities

- **Take advantage of DRAM's internal weaknesses**
  - Uses linear aspects of commodity DRAM chips
- **Applies different math: Majority/Not**
  - Algorithms must be re-worked
  - Architectures need re-configuring
- **In research institutes:**
  - ComputeDRAM: Princeton
  - SIMDRAM: ETH Zurich, U of Ill., etc.
  - Ambit: ETH Zurich, CMU, Microsoft, Nvidia

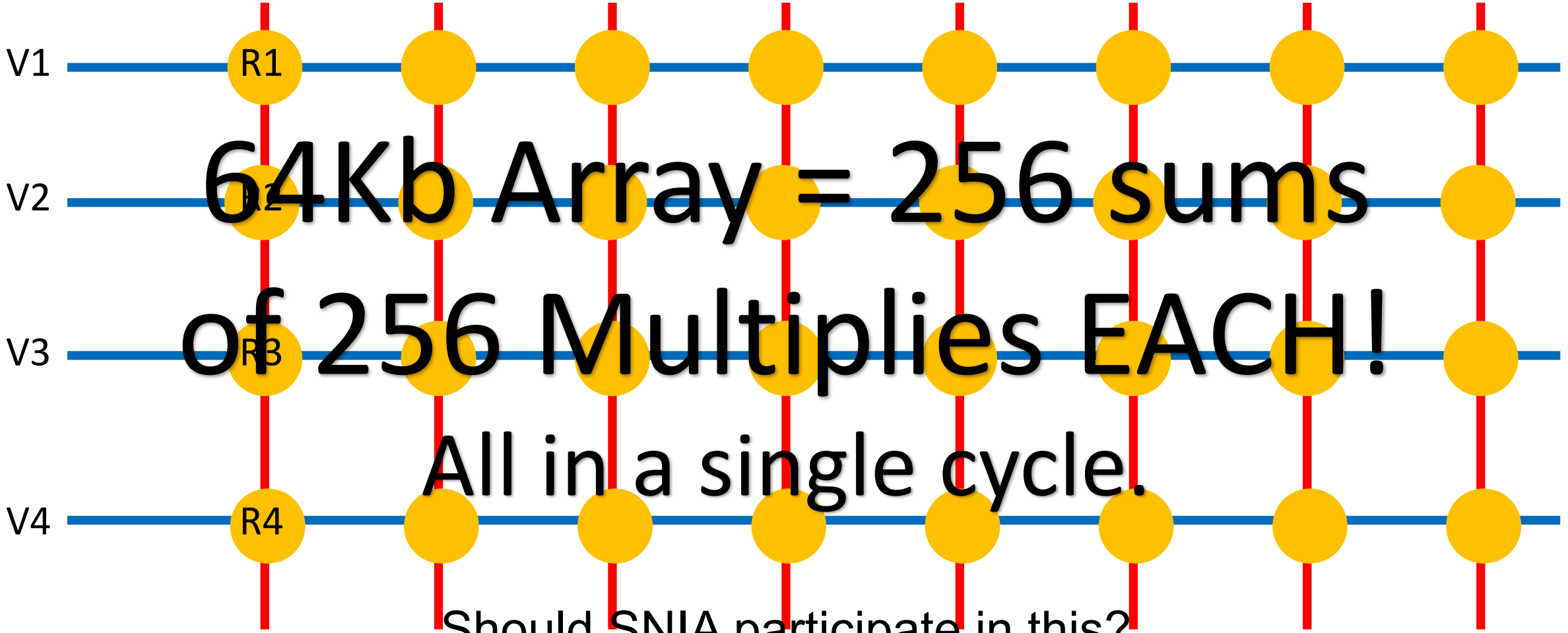
Neural Nets

- **Old idea seeing renewed interest**
- **Instant Matrix Algebra**
  - Somewhat slow because it's linear
- **Simple operation**
  - Difficult to set up
- **A good accelerator to a standard CPU**
- **Fits emerging memories well**
- **Lots of research, but no products**

5.5E-07	3.7E-07	4.2E-07	3.5E-07	5.0E-07	4.7E-07	4.4E-07	5.0E-07
3.7E-07	3.6E-07	6.3E-07	3.7E-07	4.1E-07	4.2E-07	5.3E-07	3.3E-07
4.6E-07	5.7E-07	5.4E-07	4.9E-07	4.9E-07	4.2E-07	5.6E-07	6.0E-07
4.6E-07	4.2E-07	3.6E-07	3.1E-07	2.7E-07	3.7E-07	4.4E-07	3.7E-07
3.5E-07	4.0E-07	5.8E-07	4.8E-07	6.5E-07	4.1E-07	4.0E-07	4.4E-07
4.5E-07	3.8E-07	5.4E-07	4.7E-07	5.9E-07	4.6E-07	4.7E-07	4.8E-07
3.6E-07	4.1E-07	4.5E-07	3.9E-07	5.0E-07	3.6E-07	5.6E-07	4.8E-07
3.5E-07	4.0E-07	4.3E-07	4.1E-07	3.5E-07	4.4E-07	4.6E-07	3.7E-07
4.9E-07	3.7E-07	6.0E-07	3.6E-07	3.3E-07	5.1E-07	3.9E-07	4.2E-07
4.4E-07	3.3E-07	3.3E-07	4.0E-07	3.9E-07	4.5E-07	4.3E-07	4.4E-07



Simplifying AI



64Kb Array = 256 sums of 256 Multiplies EACH!

All in a single cycle.

Should SNIA participate in this?

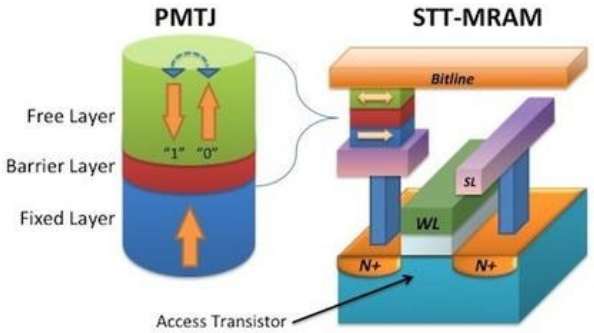
=V1*(1/R1)+V2*(1/R2)+V3*(1/R3)+V4*(1/R4)

Outline

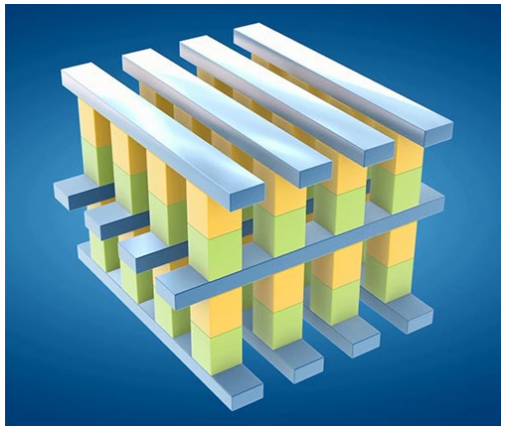
- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

Lots of Emerging Memories…

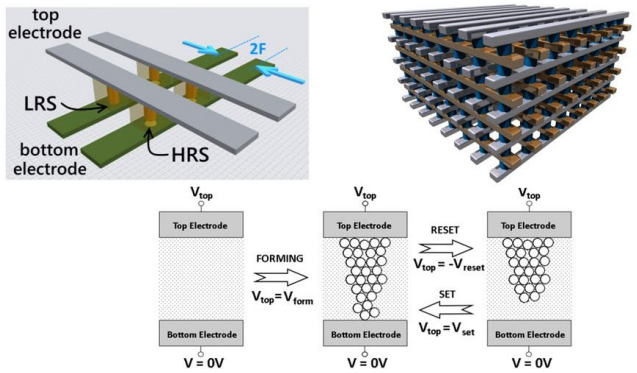
MRAM



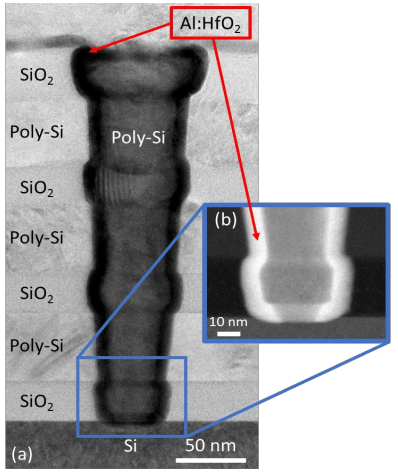
PCM



ReRAM



FRAM

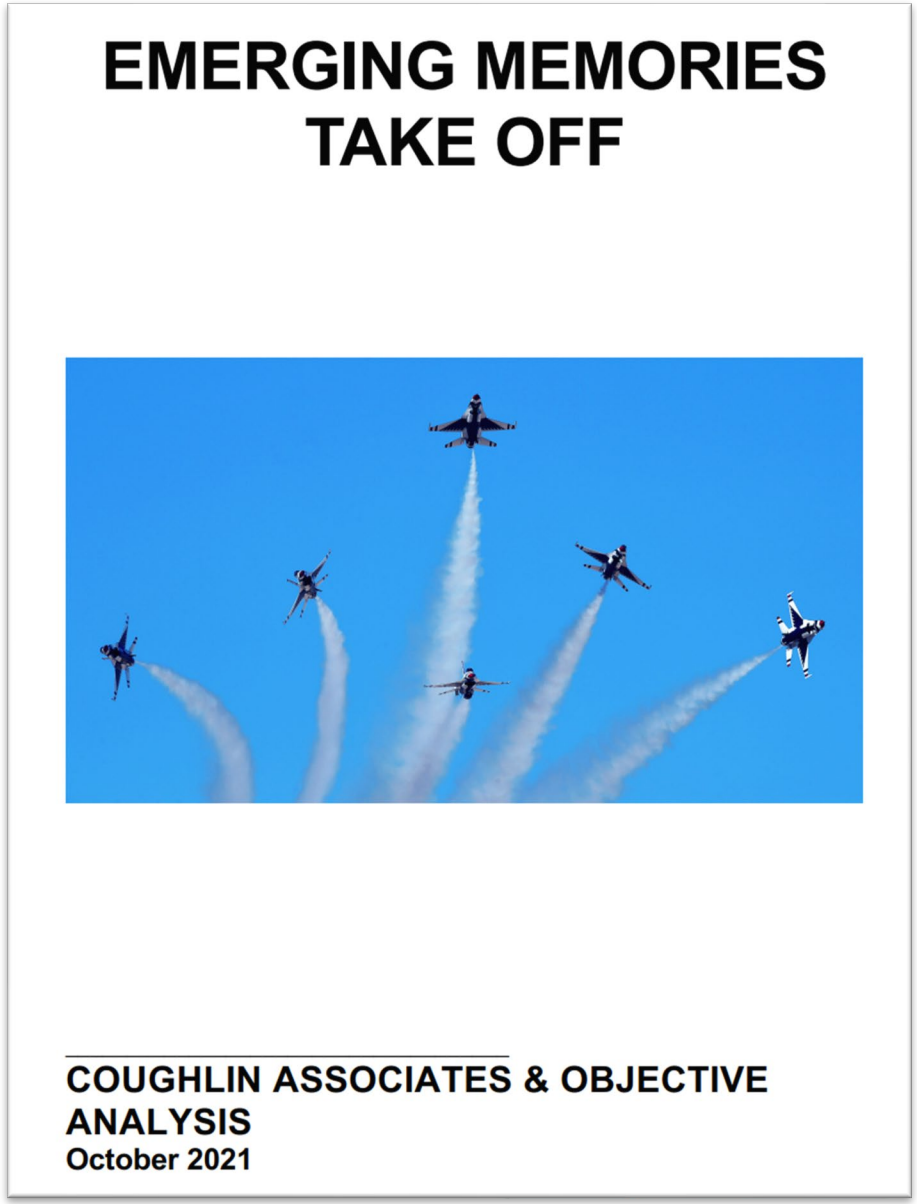


What Emerging Memories Can Offer

- **Persistence**
  - Instant On
  - Better for power-loss protection
  - Reduce power consumption
- **Small cell size**
  - Large arrays fit onto the processor die
- **Crosspoint configuration**
  - Fits neural networks well
  - Can store linear values

Hey! We Wrote a Report on These!

- **Emerging Memories Take Off**
  - In-depth coverage of everything in this presentation
  - 231 pages, 155 figures, 36 tables
  - Can be purchased on-line for immediate download
- **Two ways to order:**
  - https://Objective-Analysis.com/reports/#Emerging
  - http://www.TomCoughlin.com/tech-papers.htm



Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

Standards Are Essential

- **SNIA achieved a lot with the NVM programming model**
  - Now we need to consider persistent processor caches and registers
- **The Computational Storage TWG is well embarked for success**
  - Standards and taxonomy are progressing well
  - Processing in Memory (PIM) should follow their lead
    - **Perhaps not in SNIA**
    - **PIM interfaces will need to be standardized as was CXL**
- **Neural nets may be the next frontier**
  - It's storage, but is it storage?

Outline

- Coping with Inefficient Data Movement
- Bringing Persistence Closer to the Processor
- Memory & Storage Interfaces Changing, Growing
- Compute-in-Memory, Computational Storage
- New Algorithms Require New Architectures
- Abandoning the von Neumann Architecture
- Emerging Memories to the Rescue
- Making It All Work Together
- Q&A

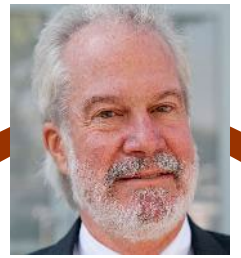
Please take a moment to rate this session.

- Your feedback is important to us.

Coughlin Associates

- https://tomcoughlin.com
- Technical and Market Analysis
- Consulting
- Reports and Newsletter
  - Emerging Memories Report
  - Digital Storage in Media and Entertainment
  - Digital Storage Technology Newsletter

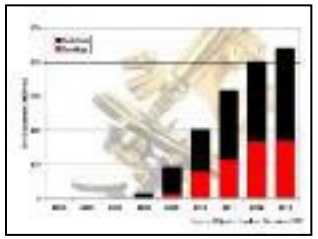
Objective Analysis

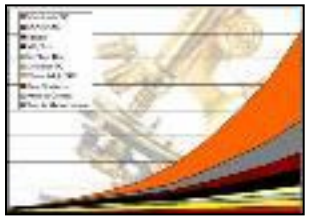


Profound Analysts

Reports & Services

Custom Consulting





Objective Analysis Semiconductor Forecast Accuracy

Year	Forecast	Actual
2008	Zero growth at best	-3%
2009	Growth in the mid teens	-9%
2010	Should approach 30%	32%
2011	Muted revenue growth: 5%	0%
2012	Revenues drop as much as -5%	-2.7%
2013	Revenues increase nearly 10%	4.9%
2014	Revenues up 20%+	9.9%
2015	Revenues up ~10%	-0.2%
2016	Revenues up ~10%	1.1%
2017	Revenues up ~20%	22%
2018	Strong start supports 10+% growth	14%
2019	Semiconductors down -5%	-12.5%
2020	Zero growth at best	6.8%
2021	Revenues grow 6% by remaining flat	26.2%
2022	Total semi still grows 6%	TBD

