The CERN Tape Archive : Preparing for the Exabyte Storage Era

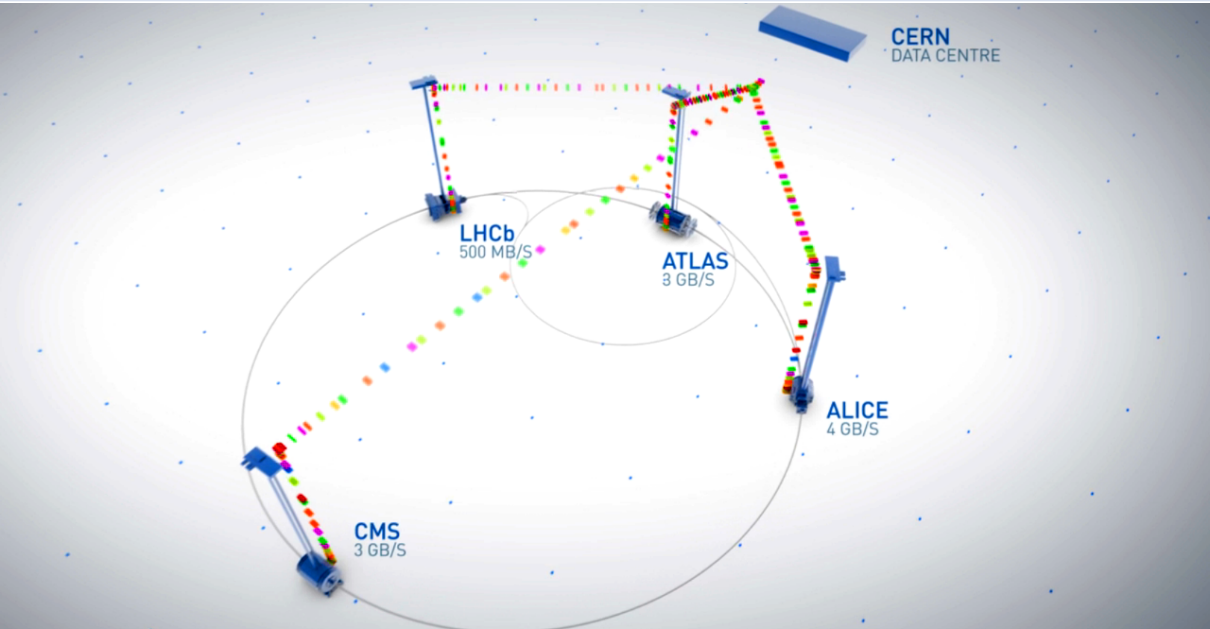
Michael Davis

CERN • IT Department • Storage Group

PASC19 • 14 June 2019

Data Management at CERN

See: Processing LHC Data (video)



Data Management at CERN

Data Storage at CERN Tier–0

Online Storage

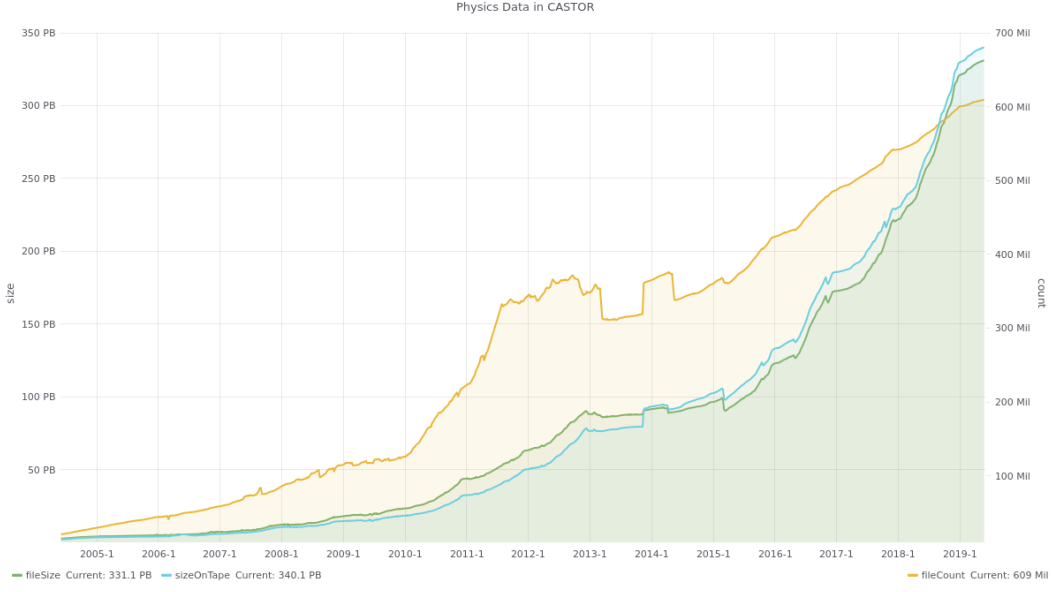
- 90 000 disks (85% HDD/15% SSD)
- Nominal capacity 280 PB
- Effective capacity is lower due to redundancy

Nearline Storage

- 29 500 tape cartridges
- 609 million files / 340 PB
- 115 PB written to tape in 2018
- 15.8 PB written to tape in November 2018
- Data retrievals exceed 1 EB/year

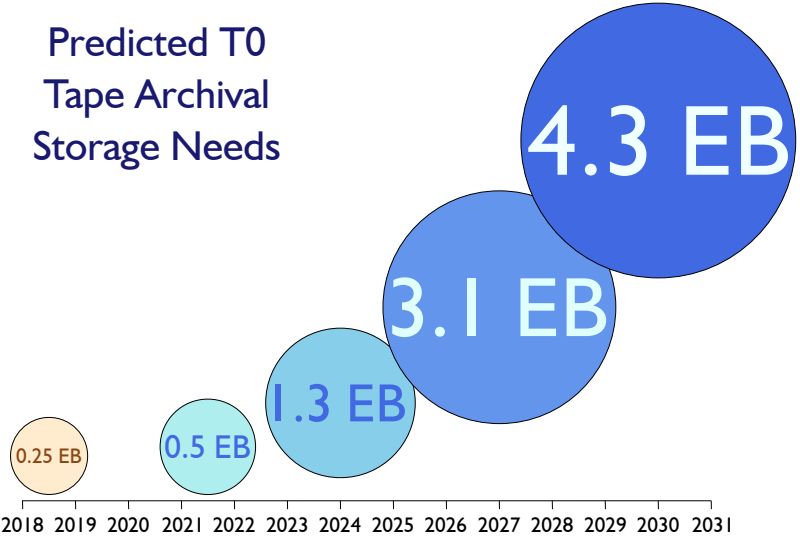
Data Management at CERN

Tape Archival Storage



Data Management at CERN

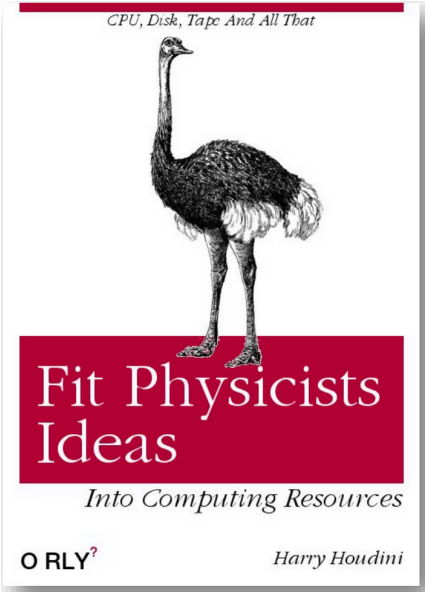
Scaling up for Run 3 and HL–LHC



See: Experiences and challenges running CERN's high capacity tape archive (2015)

Data Management at CERN

Scaling up for Run 3 and HL–LHC



CERN Tape Archival Storage

CERN Tape Archival Storage

Advantages of Tape

Energy efficiency

- CERN T0 Data Centre has an electrical capacity limit of 3.5 MW
- Disk storage requires power and cooling
- Tape cartridges only require power when mounted for read/write operations
- **Tape capacity can be increased without requiring additional power**

Reliability and Data Security

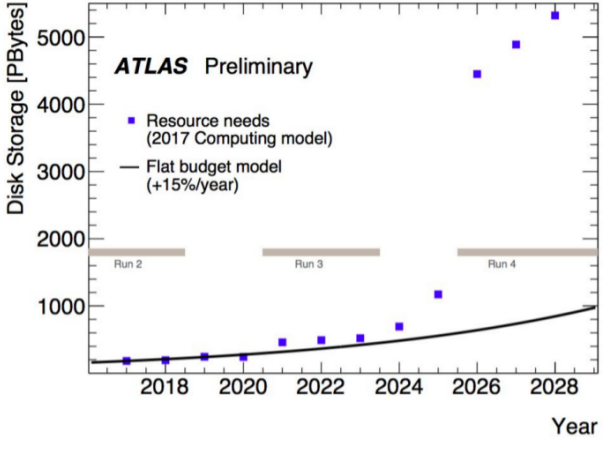
- Read after write verification
- No data loss in case of drive failure
- Immutable files
- Air-gap security

Long media lifetime (30+ years)

CERN Tape Archival Storage

Advantages of Tape

Cost

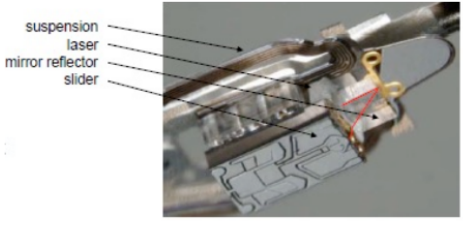


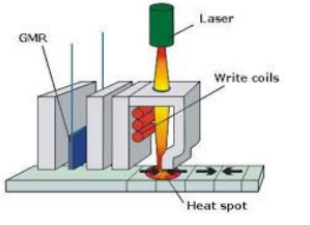
- Storage needs are increasing exponentially
- Budget is increasing linearly
- **Tape storage is 6.7× cheaper than hard disks**
  [The Clipper Group (2015)]
- Will this cost advantage continue into the future?

The Future of Magnetic Storage

Next generation of HDD technologies [The Future of Tape (2018)]

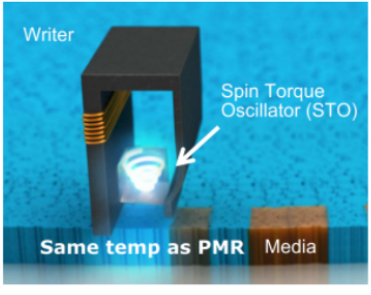
Heat Assisted Magnetic Recording (HAMR)





- **Life time of laser / near field transducer**
- **Added cost**
- **Compatibility with 8 platters?**
- **Areal Density Scaling: CAGR 15-20% ???**

Microwave Assisted Magnetic Recording (MAMR)



→Spin torque oscillator used to locally generate magnetic fields that rotate at microwave frequencies and lower the magnetic field required from the writer to write a bit

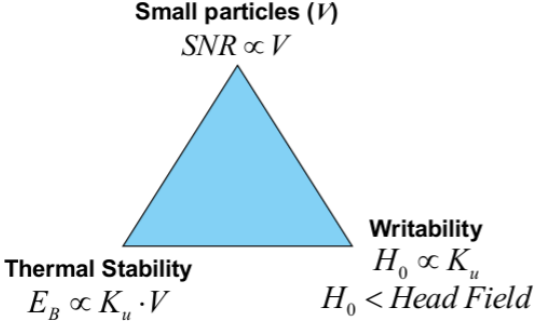
- Minimal additional cost
- Less disruptive than HAMR (more PMR-like)
- Does not use heat → better reliability than HAMR
- Still many challenges to scale:

→**projected to enable scaling at ~15% CAGR up to ~40TB capacity**

The Future of Magnetic Storage

The Superparamagnetic Limit [The Future of Tape (2018)]

Magnetic Media “Trilemma”:



- HDD areal densities have reached the superparamagnetic limit
- Tape operates at areal densities far from the superparamagnetic limit

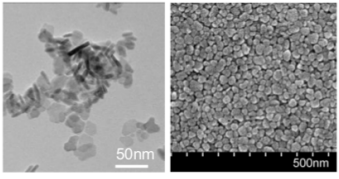
The Future of Magnetic Storage

Technologies to Continue Tape Scaling: BaFe Media—123 Gb/in^2 demo (220 TB Cartridge) [TFOT2018]

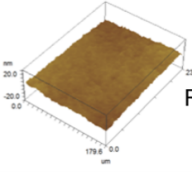
FUJIFILM IBM

Key technologies for advanced tape media

1. Fine magnetic particles (1600 nm^3) w/ high coercivity (2800Oe) → archival life

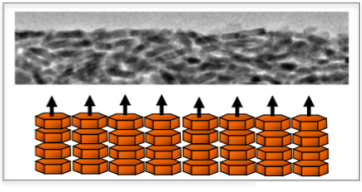


2. Smooth surface with low friction



Ra 0.9nm

3. Perpendicular orientation of magnetic particles



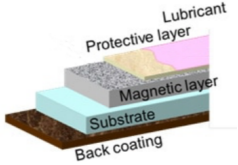
The Future of Magnetic Storage

Technologies to Continue Tape Scaling: Sputtered Media—201 Gb/in^2 demo (330 TB Cartridge) [TFOT2018]

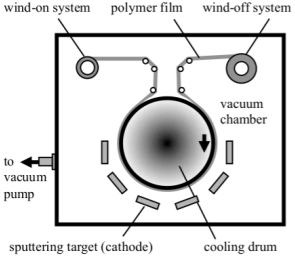
Key technologies for advanced sputtered tape media

1. Ultra-small grain size with high coercivity → archival lifetime
2. Perpendicular orientation of magnetic grains
3. Smooth surface with new lubricant bonded to tape surface
4. Advanced sputter technique for low defect density

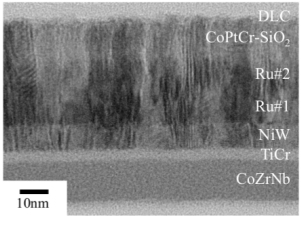
SSMD Sony Storage Media and Devices Corp. IBM



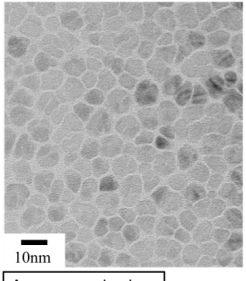
Reel-to-reel sputter coating system



TEM X-section of media stack



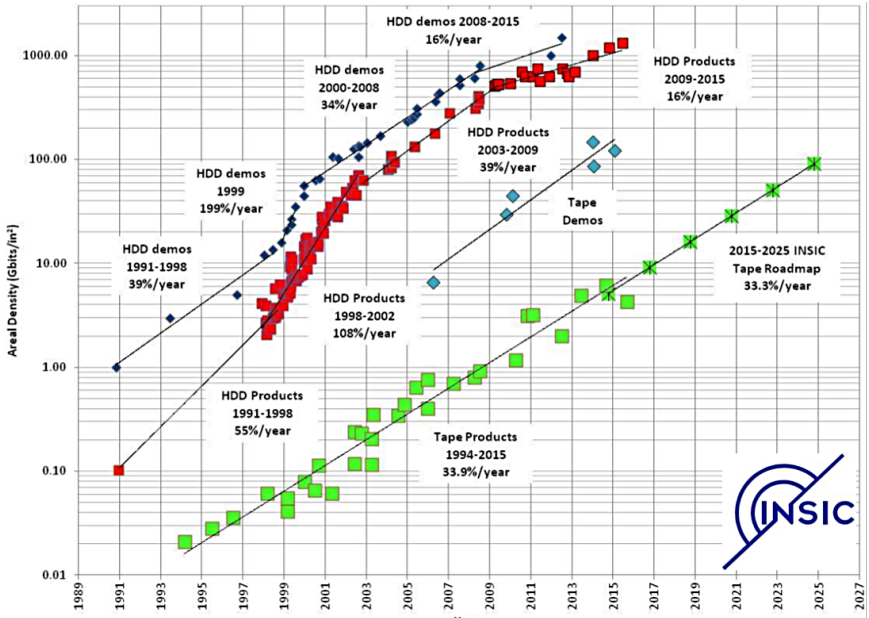
SEM Top View



Average grain size:
6.6 nm, σ = 1.2 nm

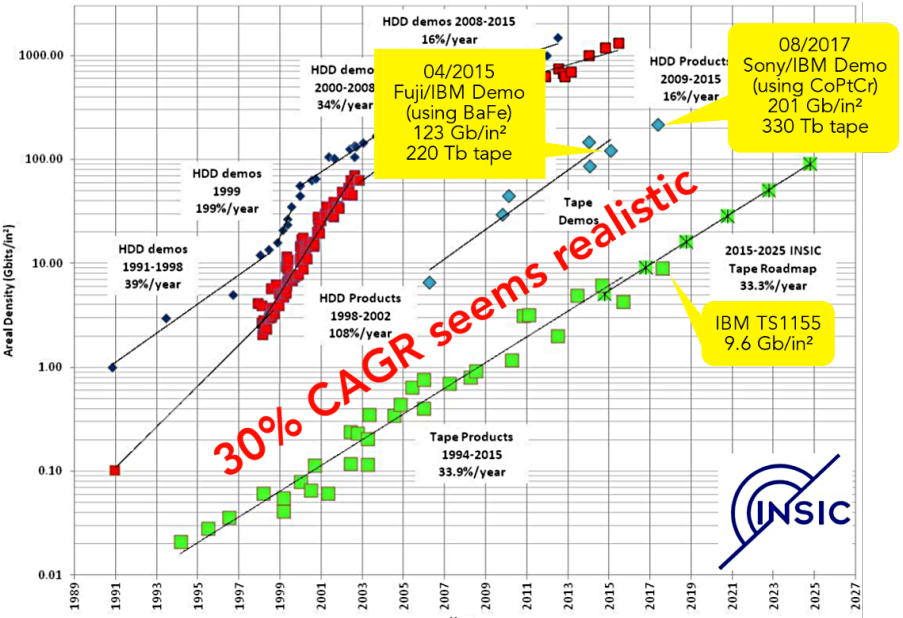
The Future of Magnetic Storage

Areal Density Scaling [INSIC (2015)]



The Future of Magnetic Storage

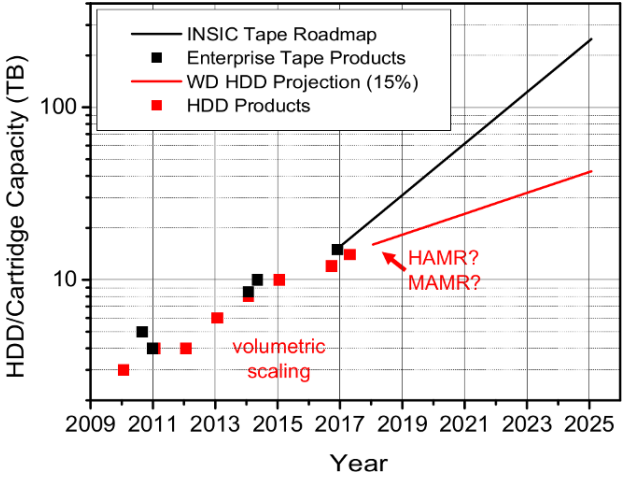
Areal Density Scaling [INSIC (2015)]



The Future of Magnetic Storage

Tape and HDD Projections

- For HDD, HAMR/MAMR are disruptive technologies which require large sustained engineering efforts. Optimistically, 15% CAGR
- For tape, 30% CAGR seems realistic



The cost advantage of tape will continue to grow:

6.7x Today

38x by 2025!

The Future of Magnetic Storage

Tape Market Risks

- Technology outlook very positive in terms of projected improvements in tape capacity

BUT:

- Oracle withdrew from tape drive manufacturing/R&D in 2017, leaving IBM as the sole provider of tape drive technology
- Only two remaining media manufacturers (Fujifilm, Sony), who are locked in a patent war which has disrupted supply of LTO–8 media
  [The Register (May 2019)]
- Price decline of tape media has slowed in recent years (≈50% of TCO)
- Tape IO speed not increasing at the same rate as capacity

The Future of Magnetic Storage

Alternatives to Tape

Cloud

- Cheap to put data in, expensive to get it out LOBSTER POT
- Danger of vendor lock-in
- Better suited to backup/compliance use case than active archive use case
- CERN does use cloud to add additional compute capacity

Other Storage Technologies

- **SSD:** Price/TB expected to be $\approx 10\times$ HDD price/TB for the forseeable future
- **Optical:** Sony and Panasonic have proof of concept demos for "tape-like" optical disc libraries, but no timeline to new products
- Other technologies (holographic, DNA) promise high density and high reliability, but no sign of any products as yet

The Future of Magnetic Storage

Summary

- Tape is the best currently-available technology for archival storage, in terms of reliability and stability over long periods of time
- CERN power and cooling constraints make it difficult to add more disk storage capacity; tape storage capacity can easily be increased
- Cost benefits over HDD are significant and look set to increase over the next decade
- **CERN is investing in tape as its primary archival storage medium for LHC Run–3 and Run–4**
- …but always with an eye to the market risks and alternatives

Tape Infrastructure at CERN

…since the 1960s

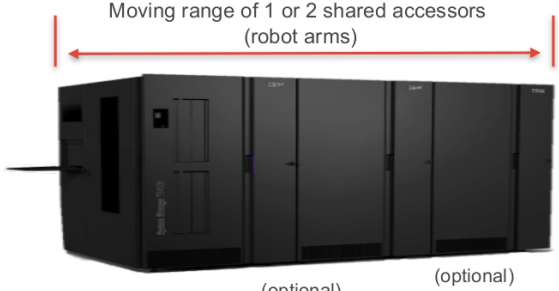
Tape Infrastructure at CERN

54 years later...

Tape Infrastructure at CERN

Tape Libraries

Tape Library Components



L frame:
Controller
Drives
Cartridges

(optional)
D frames:
Drives
Cartridges

(optional)
S frames:
Cartridges

Workload

- Write/read data rate and pattern
- Data retention policy
- Backup/restore
  - Full or partial
  - SLA: time to restore all or a percentage of data
- Active archive write/read
  - Independent requests
  - SLA: average data access time

Technology

- LTO: LTO7, LTO8
- IBM Enterprise
  - Recommended access order (RAO) function

Tape Infrastructure at CERN

Physics Archive Use Case

	IBM	Oracle
Libraries	3× TS4500	1× SL8500
Drives	46× TS1155 20× LTO–8	10× T10000D
Media	15000× JD media (15 TB) 6000× JC media (7 TB) 4500× LTO–7M media (9 TB) 160× LTO–8 media (12 TB)	3000× T2 media (8 TB)

- 10 PB disk cache
- 340 PB on tape
- 15 PB free capacity

Tape Infrastructure at CERN

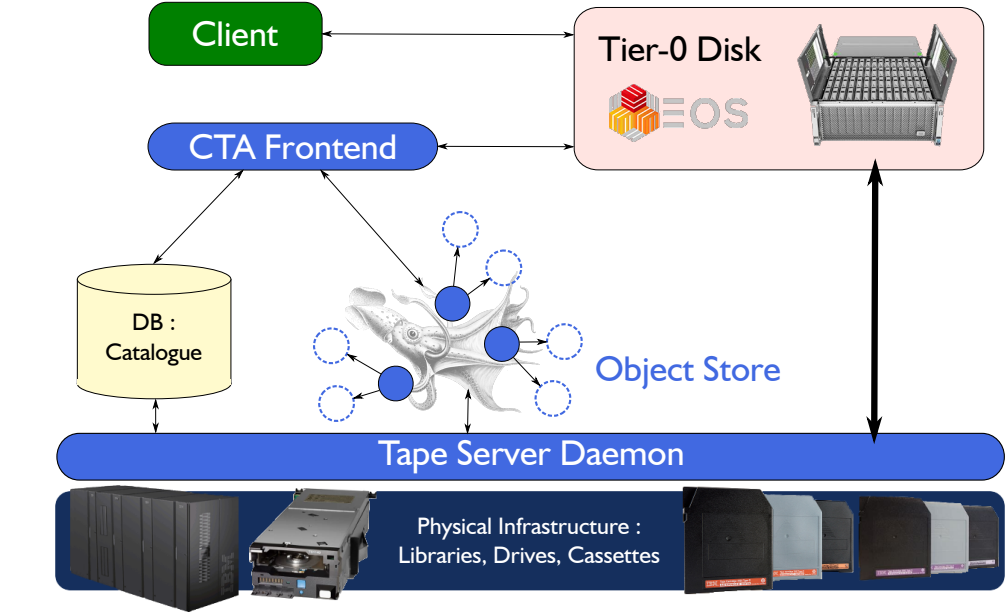
Backup Use Case

	IBM
Libraries	2× TS3500
Drives	50× TS1140
Media	200× JC media 12000× JB media

- 18× TSM 7.1.4 servers
- 8 PB on tape (2.3 billion files)

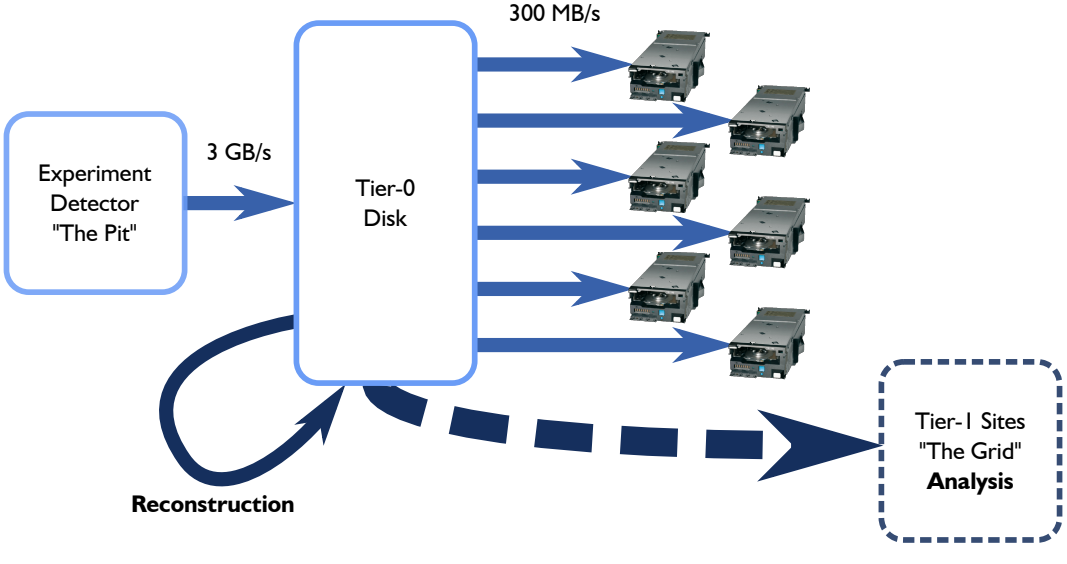
Tape Infrastructure at CERN

CTA (CERN Tape Archive)



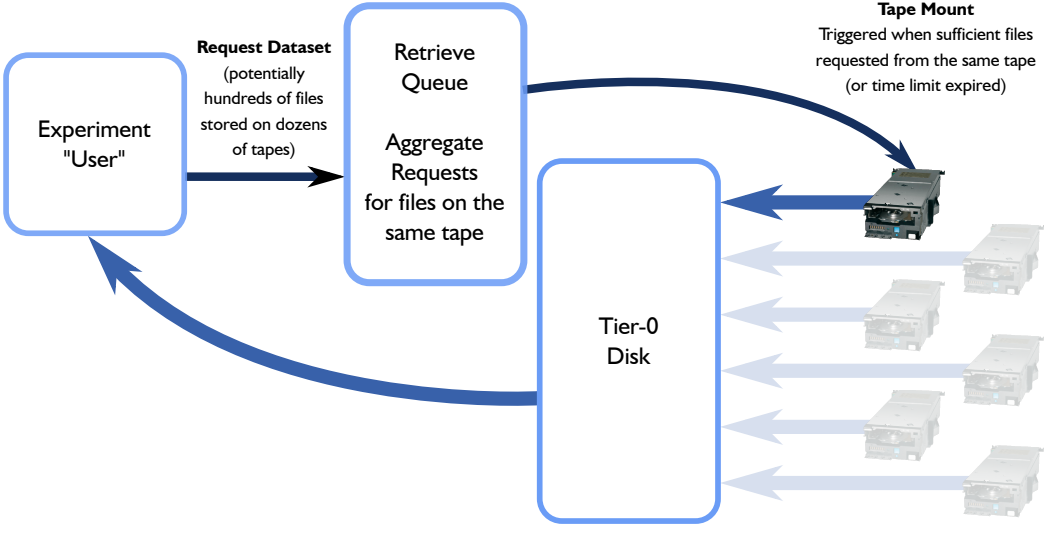
Physics Data Flows

Archival workflow



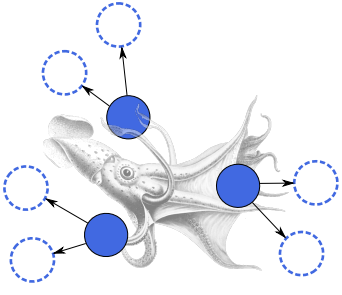
Physics Data Flows

Retrieval workflow



Scaling Up for Exabyte Storage

New Queueing System



Object Store

- Ceph RADOS Object Store for transient data
  - Archive and retrieval requests
  - Queues
- Distributed architecture
  - Multiple Frontend servers to field requests
  - Multiple redundant RADOS instances
  - Scale-out scalability
- Resilience against crashes
  - No direct Inter-Process Communication
  - Guaranteed coherency
  - Agent heartbeat, garbage collection
- No single point of failure

Scaling Up for Exabyte Storage

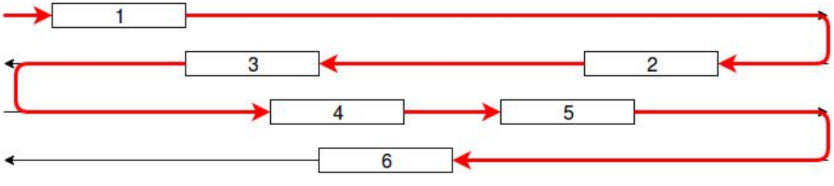
New Scheduler

CASTOR	CTA
Scheduling decisions made at time of user request.	Scheduling decisions made at time of tape mount.
Tape drive may not be available when job reaches the front of the queue.	*Tape drive allocated when job reaches the front of the queue. Reduced latency.*
High-priority jobs cannot interrupt running jobs.	High-priority jobs can preempt lower-priority jobs.
	Can switch from repack to data taking and back without operator intervention. System operates at full capacity at all times.

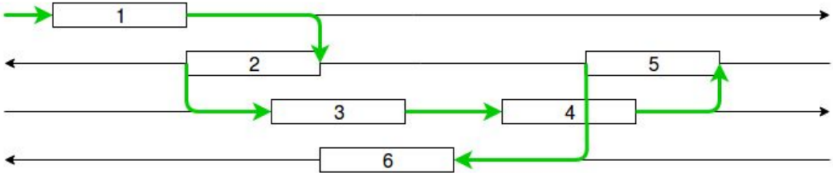
Recommended Access Order

See: Enhancing the low-level tape layer of CERN Tape Archive software (2017)

Recall Files in Sequential Order

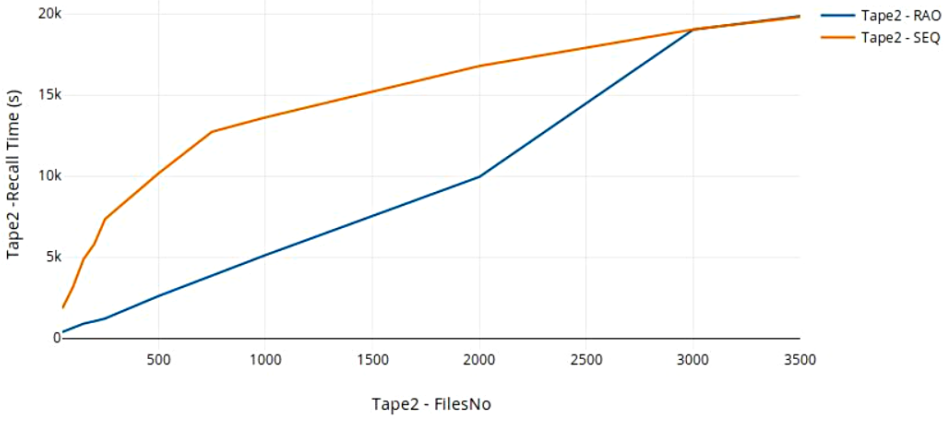


Recall Files in Recommended Access Order



Recommended Access Order

See: Enhancing the low-level tape layer of CERN Tape Archive software (2017)



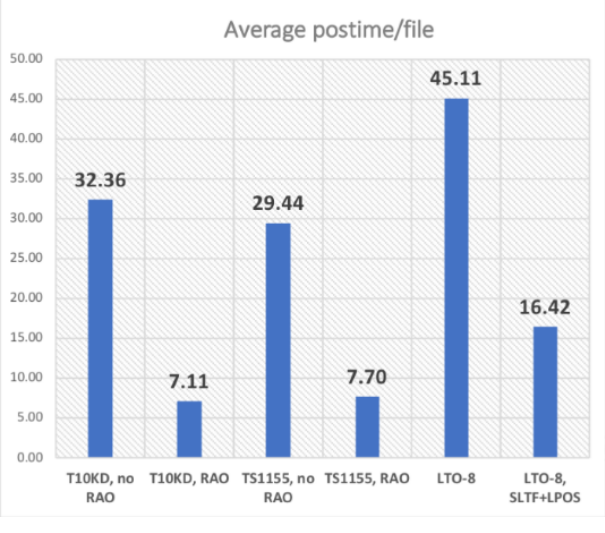
Optimising Read Performance for LTO Drives

See: LTO Experiences at CERN

- CERN introduced LTO drives in 2018, following Oracle's withdrawal from the tape drive market
- LTO–8 positioning time 5.8× slower than Enterprise, read performance 3× slower than Enterprise
  - LTO lacks a high-resolution tape directory (high speed positioning areas)
  - Resolution on LTO–8 is ½ tape wrap compared to 1/64 on TS1155 Enterprise drive
  - No drive-assisted RAO
- For larger tapes, performance bottleneck is moving from robotics/drive availability to head positioning time
- LTO media will experience significantly more wear due to long head traversals

Optimising Read Performance for LTO Drives

See: LTO Experiences at CERN



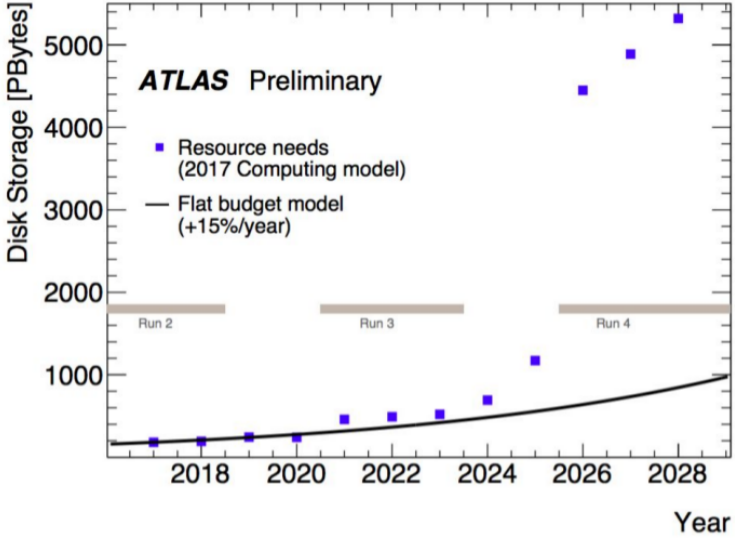
Approach

- Estimate "longitudinal position" and "wrap" of the start/end blocks of the file segments on the tape
- Define a cost function (distance, direction changes, …) for getting from block to block
- Calculate the shortest traversal of all the file segments to be recalled

Results

- Initial tests reduced media traversal distance by over 11×
- Positioning time 3× speed-up
- Optimised LTO read access will be implemented in CTA

Exponentially-increasing Storage Needs, Flat Budget



Data Carousel

[Fascinating Vintage 20 Cassette Carousel from 1972]

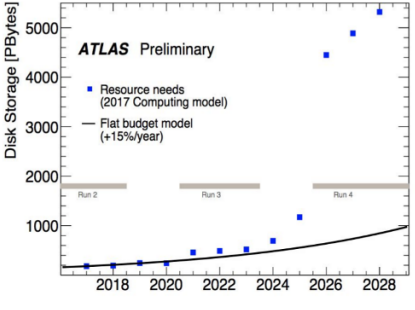
Data Carousel

Data for online analysis stored on tape [Tape Usage (2018)]

What is ‘data carousel’ and why ?

Data storage challenge of HL-LHC :

- ➔ ‘Opportunistic storage’ basically doesn’t exist
- ➔ Format size reduction and data compression are both long-term goals, require significant efforts from the software and distributed computing teams
- ➔ Tape storage is 3~5 times cheaper than disk storage, increasing tape usage is a natural way to cut into the gap of storage shortage for HL-LHC

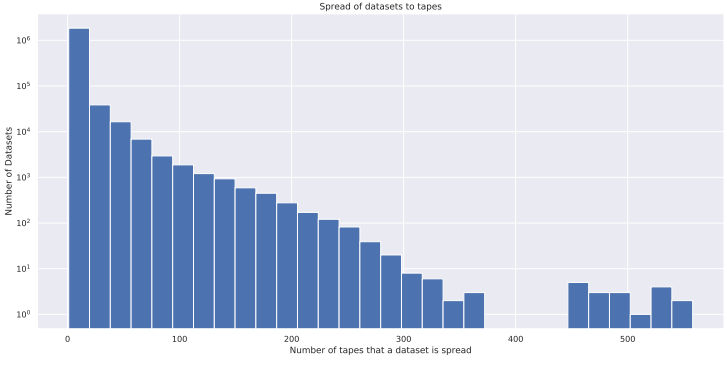


‘Data Carousel’ R&D → to study the feasibility to use tape as the input to various I/O intensive workflows.

Optimising Access to Exabyte-scale Data Archives

- User recalls of a dataset can take several hours or several days
- File recall requests pass through multiple queues (experiment data management software stack, grid/file transfer system, disk stager, tape system retrieve request queue, tape mount queue, tape positioning/read queue, …)
- Build a model of the entire system based on 10 years of tape system logfiles, to understand the complex behaviour and latencies in the system
- This model will form the basis of further optimisation efforts

Data Colocation



- On first archival, datasets are split across multiple tapes
- Repacking tapes increases fragmentation
- Colocating data on a smaller number of tapes will improve recall times…
- …but increases cost of writing
- What is the optimal behaviour? Need to create a **model** and **metrics** and design new **algorithms**

Conclusions

Storage needs are growing but budgets are flat

- The CERN physics archive is 340 PB but will soon grow to 1 EB
- Data retrievals already exceed 1 EB/year

CTA is CERN's next-generation archival storage solution

- Resilient, scalable queuing system based on RADOS Distibuted Object Store
- Enhanced scheduler with pre-emption and "late binding" of jobs
- Files accessed on tape in Recommended Access Order (RAO)

CERN is researching how to optimise data retrieval

- RAO for LTO drives
- Tape Carousel
- Modelling the tape storage system
- Data Colocation

References I

CERN.
Processing LHC Data.
YouTube video. Accessed June 2019.

Germán Cancio, Vladimír Bahyl, Daniele Francesco Kruse, Julien Leduc, Eric Cano, and Steven Murray.
Experiences and challenges running CERN's high capacity tape archive.
Journal of Physics: Conference Series, 664(4):042006, 2015.

David Reine and Mike Kahn.
Continuing the search for the right mix of long-term storage infrastructure—a TCO analysis of disk and tape solutions.
The Clipper Group Calculator, July 2015.

References II

- Mark Lantz (IBM Research).
  The Future of Tape.
  The Future of Data Protection and Retention Workshop, IBM Research, Rüschlikon, Zürich, Switzerland, November 2018.
- Information Storage Industry Consortium (INSIC).
  2015–2025 International Magnetic Tape Storage Roadmap.
  December 2015.
- The Register.
  LTO–8 tape media patent lawsuit cripples supply as Sony and Fujifilm face off in court.
  May 2019.
- Cristina-Gabriela Moraru.
  Enhancing the low-level tape layer of CERN Tape Archive software.
  Master's thesis, University Politehnica of Bucharest, September 2017.

References III

- Germán Cancio Meliá.
  LTO experiences at CERN.
  HEPiX Autumn/Fall Workshop, October 2018.
- TechMoan.
  Fascinating Vintage 20 Cassette Carousel from 1972 : Panasonic RS–296US.
  YouTube video. Accessed June 2019.
- Xin Zhao (Brookhaven National Laboratory).
  Tape Usage.
  ADC Technical Coordination Board Meeting, May 2018.


CERN