# Tape and Disk evolution for the exabyte era

Hugo Gonzalez Labrador
Storage Engineer
CERN IT
7th Rucio Workshop, SDSC, 30/09/2024

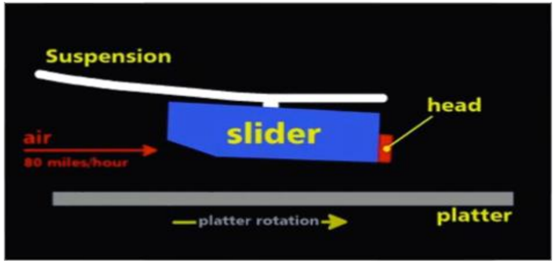
# Outline

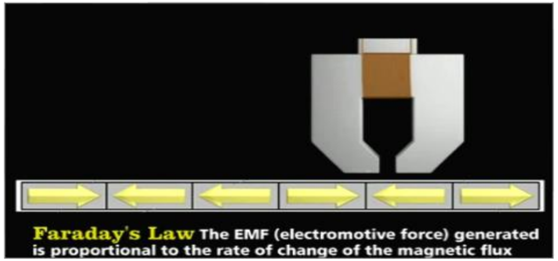
- Storage media
  - HDD (Hard Disk Drives)
  - SSD (Solid State Drives)
  - Magnetic Tapes
- Storage at CERN
- Market evolution and forecast
- Challenges for CERN

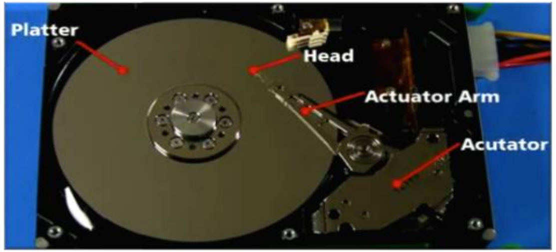
# HDDs

# Hard Disk Drives

*Basics*

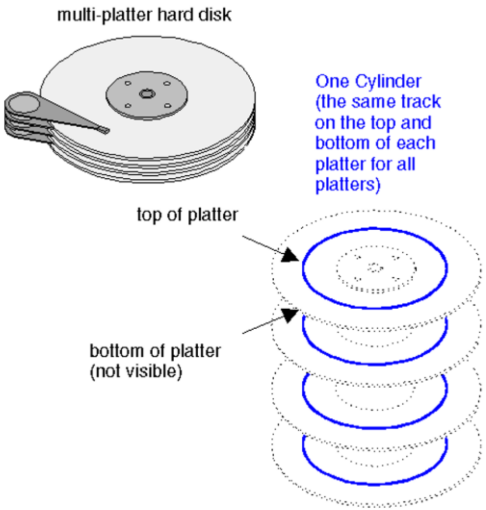






Changes in the polarization of the disk platter sections will create voltage spikes, which gets encoded to sequences of 0s and 1s: 01000111...



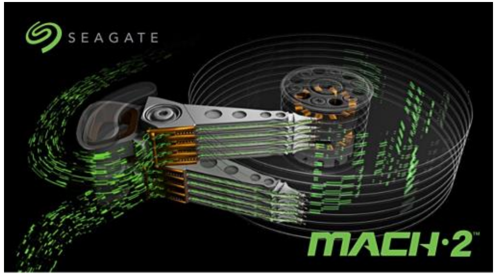
# Hard Disk Drives
## *Driving factors*

Market vendors optimize on two aspects with different efforts

- Increase performance (throughput, GB/s) -> little effort put
- Increase capacity -> most effort put

# Hard Disk Drives
## *Driving factors: Performance*

**Doble actuators**
1 actuator -> ~250 MB/s
2 actuators -> 500 MB/s
Niche market




Source: Seagate

**Write cache enabled**
Allows to cache writes to the media and transfer them in bulk to optimize performance

Data loss on power outages

Alternatives: Emergy Power Off Flush
Flash memory is used to write to the media
TODO

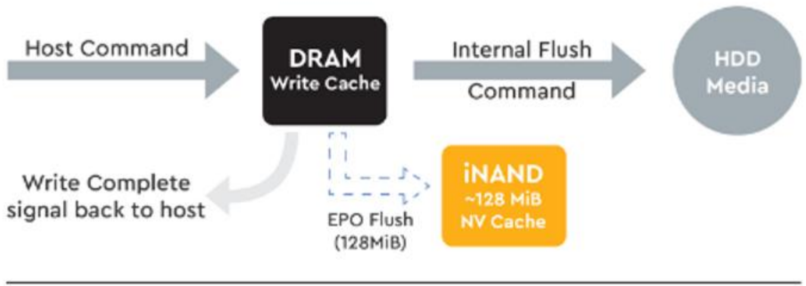


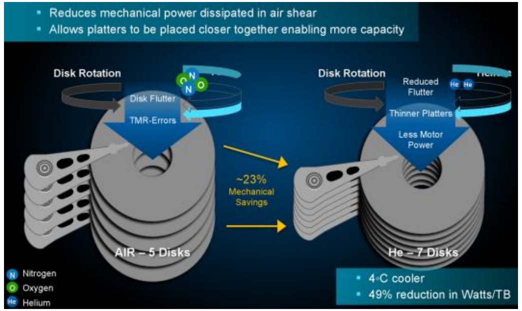

Figure 1: Emergency Power Off (EPO) Flush

Source: WD

# Hard Disk Drives

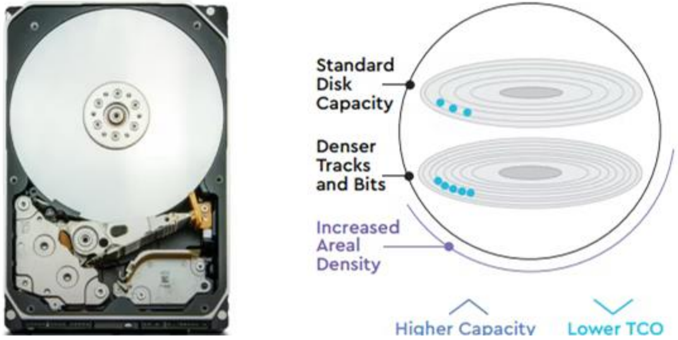
## *Driving factors: Capacity*

**Helium drives**
Replacing air with Helium
Reduces friction between platters
Cooler temperatures
Less power consumption
Thinner platters -> More platters
**Today: 10 platters**




Source: Annandtech

**Areal density**
Number of bits you can store per square inch, usually measured in Tb/inch2

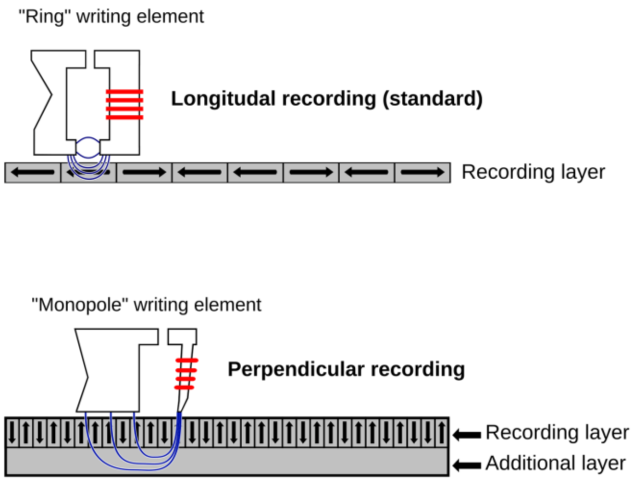



Source: WD

# Hard Disk Drives

## *Driving factors: Capacity: Areal Density: PMR and SMR*

### PMR (Perpendicular Magnetic Recording)

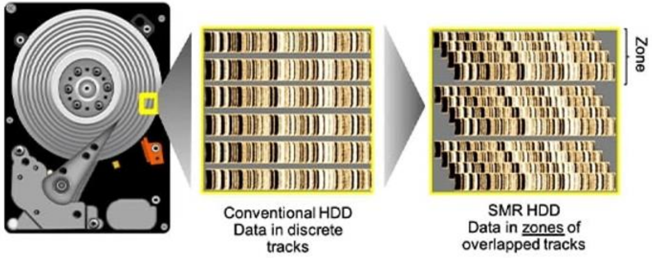



By TylzaeL - http://commons.wikimedia.org/w/index.php?title=File:Perpendicular-eng.jpg, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5734076

### SMR (Shingled Magnetic Recording)

Tracks overlap like tiles in a roof

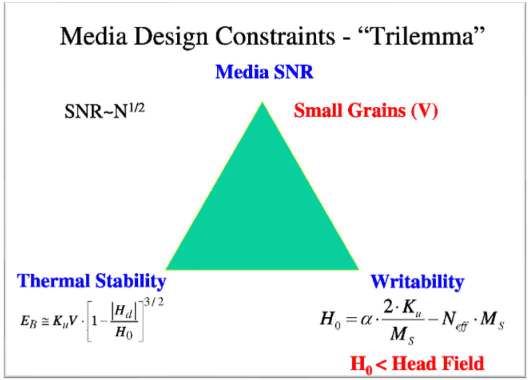
Increases density but reduces performance for random writes (zone overwrite)



# Hard Disk Drives

## *Driving factors: Capacity: Areal Density: Paramagnetic trilemma*

- To increase density we need smaller grains (less ferromagnetic molecules)
- **The head becomes smaller** to be more precise, means head generates weaker magnetic field
- As grains become smaller, they are susceptible to thermal agitation (*bitflip*)
- To increase immunity to thermal agitation, the grains need to “hold” to each other stronger, requiring a stronger magnetic field, requiring a **bigger head**




**What tricks do vendors use to avoid these constraints?**

Use another source of energy to “ease” writing in the media grain so the electromagnetic field can be weaker
These techniques are called **Energy Assisted Magnetic Recording (EAMR or ePMR)**

# Hard Disk Drives

## *Driving factors: Capacity: Areal Densi HAMR*

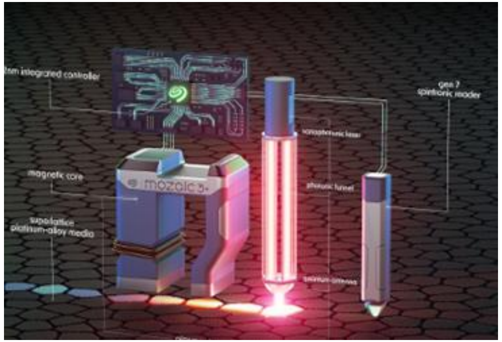
HAMR (Heat Assisted Magnetic Recording)

Uses a laser to heat media grains to Curie Point to the ferromagnetic grain loses magnetic polarisation

When reaching Curie point, a small electromagnetic field is induced to change polarisation

Heat/Write/Cool cycle is less than 1 ns

Reduces energy of required electromagnetic field to almost zero




Source: Seagate

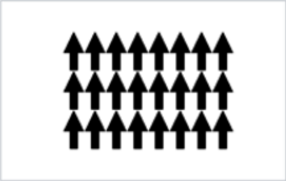


**Figure 1.** Below the Curie temperature, neighbouring magnetic spins align parallel to each other in a ferromagnet in the absence of an applied magnetic field.

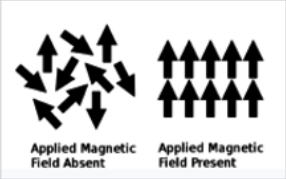



**Figure 2.** Above the Curie temperature, the magnetic spins are randomly aligned in a paramagnet unless a magnetic field is applied.

Source: Wikipedia

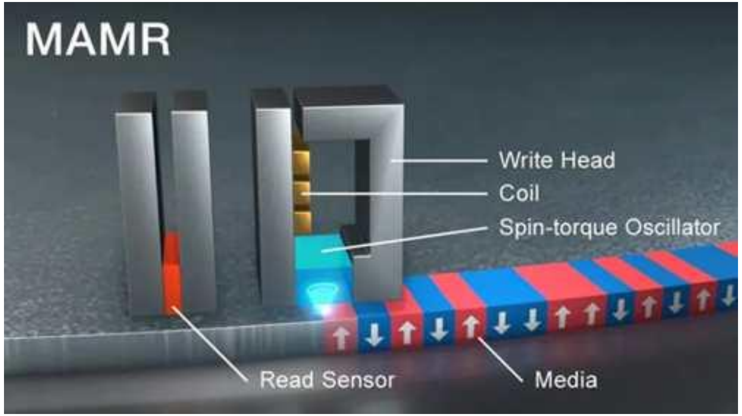
# Hard Disk Drives

## *Driving factors: Capacity: Areal Density: MAMR*

MAMR (Microwave Assisted Magnetic Recording)

Uses a microwave to “ease” the polarization of the media grains

Potentially reducing required magnetic field to ⅓

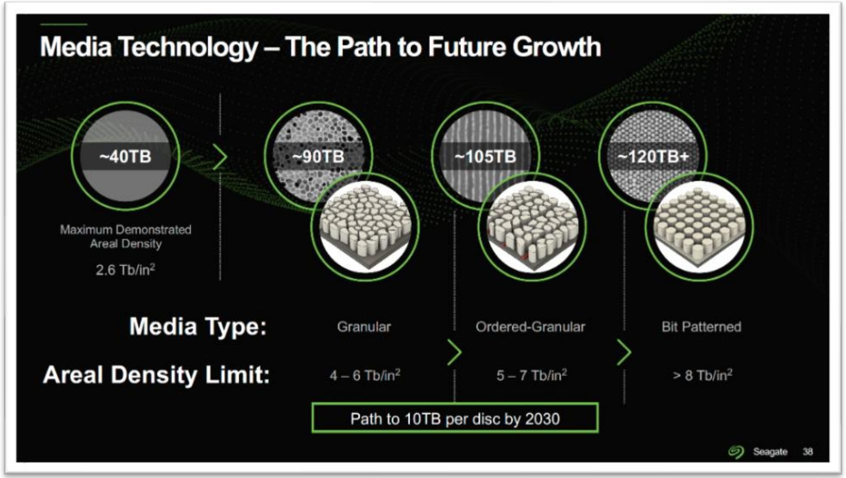



Source: storagenewsletter.com

# Hard Disk Drives

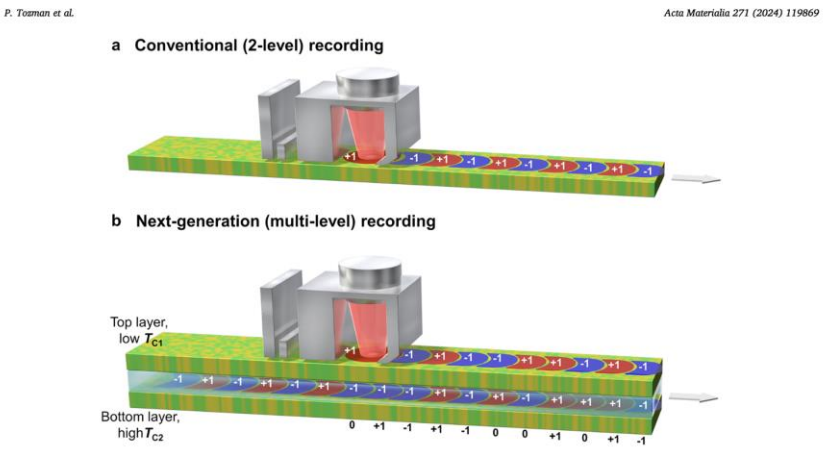
## *Driving factors: Capacity: Areal Density: Future: BPM and ML-HAMR*

BPM (Bit Patterned Media)




Source: Seagate

Multi-Layer HAMR




Source: Paper, Acta Materialia 271 (2024) 119869

# Magnetic tapes

# Magnetic tapes

**2006**

*”Tape is **dead**, Disk is Tape, Flash is Disk, RAM locality is king”*

Microsoft

**2015**

*”All cloud vendors will be using tape and will be using it at a **level never seen before**”*

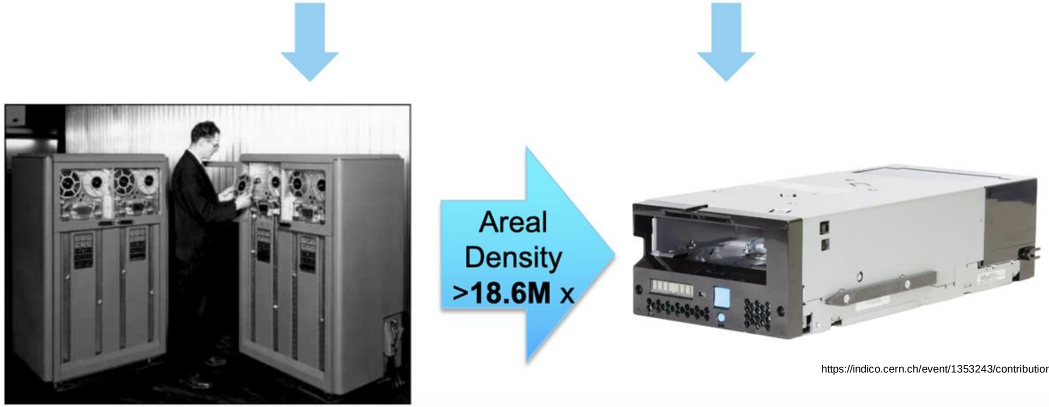
Microsoft

https://indico.cern.ch/event/1353243/contributions/5846917/attachments/2821388/4927105/Lantz_Future_of_Tape_CERN_18032024.pdf

# Magnetic tapes

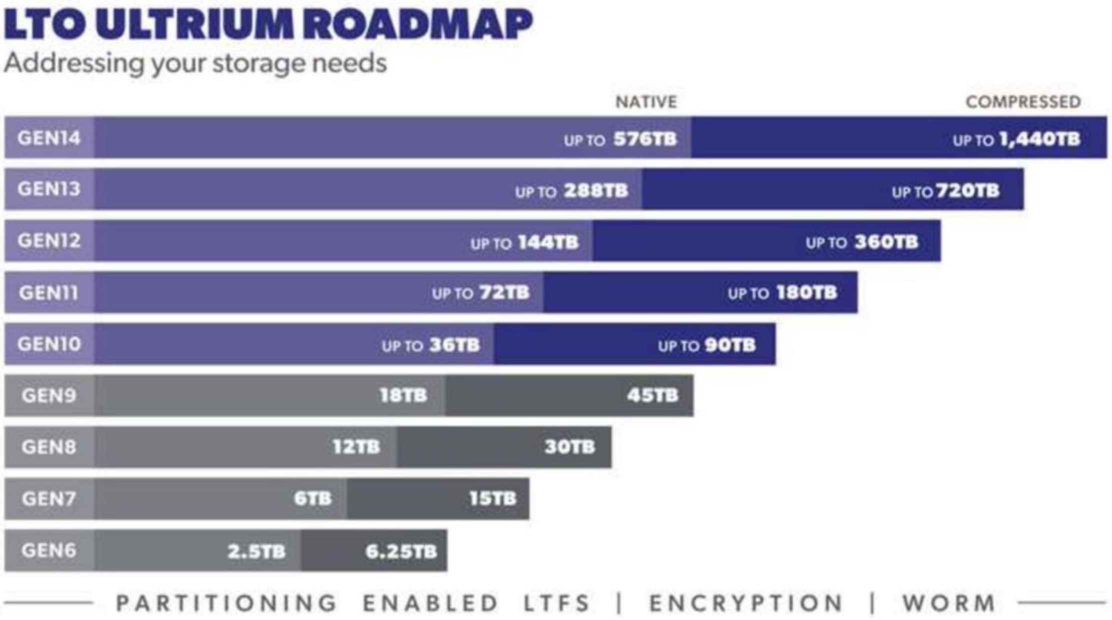
*Scaling*

| Product<br>Year | IBM 726<br>1952 | LTO9<br>2021 | **TS1170**<br>**2023** | Demo 2017<br>Sputtered Tape | Demo 2020<br>SrFe Tape |
|---|---|---|---|---|---|
| Capacity | 2.3 MB | 18 TB | **50 TB** | 330 TBytes | 580 TBytes |
| Areal Density | 1400 bit/in$^2$ | 11.9 Gbit/in$^2$ | **26.1 Gbit/in$^2$** | 201 Gbit/in$^2$ | 317 Gbit/in$^2$ |
| Linear Density | 100 bit/in | 545 kbit/in | **555 kbit/in** | 818 kbit/in | 702 kbit/in |
| Track Density | 14 tracks/in | 21.9 ktracks/in | **47 ktracks/in** | 246 ktracks/in | 452 ktracks/in |



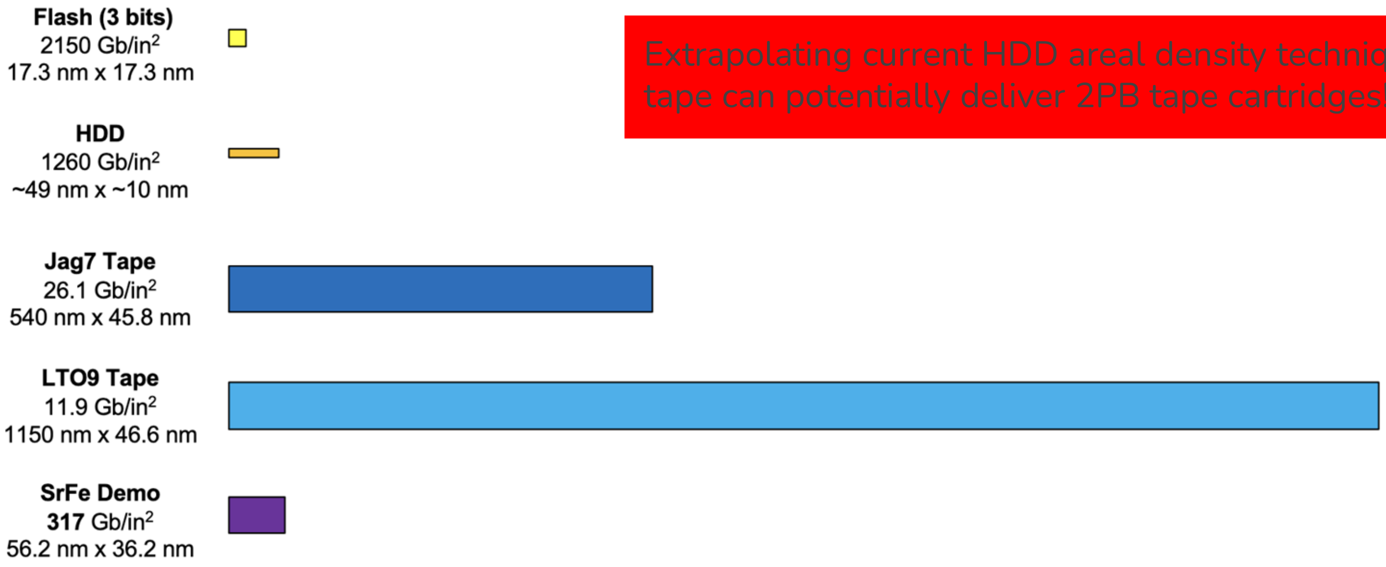

https://indico.cern.ch/event/1353243/contributions/5846917/attachments/2821388/4927105/Lantz_Future_of_Tape_CERN_18032024.pdf

# Magnetic tapes



# Density comparison across media

Extrapolating current HDD areal density techniques on tape can potentially deliver 2PB tape cartridges!




https://indico.cern.ch/event/1353243/contributions/5846917/attachments/2821388/4927105/Lantz_Future_of_Tape_CERN_18032024.pdf

→Most potential for future scaling of tape track density

# Flash: SSD and arrays

# Solid State Disks

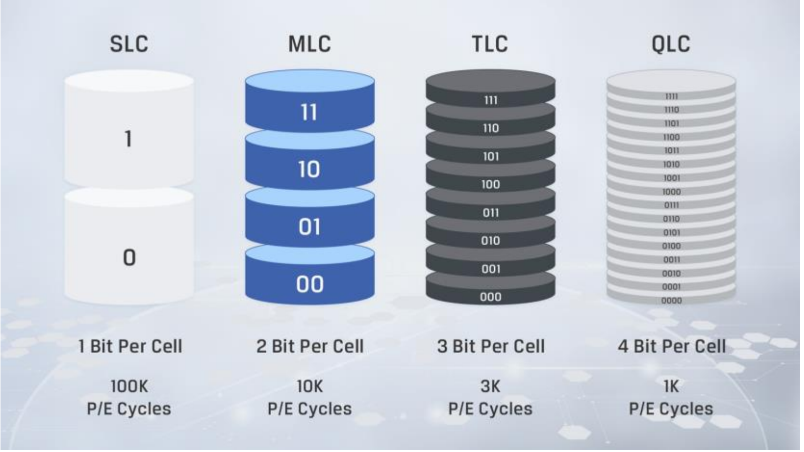
*Basics*

SSDs are made of non-volatile flash memory (NAND cells)

NAND cells can be electrically erased and reprogrammed, known as Program/Erase (P/E) cycles. Number of these cycles determines **endurance** of the device

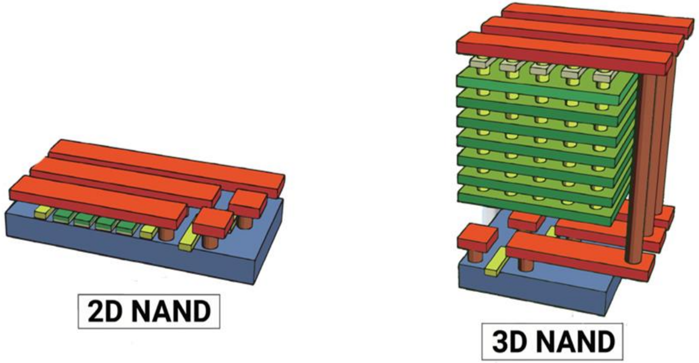
# Solid State Disks
*Basics*

NAND cell types




Source: Kingston

NAND layout




Source: theicttrends.com

# Solid State Disks
*Flash Arrays*


Flash arrays abstract the complexity of flash devices (the FTL, the leveling functions) to software, providing maximum flexibility for use-cases:

- Maximise performance (SLC)
- Maximise throughput (QLC)
- Mix

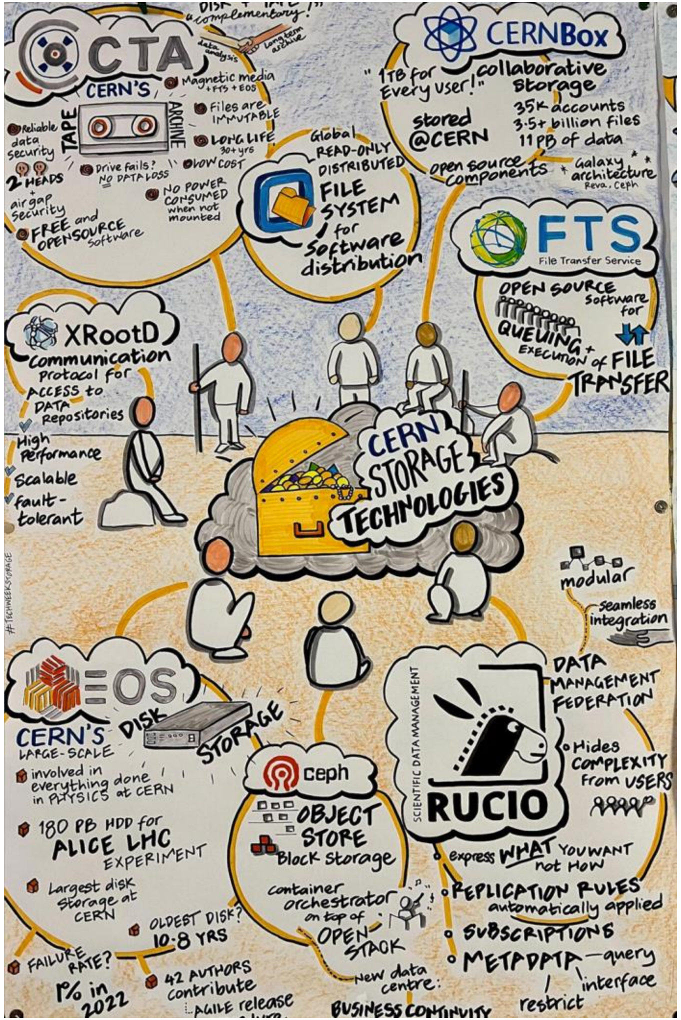
https://www.purestorage.com/knowledge/what-is-directflash-and-how-does-it-work.html?shareVideo=6330700995112

# Comparison



| | | | |
|---|---|---|---|
| *Throughput* | ~250 MB/s | 400MB/s | 3-12 GB/s |
| *IOPS* | Hundreds | - | Thousands |
| *Latency to 1st byte* | < 5 ms | Minutes | < 0.5ms |
| *Capacity* | Up to 32 TB | ~ 20/45 TB (compressed) | Up to 100 TB |
| *Price* | $$ | $ | $$$ |

# Disk and Tape Storage at CERN

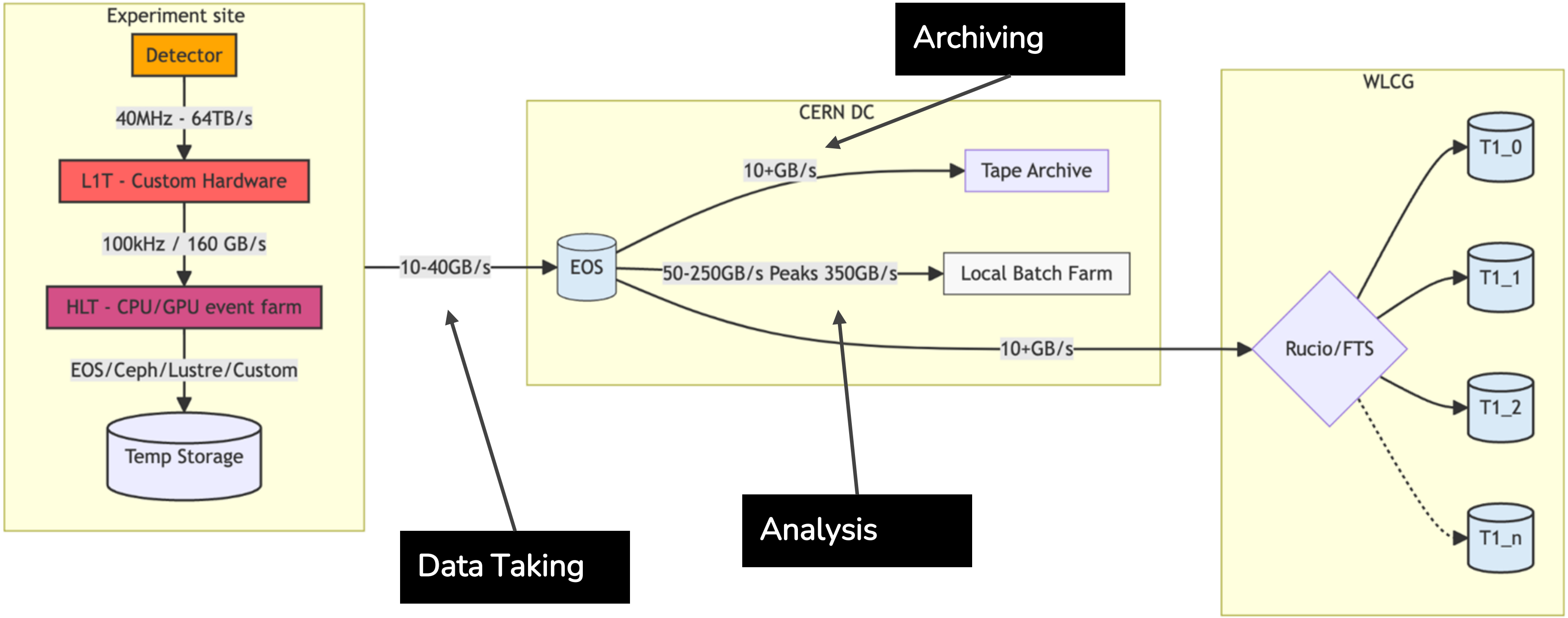



The **CERN IT Storage group** mission is to ensure coherent design, development, operation and evolution of storage and data management services at CERN for all aspects of physics, user and project data and general needs of the Laboratory.

For this presentation, we'll focus only on the two major open source systems developed in-house and used worldwide: EOS and CTA

# Data access patterns

Example for only one experiment!




Source: https://storagelecture24.web.cern.ch/#/5/1

# EOS and CTA

EOS

Disk-based system with dedicated “storage pools” with defined QoS Experiments’ Data Management frameworks manage the transitions to tape

Low-Latency namespace

POSIX-like file access

From RAID to **RAIN**

CERN Tape Archive

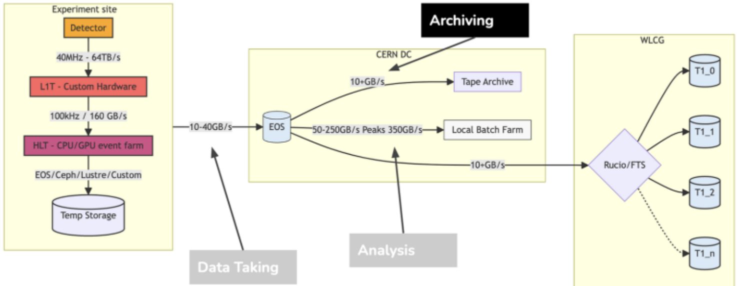
Tape-based system with fast (flash) disk buffer

Tape-backend of EOS

Supports PostgreSQL as namespace (used to be Oracle only)

Evolution of CASTOR (30+ years of tape experience)

# Why Tape?




Good fit for archiving use-cases

Reliability

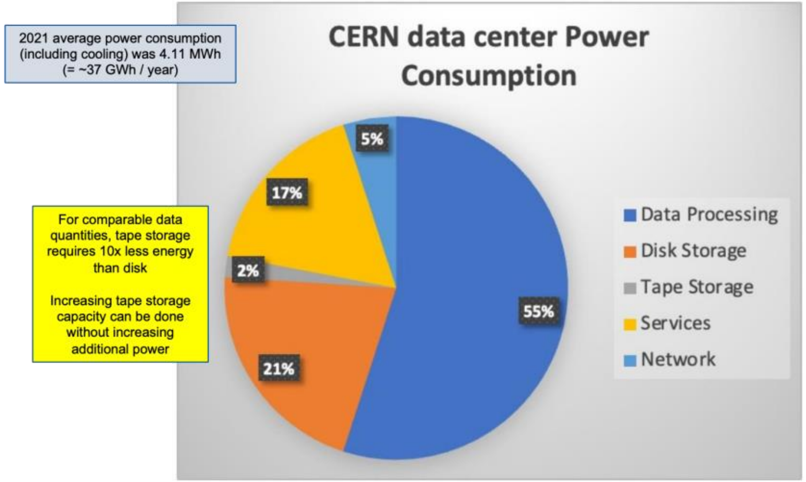
*Uncorrectable Bit Error Rate*: LTO-9 tape cartridge ($10^{-20}$) is 10 000x more reliable than typical 18 TB hard disk drive ($10^{-15}$)

Annual failures at CERN: 1% hard disks vs. 0.005% tape cartridges (~200x less)

Separation of media and data access device: No data loss if the drive fails

Long media lifetime (30+ years)

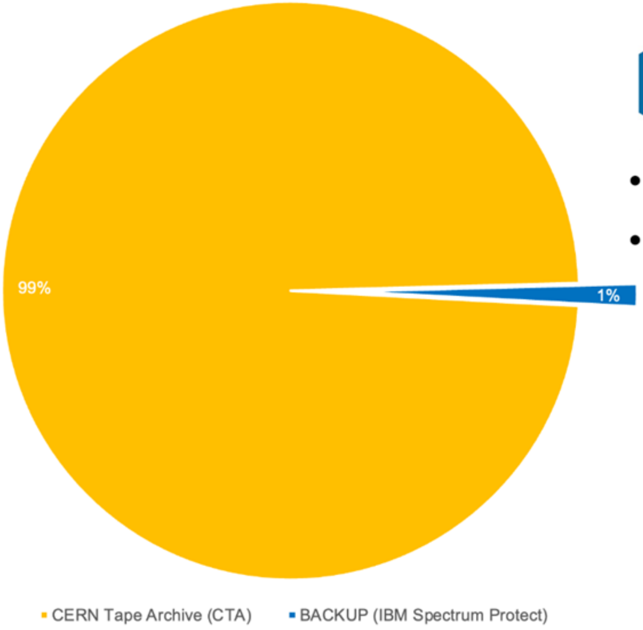
Energy efficiency



# Tape use-cases

CERN Tape Archive

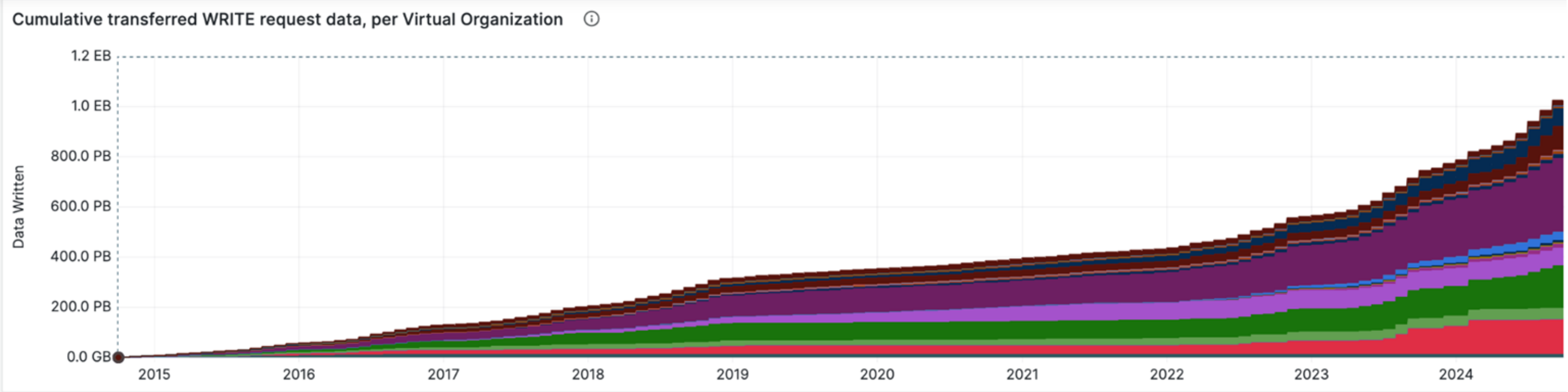
- Archive of the physics data
- Provisioned capacity: ~1.2 Exabytes

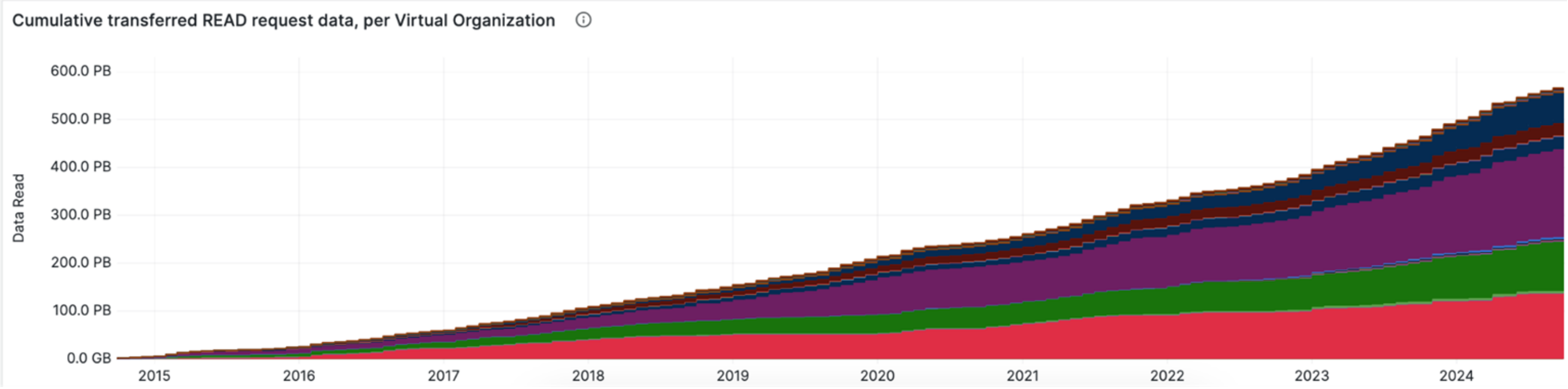



IBM Spectrum Protect

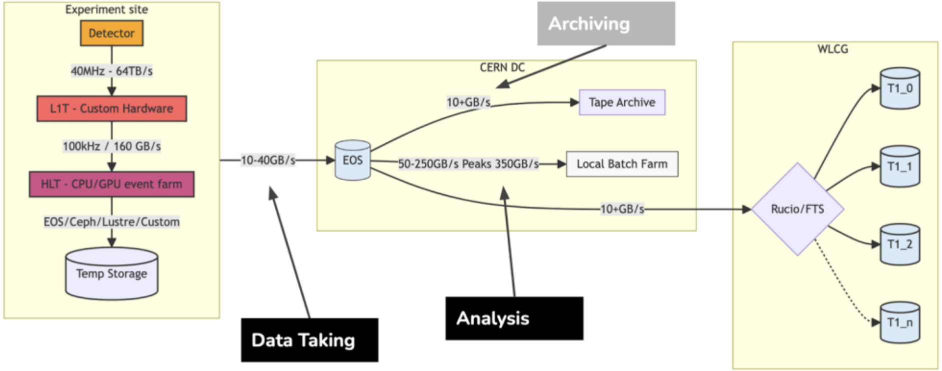
- Backup of the business data
- Licensed capacity: ~15 PB

# Cumulative writes and reads to tape



# Why Disk?




## Analysis use-case

100K clients streaming data from over 100k disks.

1-150MB/s throughput per stream

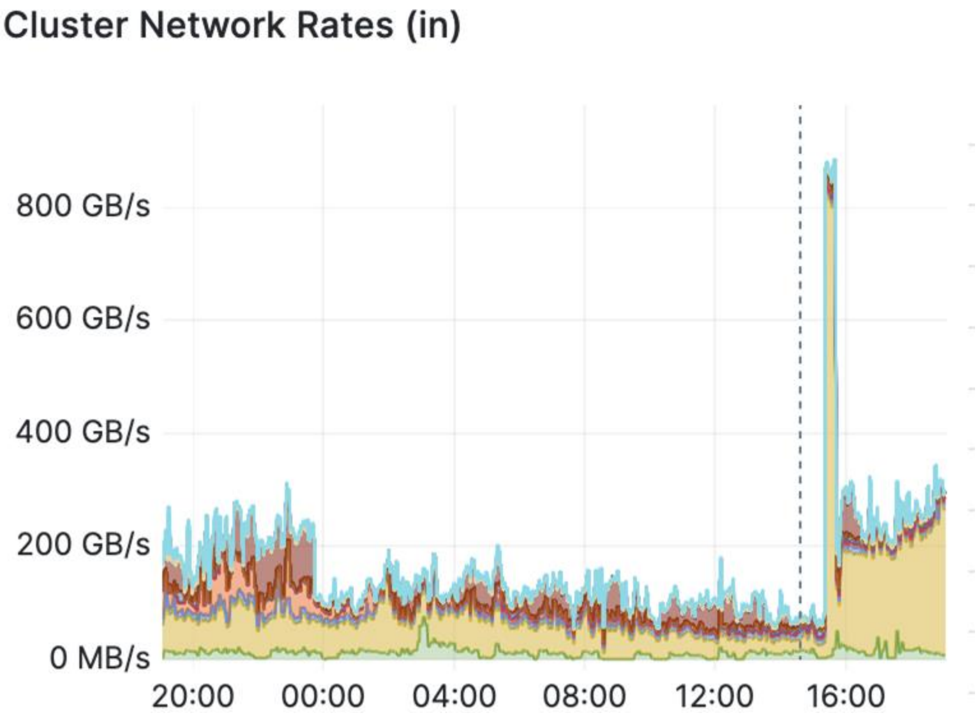
"Similar to having 100K people watching Netflix and skipping the boring parts"

## Data taking use-case

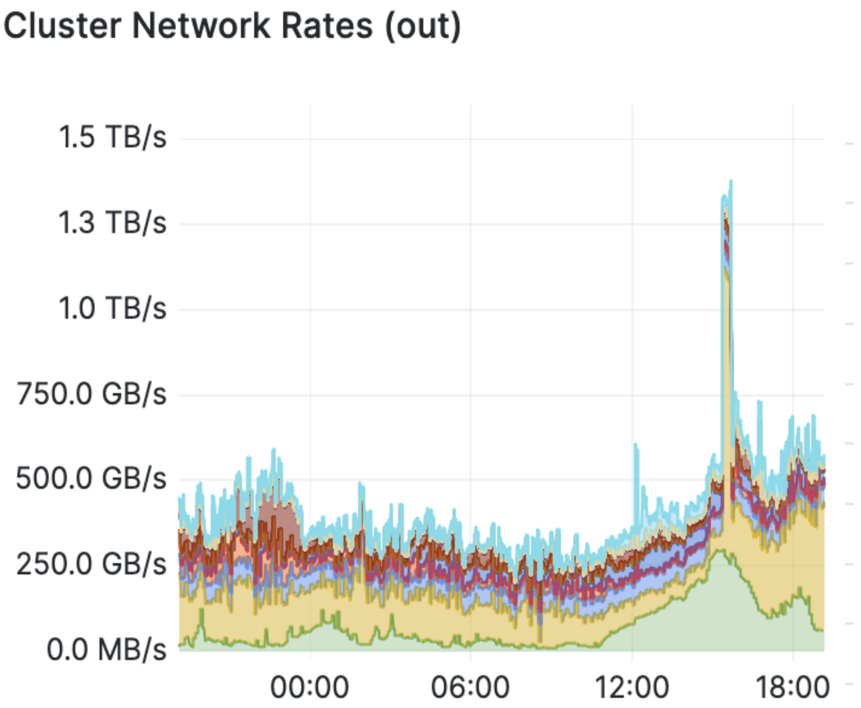
100s of clients streaming as fast as possible

0.5-1GB/s per stream

```
[gonzalhu@lxplus982 gonzalhu]$ dd if=/dev/zero of=/eos/user/g/gonzalhu/bigfile.txt bs=500M count=1
1+0 records in
1+0 records out
524288000 bytes (524 MB, 500 MiB) copied, 3.41433 s, 154 MB/s
```




WRITES




READS

EOS data rates last 24h

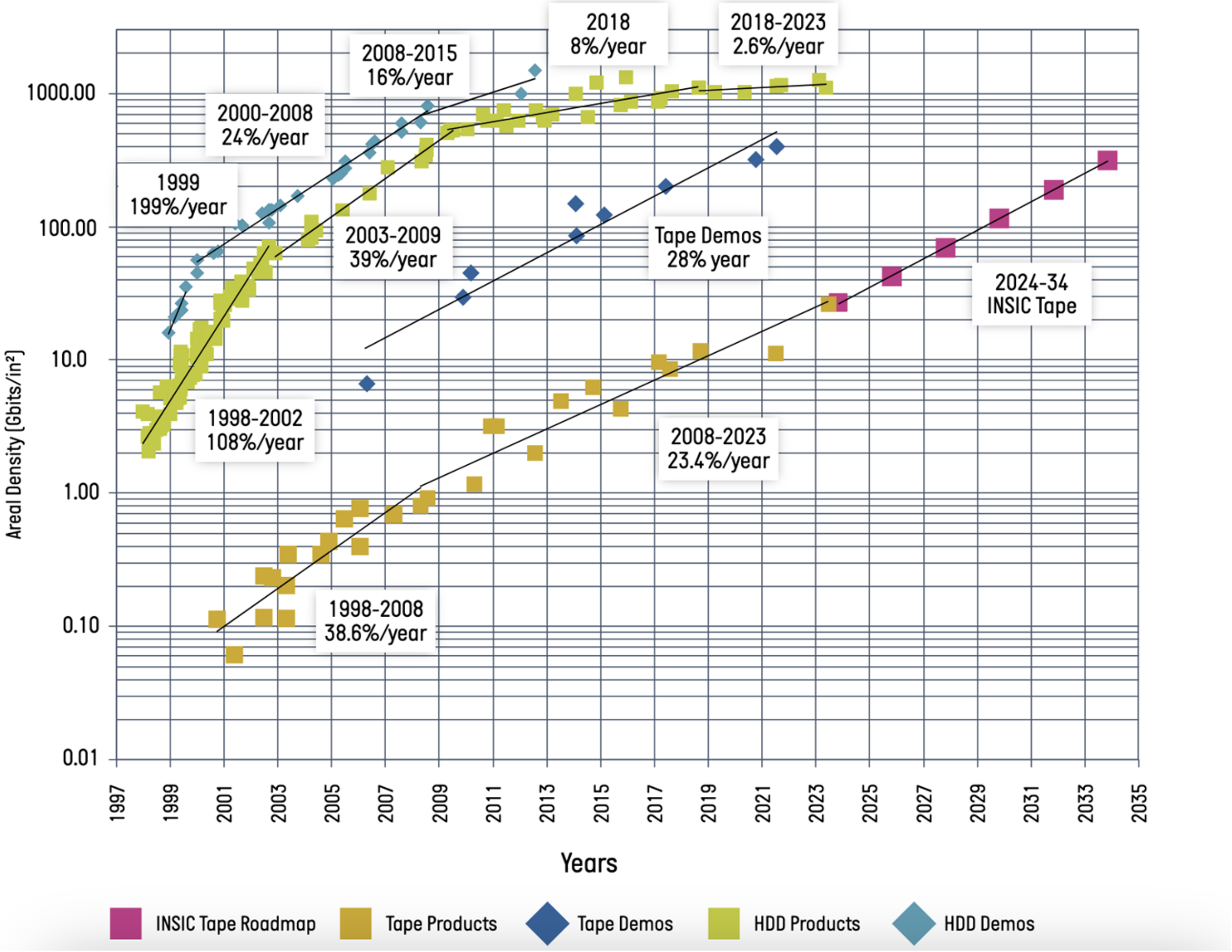
# CERN’s approach

For Tape: Introduce the latest tape technology as it becomes available
For Disk: Purchase the cheapest $/TB hard disk drives

Outcomes (rough estimates):

- tape storage is ~3x cheaper than disk
- 50% disk capacity and 50% tape capacity
- Tape: 0.6 reads per 1 write
- Disk: 5 reads per 1 write

# Market




Source: https://www.lto.org/wp-content/uploads/2024/07/INSIC-International-Magnetic-Tape-Storage-Technology-Roadmap-2024.pdf

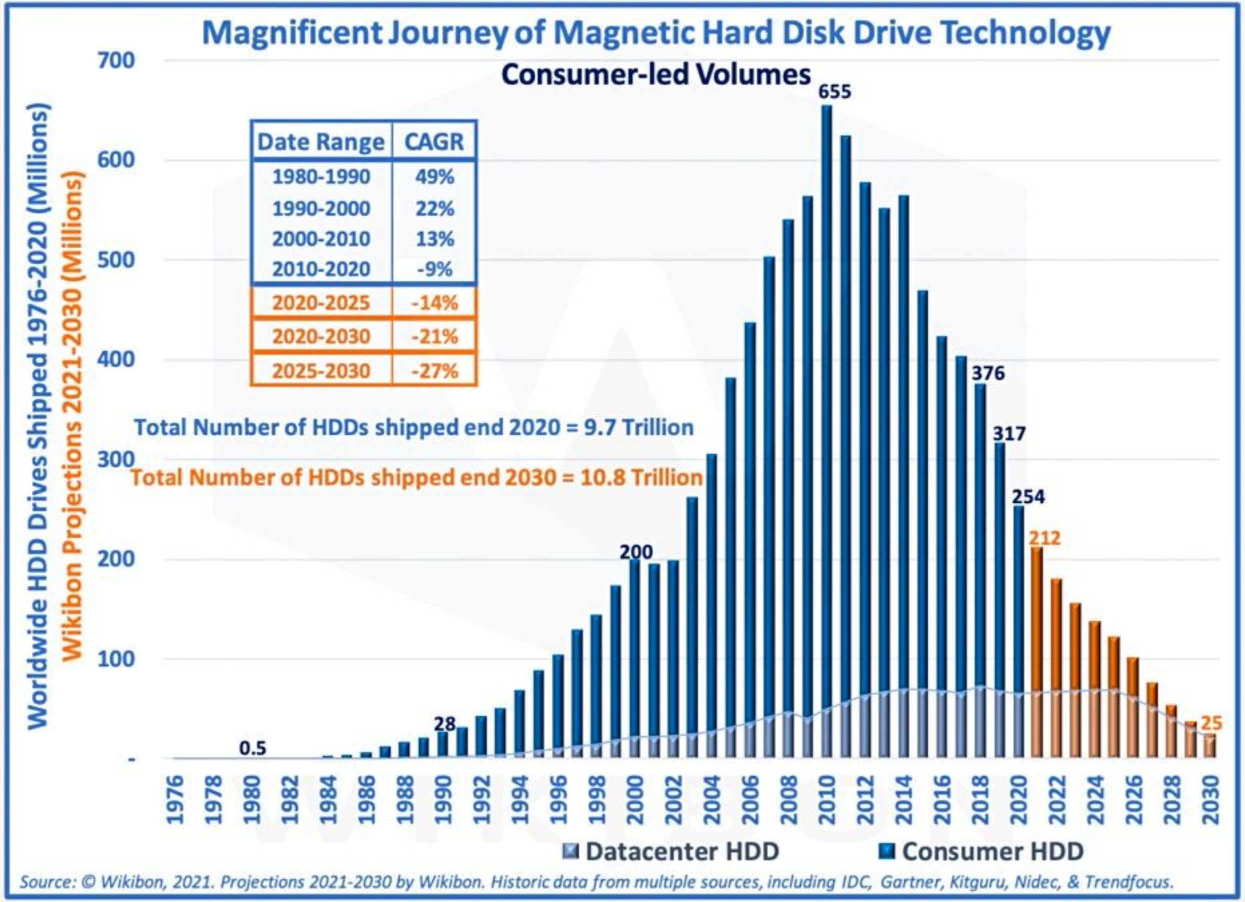
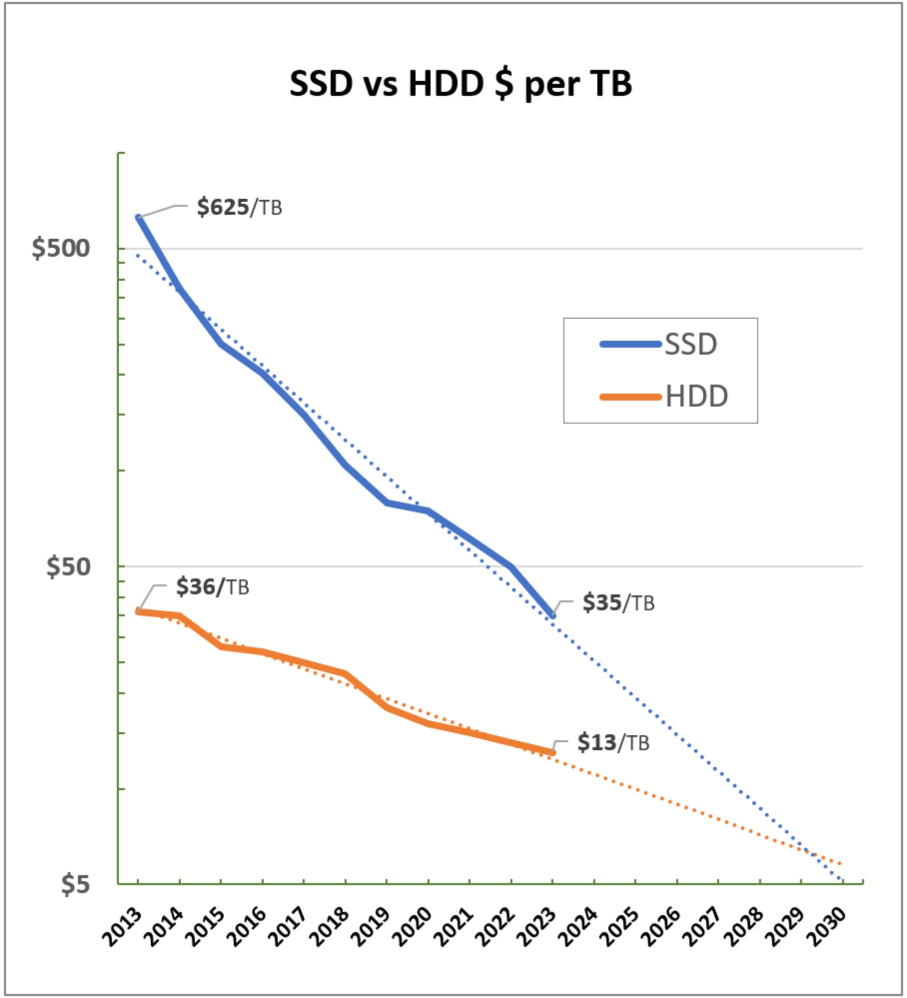



**Figure 7 – History and Wikibon Projection of HDD Shipments (millions)**

*Source: © Wikibon 2021*

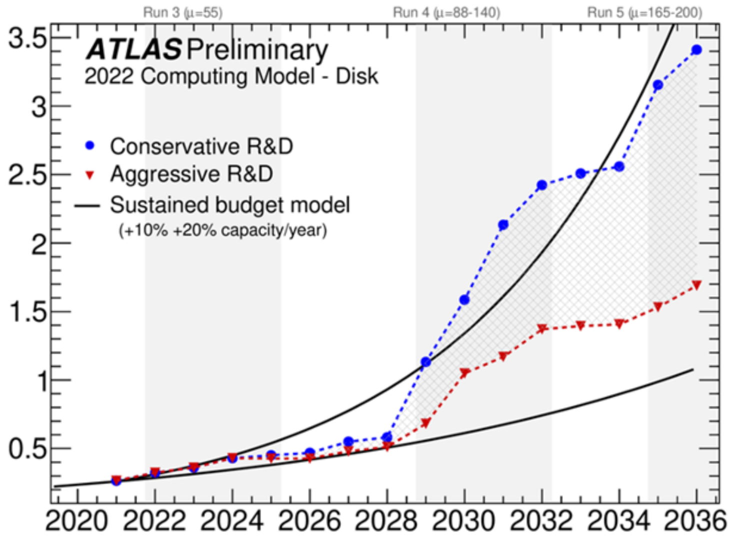


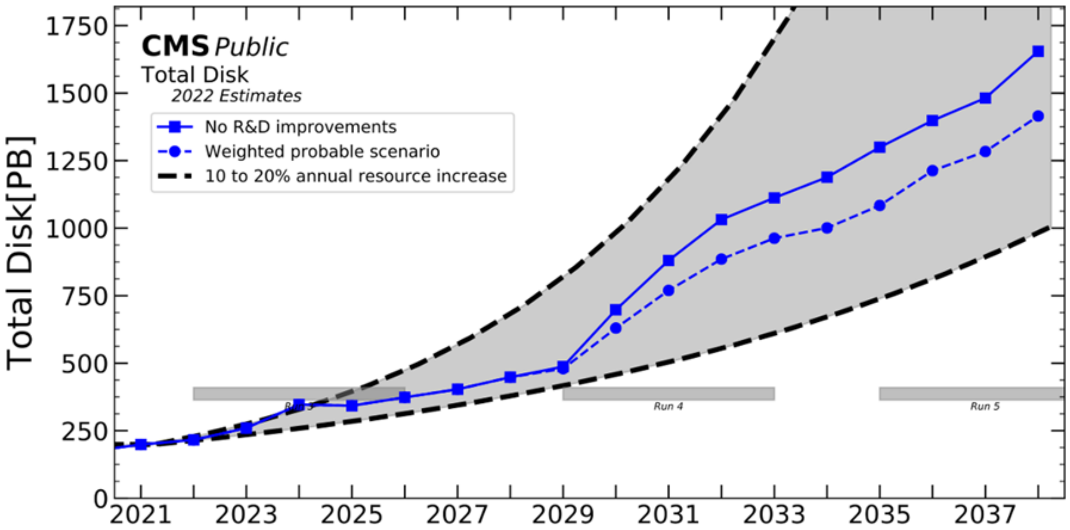

- Some experts forecast there will a crossing point where flash will be more cost-effective than HDD around 2030
- Other experts coincide that this point will not be reached by that time
- Some experts coindice that HDD market will be very small and that existing companies will create a consortium  to still benefit from this market. For example: WD created two new companies: one for HDD and one for flash
- Wrights' Law: "the more efficient vendors are making flash, the cheaper it will be"
- Some companies will probably jump before the crosspoint is met (compounded costs, including power,cooling, … are less for flash)

Source: DataHoarder

# Challenges

# ATLAS and CMS storage predictions



# Challenges for tape

### The good old days

- Previously, tape drives could
  - read current + last 2 generations, and
  - write current + last 1 generation

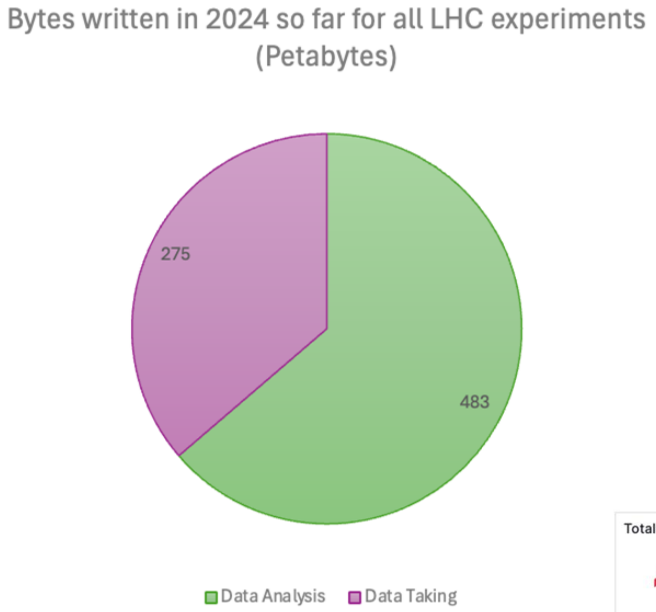
  of tape media
- Older media could be upformatted for use in newer drives

### Today that has changed

- Tape technology evolution is being driven by requirements from hyperscalers
- Emphasis on greater capacity rather than backwards compatibility
- Jump from 20 TB → 50 TB cartridges
- But NO backwards compatibility

Consequence: We need to **repack tapes** on a much more aggressive schedule than in the past

# Challenges for disk




Data taking account  for roughly 36% of all data the written and only for 5% of the write streams into the system

Data taking rates are *predictable*, analysis is not

# Challenges for disk

Experiments **pledges** are on **capacity**, performance is provided for “free”, however the disk market driven towards high density disks, which have a significant penalty in performance (throughput, IOPS).

**HDD throughput is stale at 250MB/s** and is not going to change any time soon

CERN disk infrastructure runs with ~1K disk servers accounting for 100K disks.

The number of parallel streams per disk is ~2 -> 220K parallel streams across whole cluster

With new disk servers (2024): 120disks x 24TB drives with 100Gb interfaces, Hyper-optimized $/TB, it will result in:

- ~ 300 disk servers (replica) -> 36K disks-> 72K parallel streams -> 3x times load per HDD
- ~ 185 disk servers (EC 10+2) -> 22K disks -> 44K parallel streams -> 5x times more load per disk

# How to increase disk capacity without losing performance with linear budget?