Storage technology perspectives

Vladimir Sapunenko,
Head of Data Management and Storage group,
INFN-CNAF

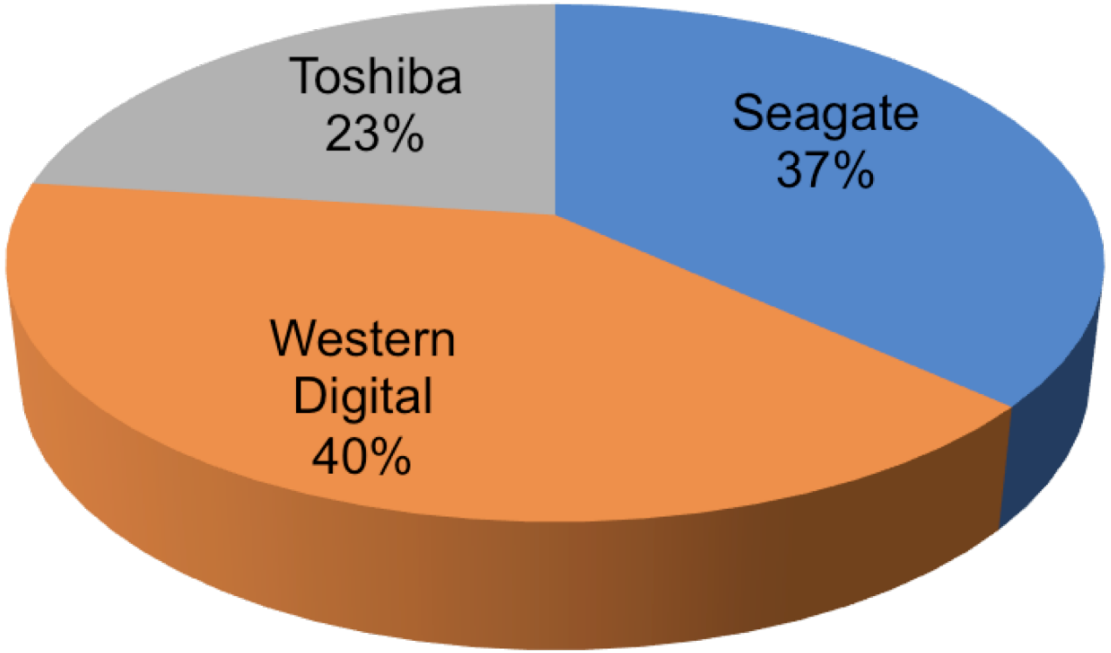
Agenda

- HDD market
- HDD technology overview
- Common trends
- Data protection
- Tape technology overview

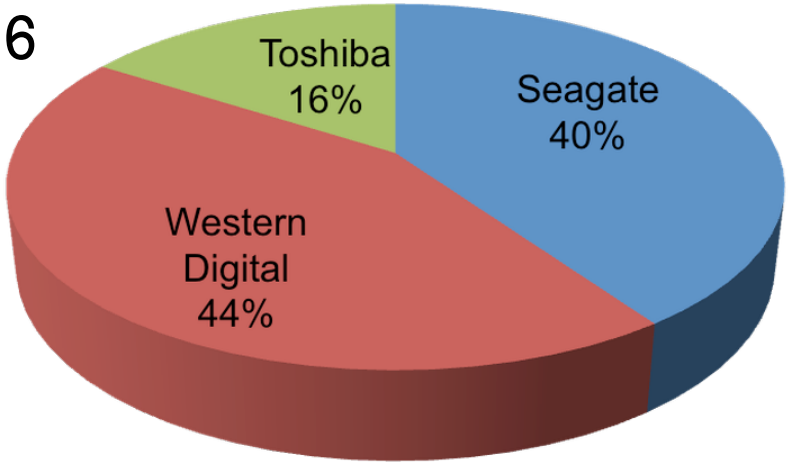
HDD unit shipment: 2018 market share

- WD – 40%
- Seagate – 37%
- Toshiba – 23% (mainly 2.5” drives)

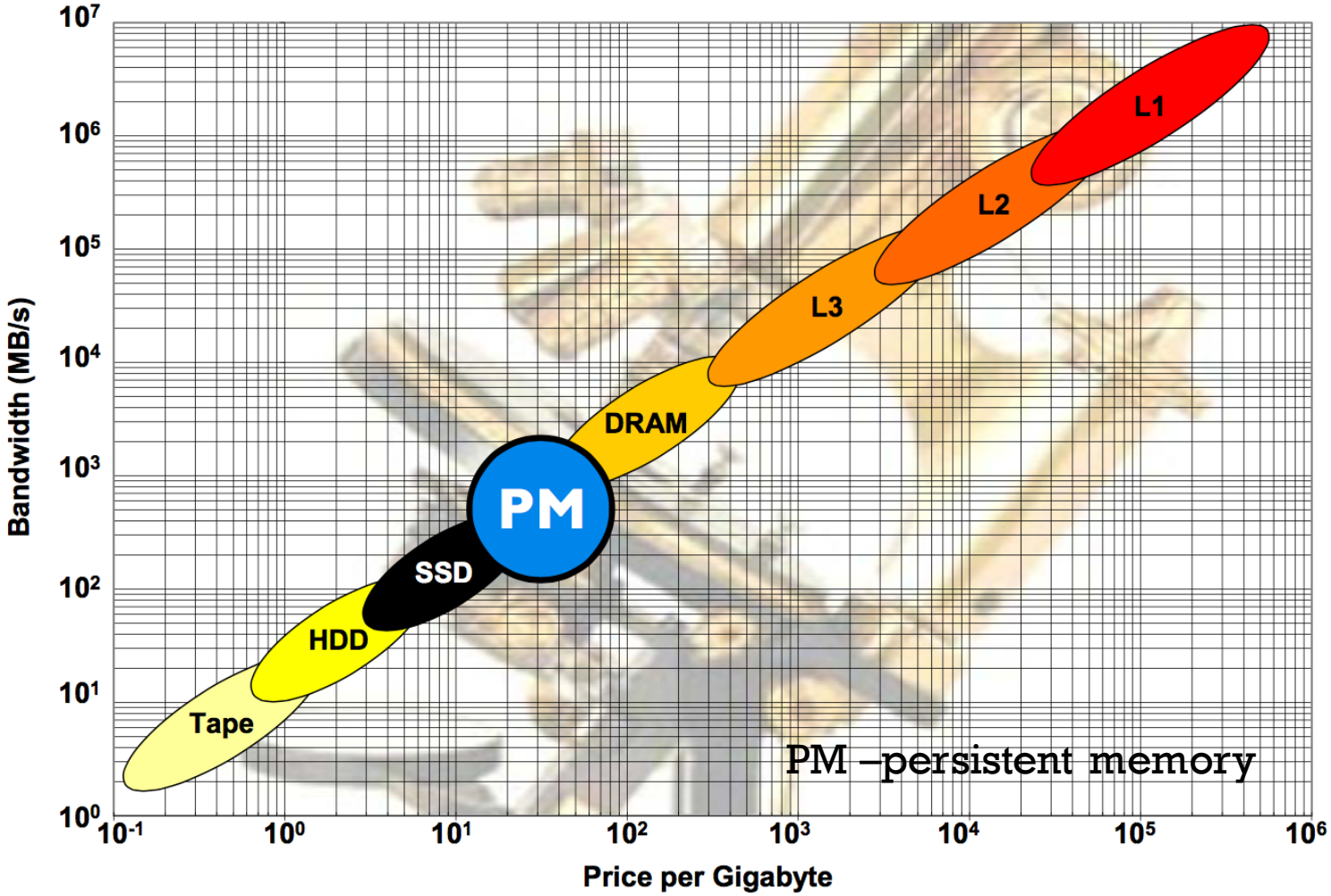
2018



2016



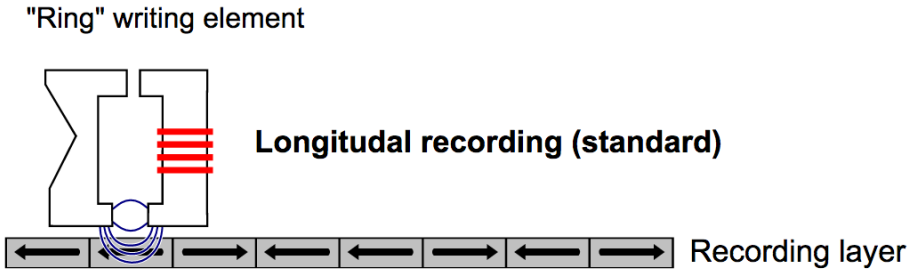
Memory price per GB

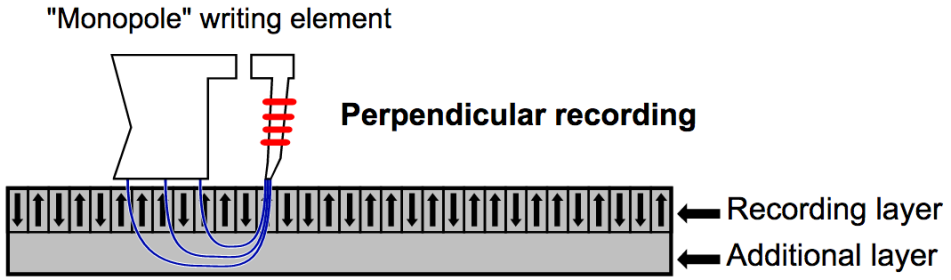


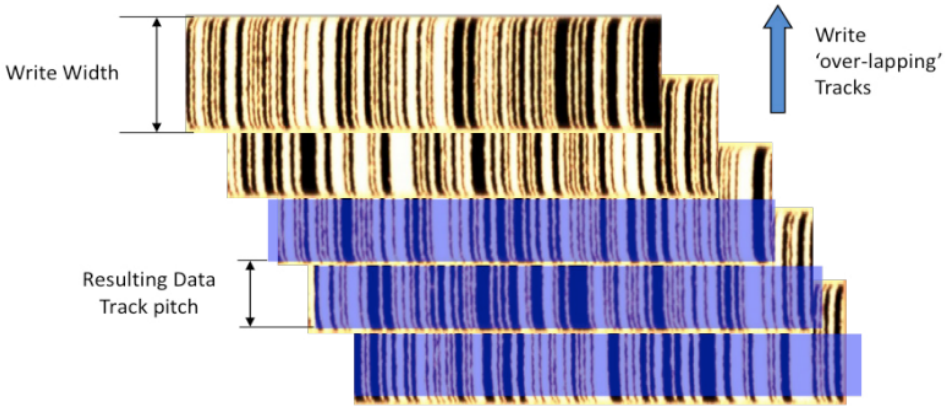
Source: ***A Close Look at the Intel/Micron 3D XPoint Memory***, Objective Analysis 2015

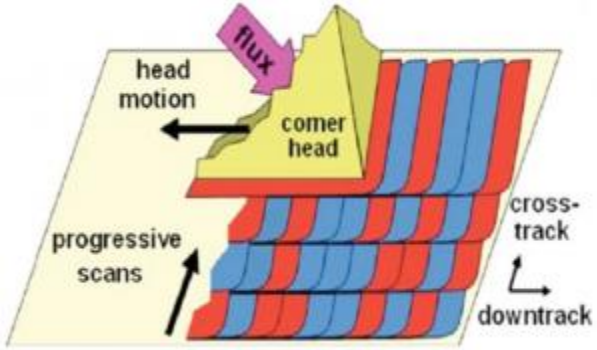
Current HDD technologies

- PMR (Perpendicular Magnetic Recording) – in use from 2005
- SMR (Shingled Magnetic Recording) - Overlapping /overwriting of tracks





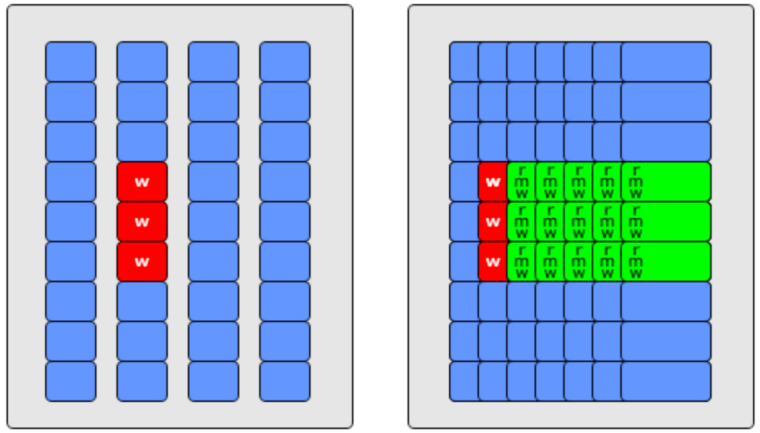




Classic HDD writes vs. read-modify-write on SMR disks

Regular hard drive:

- Wait for platter to rotate and seek head to first target sector in track
- Write three sectors in direct succession

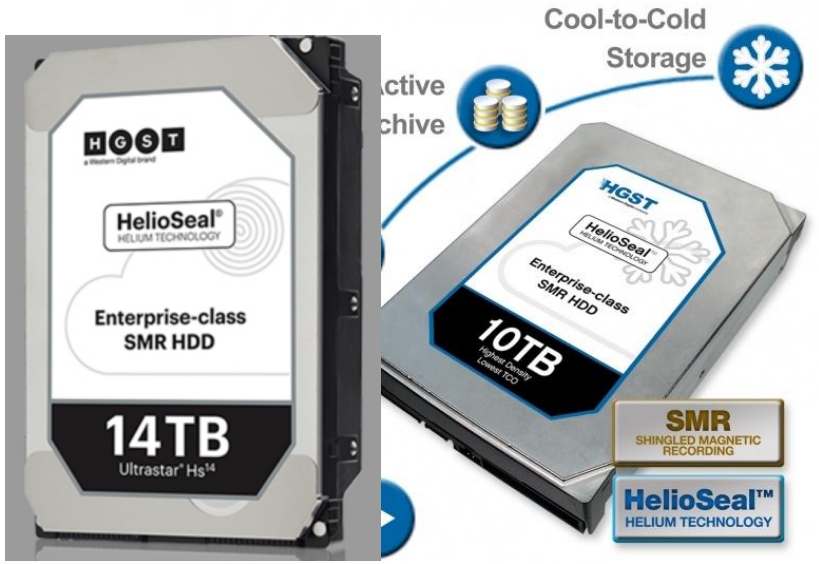


SMR hard drive:

- Wait for platter to rotate and seek head to target track + 1
- Read three sectors in direct succession, store in cache
- Wait for platter to rotate and seek head to target track + 2
- Read three sectors in direct succession, store in cache
- Wait for platter to rotate and seek head to target track + n
- Read three sectors in direct succession, store in cache
- (Repeat until we hit end of medium* or band)
- Seek head to target track
- Write original three sectors
- Wait for platter to rotate and seek head to target track + 1
- Rewrite three previously stored sectors, recalled from cache
- Wait for platter to rotate and seek head to target track + 2
- Rewrite three previously stored sectors, recalled from cache
- Wait for platter to rotate and seek head to target track + n
- Rewrite three previously stored sectors, recalled from cache
- (Repeat until we hit end of medium* or band)

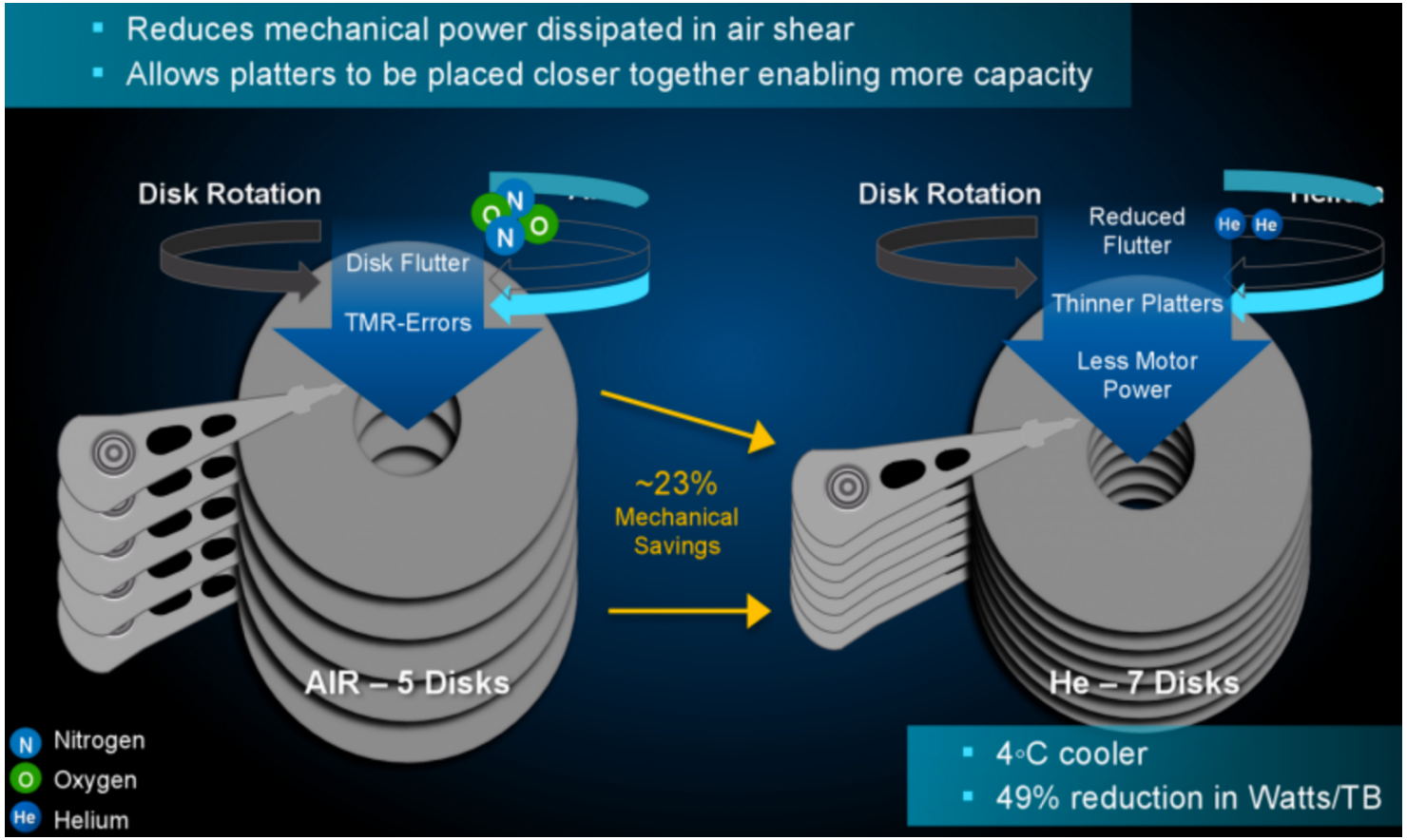
SMR Hard drives
(often marketed as "Archive" drives.)

- Archival HDDs are not designed for high performance like nearline HDDs are, which is clear from the 5,900-RPM spindle speed
- An SMR HDD must constantly shuffle incoming (and existing) data in the background due to the shingled nature of the drive. This limits its performance and consistency in several types of applications
- Traditional software or hardware RAID is simply not recommended due to the sustained write penalty that occurs during rebuild (10MB/s vs. 156MB/s, see http://www.storagereview.com/seagate_archive_hdd_review_8tb)
- Systems that address drives separately (such as object storage, erasure coding and some backup/archival implementations) are well-suited for the Archive HDD



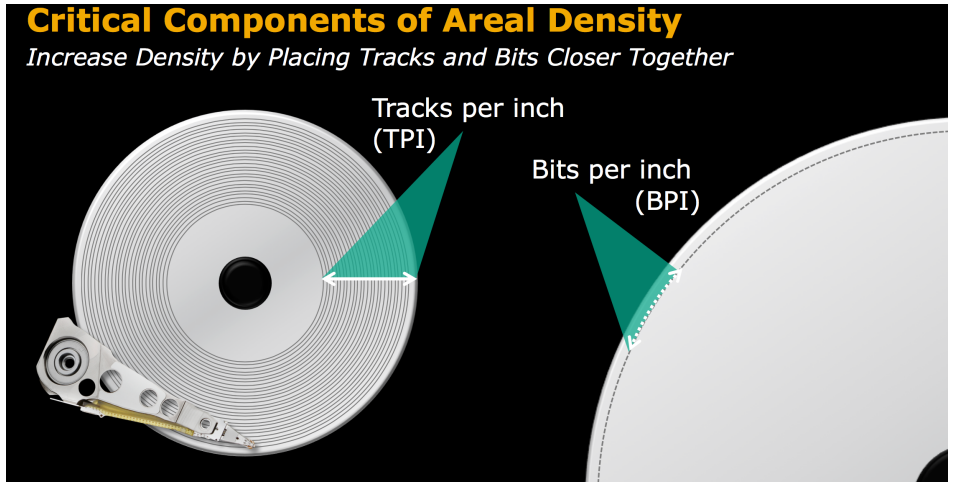
Helium filled HDDs

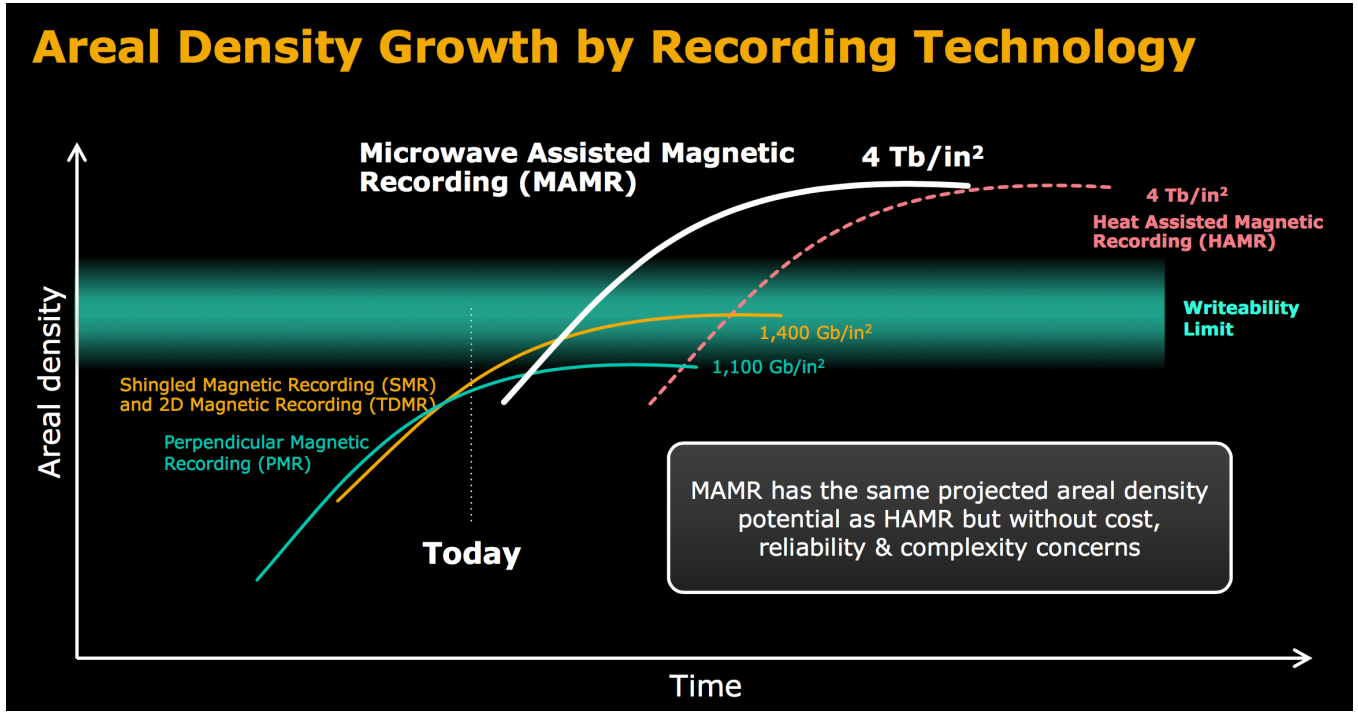
- Helium density 1/7 of air
- By the end of 2017 Seagate, Western Digital and Toshiba were all shipping He-filled drives with Toshiba announcing up to 9 disks in a conventional 3.5-inch HDD with capacity as high as 14 TB without shingled recording.



Scaling beyond PMR requires energy assisted recording

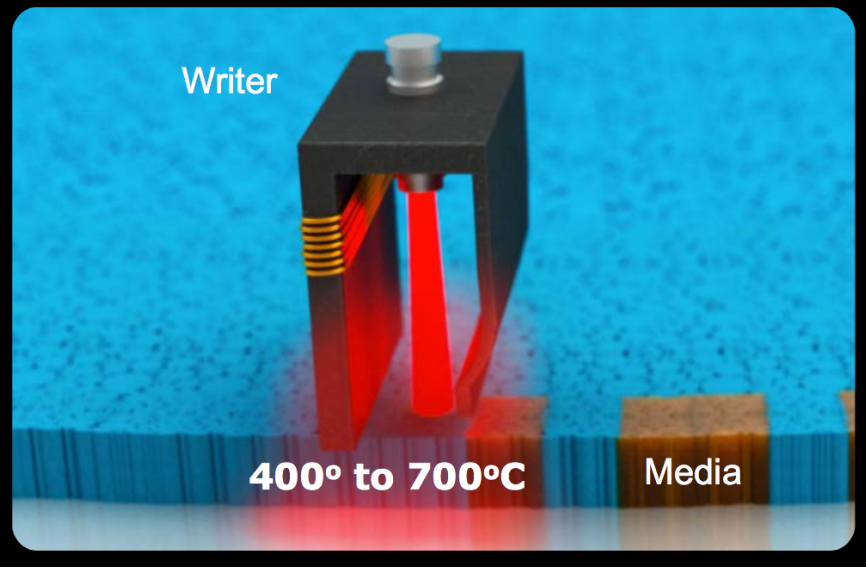
- MAMR (Microwave Assisted Magnetic Recording) – exp. 2019
- HAMR (Heat Assisted Magnetic Recording) – exp. 2019



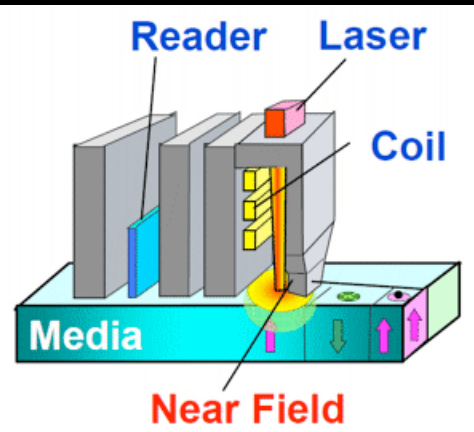


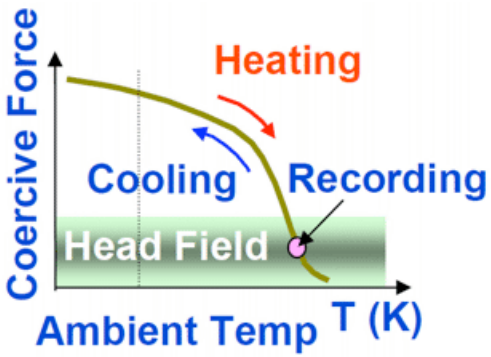
HAMR (Seagate)

How HAMR Works



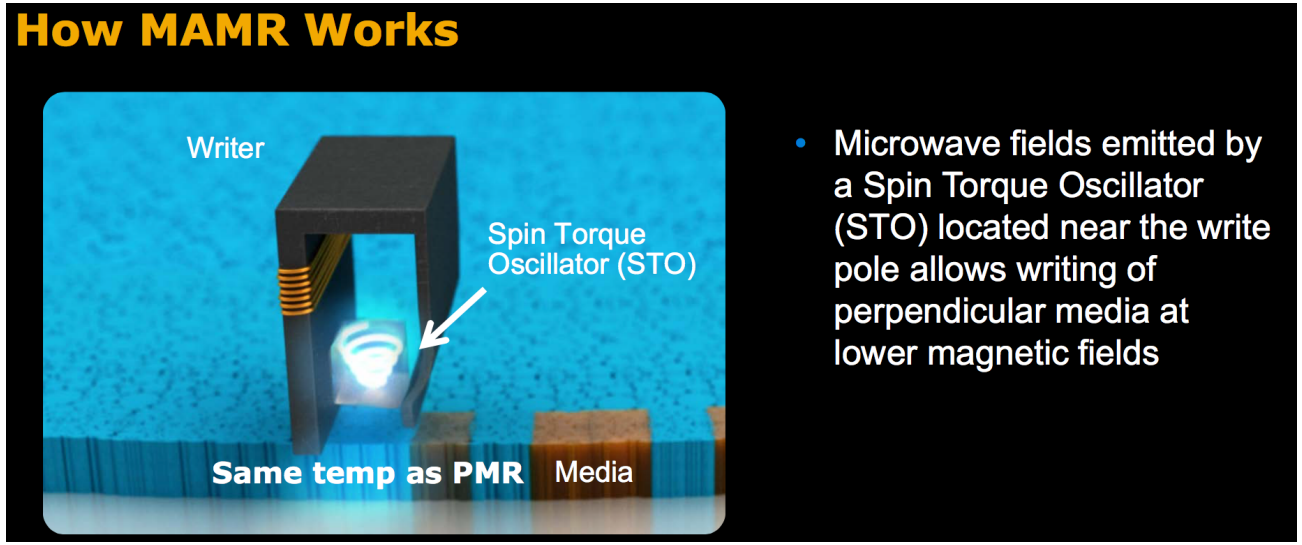
- Heat from laser lowers the energy barrier to write on media and magnets can be switched with smaller magnetic field
- When media cools, the data is harder to erase

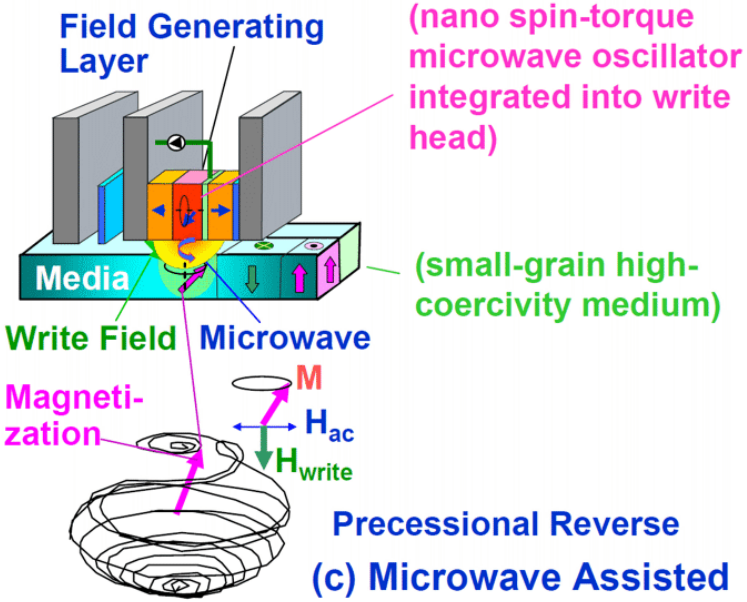


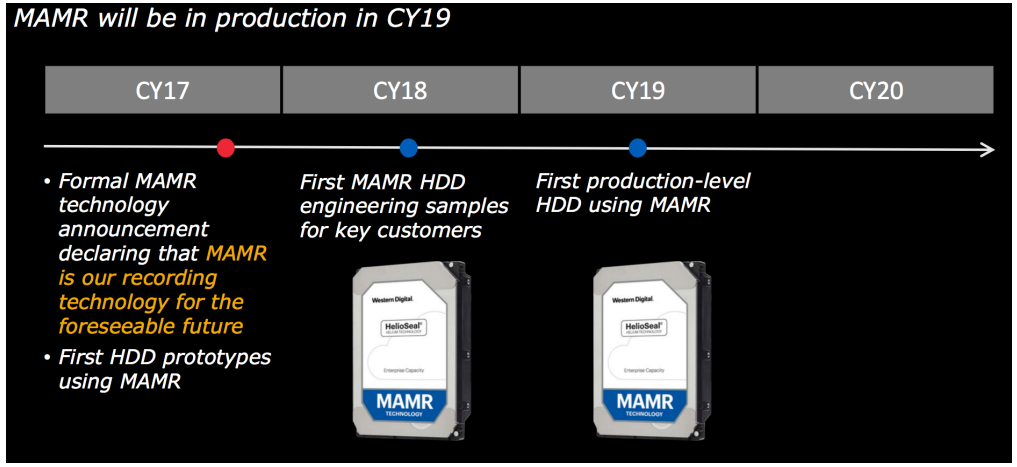


(b) Heat Assisted

MAMR (WD)


How MAMR Works
Writer
Spin Torque Oscillator (STO)
Same temp as PMR
Media
• Microwave fields emitted by a Spin Torque Oscillator (STO) located near the write pole allows writing of perpendicular media at lower magnetic fields


Field Generating Layer
(nano spin-torque microwave oscillator integrated into write head)
Media
(small-grain high-coercivity medium)
Write Field
Microwave
Magneti-zation
M
Hac
Hwrite
Precessional Reverse
(c) Microwave Assisted


MAMR will be in production in CY19
CY17
CY18
CY19
CY20
• Formal MAMR technology announcement declaring that MAMR is our recording technology for the foreseeable future
• First HDD prototypes using MAMR
First MAMR HDD engineering samples for key customers
First production-level HDD using MAMR
MAMR
MAMR

Market availability

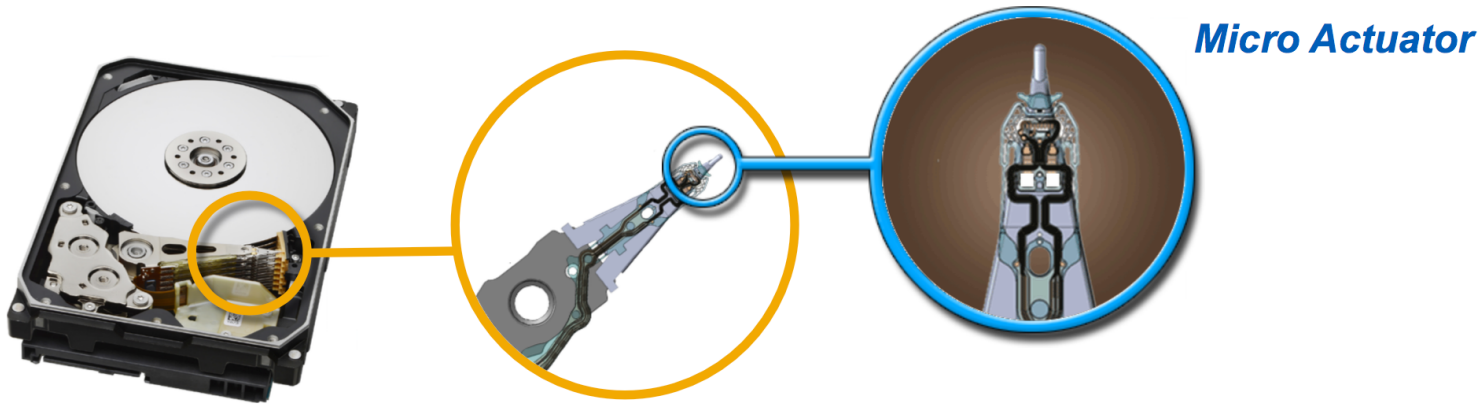
- Both vendors plan to begin volume shipments of their respective technologies in 2019.
- Seagate projects 40TB+ drives by 2023
- WD plans to pass the 40TB threshold in 2025
- WD's MAMR relies largely upon proven technologies
- Seagate claimed that it's already producing the HAMR drives on the same production lines as its existing PMR-based drives

WD: Micro Actuator

Industry's First Micro Actuator for Data Center Drives

Increasing Capacity Through Mechanical Innovation

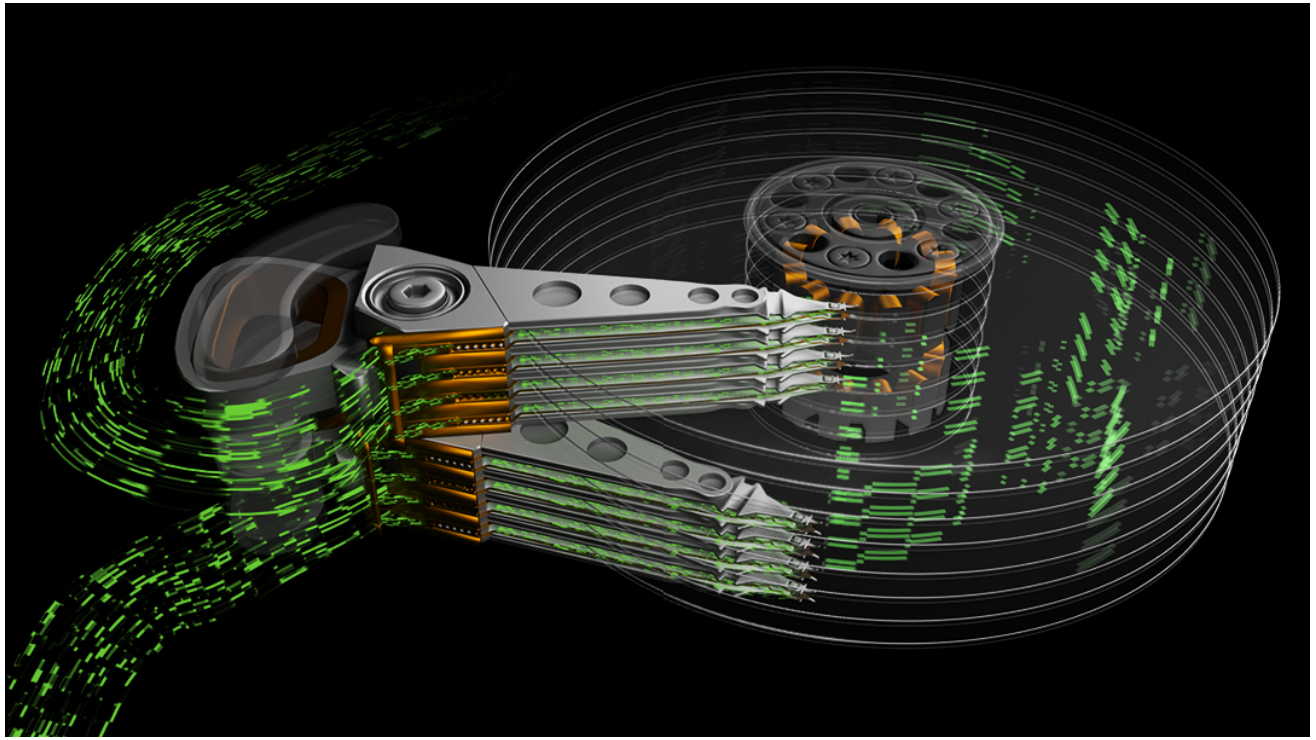
- Finer positioning and control
- Higher servo bandwidth supports enterprise-class performance and vibration specs



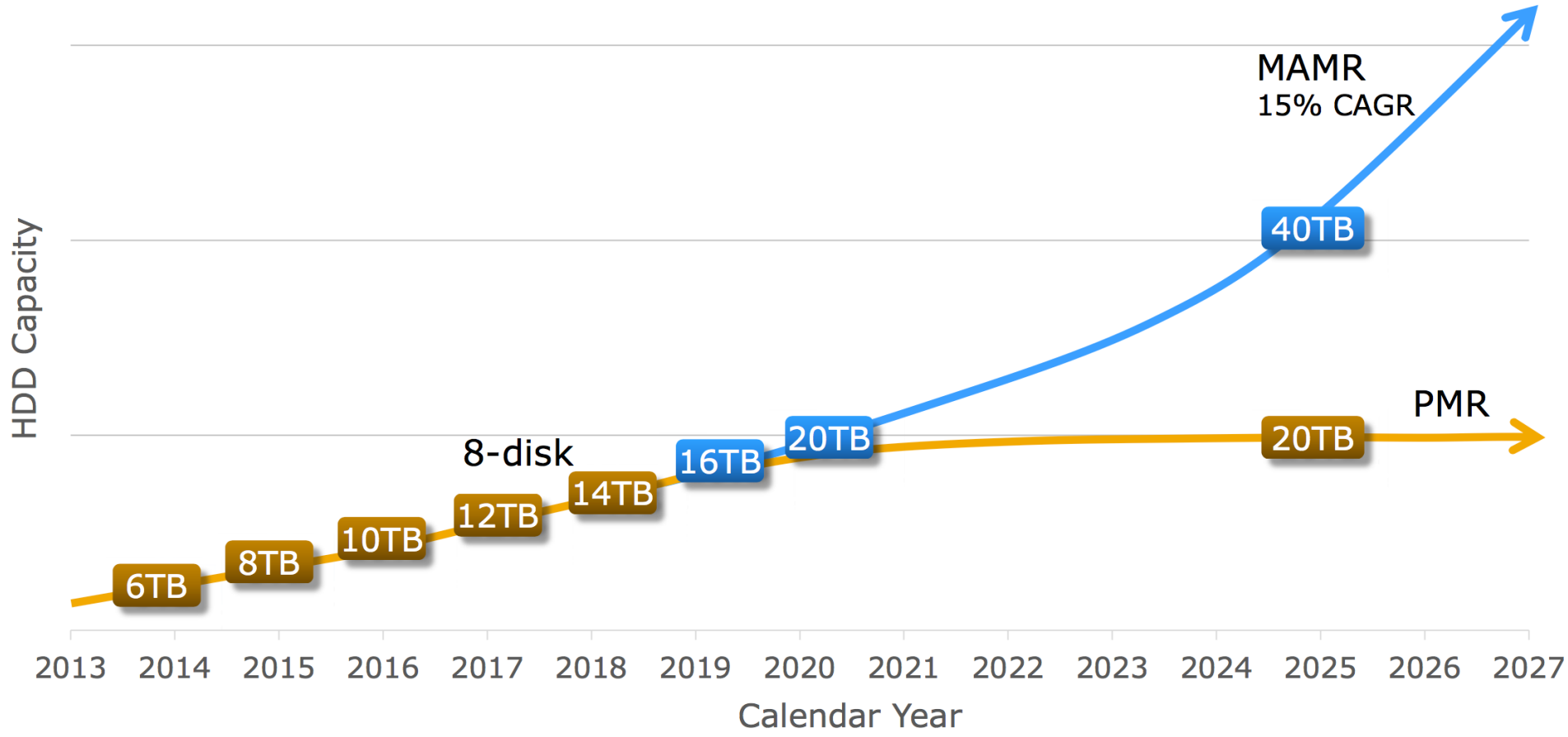
Micro actuation provides finer control and supports higher track density (>400K TPI)

Seagate: Multi Actuator Technology

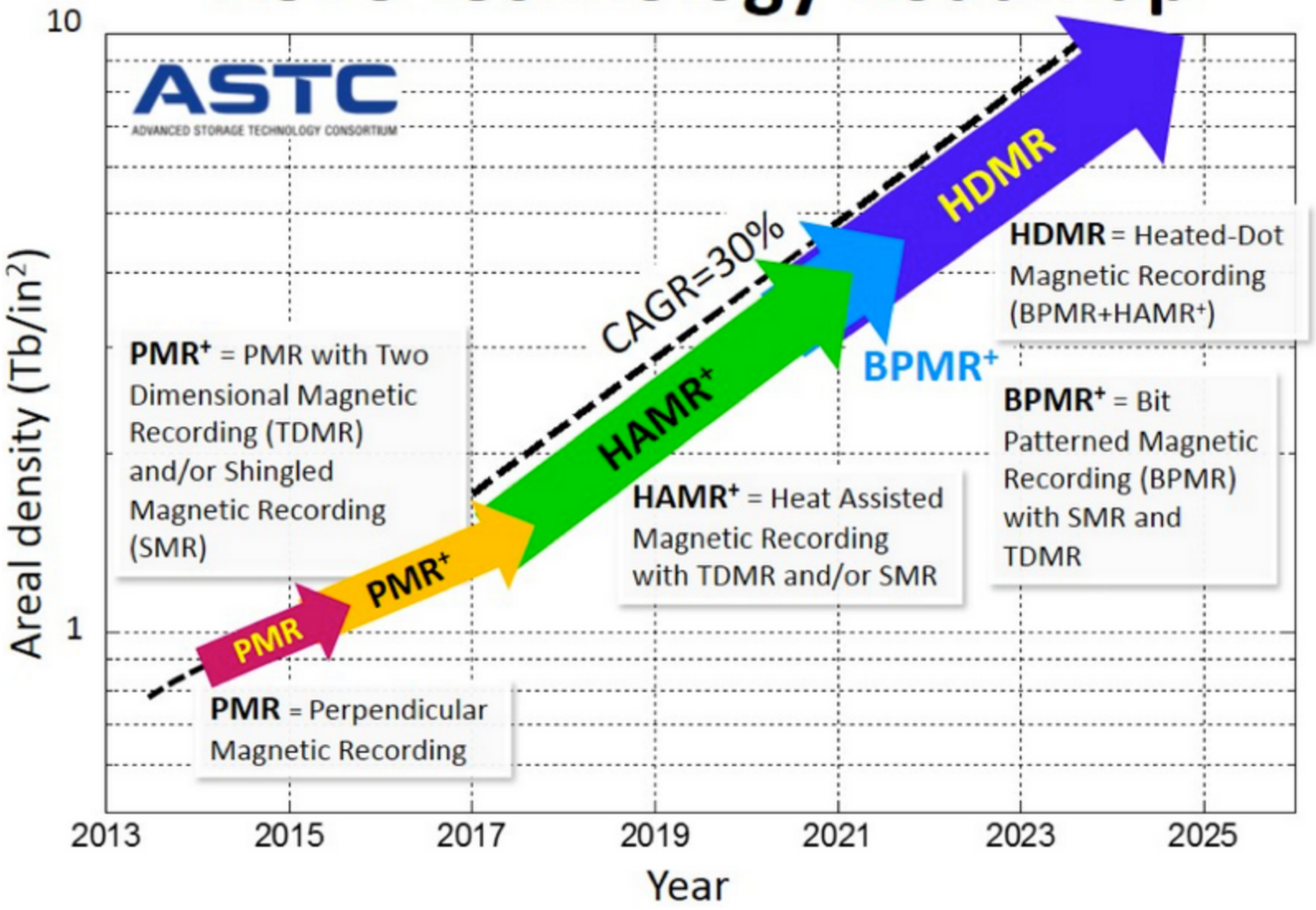
- Two HDD in one case to keep IOPS/TB constant
- The host computer can treat a single Dual Actuator drive as if it were two separate drives
- Need a special device driver



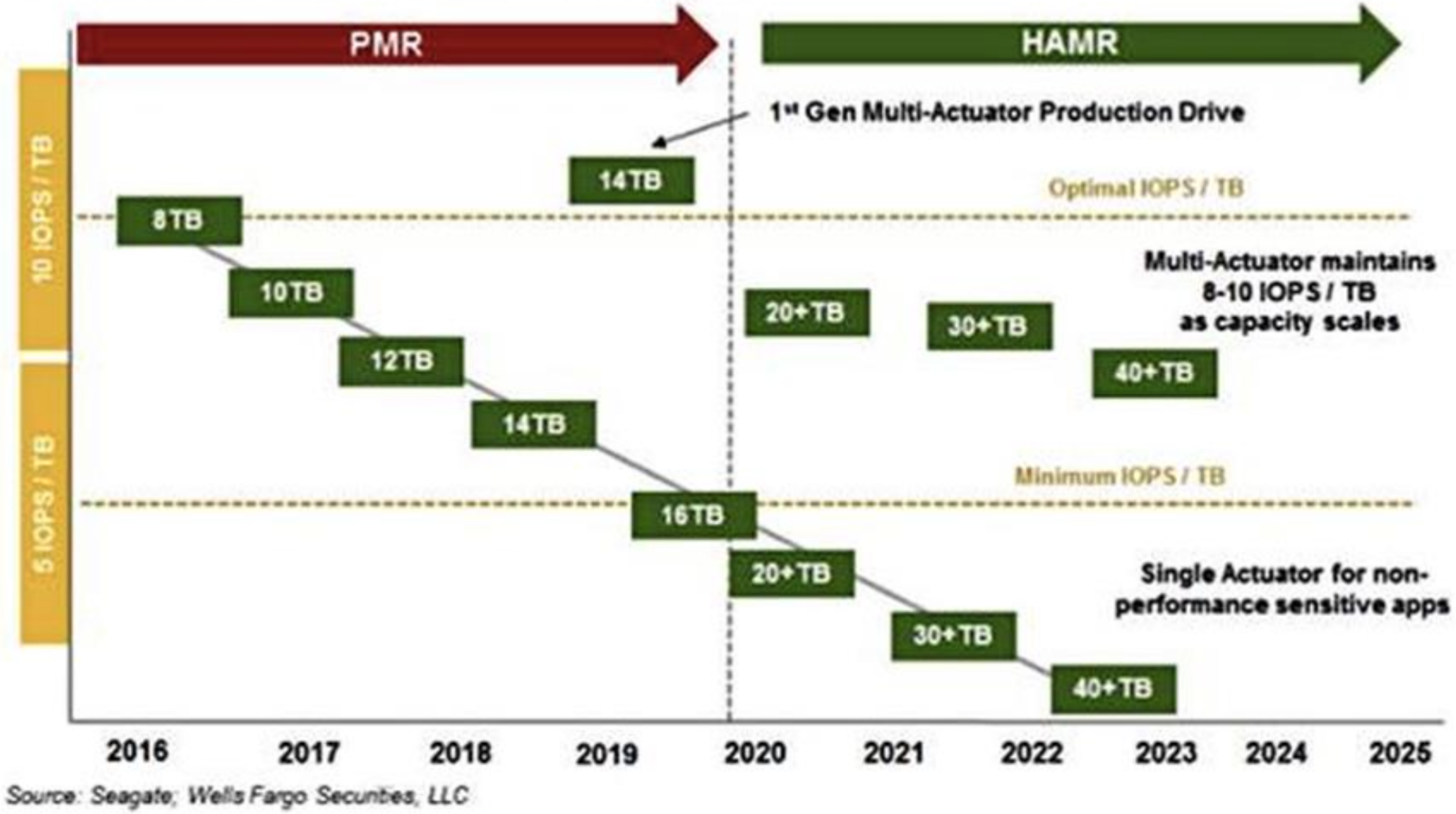
Capacity growth (by WD)



ASTC Technology Roadmap



Seagate roadmap for multi-actuator HDD



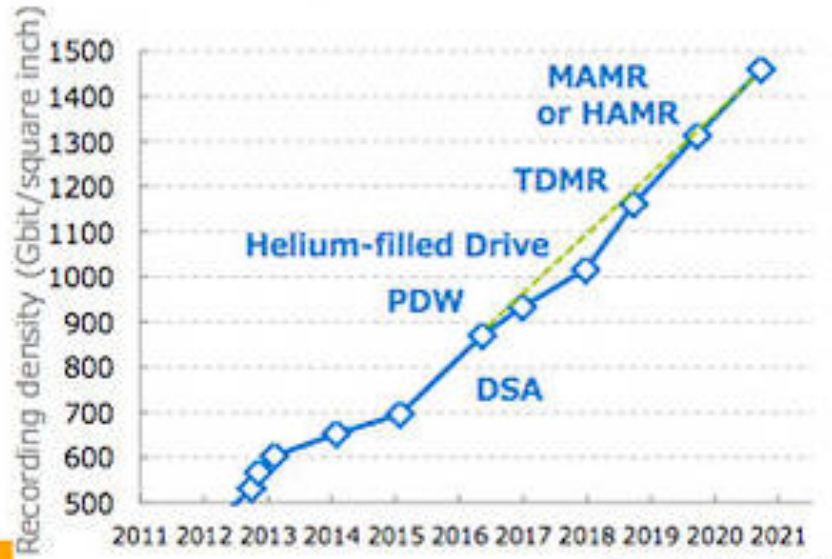
Source: Seagate; Wells Fargo Securities, LLC

Future of Nearline HDD

Recording density improving >15%/Y, Realizing better bit cost over SSD

Bit cost comparison(vs SSD)

Recording density +>15%/Y

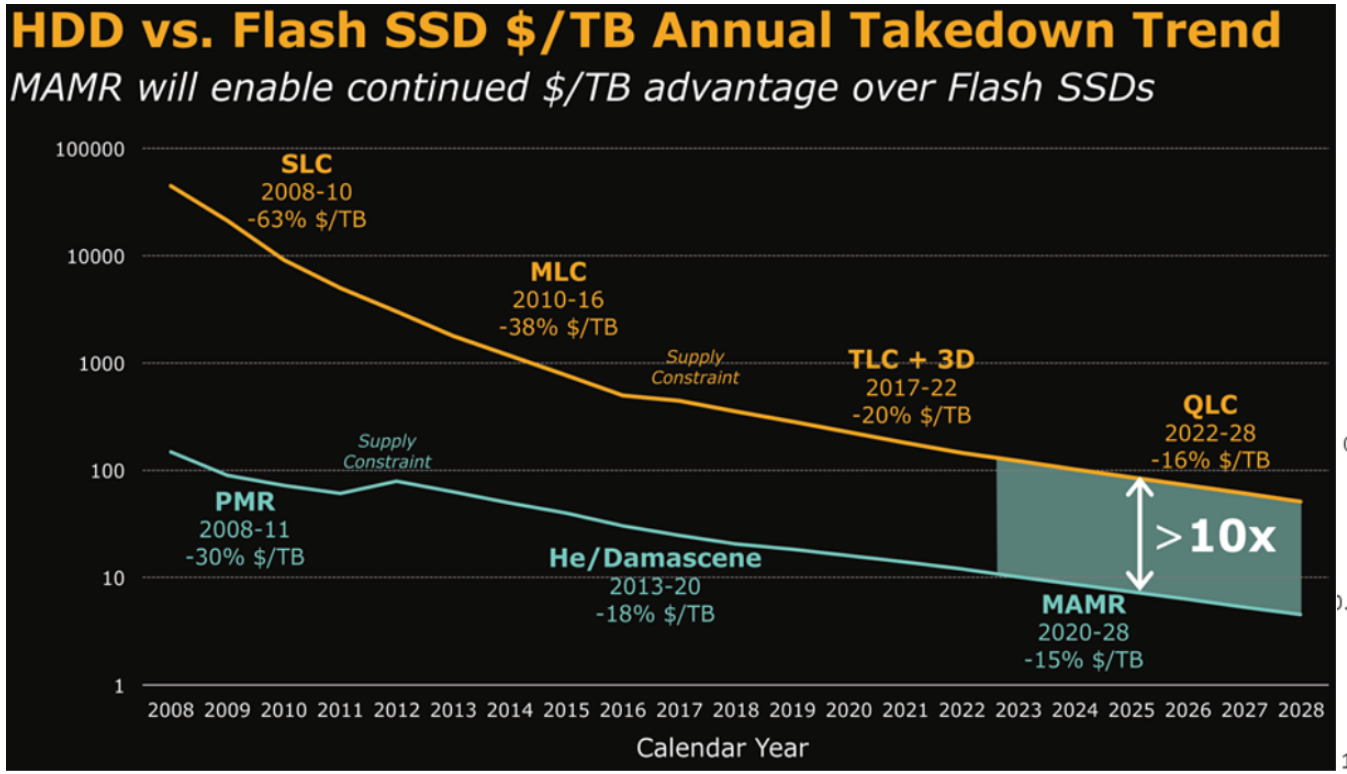


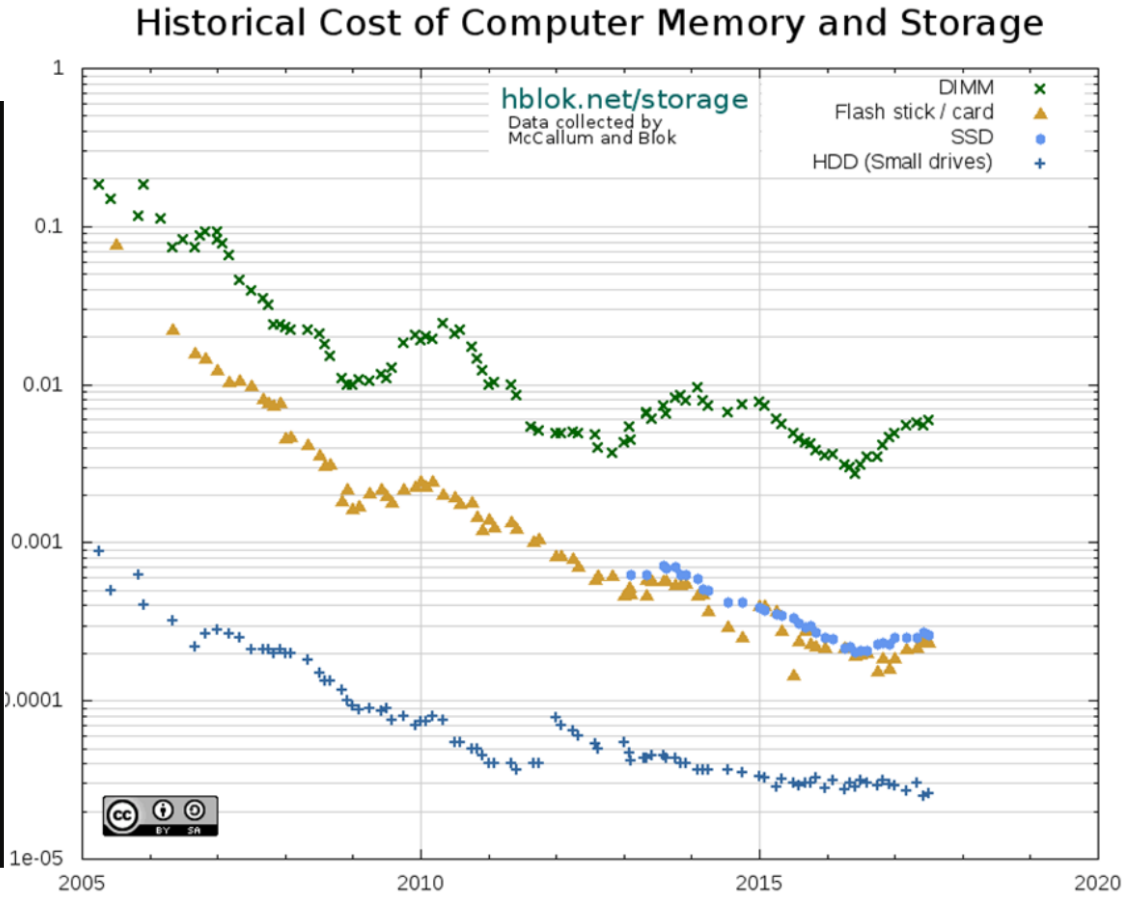
MAMR: Microwave Assisted Magnetic Recording
HAMR: Heat Assisted Magnetic Recording
TDMR: Two-Dimensional Magnetic Recording
PDW: Pattern Dependent Write Control
DSA: Dual Stage Actuator

Cope with shrinking HDD market by shifting resources to SSD business

TOSHIBA
Leading Innovation

HDD vs. FLASH


HDD vs. Flash SSD $/TB Annual Takedown Trend
MAMR will enable continued $/TB advantage over Flash SSDs
100000
10000
1000
100
10
1
SLC
2008-10
-63% $/TB
MLC
2010-16
-38% $/TB
Supply Constraint
TLC + 3D
2017-22
-20% $/TB
QLC
2022-28
-16% $/TB
Supply Constraint
PMR
2008-11
-30% $/TB
He/Damascene
2013-20
-18% $/TB
MAMR
2020-28
-15% $/TB
>10x
2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028
Calendar Year

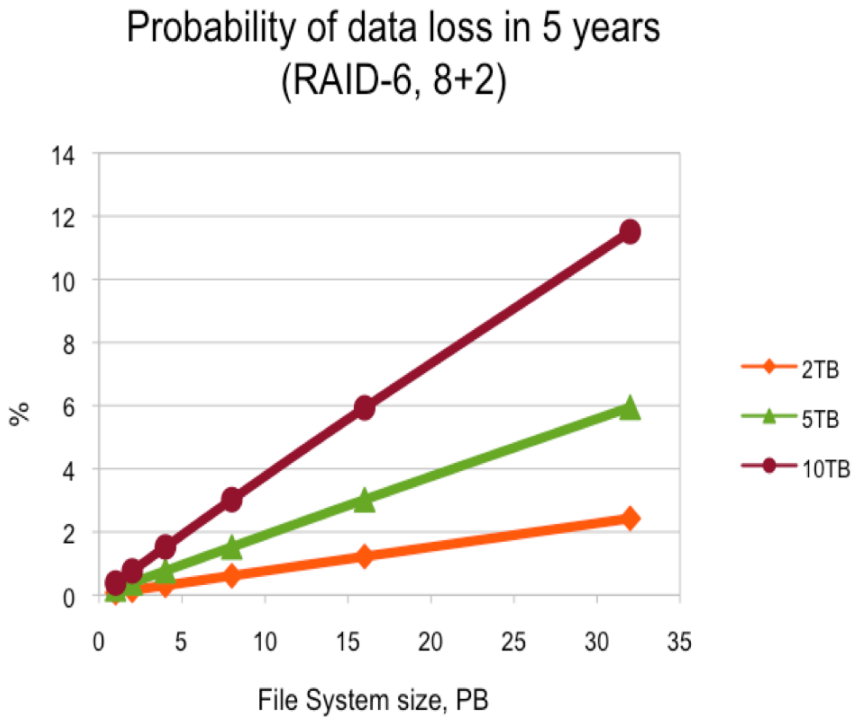

Historical Cost of Computer Memory and Storage
hblok.net/storage
Data collected by
McCallum and Blok
DIMM
Flash stick / card
SSD
HDD (Small drives)
1
0.1
0.01
0.001
0.0001
1e-05
2005
2010
2015
2020

Backblaze Average Cost per Drive Size

By Quarter: Q1 2009 - Q2 2017

Data protection: To RAID or not to RAID?

- RAID-5 for big file systems with big capacity (>4TB) disks is too risky
  - ~50% of probability of data loss in 5 years on 2PB file system
- RAID-6 showing better protection
  - Acceptable up to some extent
- Other methods
  - Replication – double o triple costs (expensive)
  - Erasure coding – high CPU demand
  - Distributed (on "de-clustered") RAID



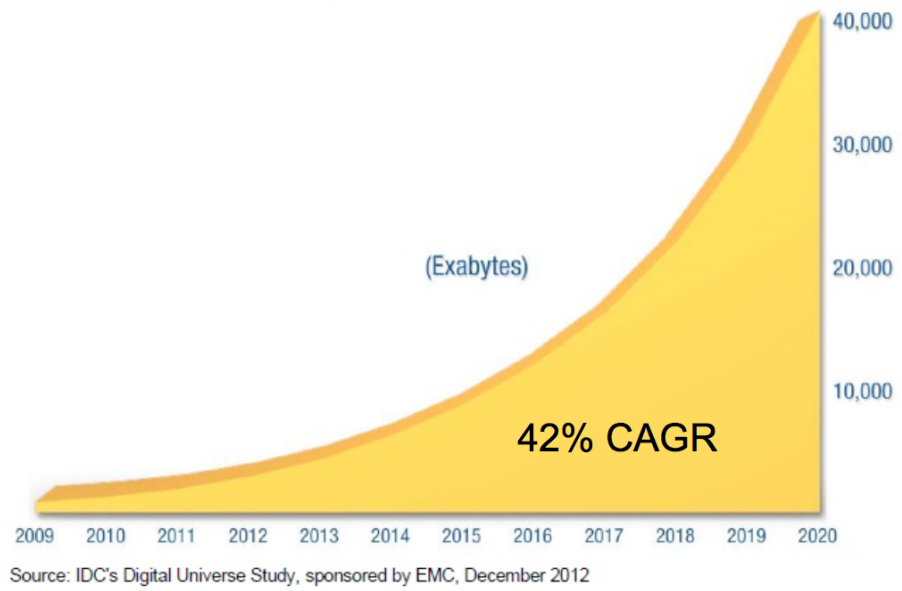
V. Sapunenko: What comes after RAID?

Distributed RAID

- Usually implemented as "floating" RAID-6 (8+2) over bigger (>10) disk pool
- Using "reserved capacity" to restore missing blocks in case of disk failure
- With a disk pool big enough recovery time becomes significantly reduced
  - Failure of 4TB disk: Disk pool of 180 disks – 3.5 hours to restore redundancy (under heavy I/O load) against 20-22 hours in traditional RAID-6
  - Failure of 6TB disk: Disk pool of 95 disks – 3 hours to restore redundancy
- Drawbacks
  - I/O performance may be affected by ~20%
  - Limits on LUN size may require creating more LUNs on the same disk pool reducing to zero I/O optimization made by file system software
- Becomes more and more widespead
  - In some cases the only possibility offered (ad es. Huawei)

Tape's Renaissance

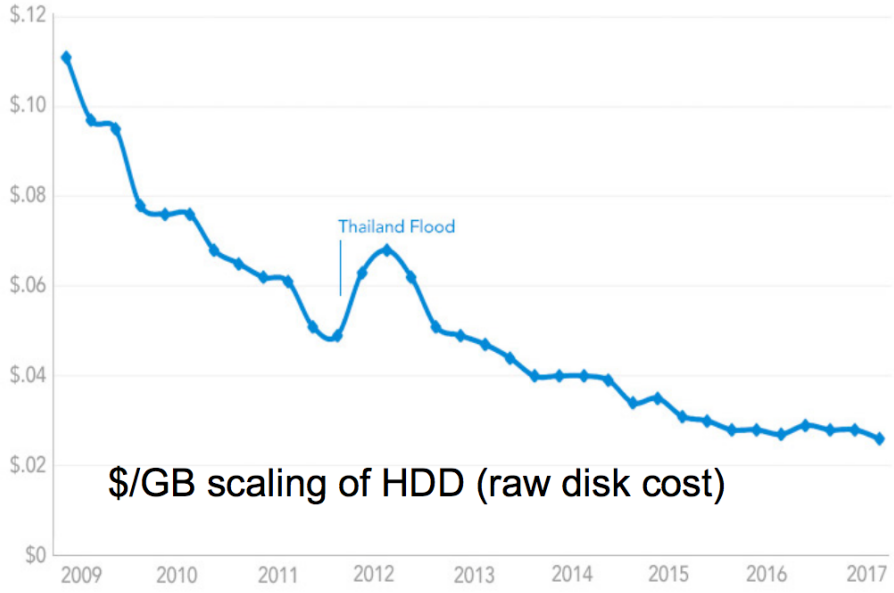
IDC Projection for Data Growth



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Backblaze Average Cost per GB for Hard Drives

By Quarter: Q1 2009 - Q2 2017



BACKBLAZE

80% of all files created are inactive – no access in at least 3 months!

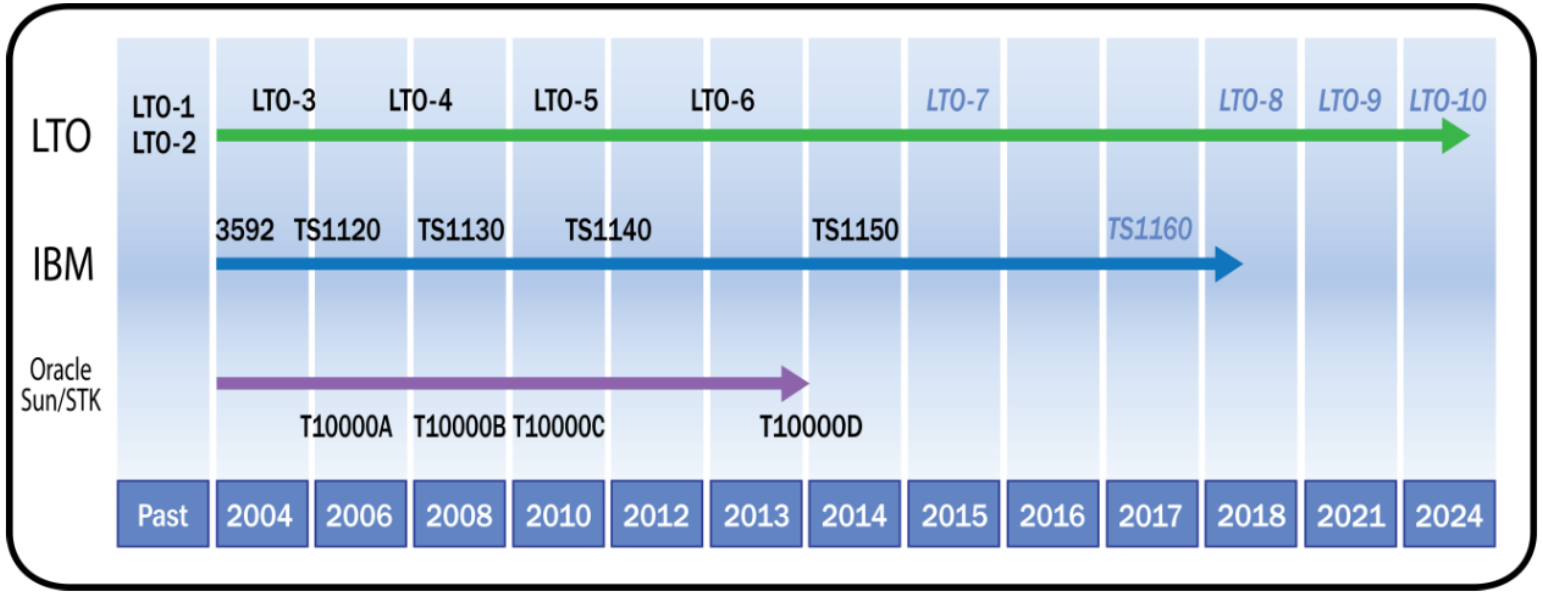
Ref: https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/

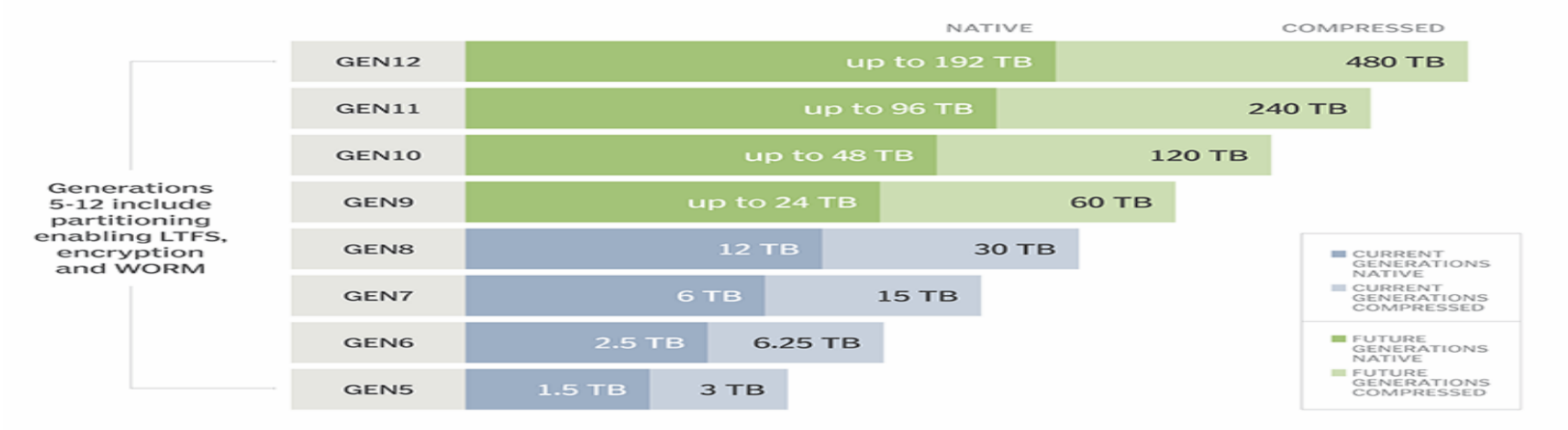
Near-line (tape) Storage

- LTO vs. Enterprise tapes
  - On May 16, 2017 the TS1155 enterprise tape drive was announced by IBM®
    - **15 TB** native capacity (45 TB compressed) offering a 50% greater capacity than the IBM TS1150 drive
    - has a data rate of 360 MB/sec.
  - On October 17, 2017 the LTO Ultrium Generation 8 tape drive (LTO-8) was announced by the LTO Program Technology Provider Companies (TPCs), Hewlett Packard Enterprise, IBM and Quantum.
    - doubles the native capacity from its previous generation to **12 TB** (30 TB compressed)
    - improves throughput rates by 20% to 360 MB/sec

Open and proprietary technologies

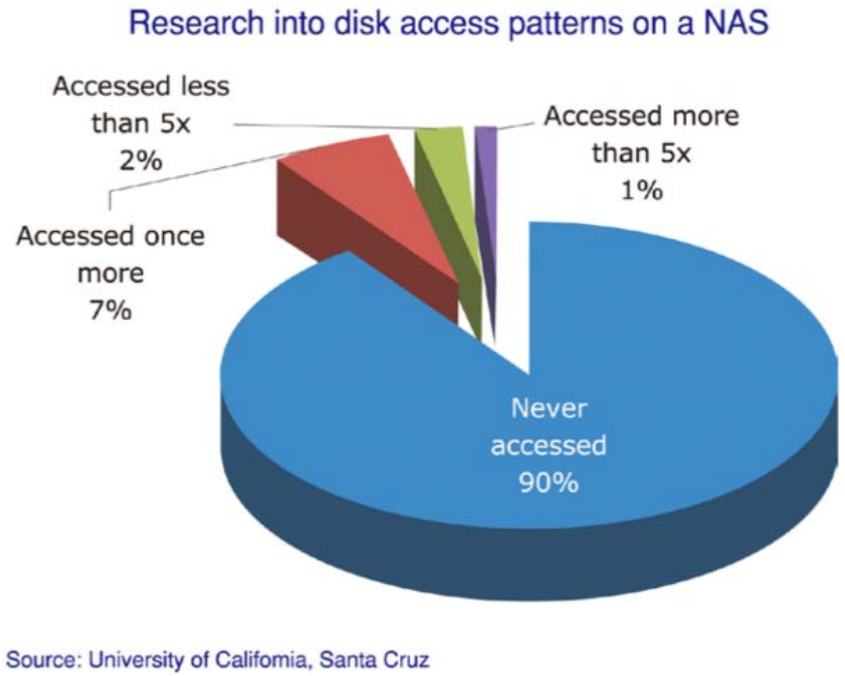
- Linear Tape Open/LTO-8
- Oracle/T10000D
- IBM/TS1150

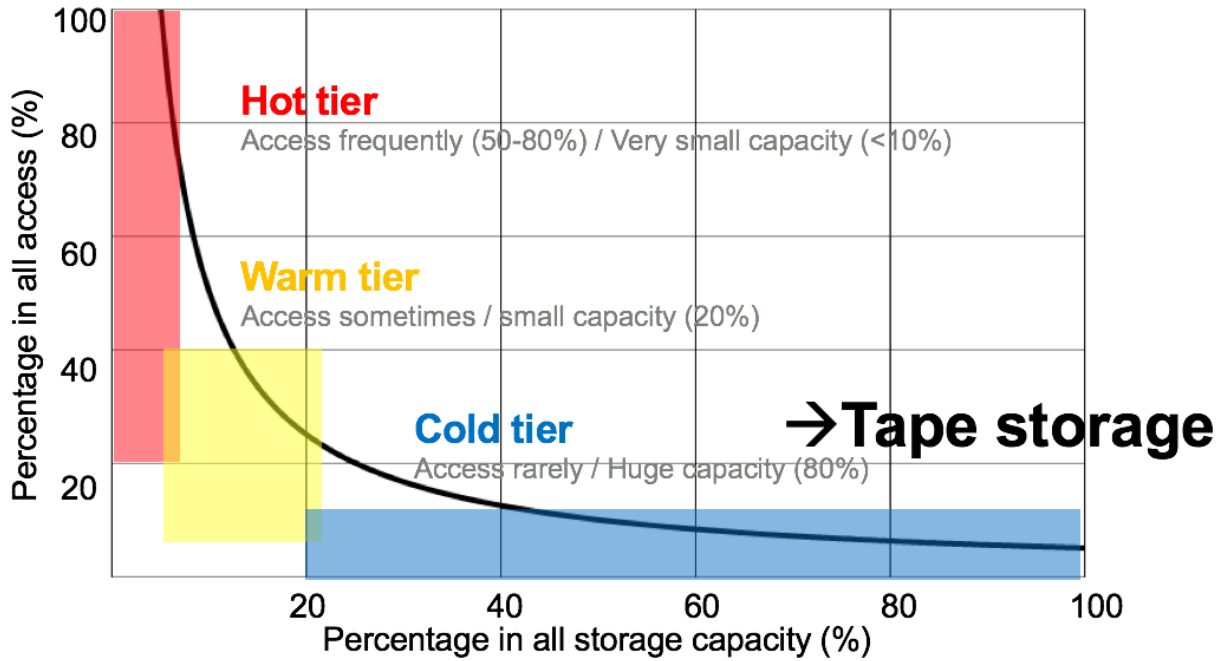




Hot and Cold data

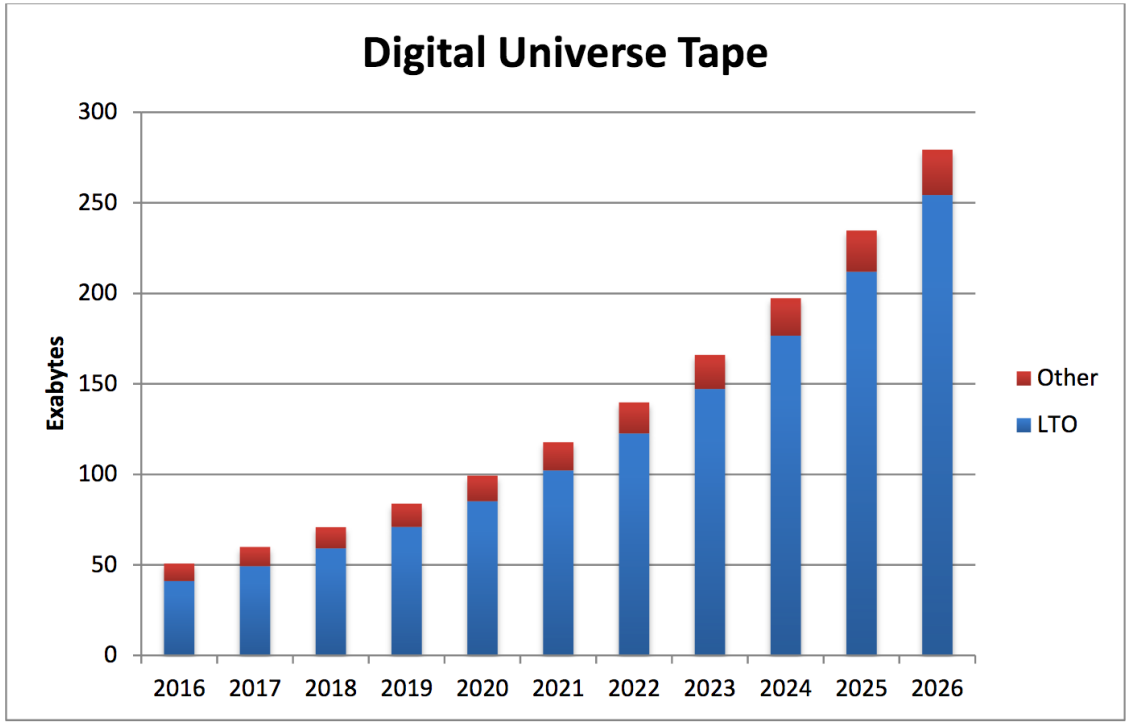
- Most data is never or very rarely accessed
- however, data must be retained for preservation to ensure compliance with legal requirements or, for future reference or for analysis





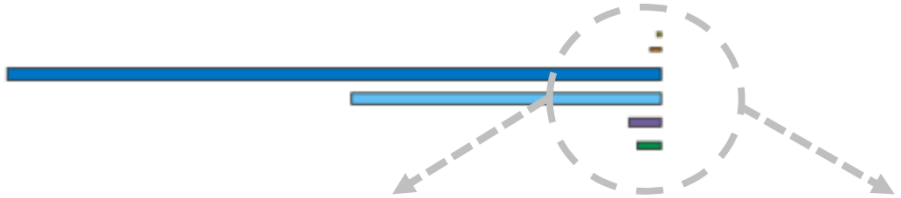
LTO vs. Enterprise tapes share

- LTO takes 95% of the market
- Enterprise tapes ~5%
- Related to different usage?
  - LTO - Write Once Read Newer (legal requirement)
  - Enterprise tapes: frequent updates (backup)
- Only IBM producing LTO and Enterprise drives
- Only two suppliers of media
  - FujiFilm
  - Sony



Storage bit cells comparison

- Scaled bit cells:



- Magnified 25x:

NAND Flash (3 bits) 2150 Gb/in^2
17.3 nm x 17.3 nm

HDD 1000 Gb/in^2
47 nm x 13 nm

LTO7 Tape ~4.3 Gb/in^2
2850 nm x 52 nm

Jag5A Tape ~9.6 Gb/in^2
1347 nm x 50 nm

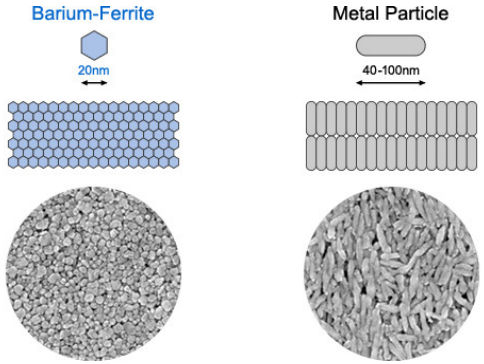
BaFe Demo 123 Gb/in^2
140 nm x 37 nm

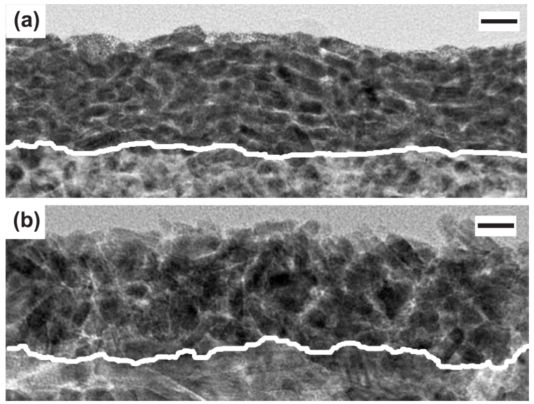
Sputtered Demo 201 Gb/in^2
103 nm x 31 nm

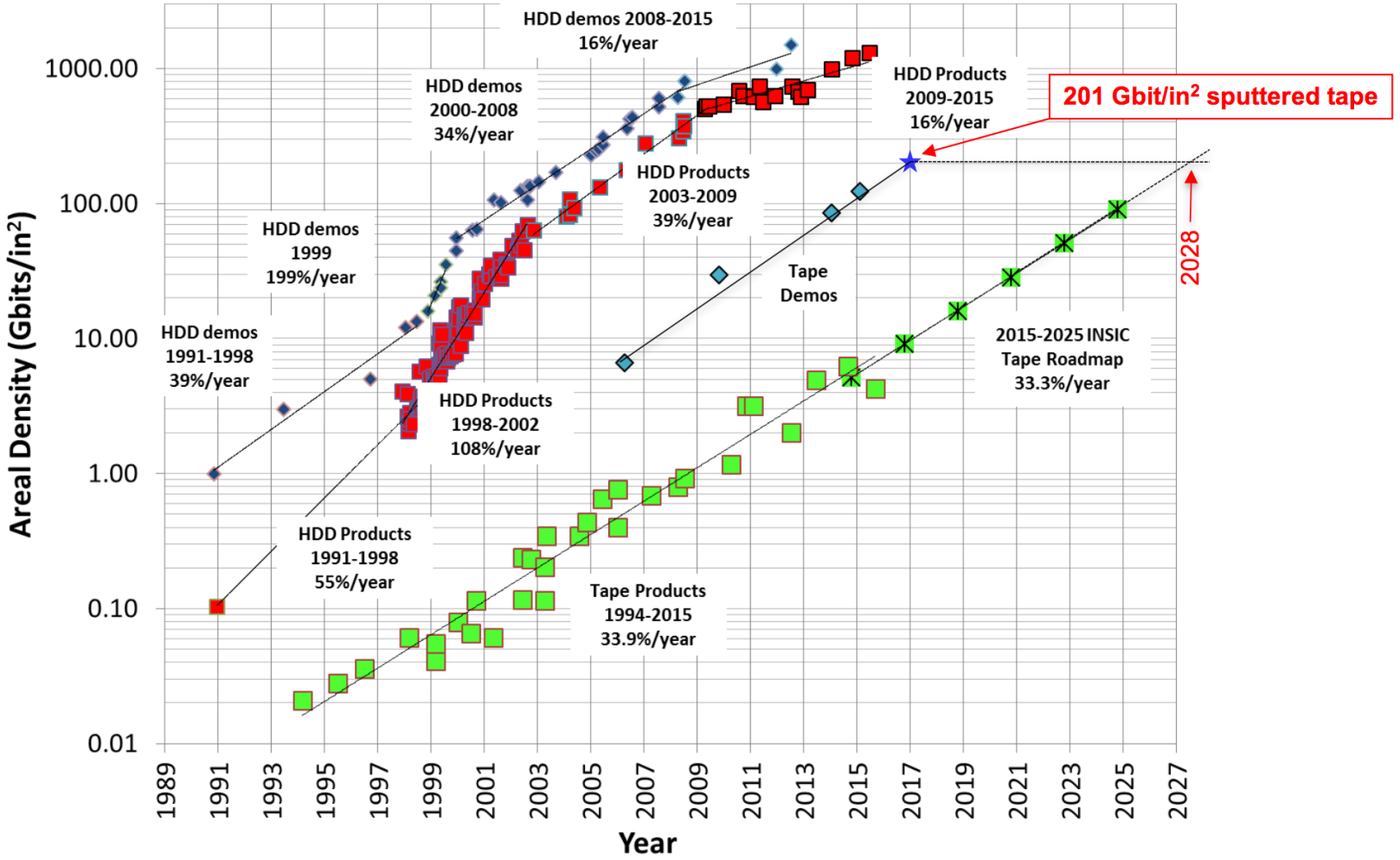
Tape and disk areal density

2015: IBM-FujiFilm demonstration of 123 Gb/in^2 on BaFe tape

2017: IBM-Sony demonstration of 201 Gb/in^2 on Sputtered Tape

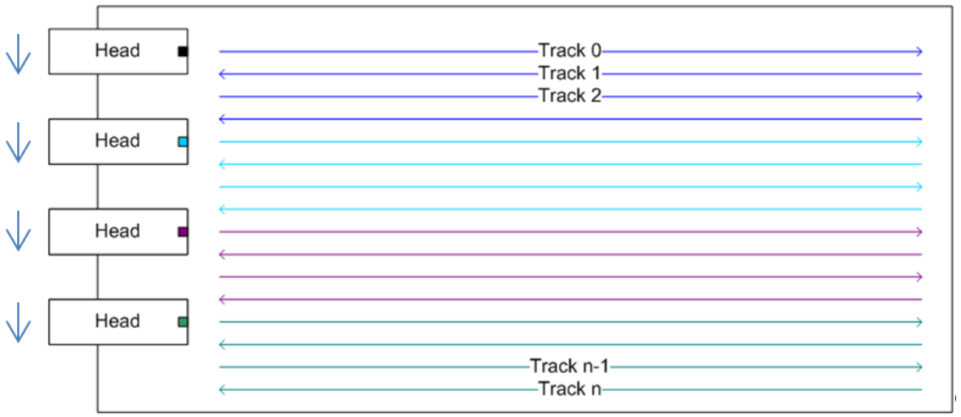




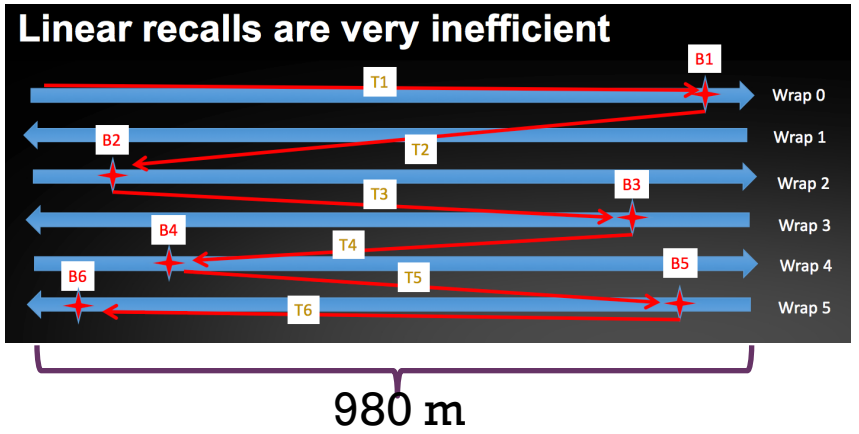


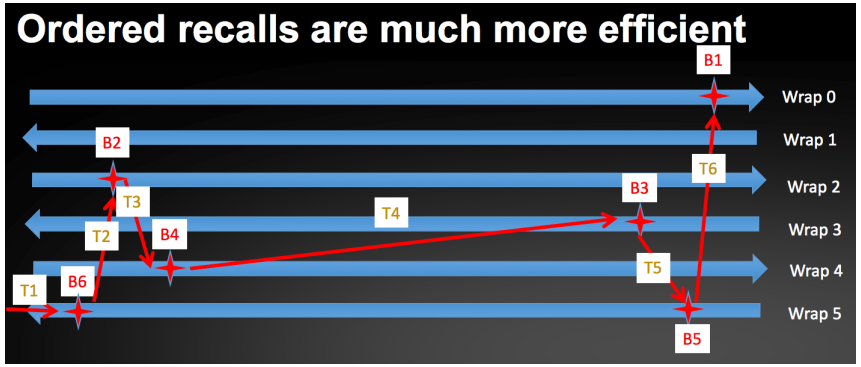
Linear serpentine technology

- Tape has a serpentine pattern
- Possible improvement in recall ops
  - Native for TS drives
  - Developed by library manufactures for LTO drives

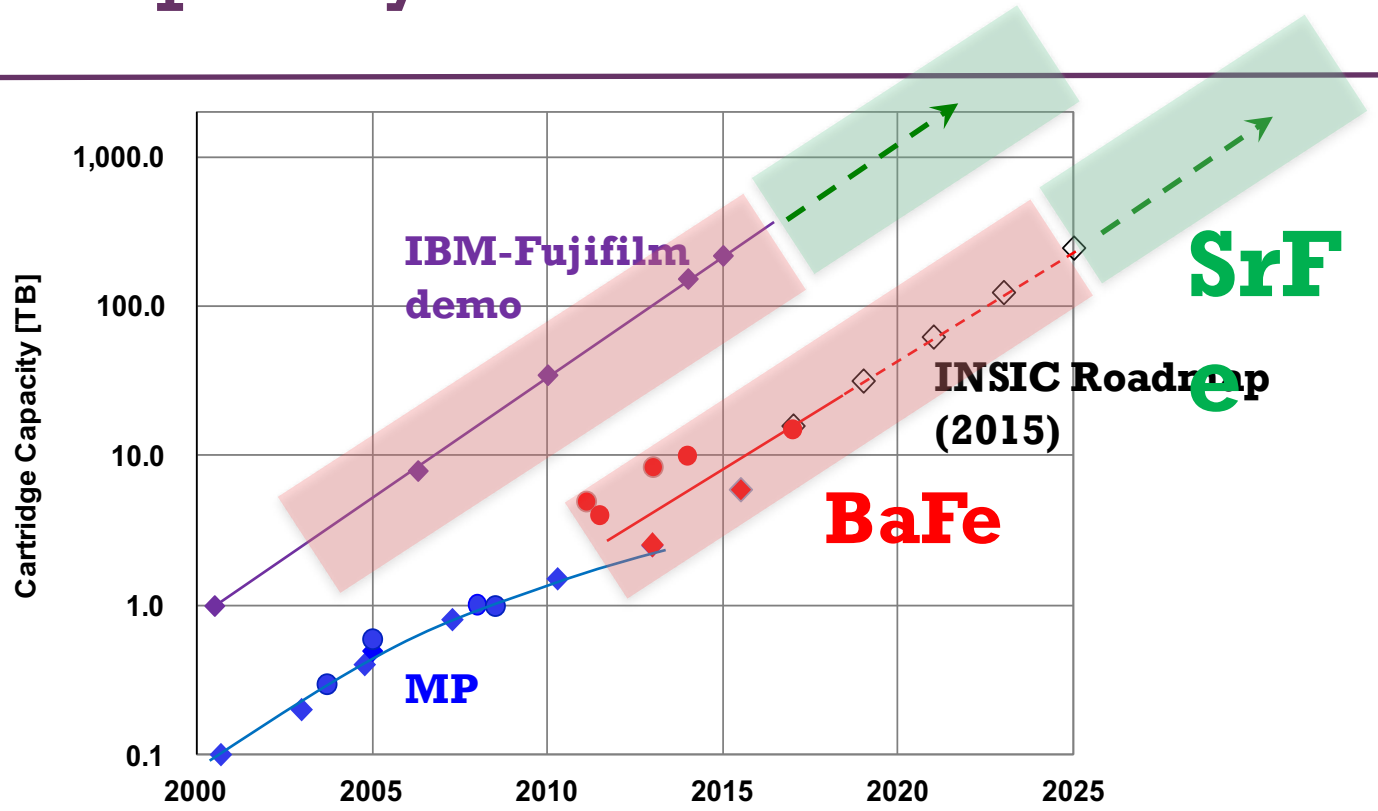


LTO7: 112 wraps
LTO8: 208 wraps





Cartridge capacity trends



BaFe can support the next 10 year's tape roadmap.

SrFe will enable to further high capacity cartridge in the future !!

Tape Technology Summary

- Tape will remain the most efficient, cost effective, and reliable technology for long term data storage
- Tape will continue to have the highest areal density/capacity growth rate
  - Evolution, not invention
  - No fundamental issues, just scaling
- Growing consensus that 1 technology will not 'win'
  - Storage solutions will feature combo of flash, disk, and tape
  - HDD will not completely replace Tape

References

- https://3dnews.ru/94045
- https://www.qstar.com/index.php/ltfs-linear-tape-file-system
- https://www.forbes.com/sites/tomcoughlin/2018/02/05/hdd-growth-in-nearline-markets
- http://www.theregister.co.uk/2018/03/21/seagate_to_drop_multiactuator_hamr_in_2020/
- https://www.theregister.co.uk/2017/12/19/seagate_disk_drive_multi_actuator/
- http://www.tomshardware.com/news/seagate-wd-hamr-mamr-20tb,35821.html
- https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte
- https://hblok.net/blog/posts/2017/12/
- https://www.anandtech.com/show/10470/the-evolution-of-hdds-in-the-near-future-speaking-with-seagate-cto-mark-re
- https://www.extremetech.com/computing/256961-western-digital-launches-worlds-first-14tb-hard-drive
- http://www.npd.no/Global/Norsk/3-Publikasjoner/Presentasjoner/24-august-2017/IBM_Mark%20Lantz_Future%20of%20tape.pdf
- https://spectralogic.com/wp-content/uploads/white-paper-digital-data-storage-outlook-2017-v3.pdf
- https://indico.cern.ch/event/160737/contributions/1407837/attachments/184854/259795/hepix2012_after_raid_v2.pdf

Backup slides

Tape vs. disk

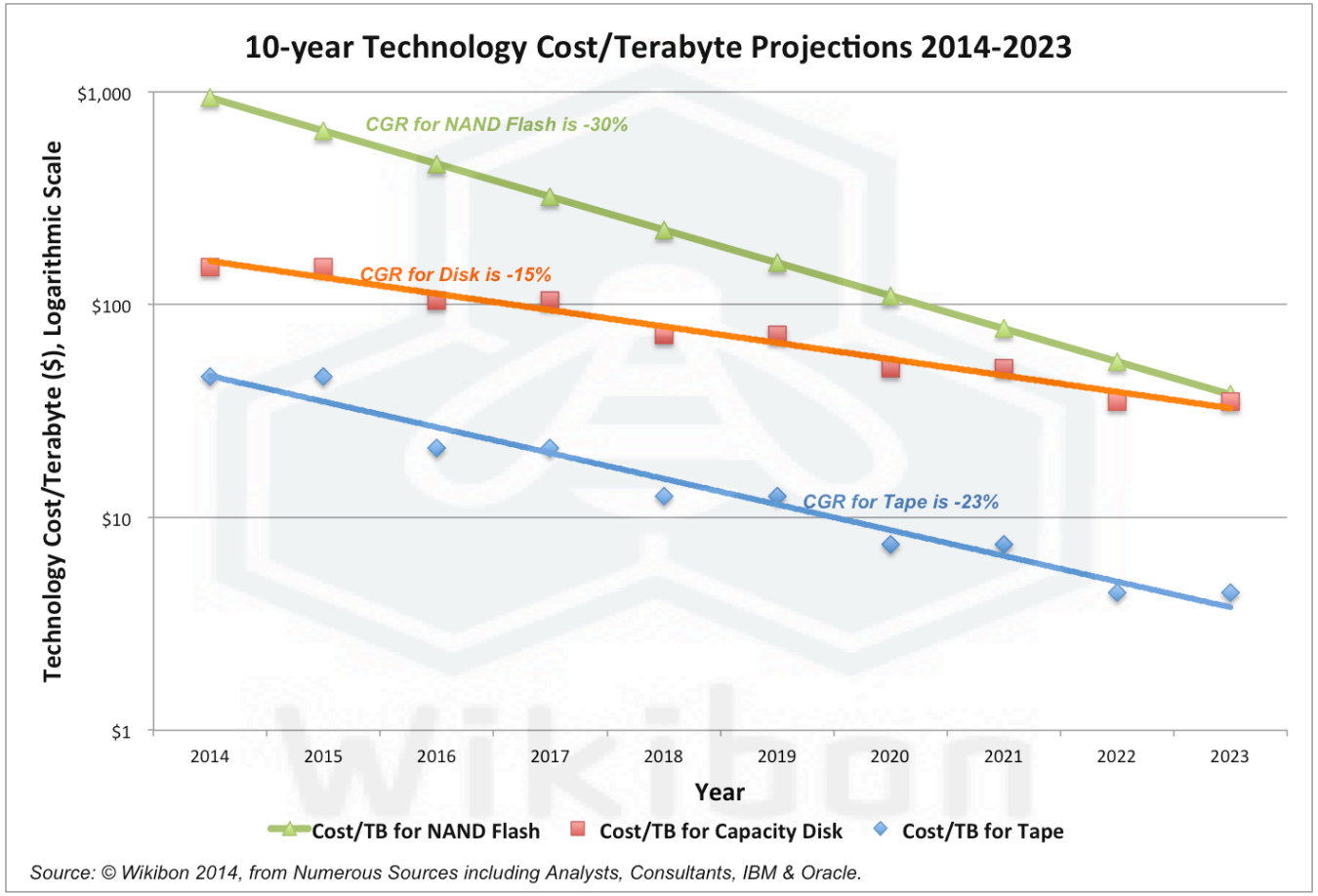
	Tape			HDD	Optical Disk
	LTO6	LTO7	TS1150	ST8000AS0002	Blu-ray(16×)
Capacity	2.5 TB	6.4 TB	10 TB	8 TB	0.1 TB
Transfer rate	160 MB/s	315 MB/s	360 MB/s	150 MB/s	72 MB/s
Error rate	1.E-17	1.E-17	1.E-20	1.E-14	-
Access time (including media mount time)	~ minute			a few ms	~ minute
Media life	30 years or more			~3 years	50

http://www.lto.org/

http://edge.spectralogic.com/index.cfm?fuseaction=home.displayFile&DocID=2513

http://www.seagate.com/www-content/product-content/hdd-fam/seagate-archive-hdd/en-us/docs/archive-hdd-dS1834-3-1411us.pdf

10-year Storage Technology Cost Projections 2014-2023 (Cost/Terabyte) Source: Wikibon, 2013

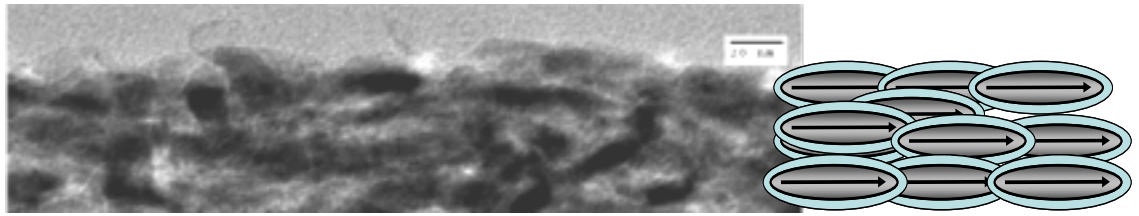


Perpendicular Orientation Technology

FUJIFILM

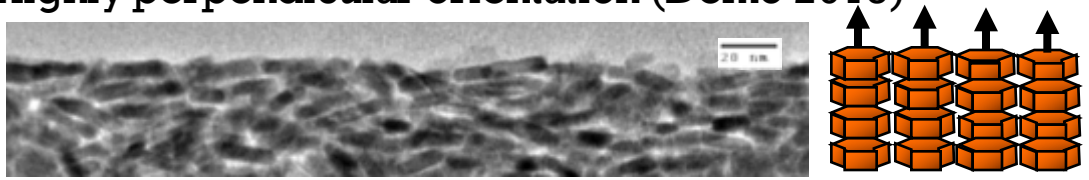
Particle orientation

Longitudinal orientation (MP tape)



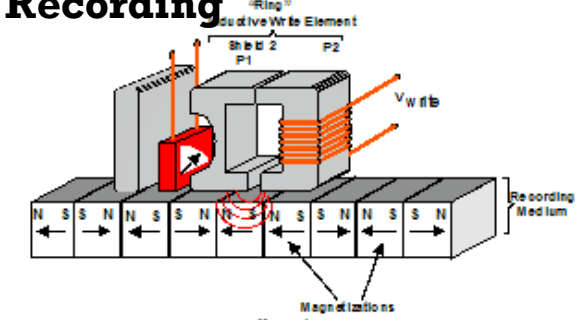
Random orientation (Current BaFe tape)

Highly perpendicular orientation (Demo 2015)

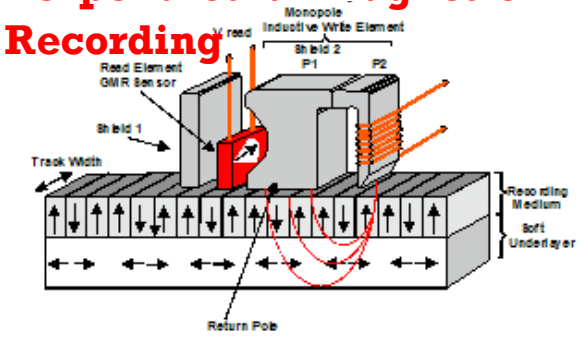


Recording system

Longitudinal Magnetic Recording



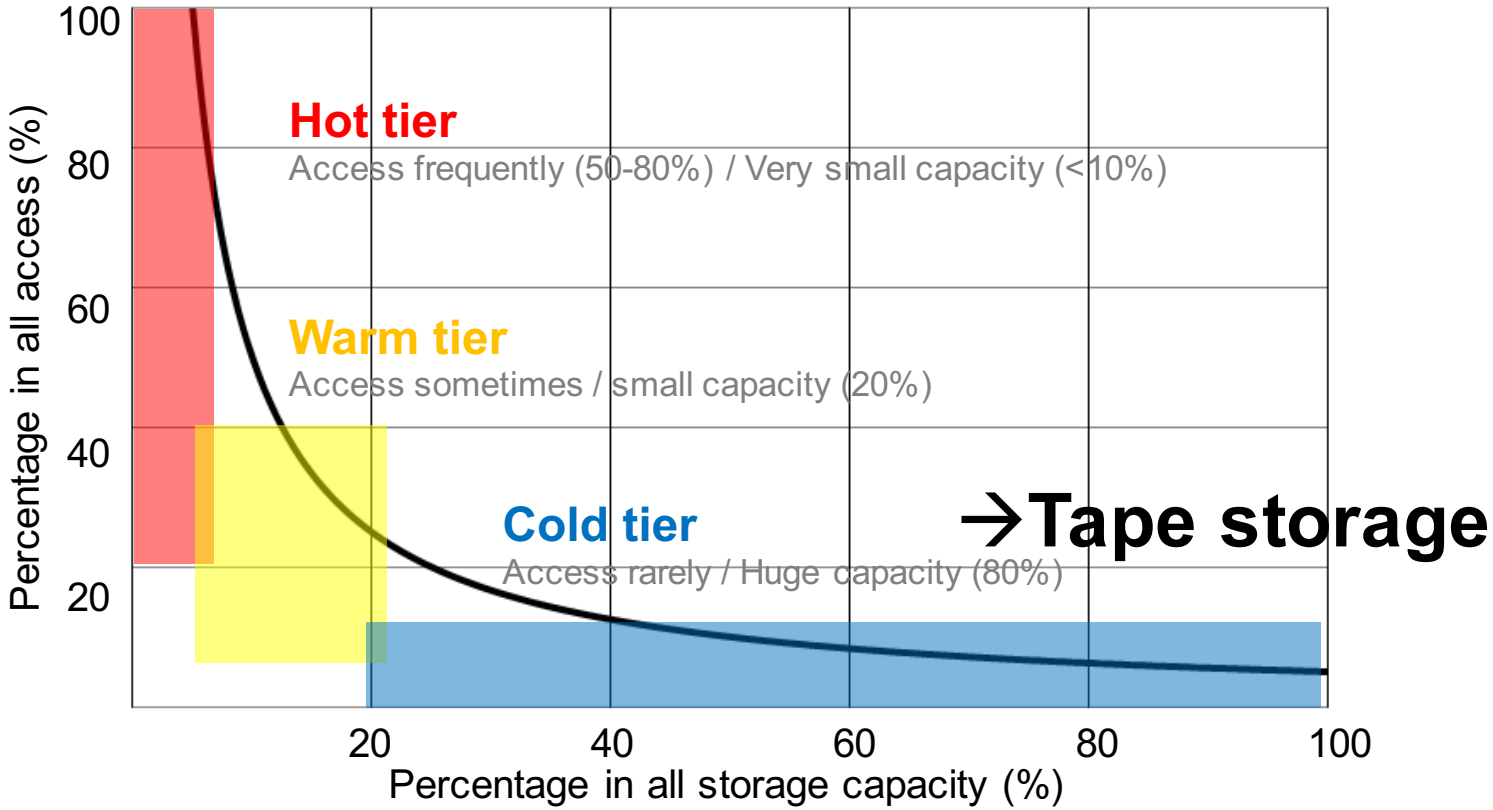
Perpendicular Magnetic Recording



- BaFe particles can be oriented in perpendicular direction.

→PMR, which contributed to increase capacity of HDD can be applied in the tape storage system.

New Role of Tape as Cold Data Storage



- Most data is very rarely accessed, however, data must be retained for preservation to ensure compliance with legal requirements or, for future reference to analyze business opportunities.**

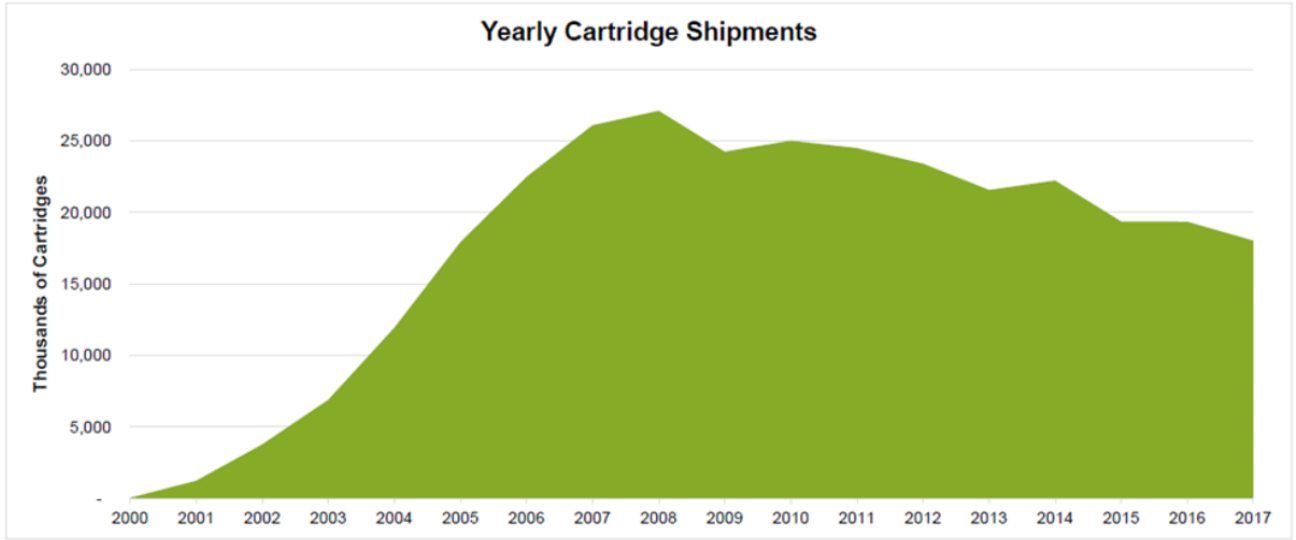
  →Storage for COLD data has become a HOT topic

- But budget is limited.

  →Reliable yet inexpensive storage media is required.

*90% data in NAS is never accessed. (Source: University of California, Santa Cruz)

**Retention of 20 year or more is required by 70%. (Source: SNIA-100 year archive survey)



LTO tape market domination >95%
Enterprise tapes 4%

44 EB of tape media in 2017 compared to 750 EB HDD
Linear increase in EB sold per year

Declining media shipment since 10 years

factor 2 decrease in #drives sold over the last 4 years

Only two suppliers of media: Fujifilm and Sony
Fujifilm only supplier in the US (patent 'war')

Only IBM left for LTO and Enterprise drives

Total Capacity Shipped: Calendar Year