HEPiX TechWatch WG: Storage

Germán Cancio (CERN), Martin Gasthuber (DESY), Kai Leffhalm (DESY), Shigeki Misawa (BNL), Harvey Newman (Caltech), Vladimir Sapunenko (INFN), Loïc Tortay (IN2P3)

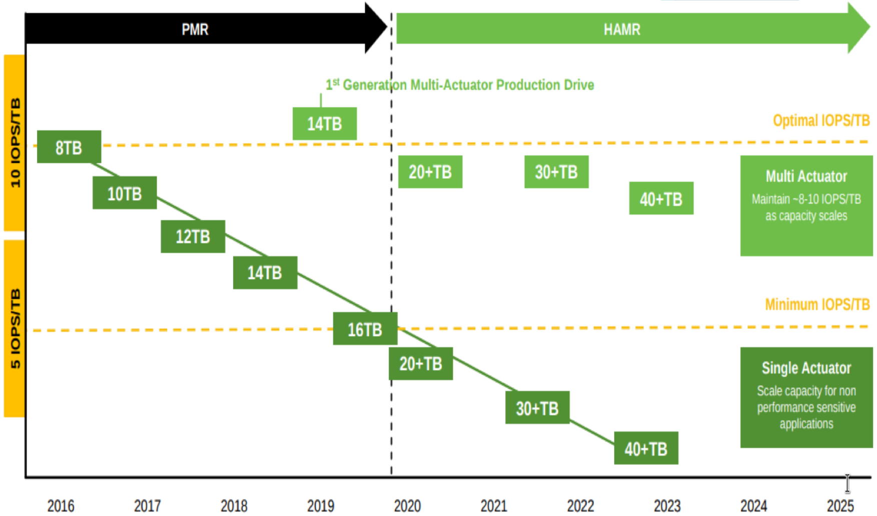
Overview

Update/outlook on technology, market and HEP impact + relevance:

- Hard Disks
- Solid-State Storage
- Tape
- Optical / Others

Hard Disks (I) Technology

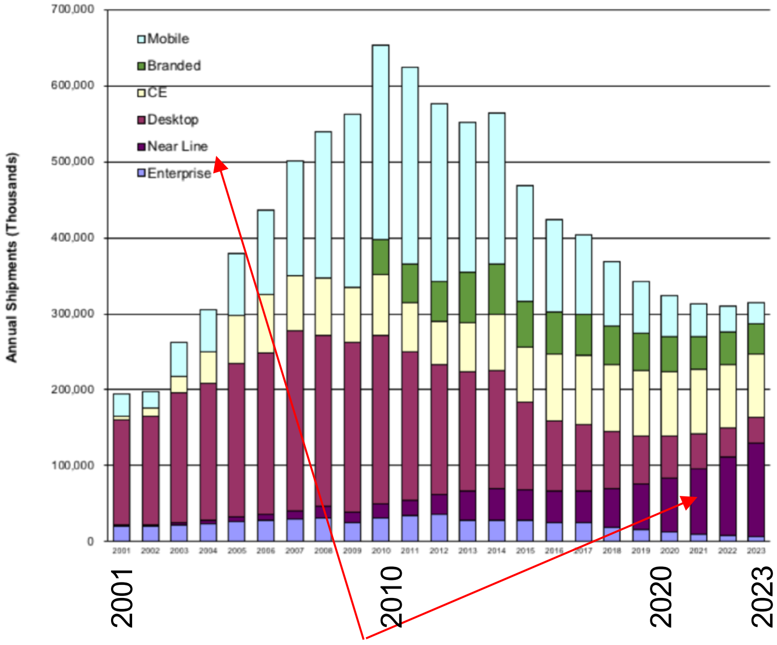
- Problems with existing HDD technology
  - Perpendicular magnetic recording at areal density limit
  - IOPS per TB continues to fall
  - Bandwidth not keeping up with drive capacity



- New technology expected soon
  - Energy Assisted Magnetic Recording (MAMR & HAMR) should allow 40TB drives by 2025
    - 2019 - 16TB MAMR drives from WD and HAMR drives from Seagate
  - Dual Actuator drives double disk IOPS and drive bandwidth
    - 2019 - 14TB Dual actuator drives from Seagate

Hard Disks (II) Market

- Total HDD Market is shrinking
- Sole growth market is in near-line (capacity) HDD used by Cloud and HEP.
- Shrinking market introduces risks
  - Higher costs -> Reduced economies of scale
  - Production risk -> Fewer factories
  - Technology risk -> Insufficient revenue to finance continued R&D
- Narrowing HDD/SSD price gap causing more HDD to be replaced by SSD (eg HSM buffers)

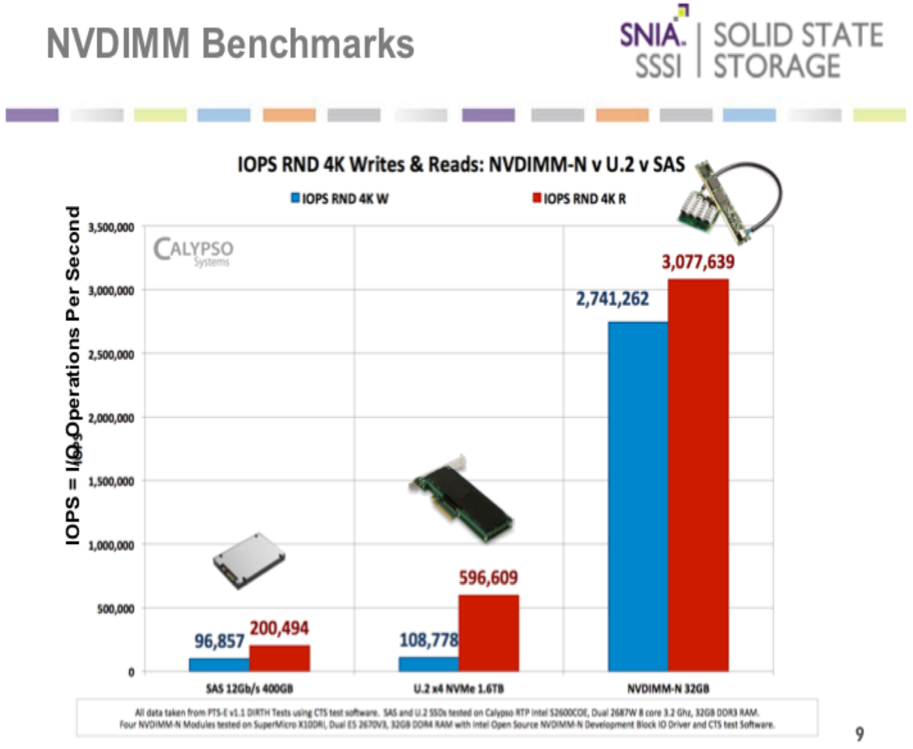


HDD type used for Cloud + HEP

Solid-State Storage (I) - Current State

3D NAND Flash

- Continues to be the underlying non-volatile memory technology of choice
- Capacity increasing through additional layers (96 layers now, roadmaps out to a few hundred layers, >200 cost incr.)
- Additional capacity also obtained by increasing bits per memory cell
 SLC (1 bits) -> MLC (2 bits) -> TLC (3 bits) -> QLC (4 bits) > ???
- Challenge - endurance and retention - increased ECC overhead



NVDIMM Benchmarks

IOPS RND 4K Writes & Reads: NVDIMM-N v U.2 v SAS

All data taken from PTS-E v1.1 DIRTH Tests using CTS test software. SAS and U.2 SSDs tested on Calypso RTP Intel S2600COE, Dual 2687W 8 core 3.2 Ghz, 32GB DDR3 RAM.
Four NVDIMM-N Modules tested on SuperMicro X10DRI, Dual E5 2670V3, 32GB DDR4 RAM with Intel Open Source NVDIMM-N Development Block IO Driver and CTS test Software.

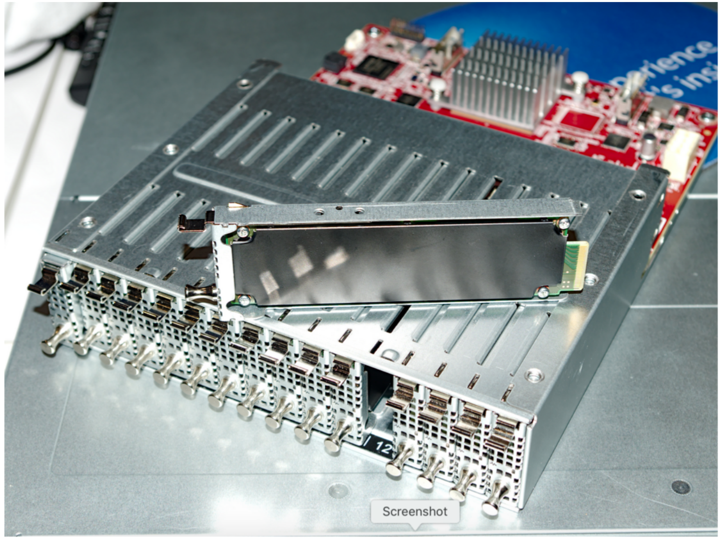
Solid-State Storage (II) - Current State

Solid State Disks (SSD)

- SAS/SATA software stack and hardware severely limits IOPS and rates (i.e. incl. Linux SCSI driver - sd)
- **NVMe** + **PCI-e** new **interconnect** and **protocol** replacing SATA/SAS, significantly alleviating I/O bottleneck
- NVDIMM (persistent memory) technology effectively eliminates all I/O bottlenecks (connected to memory bus)
- New form factors **EDSFF** (“ruler”) and U.2 in addition to existing 2.5” HDD, M.2, and PCI-e card form factors,

Flash Systems

- All Flash Arrays (AFA) common (analogous to HDD HW RAID arrays, but all flash instead of disk)
- Mostly SAS attached, although other interconnects also available. NVMe-oF will come to replace SAS.

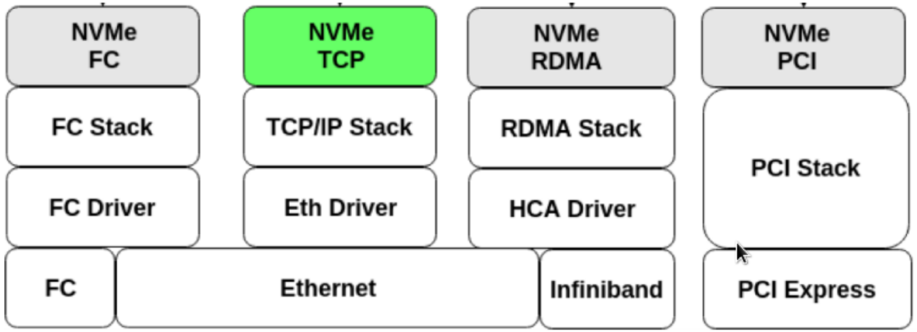


36 cards (“disks”) in 1U (19”)

Solid-State Storage - NVMe/NVMe-oF

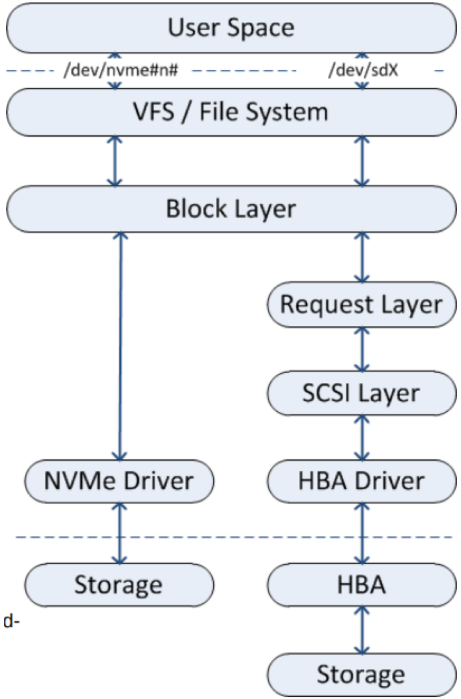
NVMe and NVMe over Fabrics (NVMe-oF) is the center of industry attention and activity

1. NVMe (NVM express) eliminates multiple software layers in the OS stack.
2. NVMe-oF extends NVMe interface to other interconnects (PCI-e, IB, FC, DC Ethernet, …)
3. NVMe being expanded (e.g., enclosure management, multi-path, device management, …)
4. Aiming to be “lingua franca” for high performance solid state storage - unleash SSD potential
5. Allows for more radical solid state storage architectures/systems.



NVMe FC
NVMe TCP
NVMe RDMA
NVMe PCI
FC Stack
TCP/IP Stack
RDMA Stack
PCI Stack
FC Driver
Eth Driver
HCA Driver
FC
Ethernet
Infiniband
PCI Express

Source: Kam Eshghi (Lightbits Labs)



User Space
/dev/nvme#n#
/dev/sdX
VFS / File System
Block Layer
Request Layer
SCSI Layer
NVMe Driver
HBA Driver
Storage
HBA
Storage

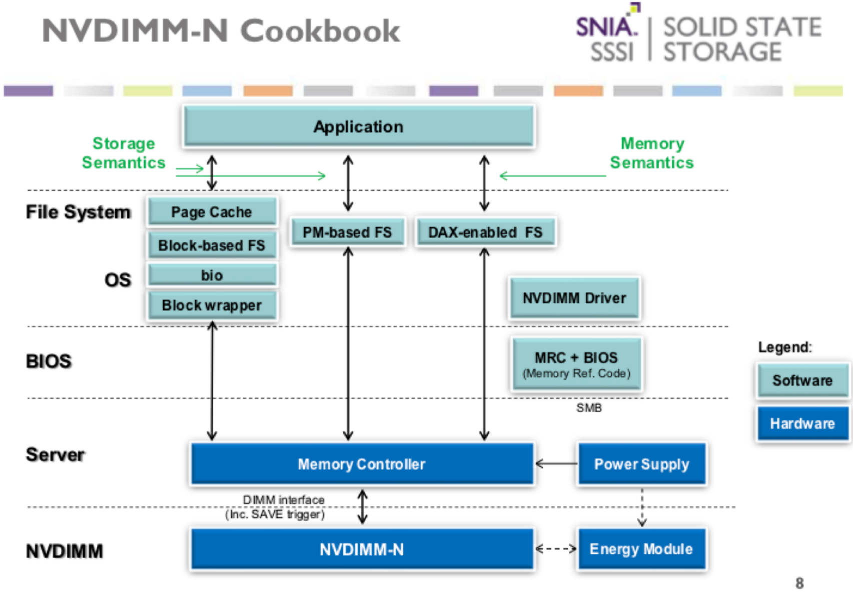
Source: K. Bush, Intel, 2014 Flash Memory Summit

Solid-State Storage - NVDIMM

Persistent Memory (NVDIMM)

- Non volatile memory on the CPU memory bus (DDR4/DDR5)
- DDR4/DDR5 DIMM physical form factor
- CPU and system support required.
- Higher density and lower costs than DRAM is expected
- DRAM-like access latencies are expected (~ x3 higher)
- Designed for “ultimate” I/O performance.
- Programming models and usage are hot-topics in academia and industry - high market expectations
- Enables new computing models which could benefit caching, (DAQ, burst buffer, in-memory DB,
- 3D NAND flash (with hybrid DRAM) and 3D XPoint memory expected to be memory technology of choice for NVDIMM.

Hardware and software components needed for Persistent Memory are becoming available.
NVDIMMoF in discussion (what we had with SGIs NUMA machines years ago)

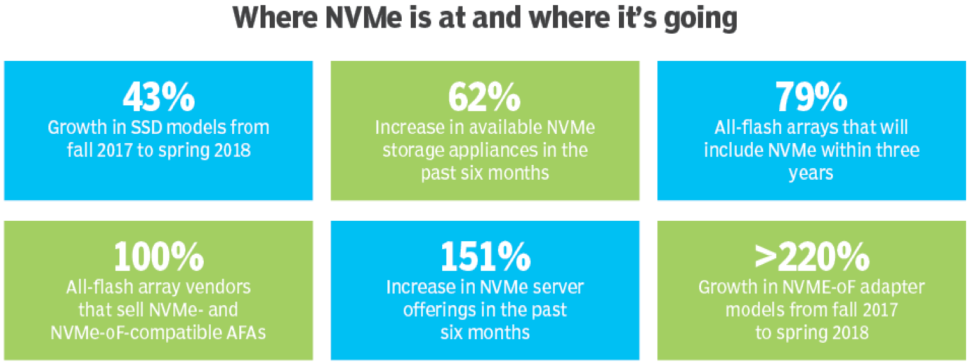


Solid-State Storage - the market

- 3D NAND Flash is the dominant underlying memory technology
- 3D XPoint memory aims to be a major player in the NVDIMM market - (significant delays in productization (i.e. HPs Memristor several years ago, 3D XPoint - aim GA 2017 / legal/IP issues, huge invests required)
- Market in 2018 - ...flash vendors overall revenue decline, bit shipments saw a 40% increase for 2018.
- NAND Flash market will remain in oversupply 2019 - slowing down capacity expansion

Amazon,com: Intel 660p M.2 2280 **2 TB** NVMe PCIe 3.0 x4 3D NAND - $194.

“NVM EXPRESS ECOSYSTEM” G2M Research, May 2018



Where NVMe is at and where it’s going

43% Growth in SSD models from fall 2017 to spring 2018	**62%** Increase in available NVMe storage appliances in the past six months	**79%** All-flash arrays that will include NVMe within three years
100% All-flash array vendors that sell NVMe- and NVMe-oF-compatible AFAs	**151%** Increase in NVMe server offerings in the past six months	**>220%** Growth in NVME-oF adapter models from fall 2017 to spring 2018

Solid-State Storage - the market (II)

NVMe/NVMe-oF

- NVMe SSDs
  - The market expected to grow from ~$2B in 2017 to $9 billion in 2022.
- All-flash Arrays (AFAs)
  - More than 60% of AFAs will utilize NVMe storage media by 2022, and more than 30% of all AFAs will use NVMe over Fabric (NVMe-oF) in either front-side (connection to host) and/or back-side (connection to expansion shelves) fabrics in the same timeframe.
- NVMe Storage Appliances
  - In their bid to match the performance of AFAs, nearly 80% of storage appliances will provide storage bays for removable NVMe drives (either U.2 or ESDFF) by 2022.
- NVMe-oF adapters
  - 1.75 million units sold by 2022, with the bulk of the adapters being smart (NVMe-oF offload)

Tape: Tech status and outlook

- Enterprise: TS1160 released in Q4 2018
  - 400MB/s (+11% over previous gen), 20TB on new JE media (+33% over previous gen)
- LTO: Gen-8 released in Q4 2017
  - 360MB/s (+20% over LTO-7), 12TB (+100% over previous gen).
  - Can use LTO-7 media @ 9TB instead of 6TB, but LTO-9 support for that format unclear
  - LTO-9 expected by EOY 2020. Roadmap claims 24TB but more realistically ~18-20TB
- Libraries (>10K slots): IBM, Oracle, Spectralogic and Quantum
  - Concerns about Quantum and Oracle future strategy (and support model/pricing for the latter)
- R&D
  - 123Gb/sqin on BaFe (~220TB tape) and 201Gb/sqin (330TB, CoPtCr media, but material/manufacturing costs expected to be high). Allows for 8-9 years of headroom at 30% density CAGR
  - Streaming read/write performance predicted to be 15-20% CAGR. But no significant improvements expected on seek times (random access) as limited by mechanics

Tape: Market status and outlook

Tape drives: LTO and IBM Enterprise only remainers after Oracle's retirement. LTO drive technology, R&D and manufacturing seemingly dominated by IBM. Will the two product lines be merged eventually?

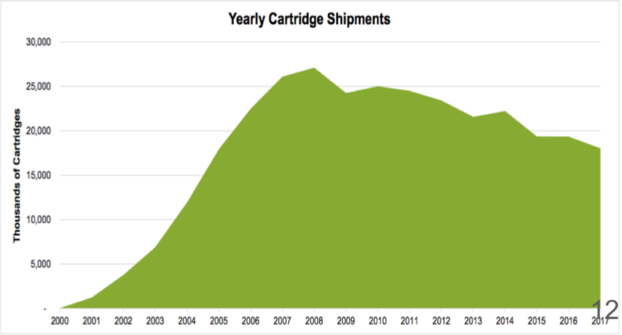
Tape media: Fujifilm entangled in ongoing patent war with Sony wrt LTO-8 media. LTO-7 media available at interesting prices (~5-6CHF/TB), not the case for LTO-8

Market continues a ~10-year sustained contraction

Revenues estimated $0.7B

-> HEP risk factor:
limited competition, contracting market

Unit Shipments: Calendar Year



Optical and others

- Archival Disk: 300GB capacity
  - Manufactured by **Sony** and **Panasonic**, agreement with Mitsubishi
  - 140MB/s write, 280MB/s read
  - Bundled by Sony in **cartridges** with **11 disks** (ODA - Optical Disk Archive)
  - Roadmap to 1TB announced in 2015. According to company sources, testing for 500GB media is ongoing but no release date yet
  - No (public?) prices available
- Large-scale libraries from Panasonic (Freeze-Ray) and Sony
  - Sony **Everspan** (up to 64 drives and 180PB) has been **retired** shortly after announcement. New library in plans (later 2019?) in collaboration with Qualstar (up to 47PB)
- Difficult to find information about any existing large-scale customers
  - Seemingly none in HEP
- Other archival technologies (holographic, nanophotonic,DNA?) are very far from production

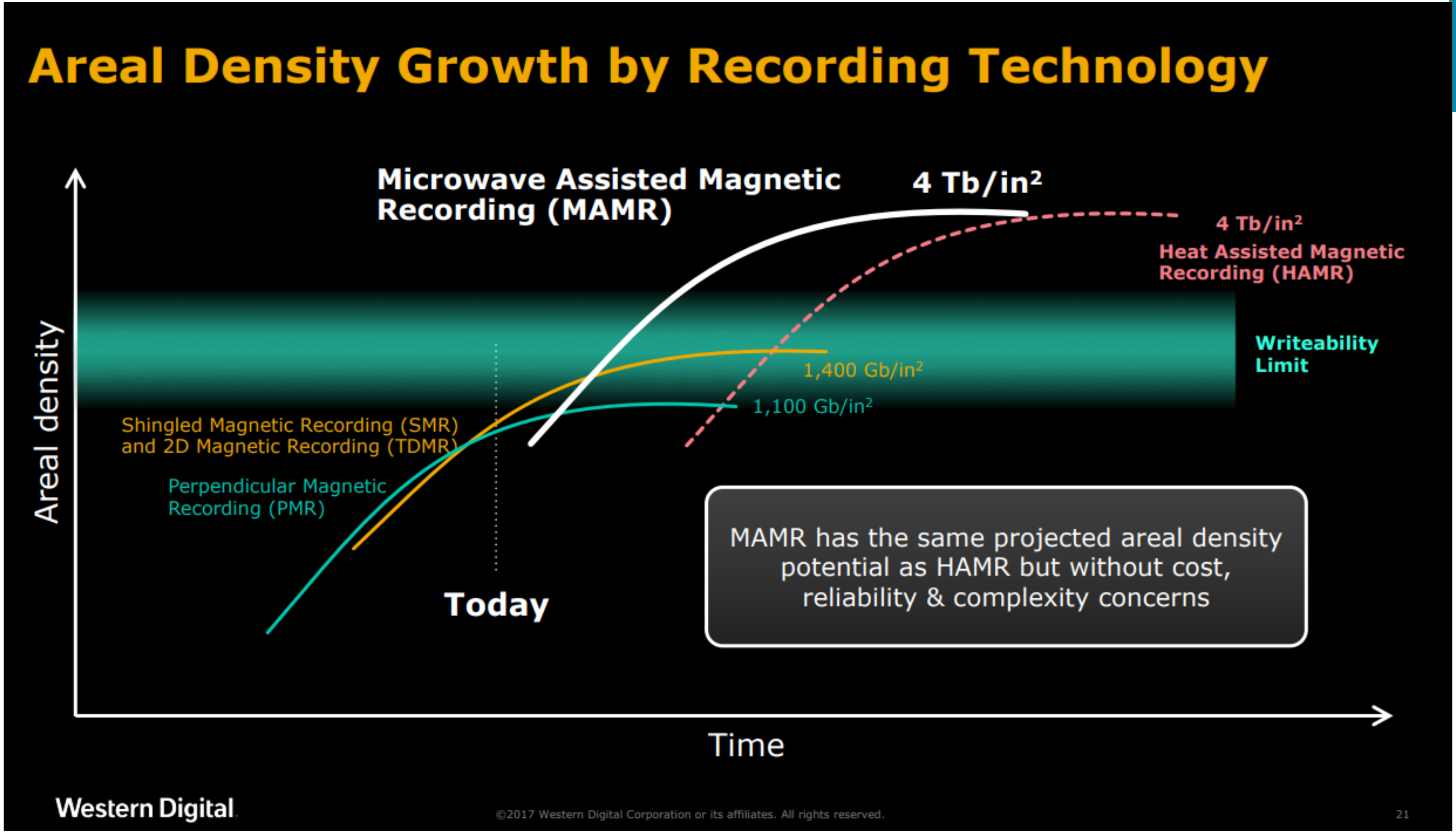
Summary

- Use of solid state storage is expanding beyond SATA SSD.
  - Flash replacing HDD in more domains as $/GB differential continues to narrow
  - NVMe/PCI-e enables orders of magnitude decrease in access latency and increase in BW
  - NVMe-oF is fostering a renaissance in flash storage system architectures
  - NVDIMM is enabling radical restructuring of applications.
- HDD vendors, with the release of H/MAMR drives, should be able to protect their position in near-line storage for a few years. However, overall revenue will continue to decline as other HDD markets continue to collapse.
- Tape still key component for large-scale archival, although risks increasing due to shrinking market and lack of competition.
- Optical is still (yet?) not reviving, other technologies a long way ahead

References

- Storage Technology perspectives, WS CCR 2018
- HEPiX Techwatch Storage WG document: link

Backup slides

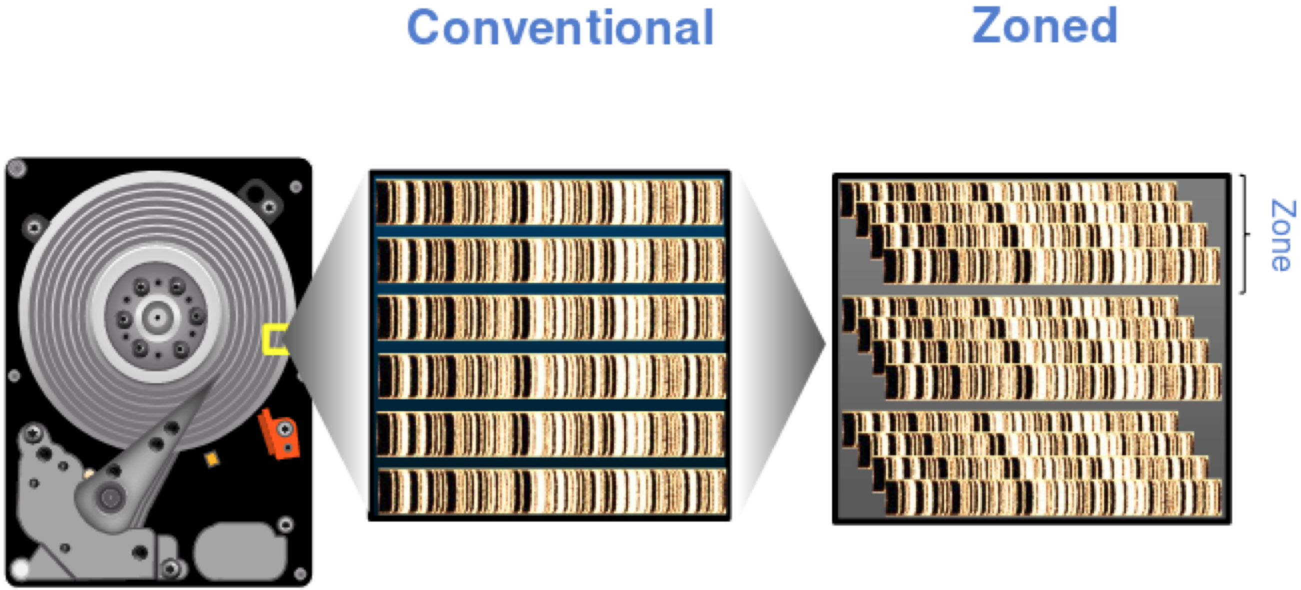


Areal Density Growth by Recording Technology

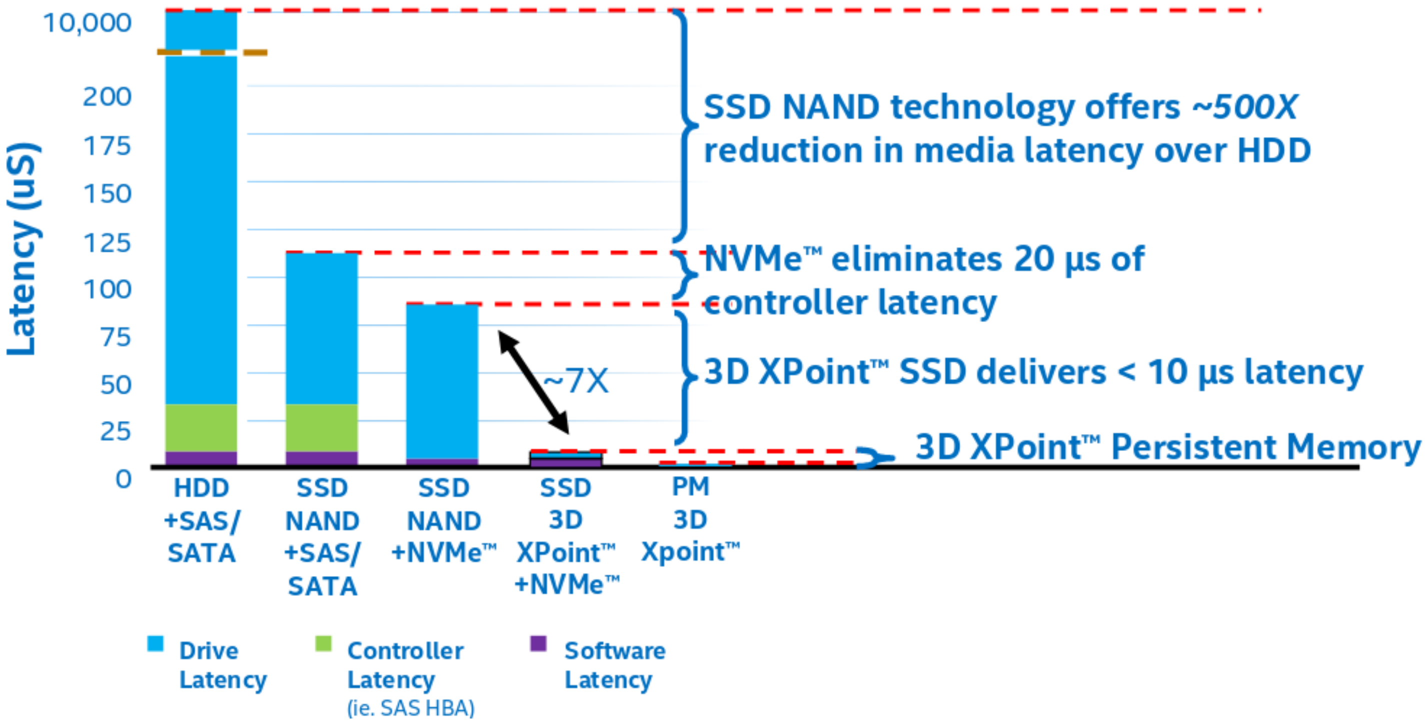
MAMR has the same projected areal density potential as HAMR but without cost, reliability & complexity concerns

Western Digital

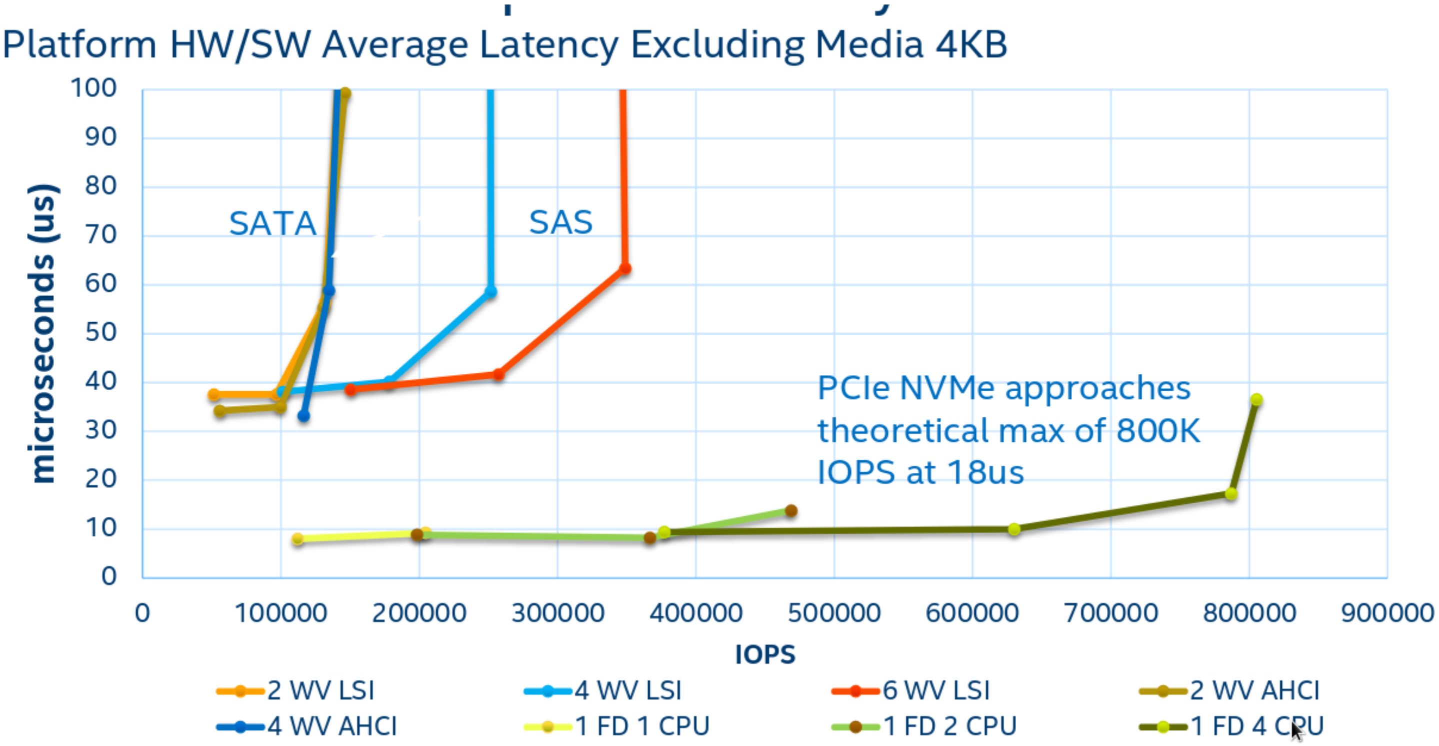
SMR zoned



Latency comparation (including media)



Latency comparation (excluding media)



Platform HW/SW Average Latency Excluding Media 4KB

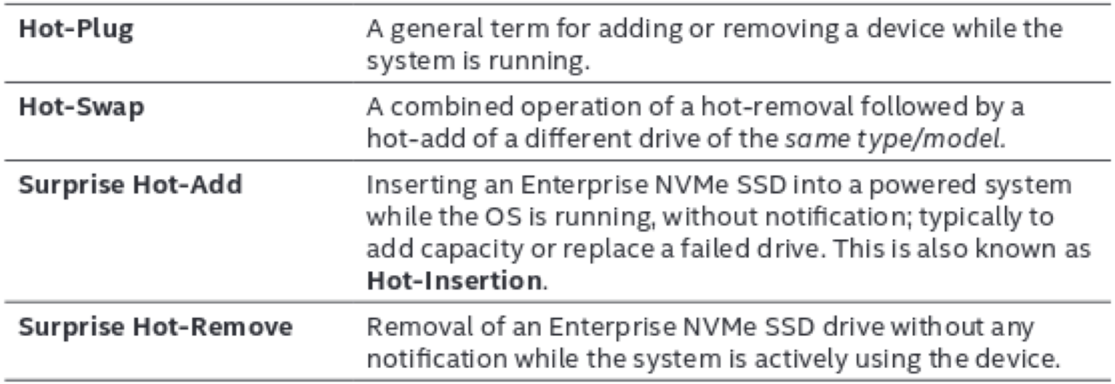
SATA

SAS

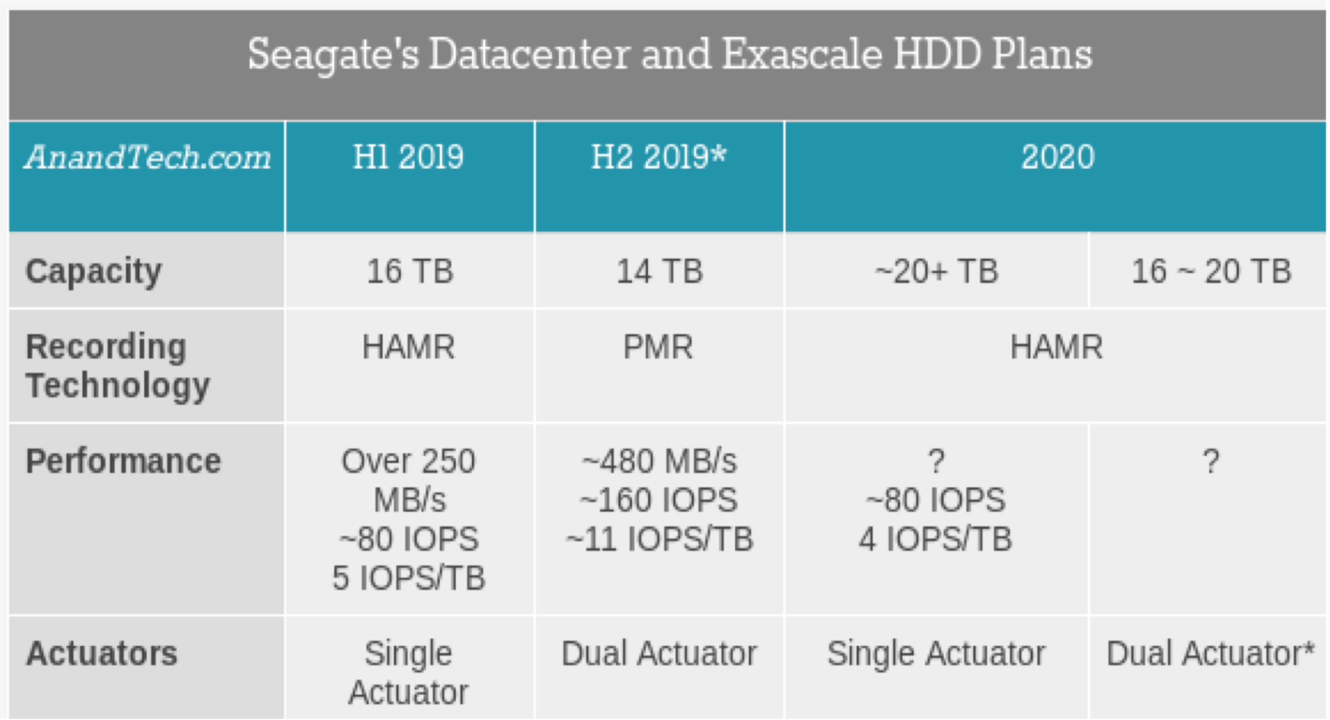
PCIe NVMe approaches theoretical max of 800K IOPS at 18us

Hot-Plug support of Intel NVMe SSDs

White Paper: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/333596-hot-plug-capability-nvme-ssds-paper.pdf



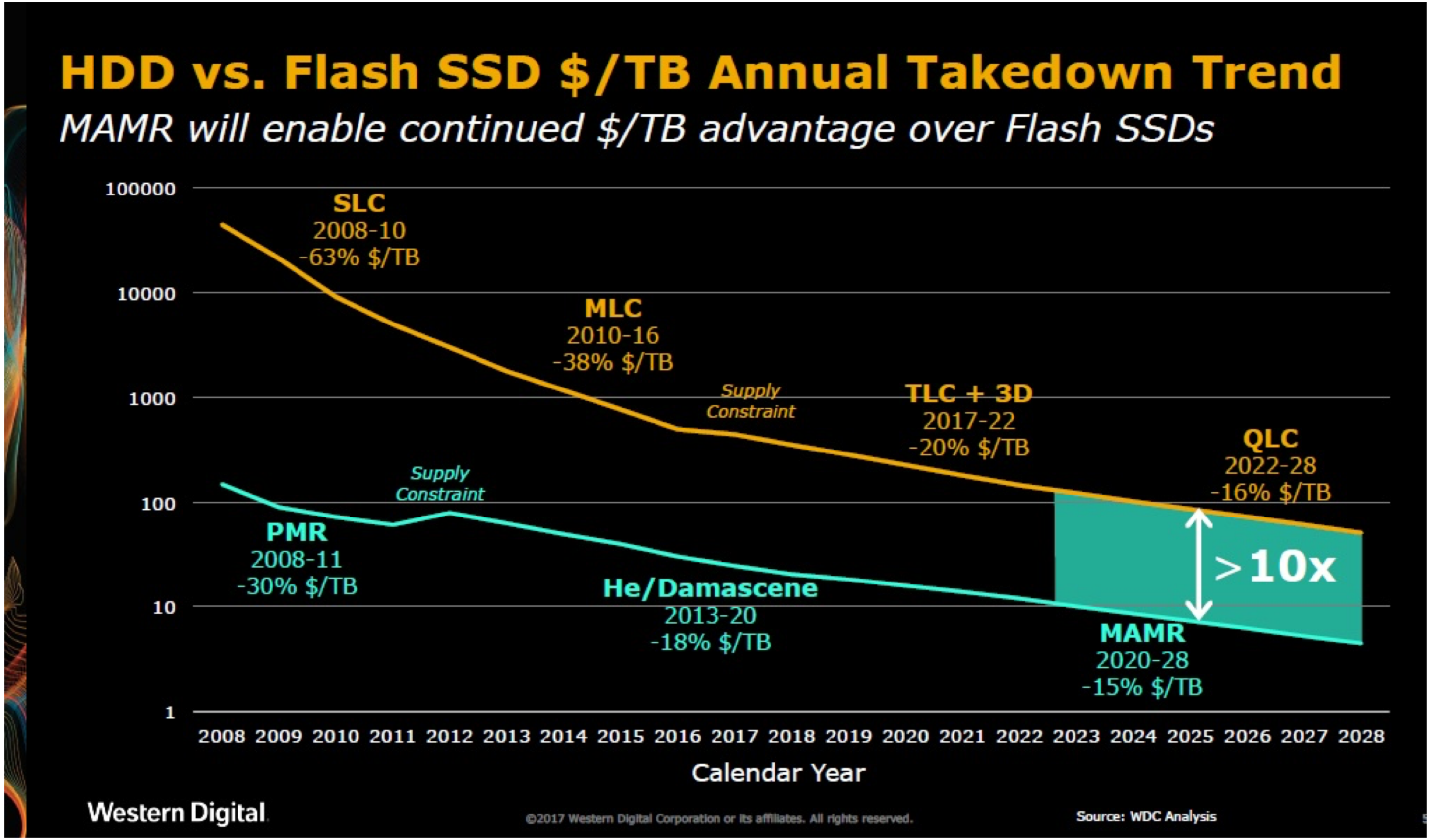
Hot-Plug	A general term for adding or removing a device while the system is running.
Hot-Swap	A combined operation of a hot-removal followed by a hot-add of a different drive of the *same type/model*.
Surprise Hot-Add	Inserting an Enterprise NVMe SSD into a powered system while the OS is running, without notification; typically to add capacity or replace a failed drive. This is also known as **Hot-Insertion**.
Surprise Hot-Remove	Removal of an Enterprise NVMe SSD drive without any notification while the system is actively using the device.



Seagate's Datacenter and Exascale HDD Plans

AnandTech.com	H1 2019	H2 2019*	2020	
Capacity	16 TB	14 TB	~20+ TB	16 ~ 20 TB
Recording Technology	HAMR	PMR	HAMR	
Performance	Over 250 MB/s ~80 IOPS 5 IOPS/TB	~480 MB/s ~160 IOPS ~11 IOPS/TB	? ~80 IOPS 4 IOPS/TB	?
Actuators	Single Actuator	Dual Actuator	Single Actuator	Dual Actuator*

Seagate says it has manufactured more than 40,000 HAMR hard drives and started shipping out test batches to storage vendors last year.



HDD vs. Flash SSD $/TB Annual Takedown Trend

MAMR will enable continued $/TB advantage over Flash SSDs

Western Digital

Source: WDC Analysis