



Flash Memory Summit

Trends and Applications for Emerging Memory Technology

Tom Coughlin

Coughlin Associates, Inc.

<https://tomcoughlin.com>

Jim Handy

Objective Analysis

Jim.Handy@objective-analysis.com

Outline



Flash Memory Summit

- Why we need emerging memories
- Changes to the computing model
- Meet the memories
- Real-life applications
- Outlook
- Summary
- References



Why we need emerging memories

Big Data Growth Ahead

- Projections of big data generation growth
- Greater use of AI in devices, at the network edge and in data centers
- Power consumption as well as latency and data rates are other important considerations in future system design
- This leads us to look at the memory technologies which will enable handling this data tsunami

Why Emerging Persistent Memories are Necessary

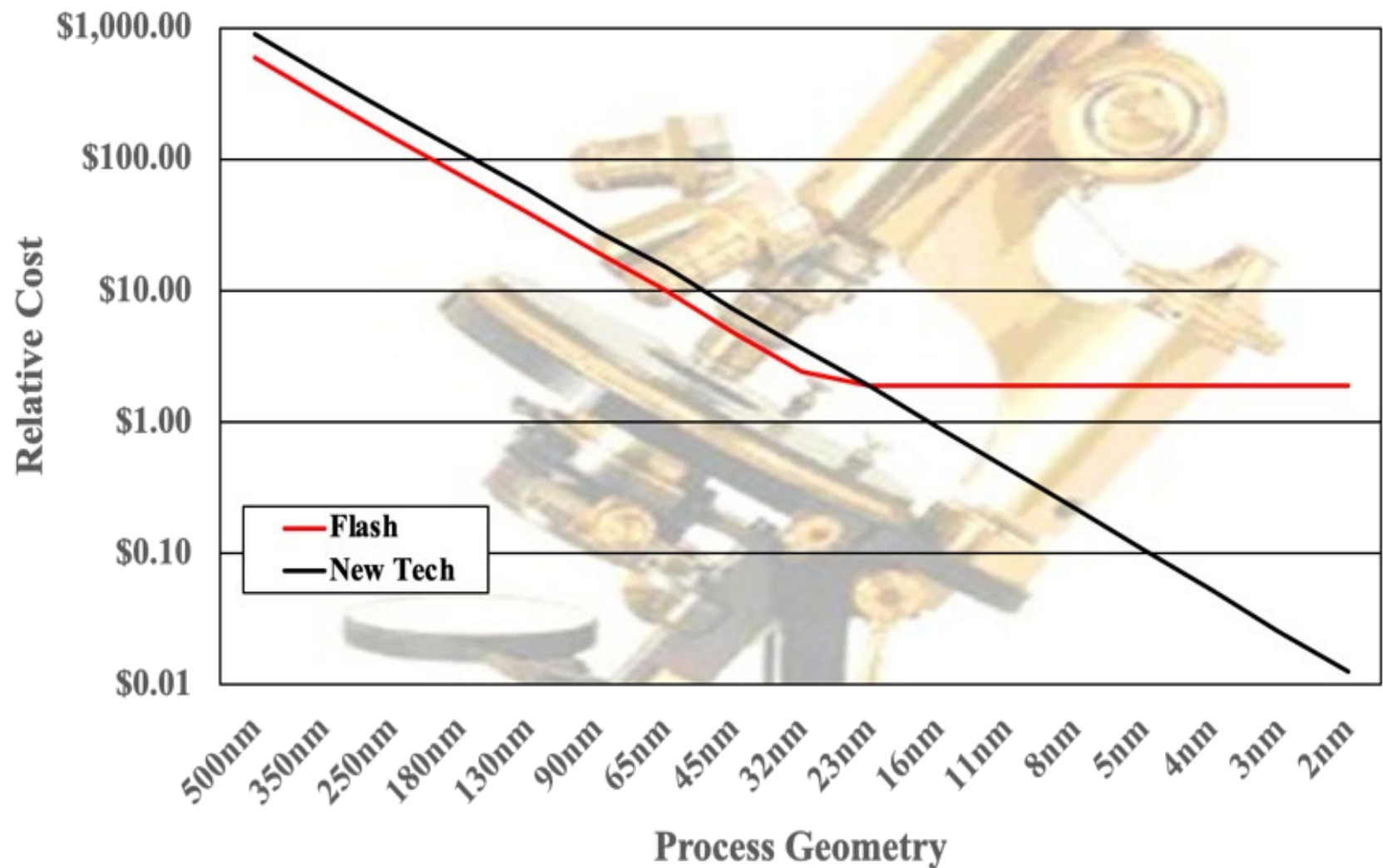


Flash Memory Summit

- Flash can't scale with process advances
 - NAND flash went 3D at 15nm
- NOR scaling stops with FinFET
 - 28nm & smaller processes need something new
- SRAM scaling may stop at 14nm
- DRAM scaling has slowed

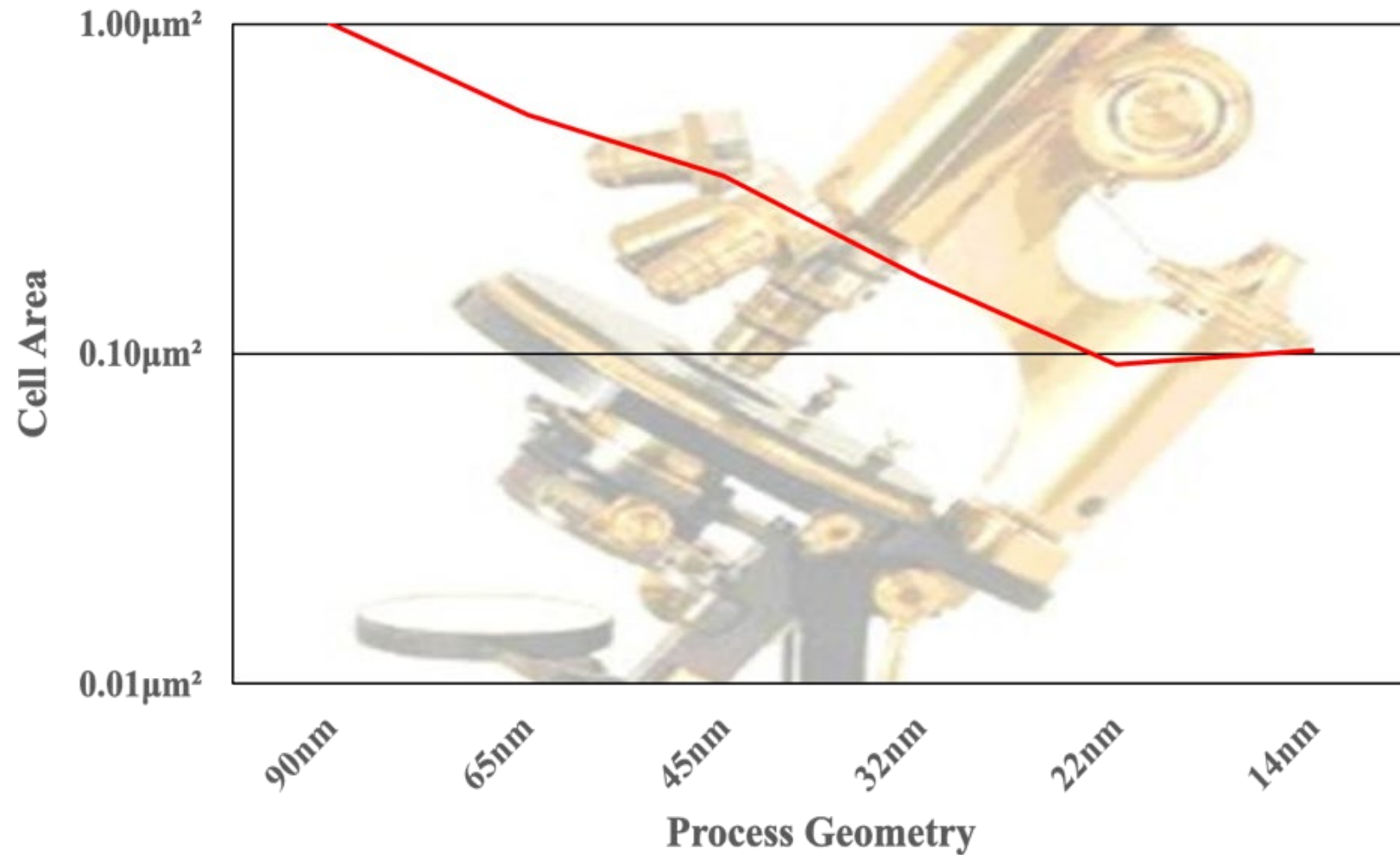
Low power high density memory needed for embedded and data center applications

NOR Flash Scaling Ends at 28nm

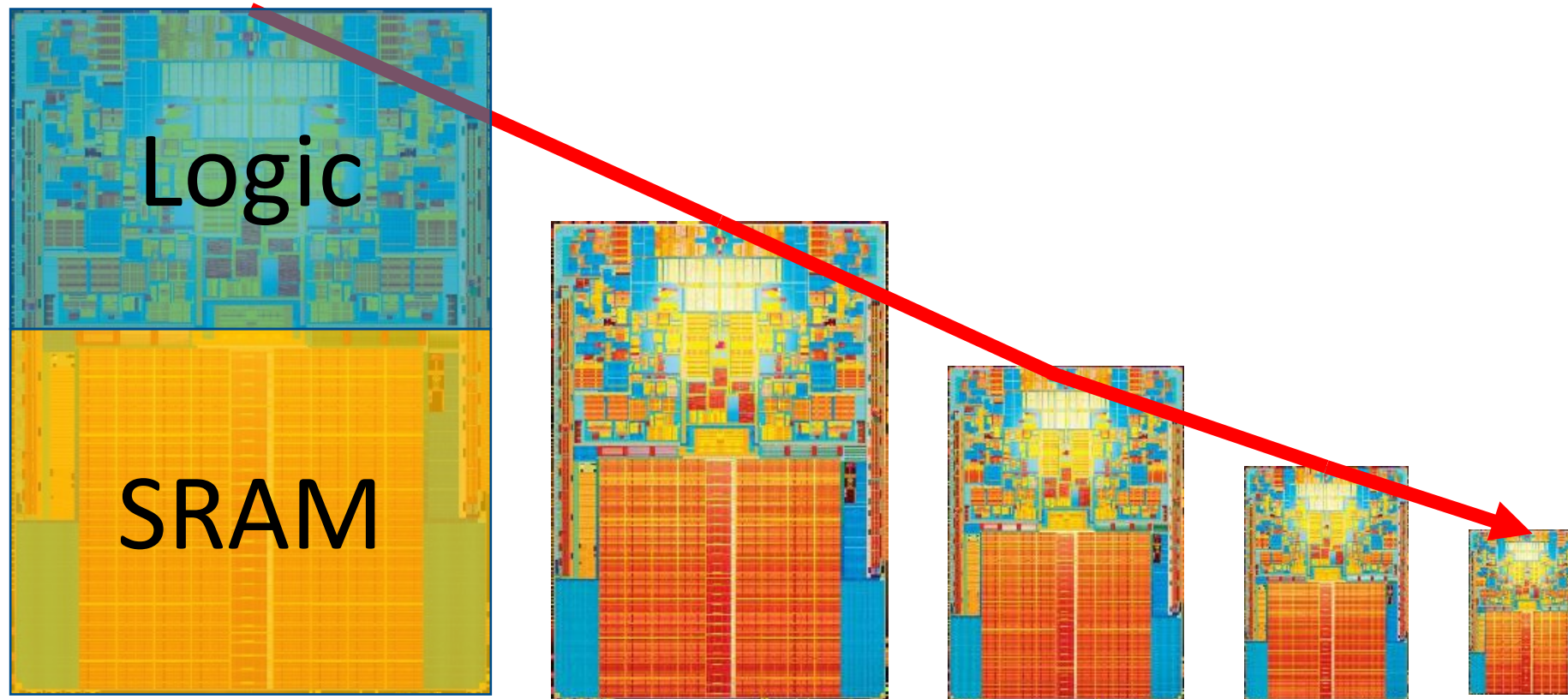


© 2022 Coughlin Associates

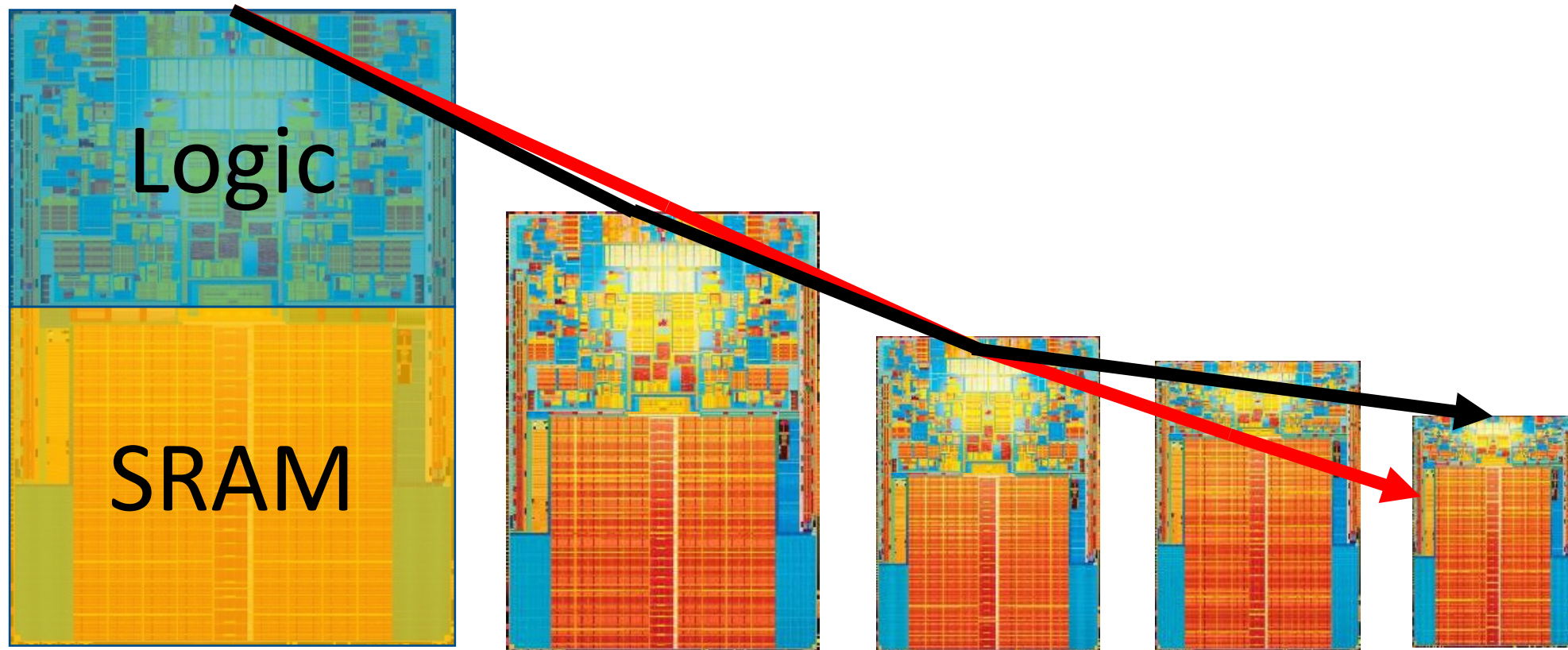
SRAM's In Trouble Too



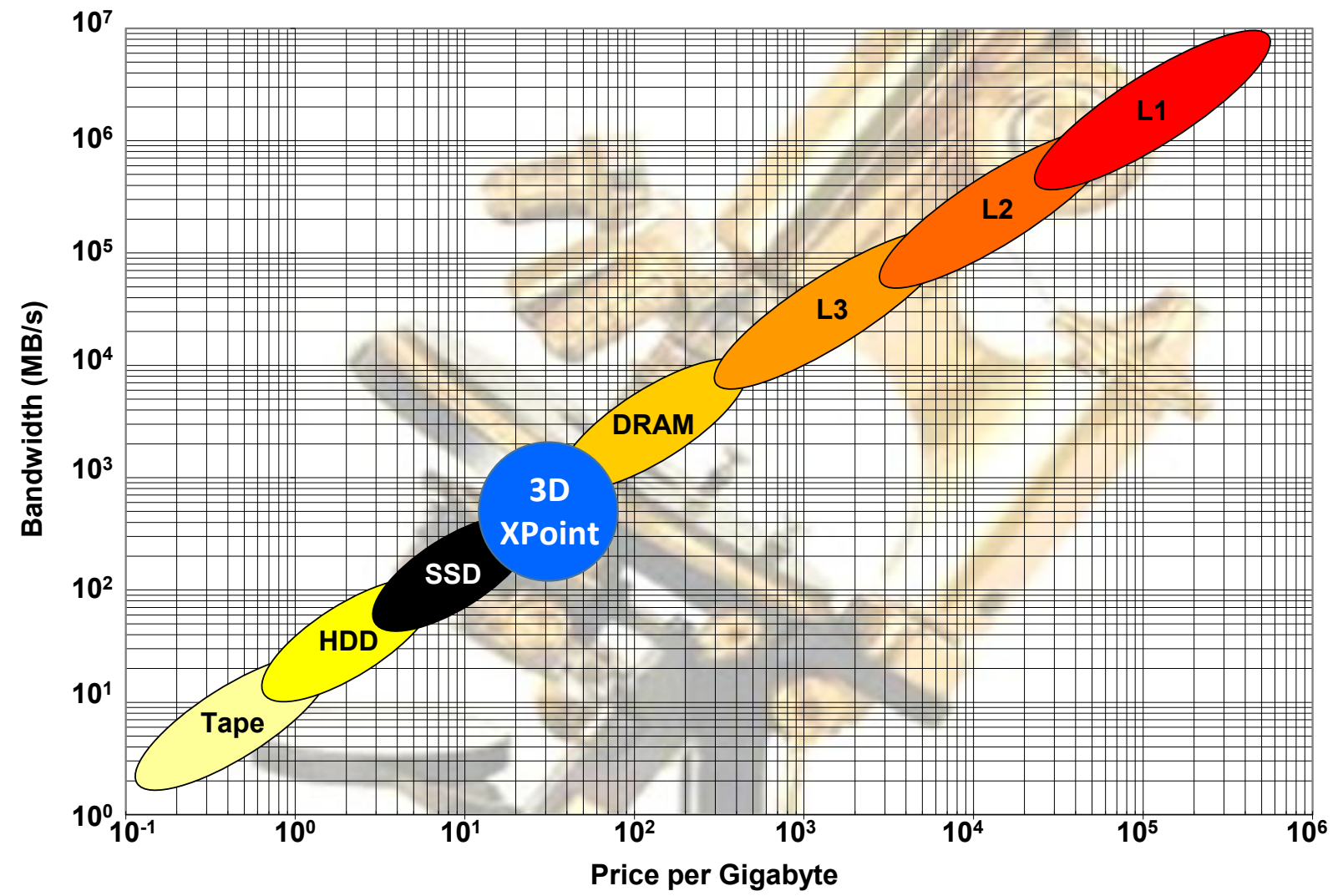
Putting Things in Perspective



Putting Things in Perspective



Key Issue: Pricing!



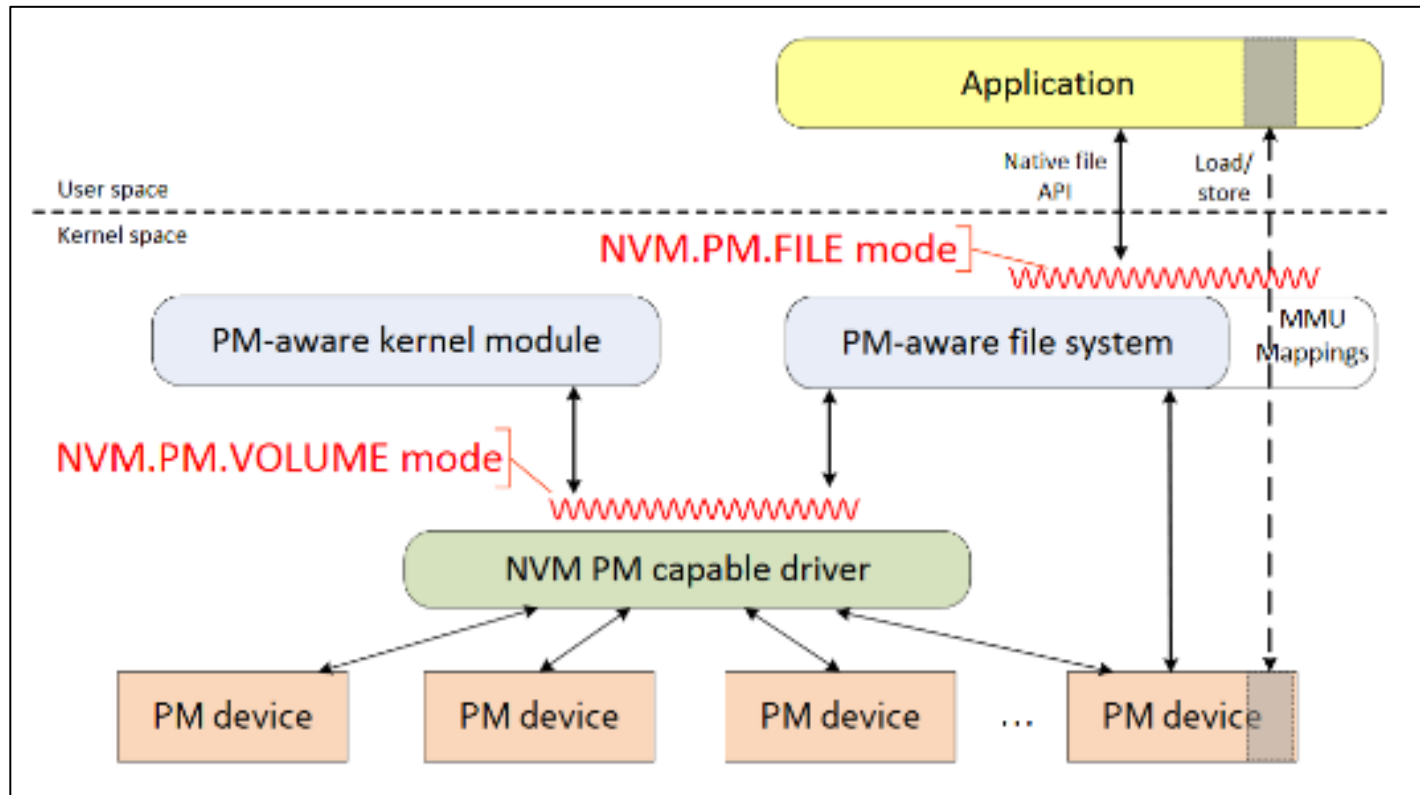
© 2022 Coughlin Associates



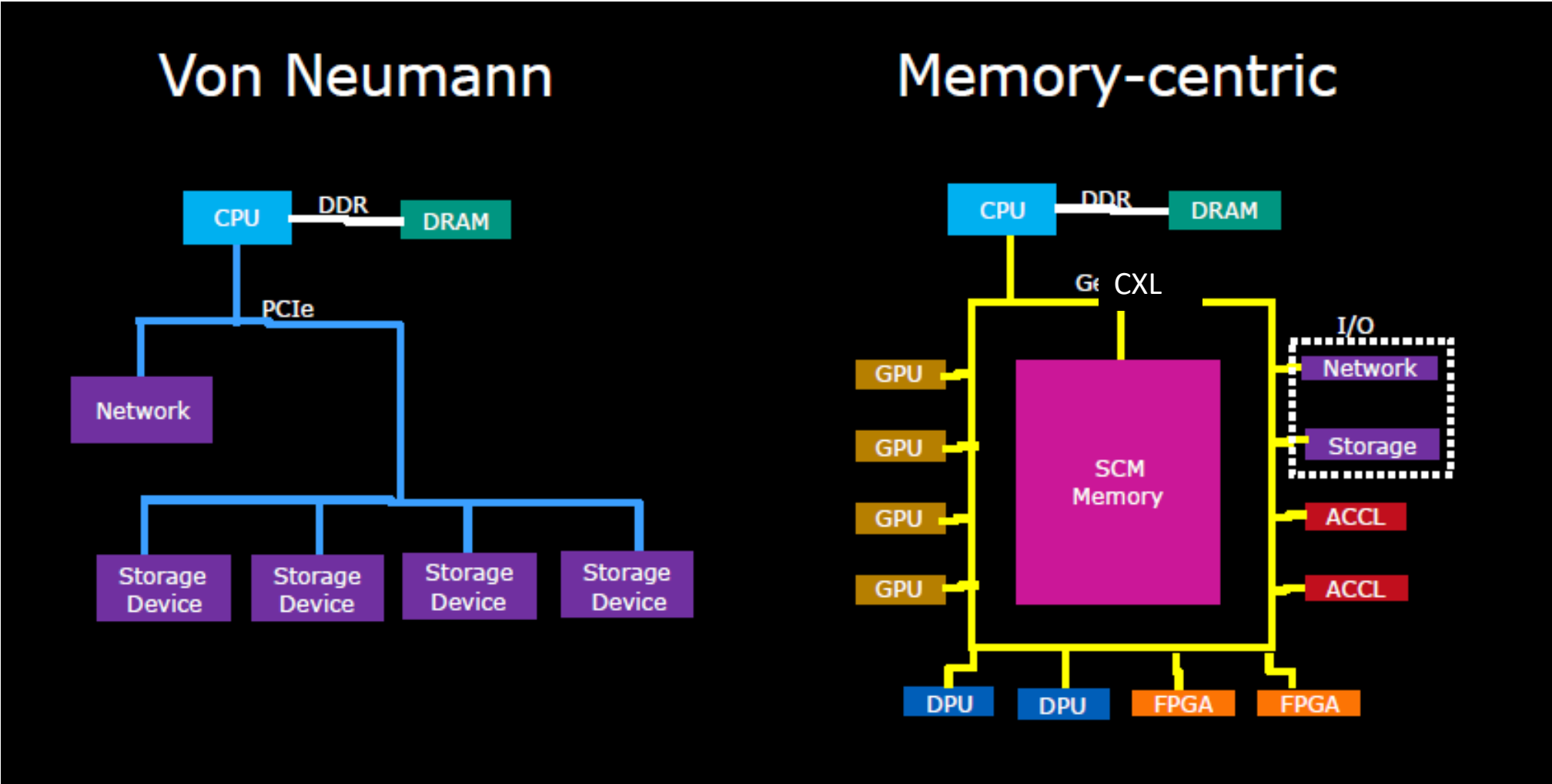
Changes in the computing model

Software Support for Persistent Memories

- SNIA's Persistent Memory Programming Model
 - <https://www.SNIA.org/PM>
- Development of NVMe
- NVMe over Fabric (NVMe-oF)
- RDMA, ROCE—remote memory-like data access
- New computing/connection models

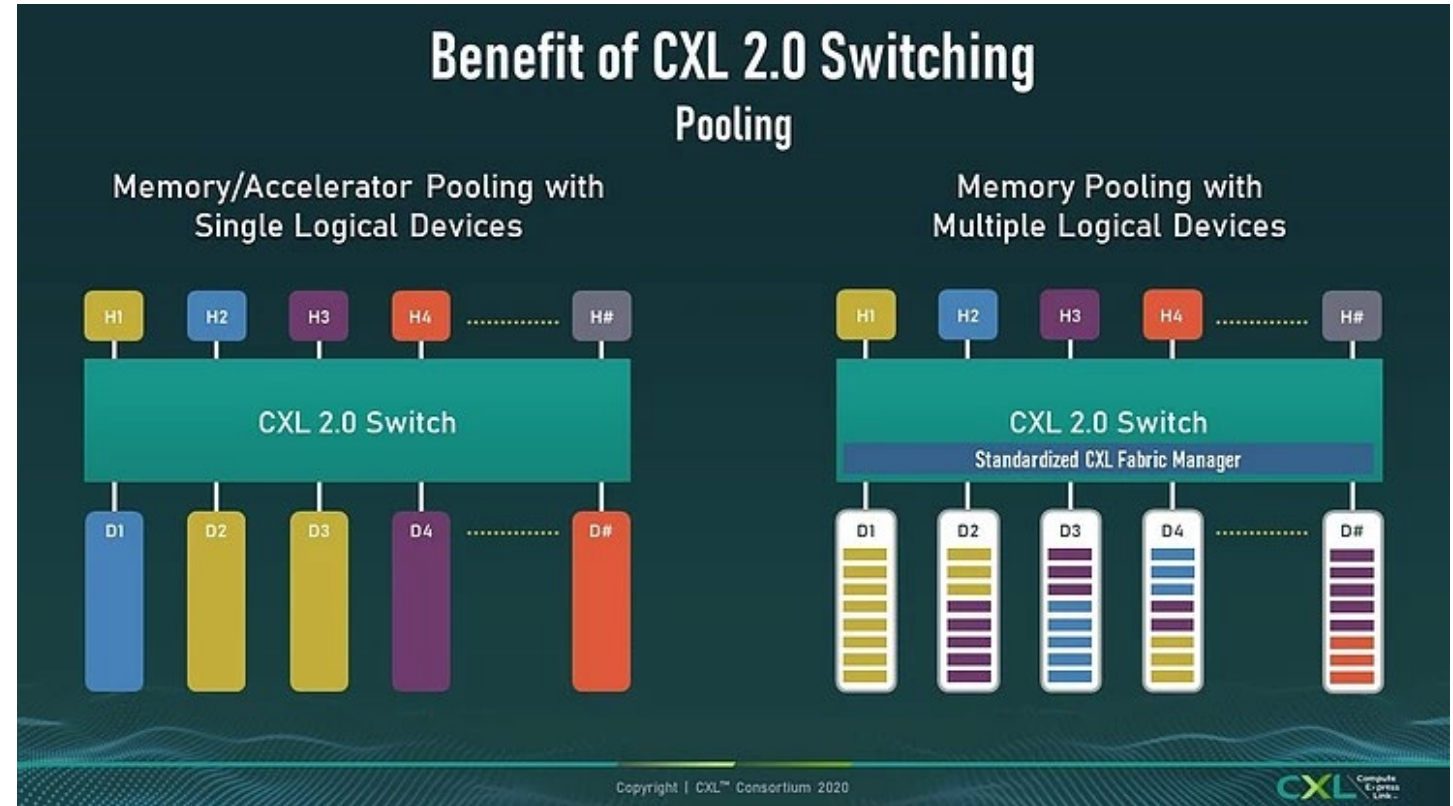


Moving from a von Neumann to a Memory-Centric Compute Model



Current Thinking on Future Memory Channels

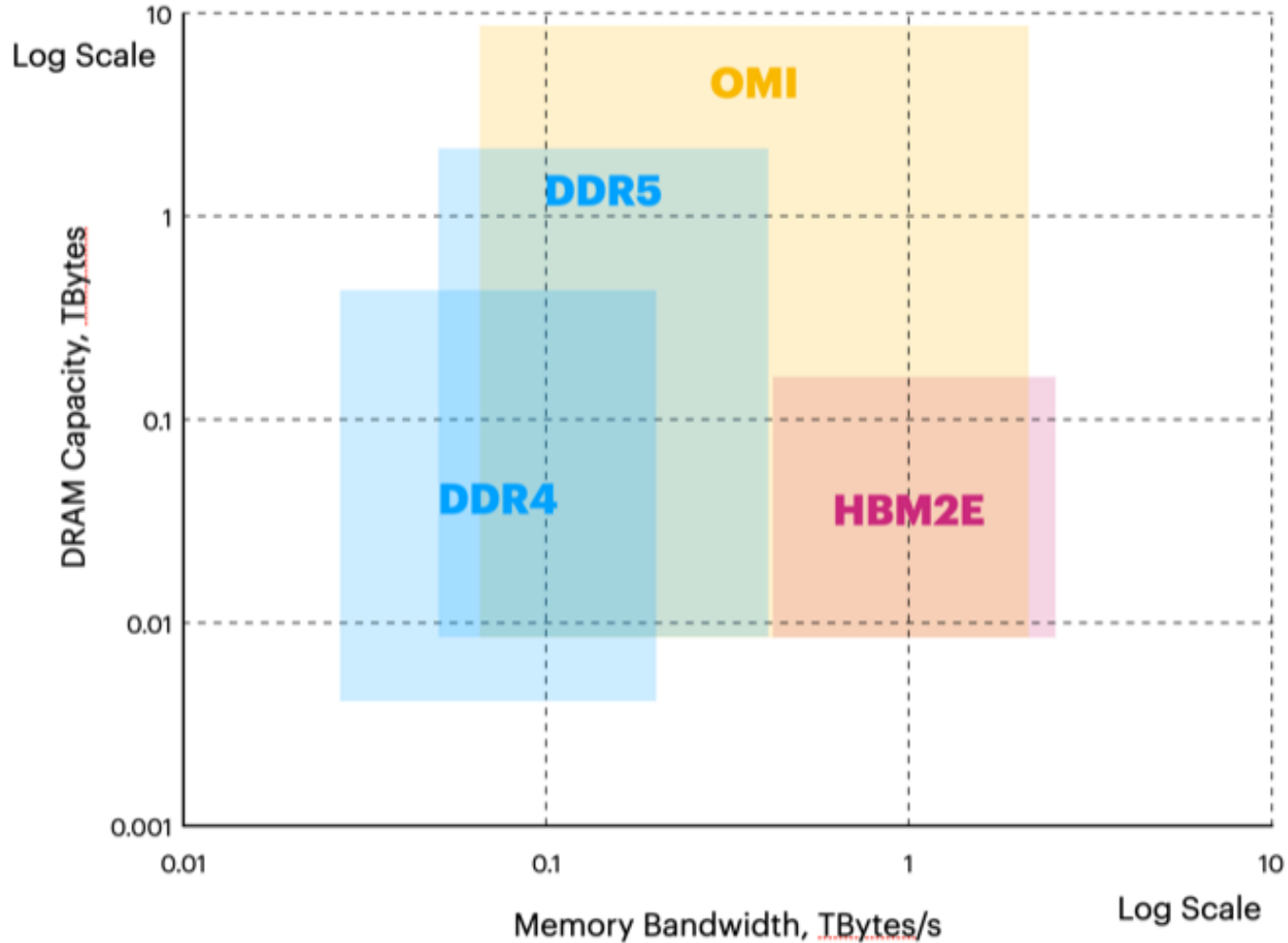
- CXL enables pooling of non-heterogenous memory
- CXL supports accelerators near memory (in memory as well)
- DDR channels will continue use for low latency memory
- Increasing data rates may drive a low latency rather than parallel memory interface
- GenZ for connecting storage boxes and racks together



Memory Interconnect Types

- CXL for “Far Memory”
 - Pools heterogeneous memories
 - Mixed latencies and data rates
- Gen-Z to connect storage boxes and racks
- The DDR interface will stay with us for “Near Memory”
 - DDR good for smaller systems
 - HBM fast but restrictive and costly
 - OMI for both high speed and large capacities

Approaches to Near Memory



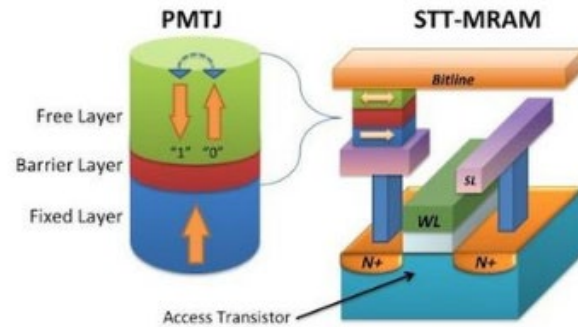
© 2022 Coughlin Associates



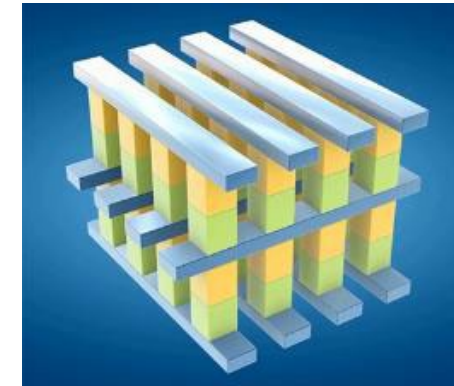
Meet the memories

Candidates for Persistent Memory

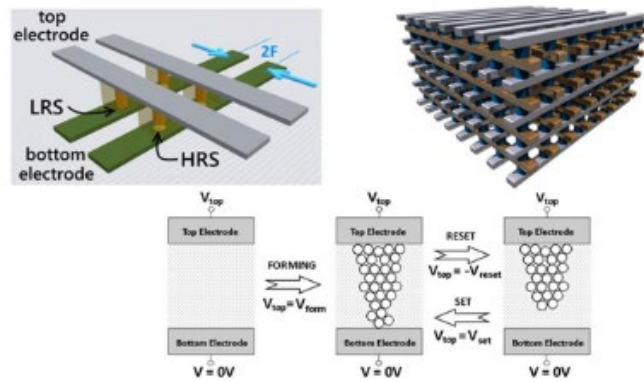
MRAM



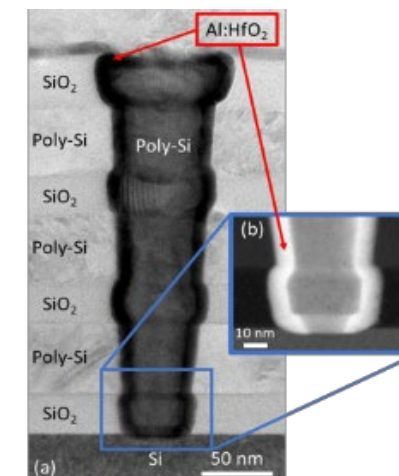
PCM



ReRAM



FRAM



MRAM



Flash Memory Summit

- Everspin
 - Partnership with Global Foundries
 - Used in IBM's FlashCore modules in FlashSystem arrays
- Renesas (Formerly IDT)
 - 8Mb, SPI
- Avalanche is shipping some MRAM for mil/aero applications
- Leading foundries starting to ship embedded MRAM
 - GlobalFoundries
 - TSMC
 - Samsung
 - Others



Source: TechInsights



PCM (3D XPoint)

- The second-oldest emerging memory (1970)
- Intel Optane products
 - NVMe shipped in 2017
 - DIMMs in 2018
 - In 2021 Intel stopped production of consumer NVMe Optane SSDs
- Micron Abandoning 3D XPoint
 - Sold Lehi fab



ReRAM



- Adesto has shipped CBRAM chips for several years
 - Dialog Semiconductor acquired Adesto June 2020
 - Will license CBRAM technology to GLOBALFOUNDRIES
 - GF will first offer as an embedded, option on its 22FDX platform
 - GF Plans to extend to other platforms.
- Cerfe Labs (Arm spin-out)
 - Correlated electron materials (CeRAM)
 - Licensed from Symetrix.
- Others (Mitsubishi, Fujitsu, Panasonic, Winbond, Honeywell,...)
- Foundry support (GLOBALFOUNDRIES, TSMC, others)

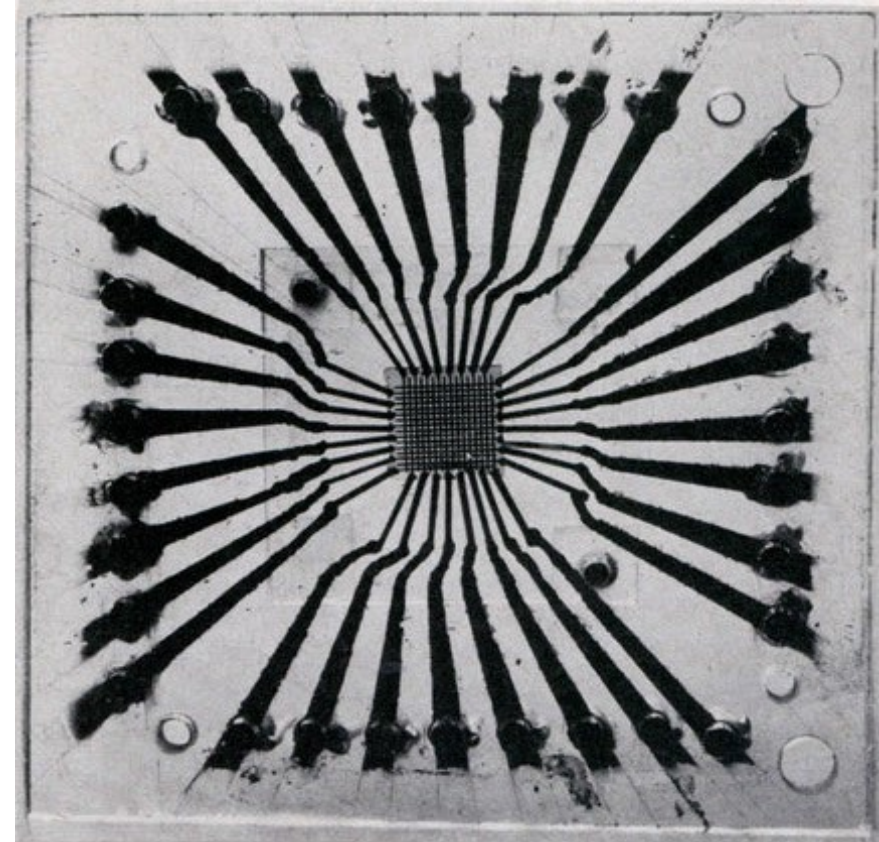


FRAM



Flash Memory Summit

- The oldest emerging memory (1955)
- The highest-shipped emerging memory
- Finding new life with new materials



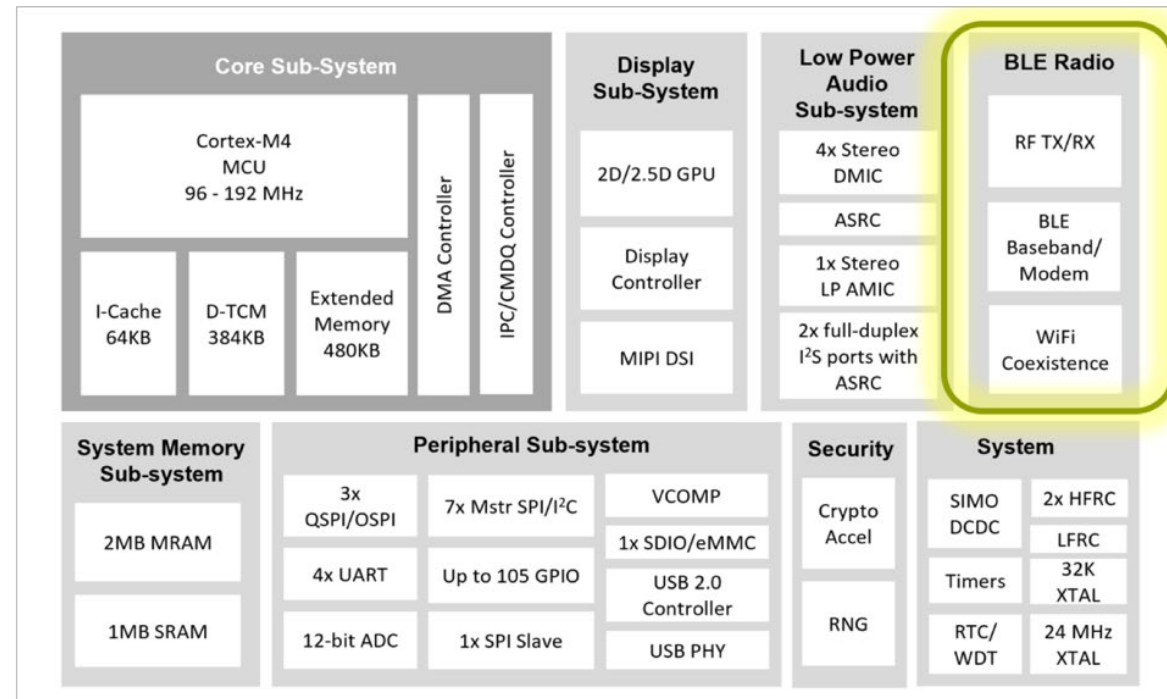
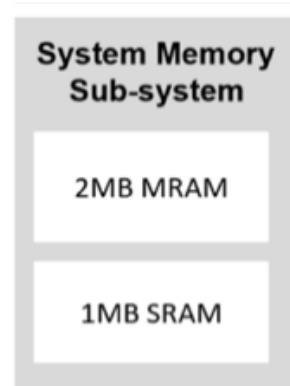
© 2022 Coughlin Associates



Real life applications

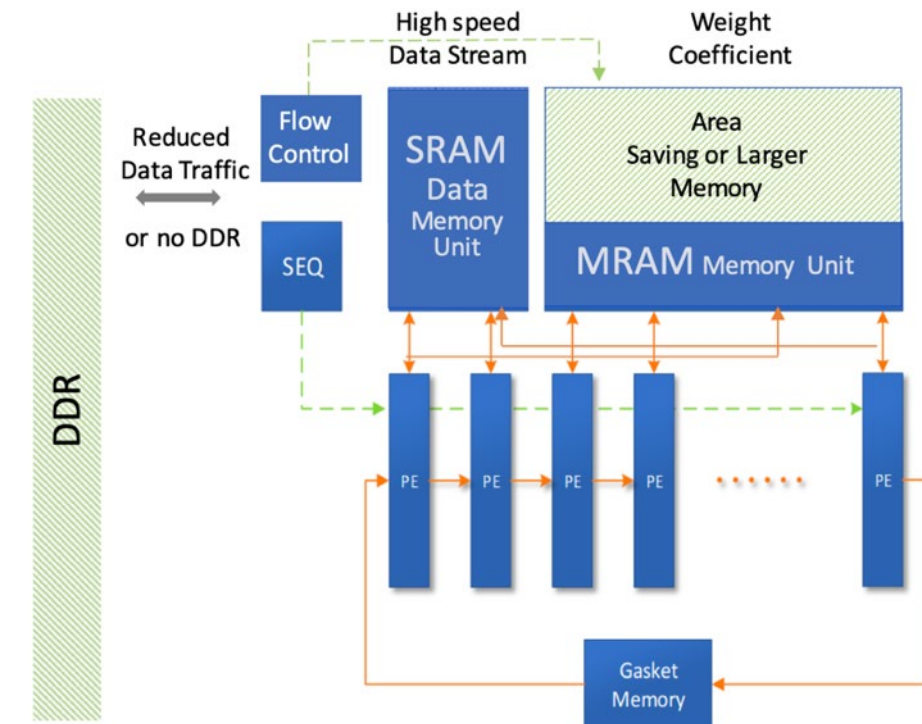
MRAM IoT SoC – Ambiq Apollo

- SoC for intelligent endpoint IoT devices
- Ultra-low battery power
- Serves as both an application processor and a coprocessor



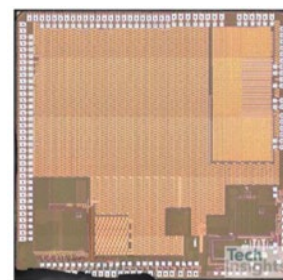
MRAM DNN Accelerator Chip – NuMem

- Used by NASA
- 1-32 processing engines, 32-1024 ALUs per chip
 - Efficient for matrix multiplication, convolution, etc.
- Radiation hard, high endurance MRAM
 - Nonvolatility reduces energy requirements
- Numerous Applications:
 - Sensor fusion for super resolution
 - Terrain Mapping for Depth and Terrain Classification
 - Navigation systems: Object Detection & Tracking



- **MRAM GPS Receiver**

- CXD5605 GPS Receiver
- Used in Huawei GT 2 Smartwatch
- 8Mb Embedded MRAM
- Samsung 28 FD-SOI Process



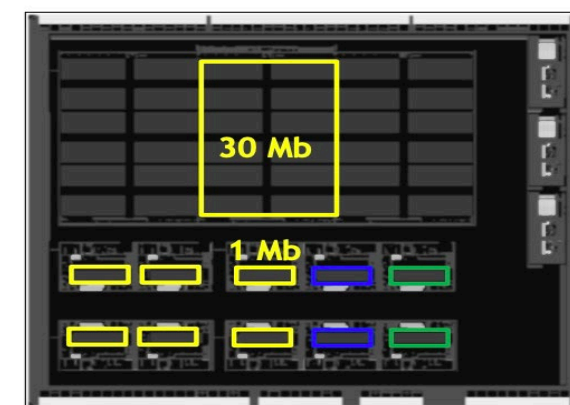
Source: TechInsights



- **MRAM memory buffer embedded in a 3D stacked imaging device for a camera**

- Three different types of MRAM depending upon the data capacity and data retention requirements
- The 30Mb MRAM cells on the right are the lower retention time buffer

Source: Sony Corp., Used with Permission



Yellow: x1.0
Blue : x1.07
Green:x1.15

Other Emerging Memory Products

- NXP MRAM MCU
- STMicro PCM MCU
- TI FRAM MCU
- Fujitsu FRAM MCUs

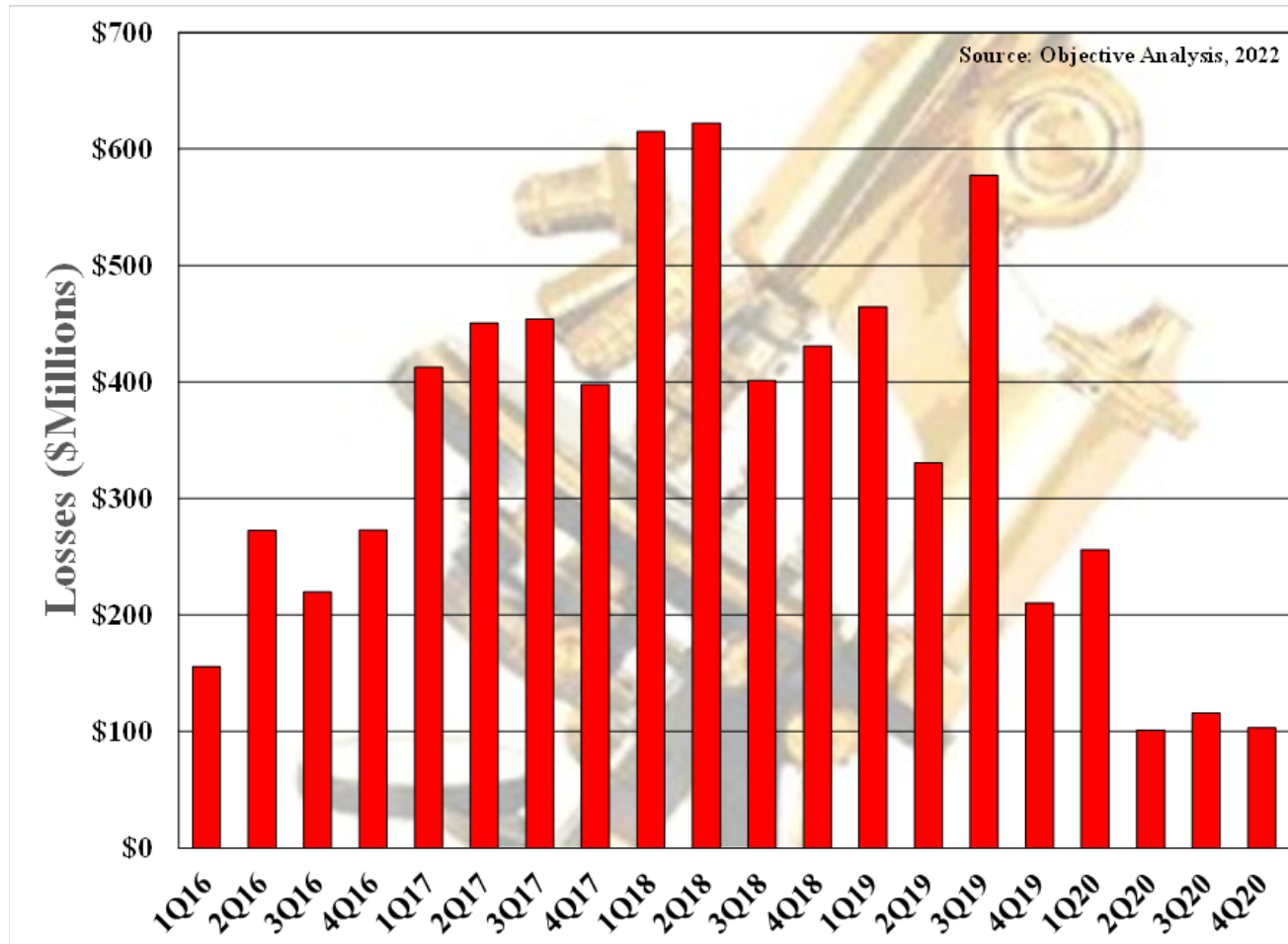




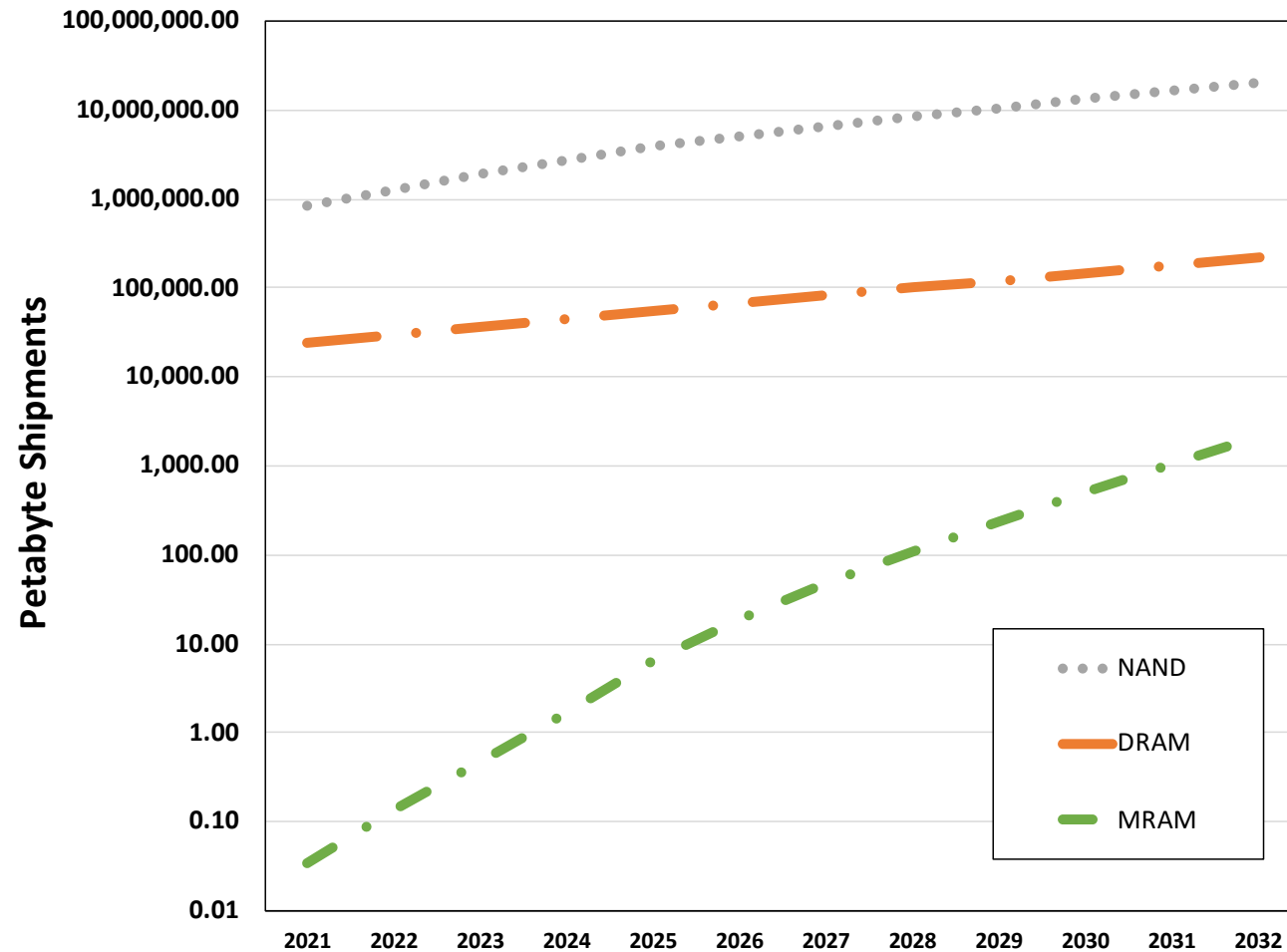
Outlook

The Importance of Economies of Scale

- Intel Optane losses total over \$7 billion!



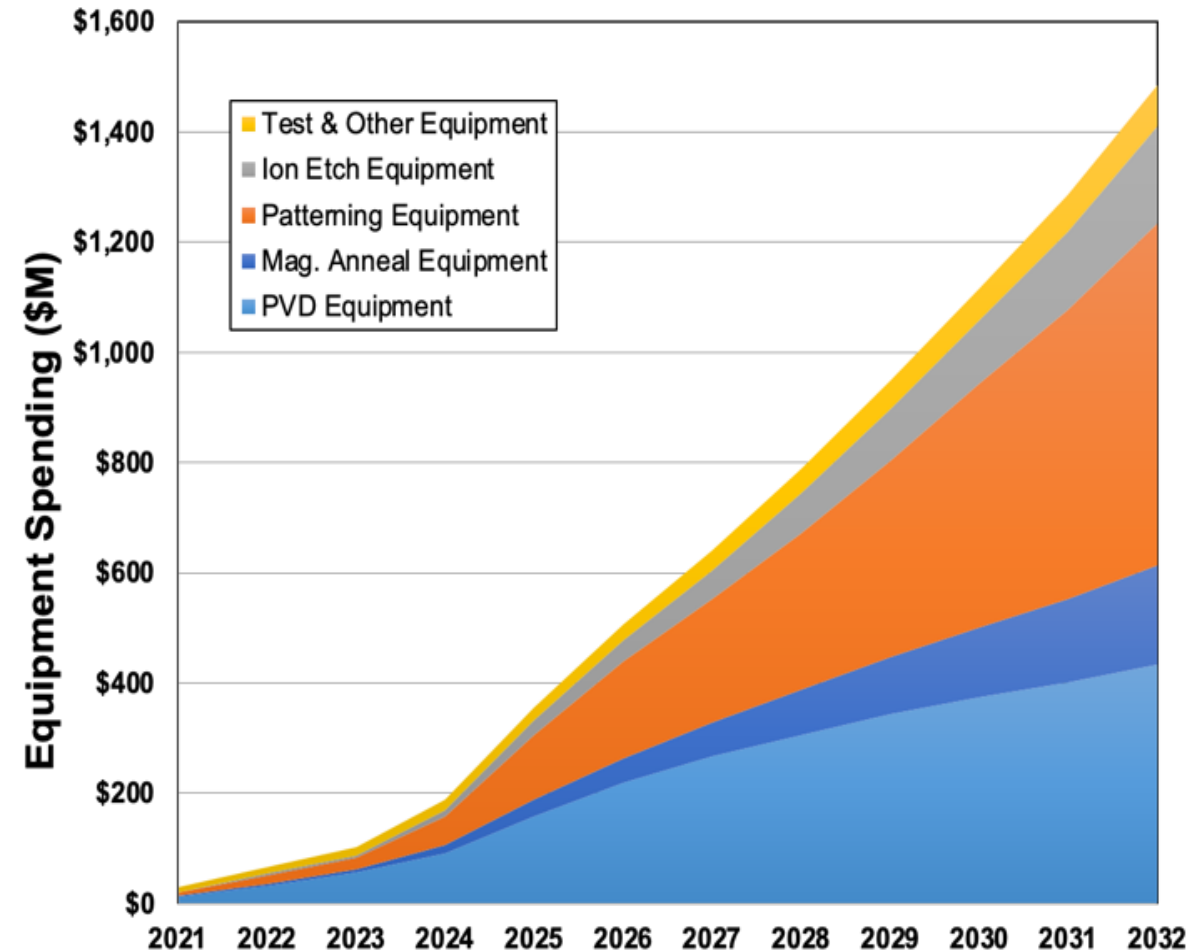
Growth in New Memory Shipments



- DRAM & NAND have staying power
- MRAM is doing well today
 - Also gaining acceptance as embedded memory in SoCs
- MRAM Revenue could exceed \$43B by 2032

Growth in MRAM Production Equipment

- The increasing demand for nonvolatile memory based upon MRAM, STT MRAM and SOT MRAM will cause total MRAM manufacturing equipment revenue to rise to a baseline projected spending of over \$1.4B.





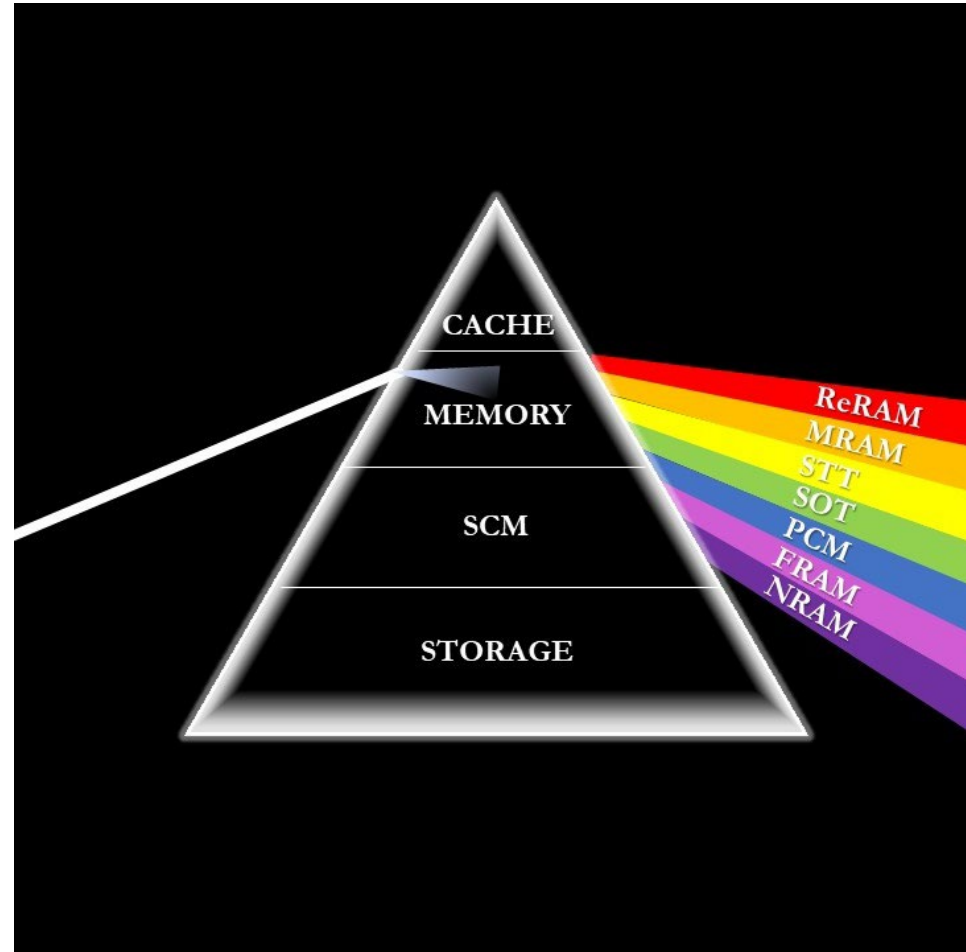
Summary

- NOR flash and SRAM have stopped scaling
- New non-volatile memory types will fill the void
- This will lead to new memory-centric computer architectures
- The storage/memory hierarchy will change
- There are four leading candidates: MRAM, PCM, ReRAM, and FRAM
- Leading foundries already support these new memories
- New memories are in use today
- MRAM revenues should exceed \$43B by 2032

New Report: Emerging Memories Enter the Next Phase



Flash Memory Summit



Coughlin Associates

Available Soon!

<http://www.tomcoughlin.com/techpapers.htm>
<https://Objective-Analysis.com/reports/#Emerging>

© 2022 Coughlin Associates

- SNIA's Persistent Memory Programming Model, <https://www.SNIA.org/PM>
- Computer Express Link 2.0 Specification: Memory Pooling, CXL Brighttalk, March 23, 2020
- OMI White Paper, <https://objective-analysis.com/uploads/2021-04-18%20Objective%20Analysis%20White%20Paper%20-%20The%20Future%20of%20Low-Latency%20Memory.pdf>
- Emerging Memories Take Off, Coughlin Associates and Objective Analysis, <https://tomcoughlin.com/tech-papers/>



Thanks

<https://tomcoughlin.com>
tom@tomcoughlin.com

LinkedIn:

<https://www.linkedin.com/in/thomas-coughlin-41a65/>

