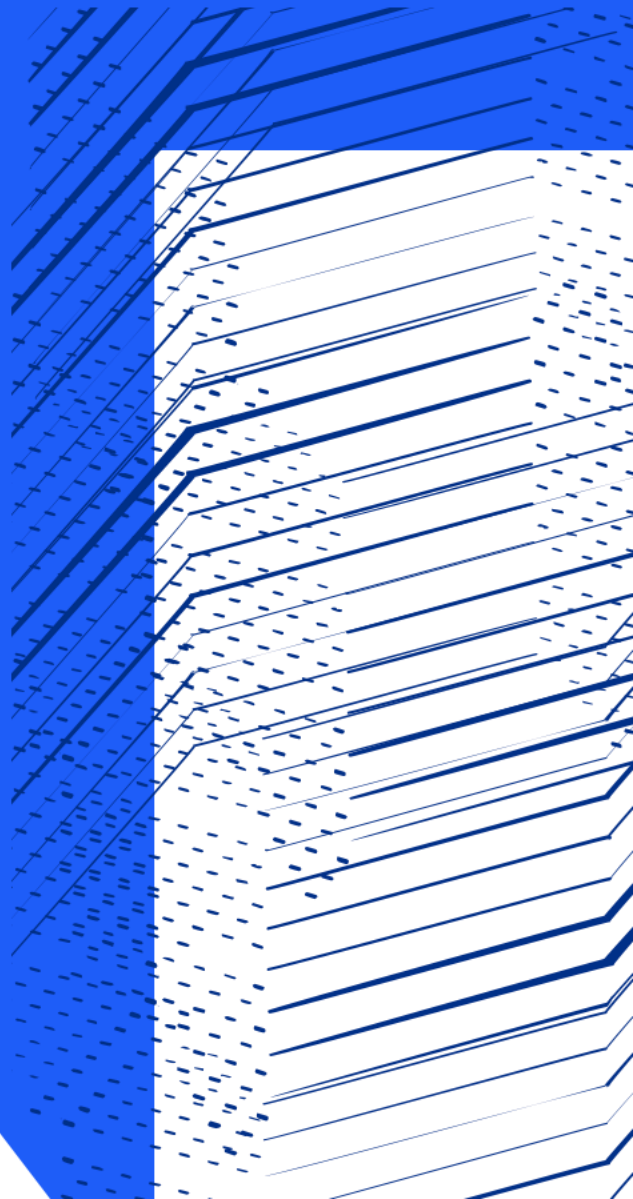
Science and Technology Facilities Council

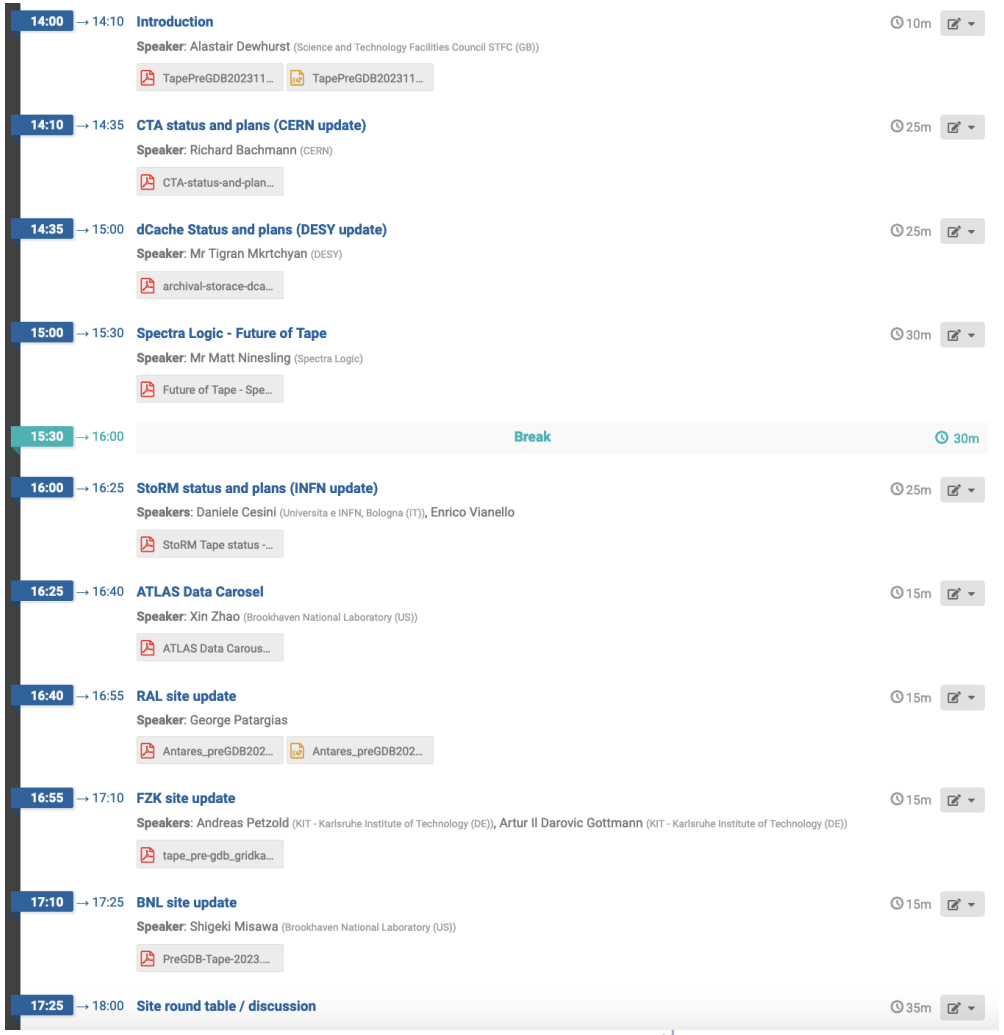
Tape Evolution pre-GDB report

Alastair Dewhurst



Pre-GDB on Tape Evolution

- Previous meeting in February 2021.
- Well attended:
  - 20 – 30 in person
  - 40+ people on zoom
- Talk from industry
- ATLAS developments
- Updates from the storage developers.
- Site reports



Tape Fundamentals

- Not everyone is an expert in tape.
- Principles:
  - Tape systems will always prioritize writing data to tape.
  - The tape drives which read/write the data to tape are the bottleneck in the system and need to be used efficiently.
  - Tape systems are designed for long term storage with infrequent access, we need to use them wisely.
- A lot of discussions are about recalling data

Spectra Logic – Future of Tape

- IBM is the one remaining company that develops new tape technology.
  - We need to watch developments carefully.
- IBM recently released (August 2023) the new TS1170 drives and media.
  - 50TB per tape (increase from 20TB in the previous generation)
  - 400MB/s read/write (same performance as previous generation)
  - Increased environmental specifications to make this work.
    - Max of 50% humidity down from 80%.
- Price per TB for Tape continues to remain well ahead of HDD.
  - E.g. for LTO-8 media <$5 / TB

Storage Technology Roadmap

- Tape has a much larger surface area that HDD.
  - LTO-9 tape is 1,035 meters long and ½ inch wide – 20,374 square inches
  - HDD is 3.5 inch in diameter with 10 platters – 96 square inches
- This difference allows for a much higher capacity with standard magnetic recording technologies using tape while disk has already hit the superparamagnetic limit with conventional technologies.

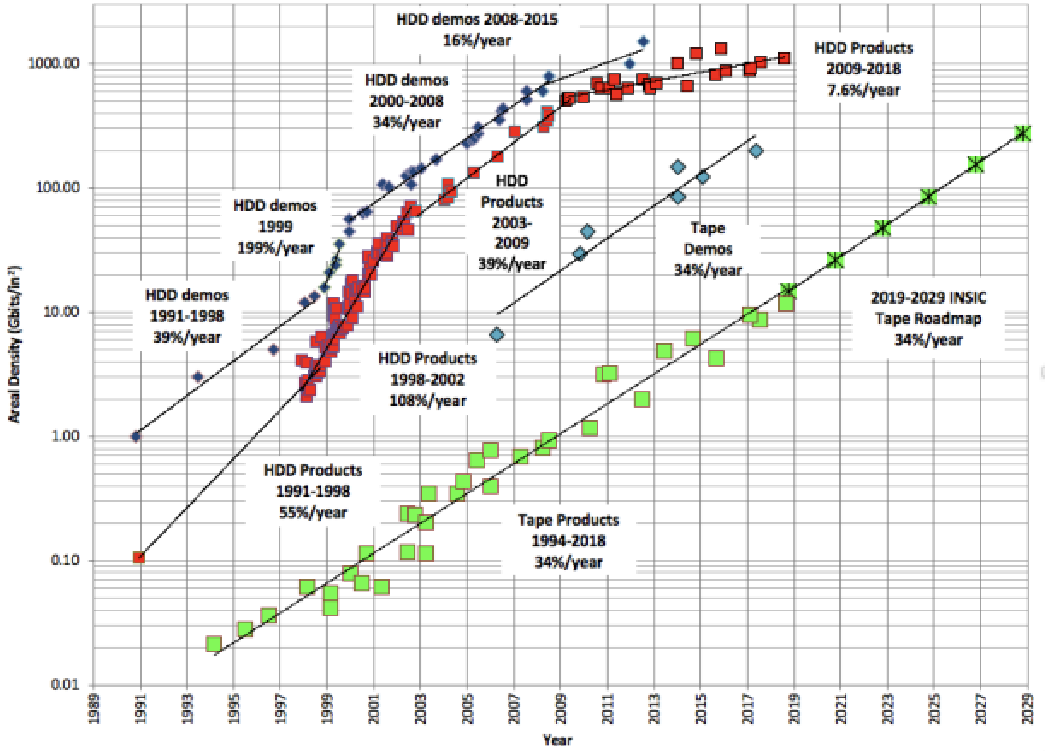


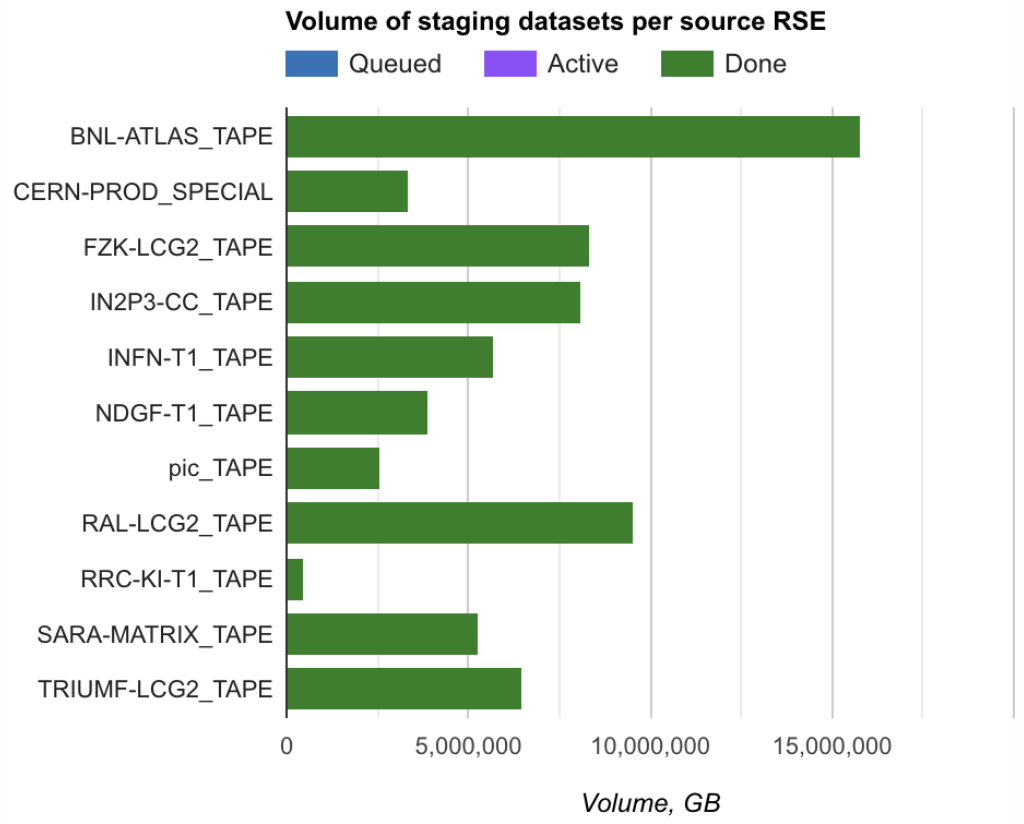
Figure 1: Areal Density Trends. Hard Disk Drive, Tape Product and Tape Technology Roadmap

INSIC 2019 Technology Roadmap
INSIC Technology Roadmap 2019 - SM

ATLAS Developments

- ATLAS Data Carousel has been in production since 2021.

- New work focuses on DAOD on demand.
  - Is it better to archive or re-creare rarely used DAOD?
- Smart writing
  - Trying to co-locating files on the same tape that will be recalled together.
  - Demonstrator at FZK showing factor of 2 performance improvement.
- Archive Metadata.



DAOD-on-demand HL-LHC demonstrator (2/2)

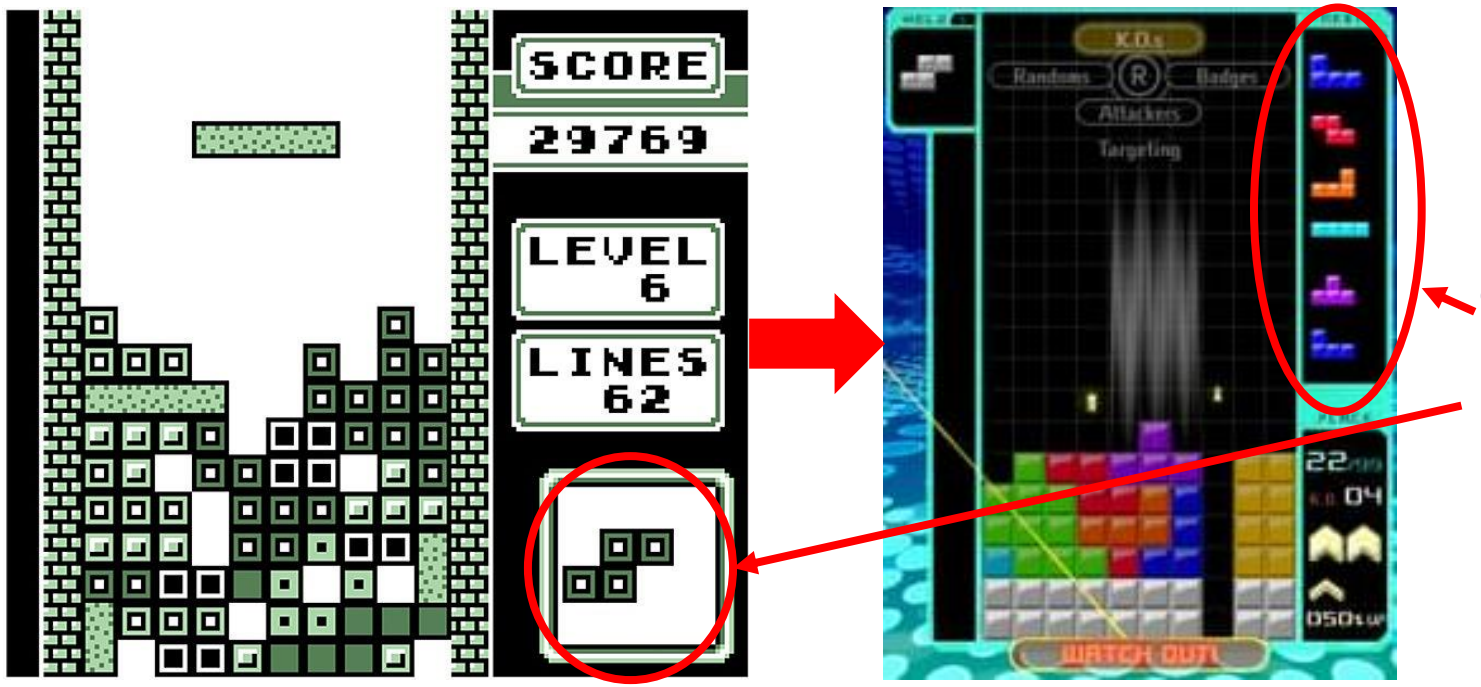
- Tests done so far on both scenarios, using data sample from the recent ATLAS data deletion campaign, at two Tier-1s (FZK and RAL).
- Preliminary results
  - Comparison of TTC (Time To Completion) among different scenarios

Data type	# datasets	#files	Size (GB)	Action	<TTC> per dataset (h)	Source (tape) site	Time stamp
AOD	13	31627	107047	Staging	19 +/- 9	FZK/RAL	July~Sep 2023
DAOD	11	1555	7284	Staging	3 +/- 4	FZK/RAL	July~Sep 2023
DAOD	5	1158	5459	recreation	7 +/- 3	N/A	July~Sep 2023

- Bulk mode tests are ongoing, of which the results will be used to estimate both the TTC at scale and the extra load on the tape resources.

Archive metadata

- There is ongoing work to allow additional metadata to be sent to tape services allowing them to optimize their service.
- Details will be presented by Julien at the Data Challenge 24 workshop.



VOs will be able to supply additional metadata regarding the datasets they are writing.

This will allow sites to improve tape utilization.

Science and Technology Facilities Council

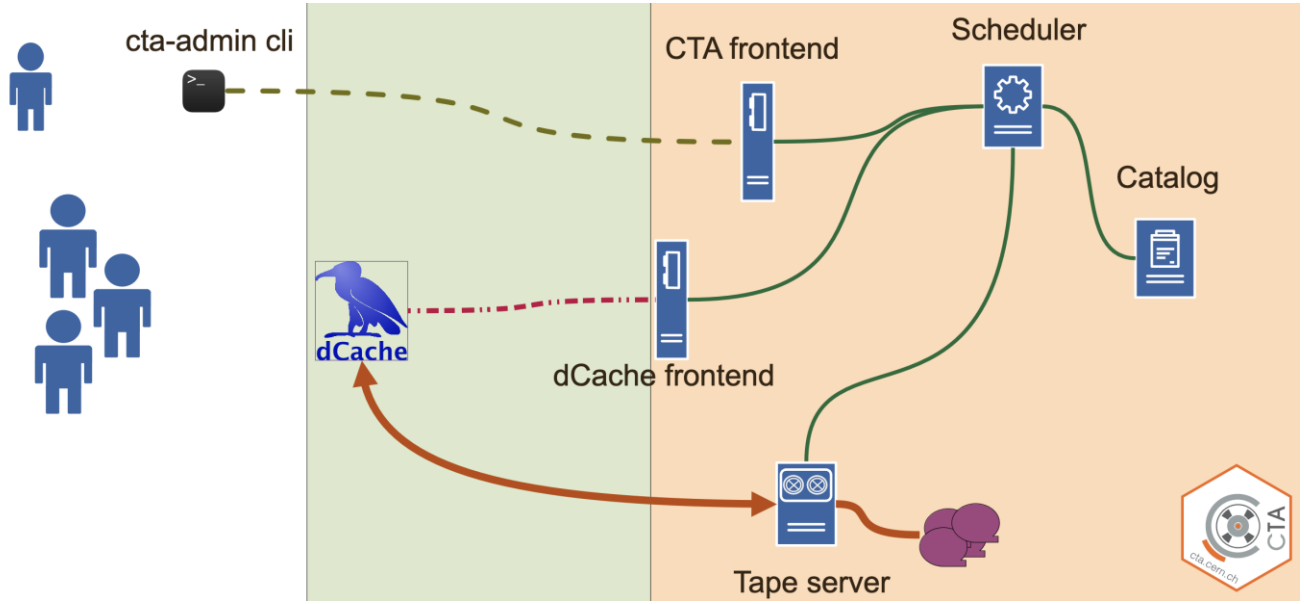
Storage providers

CTA + EOS

- Recent developments:
  - http REST API available for VOs since Q1 2023.
  - EOS 5 now available.
    - CERN will upgrade LHC experiments now heavy ion run complete.
  - cback – backup orchestrator using CTA.
- Future plans:
  - gRPC
  - Move to Alma 9
  - Addition of Archive metadata
  - Improving repacking and monitoring
  - Schedule separation and migration to PostgreSQL

dCache

- dCache = disk cache in front of tape.
- dCache can be used with a variety of backends:
  - CTA, HPSS, TSM, Enstore, DMF etc
- Seamless integration with dCache is merged into upstream CTA code.

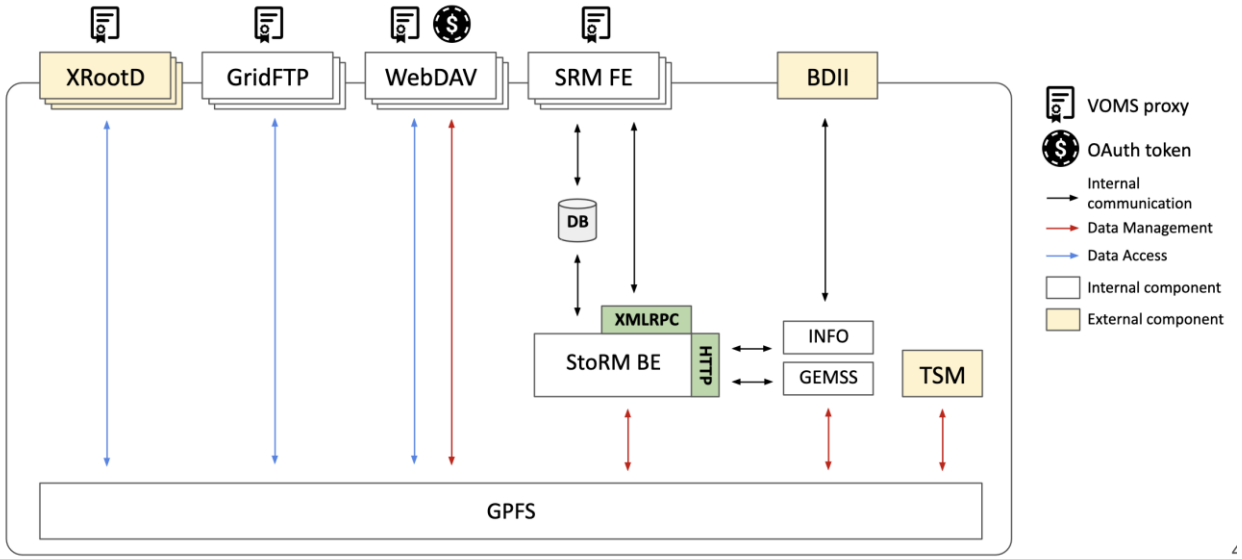


StoRM

- Extensive talk from StoRM developers.

- StoRM Tape basics
  - GEMSS component
  - Current data life cycle within a tape-enabled storage area
- StoRM Tape REST API
  - The WLCG Tape REST API specification
  - NGINX and OPA deployment roles
  - OPA authorization example
  - Testing tools
  - Ongoing developments

StoRM: a typical deployment architecture



Science and Technology Facilities Council

Site Reports

CERN

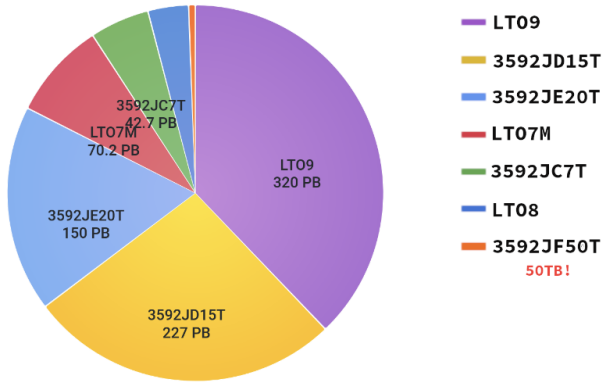
Hardware Inventory

6 Libraries:

- 4x IBM TS4500
- 2x Spectra Logic TFinity

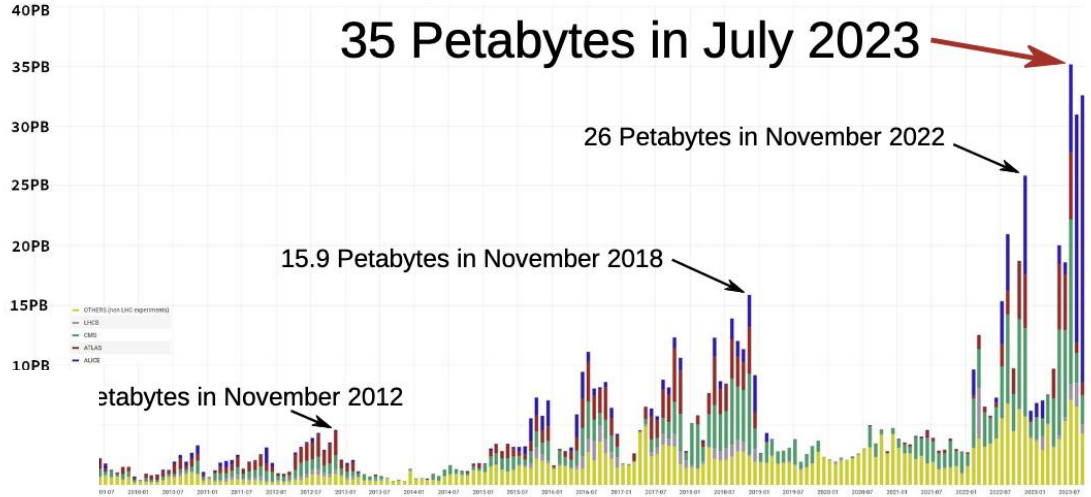
184 Drives:

- 1:1 drive to *Tape Server* mapping
- *New:* IBM TS1170

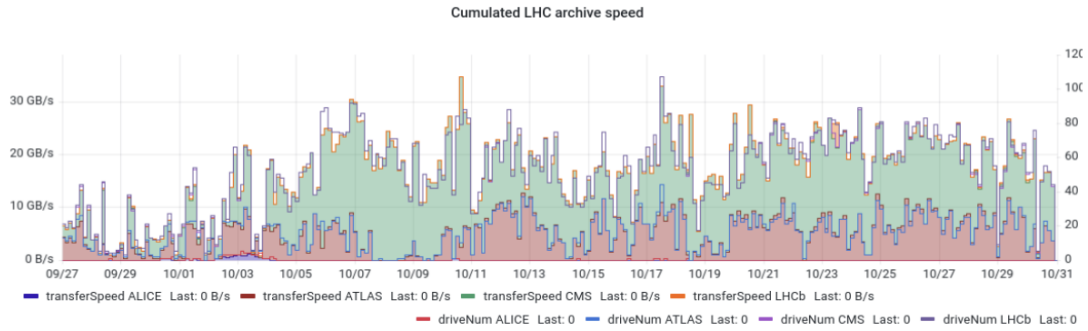


New CERN Record

Data archived to tape storage each month since 2008



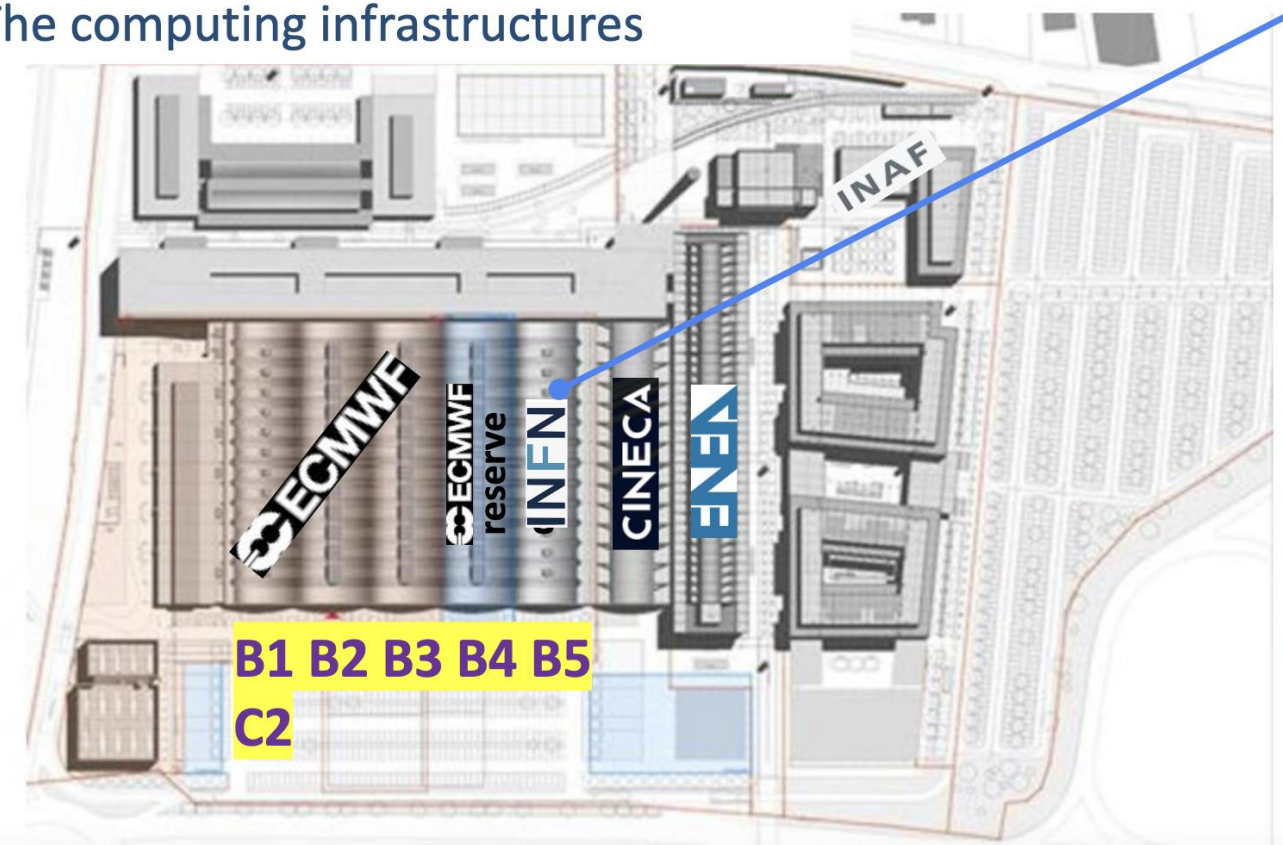
2023 Data Taking —Heavy Ion



INFN – An audience with the Pope

What can the Tecnopolo host?

The computing infrastructures



Each of the 6 “botti” (barrels) is ~5000m^2 of usable IT space

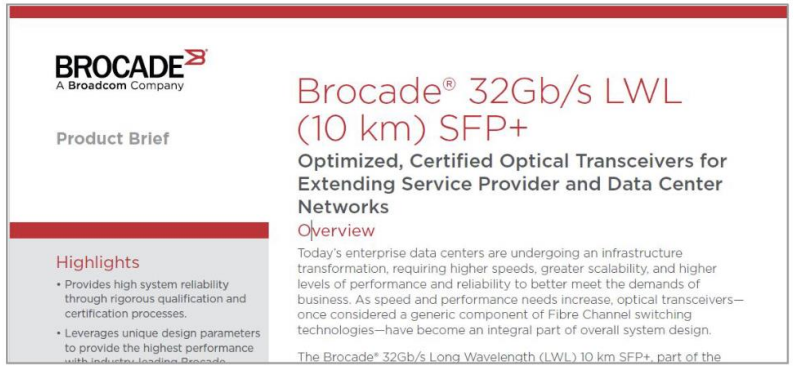
Same architect and design of the ”Sala Nervi” in the Vatican

36

INFN

Metropolitan Tape Area Network

- 2 libraries at CNAF
- 1 new library at the Tecnopole
- About **7 km** of fiber to connect the 2 datacenters
  - **yellow** + **red** paths
- 2 fiber pairs dedicated to extend the fiberchannel TAN
  - Brocade optics for 10km distance


BROCADE
A Broadcom Company

Product Brief

Brocade® 32Gb/s LWL (10 km) SFP+

Optimized, Certified Optical Transceivers for Extending Service Provider and Data Center Networks

Overview

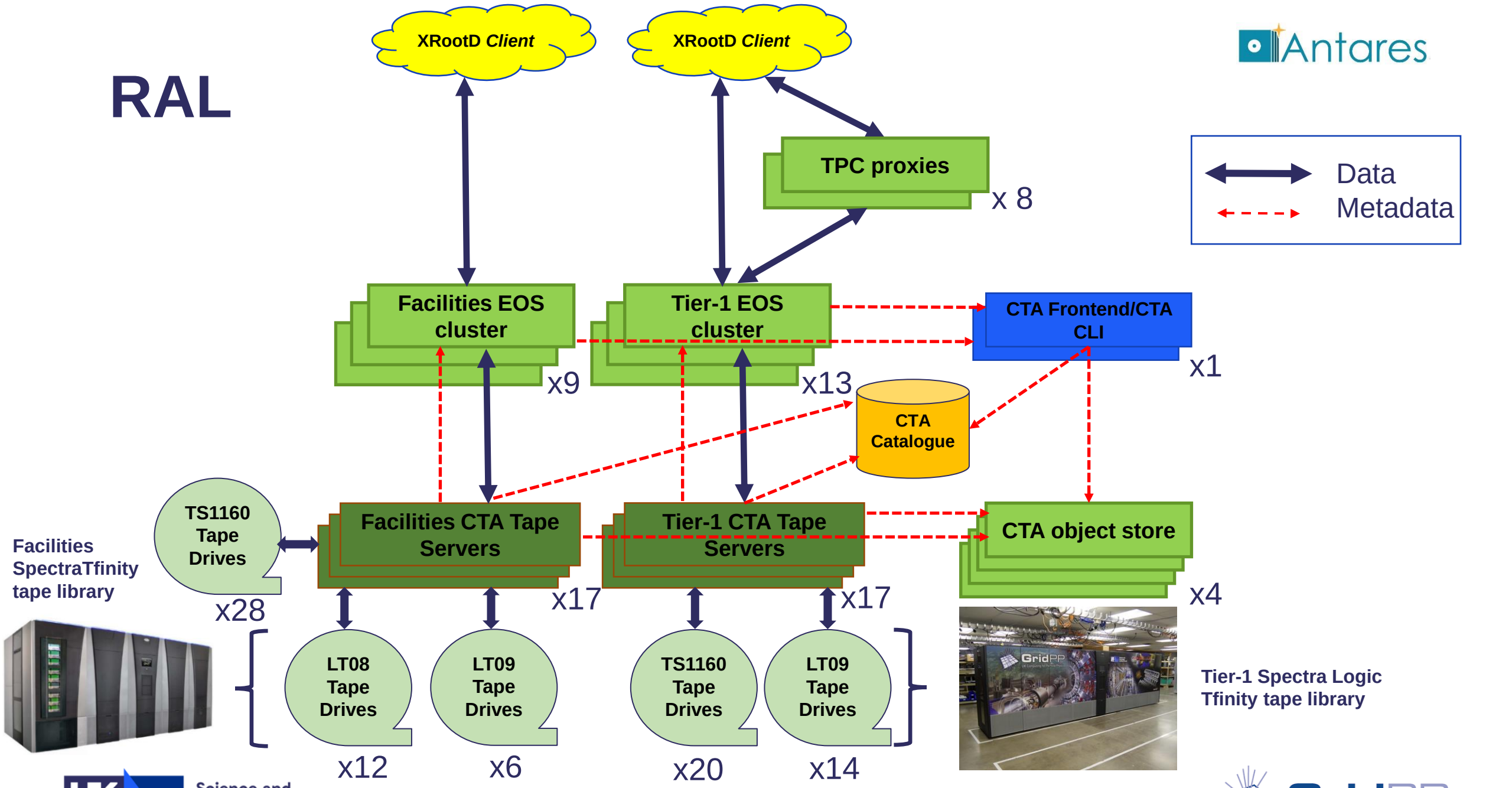
Highlights

- Provides high system reliability through rigorous qualification and certification processes.
- Leverages unique design parameters to provide the highest performance

Today's enterprise data centers are undergoing an infrastructure transformation, requiring higher speeds, greater scalability, and higher levels of performance and reliability to better meet the demands of business. As speed and performance needs increase, optical transceivers—once considered a generic component of Fibre Channel switching technologies—have become an integral part of overall system design.

The Brocade® 32Gb/s Long Wavelength (LWL) 10 km SFP+, part of the

RAL



Facilities SpectraTfinity tape library

Tier-1 Spectra Logic Tfinity tape library

Science and Technology Facilities Council

GridPP UK Computing for Particle Physics

FZK – Putting the HP in HPSS!

Recalling files from HPSS

Main goal: recall files efficiently from tapes for O(50k) requests

- Best for tapes: mount only once and read from front to end
- Best for experiments: obtain files at stable rates of O(1GB/s)
- Experiments recall large fractions of datasets during recall activities

→ Optimize based on these boundary conditions:

- full aggregate recall (FAR) in HPSS
  → faster reading of files on a tape from the same aggregate
- recommended access order (RAO) in HPSS
  → multiple aggregates are recalled in most efficient order from a tape
- number of used drives per experiment configurable
  → remaining flexible w.r.t. the load on HPSS

Deployed in an adapted dCache ENDIT-Provider and dedicated ENDIT-HPSS interface
→ technical details to be published in CHEP 2023 proceedings

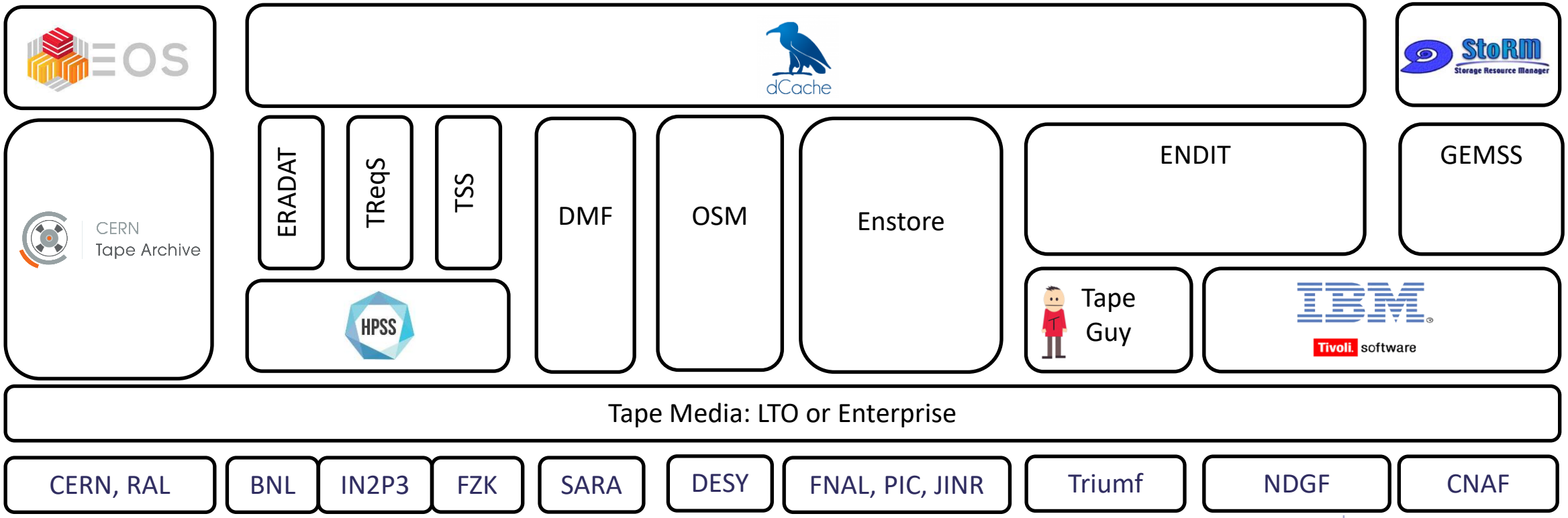
BNL

Data Center Migration

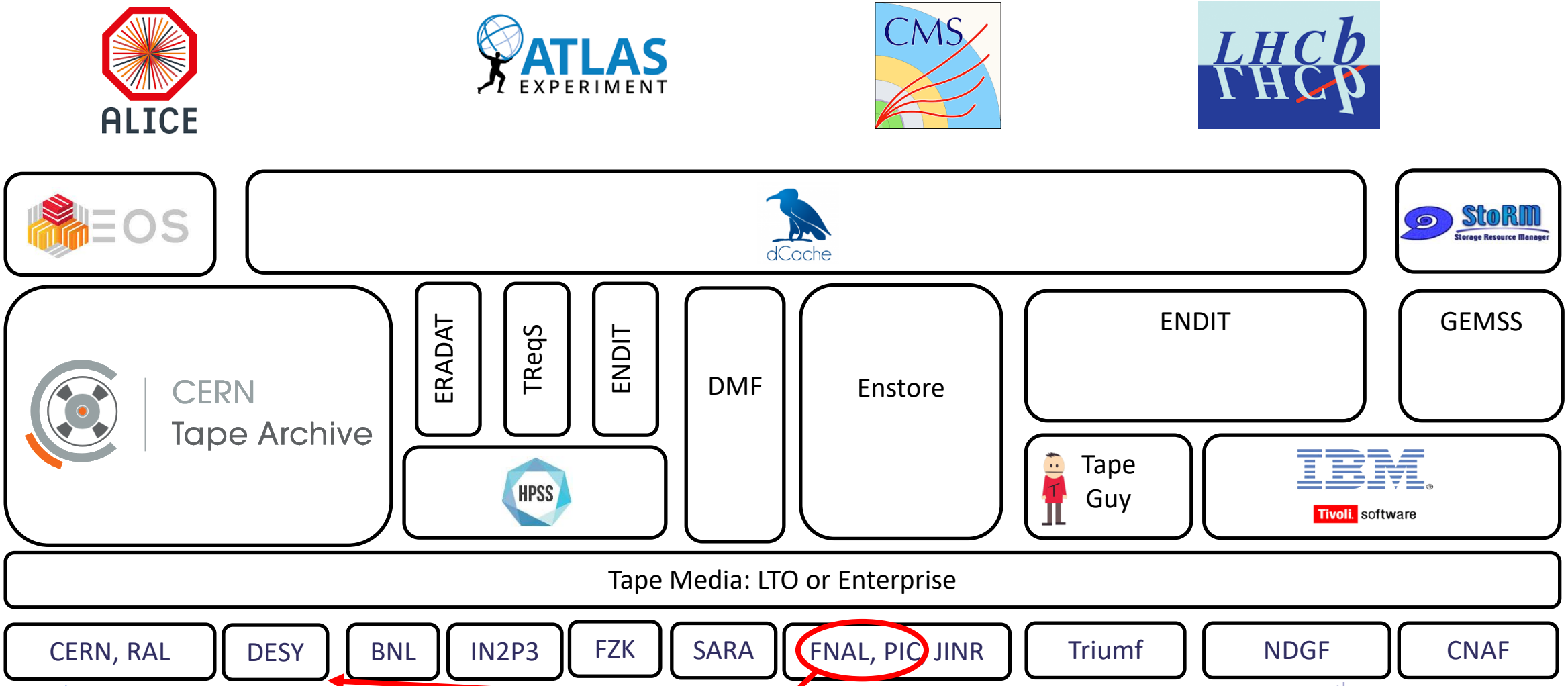
- Tape operations split between data centers
- Bldg 515 - Original “legacy” data center
  - Hosts data primarily from before run 3
  - 3 ATLAS Oracle SL-8500 libraries
  - ~11K LTO-7, 6K LTO-6 tapes with ATLAS data
- Bldg 725 - New, energy efficient and highly available data center
  - Hosts data from Run 3
  - HPSS core server
  - ATLAS HPSS disk cache
  - ATLAS IBM TS-4500 libraries
  - LTO-8 tapes containing new data
  - ATLAS LTO-8 tape drives

* High Performance Storage System - hpss-collaboration.org

Tape Summary (2021)



Tape Summary (2023)



Science and
Technology
Facilities Council

Questions?