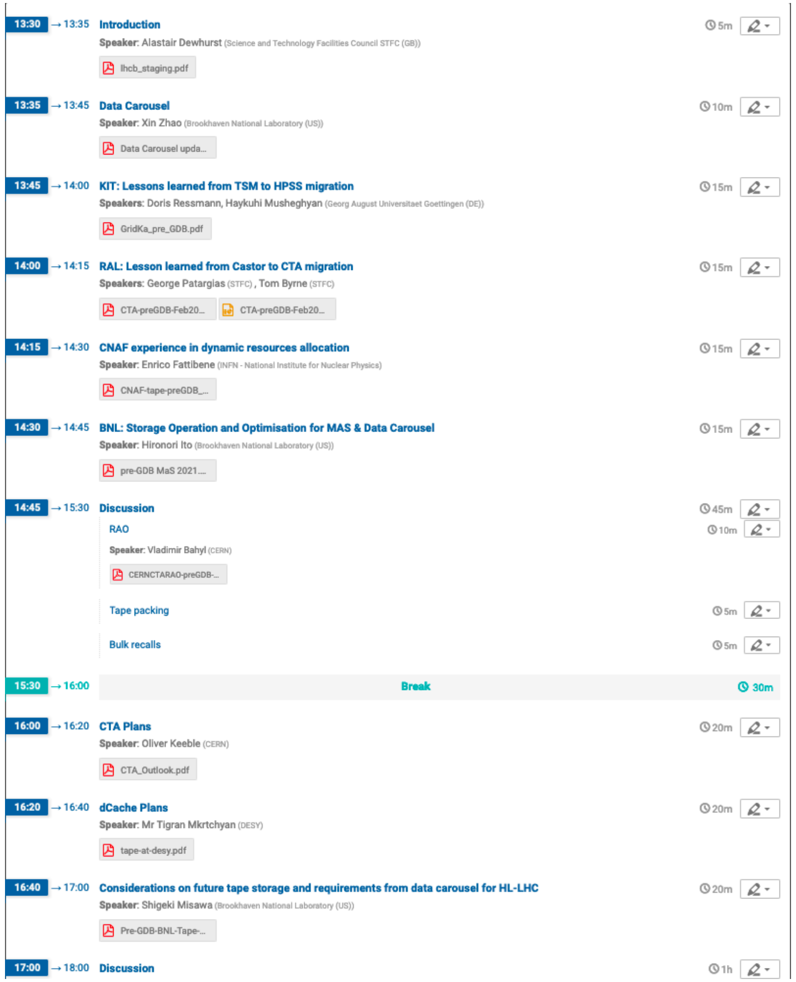
Science and Technology Facilities Council

Tape Evolution pre-GDB report

Alastair Dewhurst

Pre-GDB on Tape Evolution

- Well attended, in excess of 70 people connected.
- Focus was on the underlying tape systems.
    - Accessing Tape systems is well covered by DOMA-TPC.
- To provide some context I requested some input from VO about their workflows.
- First session was about Tape optimizations.
- Second session covered longer term goals.



13:30 → 13:35 **Introduction** 5m
Speaker: Alastair Dewhurst (Science and Technology Facilities Council STFC (GB))
lhcb_staging.pdf

13:35 → 13:45 **Data Carousel** 10m
Speaker: Xin Zhao (Brookhaven National Laboratory (US))
Data Carousel upda...

13:45 → 14:00 **KIT: Lessons learned from TSM to HPSS migration** 15m
Speakers: Doris Ressmann, Haykuhi Musheghyan (Georg August Universitaet Goettingen (DE))
GridKa_pre_GDB.pdf

14:00 → 14:15 **RAL: Lesson learned from Castor to CTA migration** 15m
Speakers: George Patargias (STFC) , Tom Byrne (STFC)
CTA-preGDB-Feb20... CTA-preGDB-Feb20...

14:15 → 14:30 **CNAF experience in dynamic resources allocation** 15m
Speaker: Enrico Fattibene (INFN - National Institute for Nuclear Physics)
CNAF-tape-preGDB_...

14:30 → 14:45 **BNL: Storage Operation and Optimisation for MAS & Data Carousel** 15m
Speaker: Hironori Ito (Brookhaven National Laboratory (US))
pre-GDB MaS 2021....

14:45 → 15:30 **Discussion** 45m
RAO 10m
Speaker: Vladimir Bahyl (CERN)
CERNCTARAO-preGDB-...
Tape packing 5m
Bulk recalls 5m

15:30 → 16:00 **Break** 30m

16:00 → 16:20 **CTA Plans** 20m
Speaker: Oliver Keeble (CERN)
CTA_Outlook.pdf

16:20 → 16:40 **dCache Plans** 20m
Speaker: Mr Tigran Mkrtchyan (DESY)
tape-at-desy.pdf

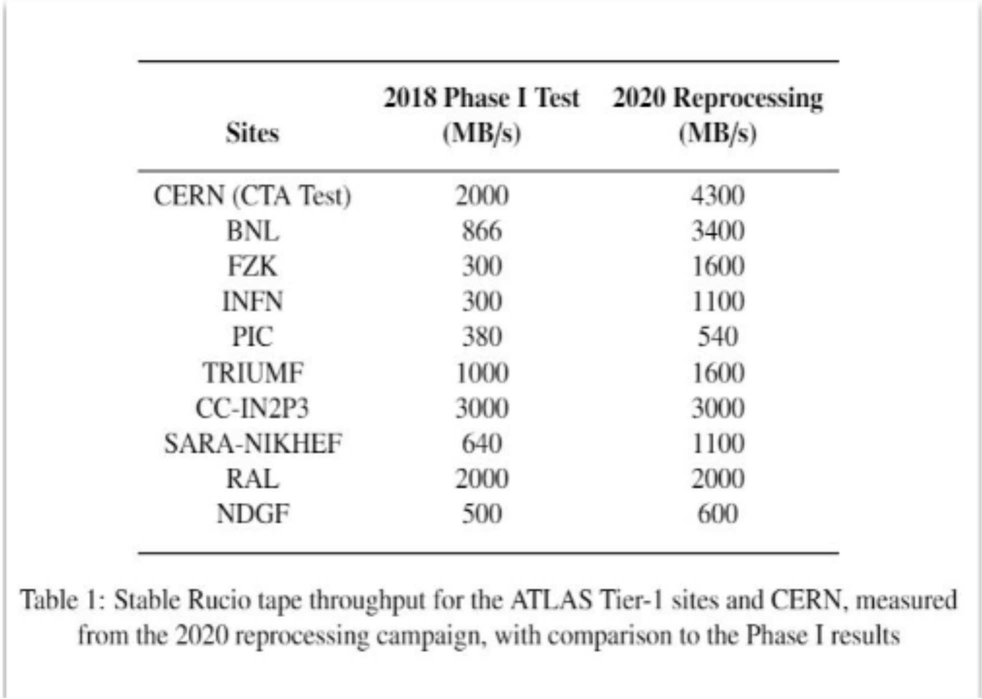
16:40 → 17:00 **Considerations on future tape storage and requirements from data carousel for HL-LHC** 20m
Speaker: Shigeki Misawa (Brookhaven National Laboratory (US))
Pre-GDB-BNL-Tape-...

17:00 → 18:00 **Discussion** 1h

Data Carousel

- ATLAS Data Carousel has resulted in significant performance improvements since 2018.

- During recalls tape drive throughput is ~30% maximum.
- We want to maximize use of hardware.
  - Can we increase this to 70%?
- Writing larger files can help (~20%)
- The way files are stored on tapes can also help (~20%).

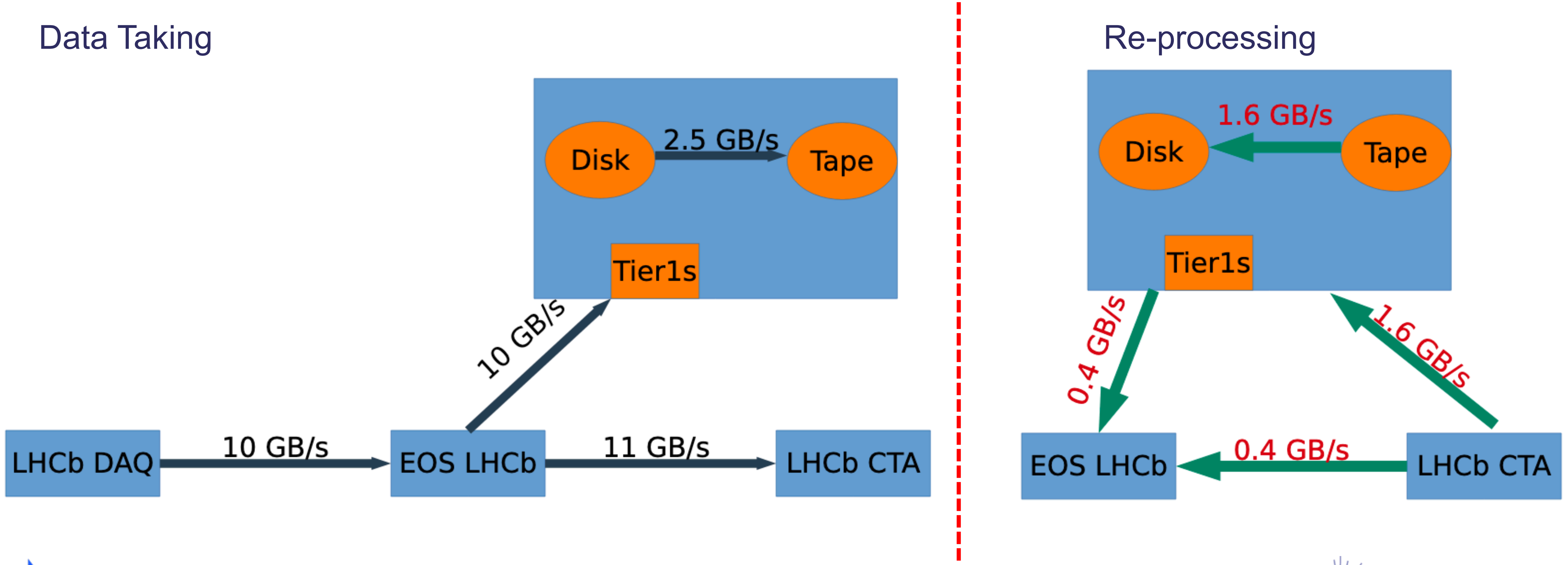


Sites	2018 Phase I Test (MB/s)	2020 Reprocessing (MB/s)
CERN (CTA Test)	2000	4300
BNL	866	3400
FZK	300	1600
INFN	300	1100
PIC	380	540
TRIUMF	1000	1600
CC-IN2P3	3000	3000
SARA-NIKHEF	640	1100
RAL	2000	2000
NDGF	500	600

Table 1: Stable Rucio tape throughput for the ATLAS Tier-1 sites and CERN, measured from the 2020 reprocessing campaign, with comparison to the Phase I results

LHCb planned usage

- High level requirement to be able to complete data re-processing in 4 months.

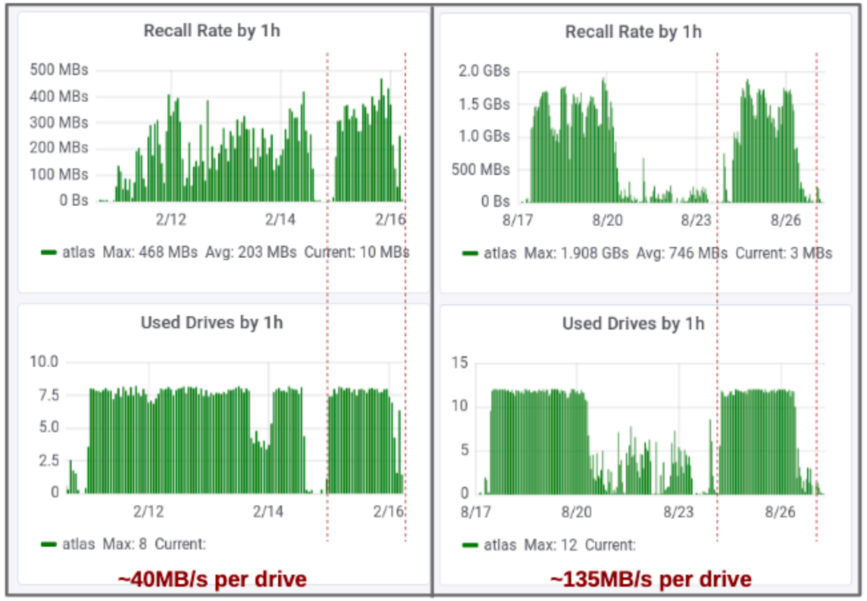


KIT - TSM to HPSS Migration

- Migrating their data from an Oracle SL8500 to Spectra Logic TFinity Library.
- Migrating from TSM to HPSS as the underlying tape system.
- Improved their scheduling software.
- Increased number of concurrent requests from 2k to 30k.
- Eliminated bottlenecks with new hardware and network.
- Recall rate is improved by a factor of 3 per drive.

Before Jan 2020

After Jan 2020

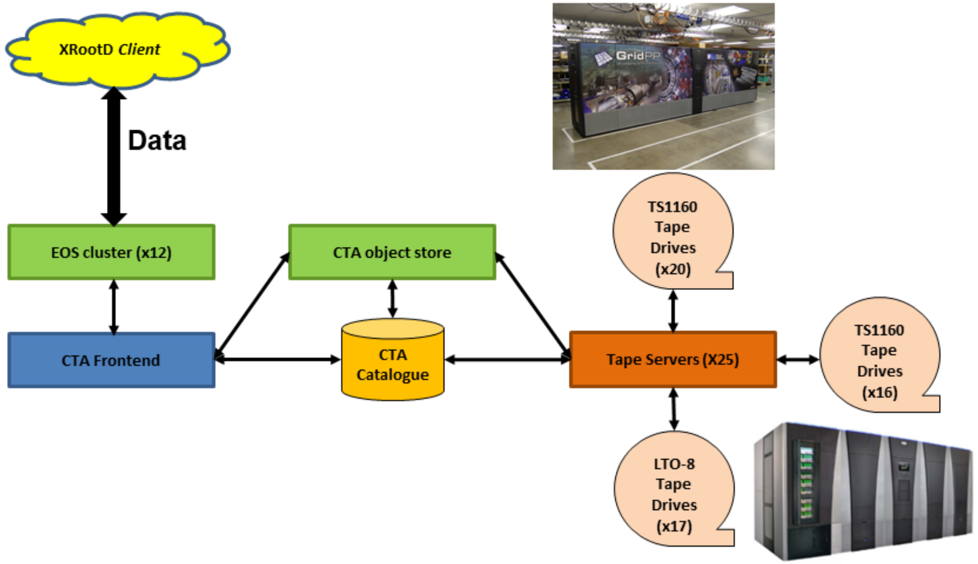


~40MB/s per drive

~135MB/s per drive

RAL - CTA and Tape Migration

- RAL provided an update on progress to migrate to CTA.
- RAL has also been migrating data from SL8500 to Tfinity Library.


XRootD *Client*

Data

EOS cluster (x12)

CTA Frontend

CTA object store

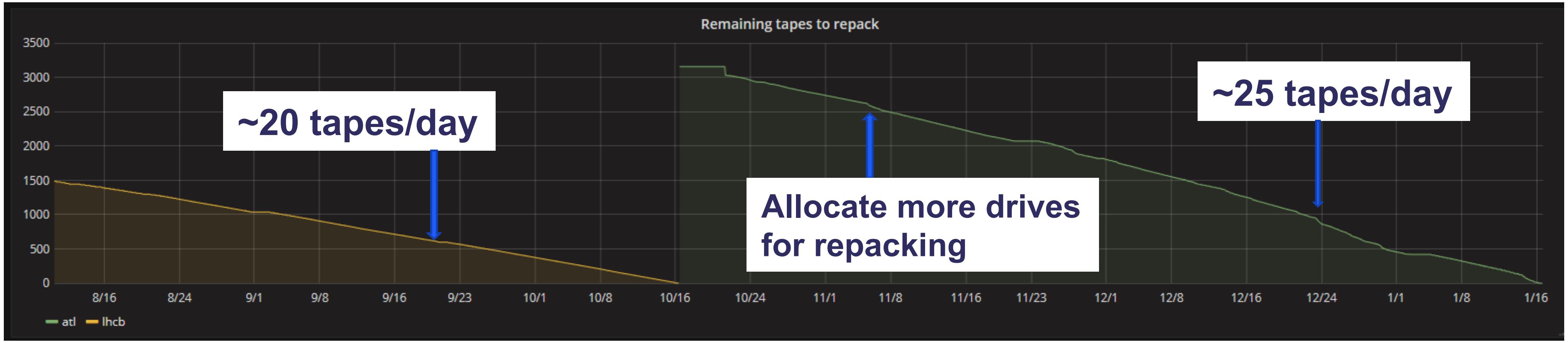
CTA Catalogue

Tape Servers (X25)

TS1160 Tape Drives (x20)

TS1160 Tape Drives (x16)

LTO-8 Tape Drives (x17)



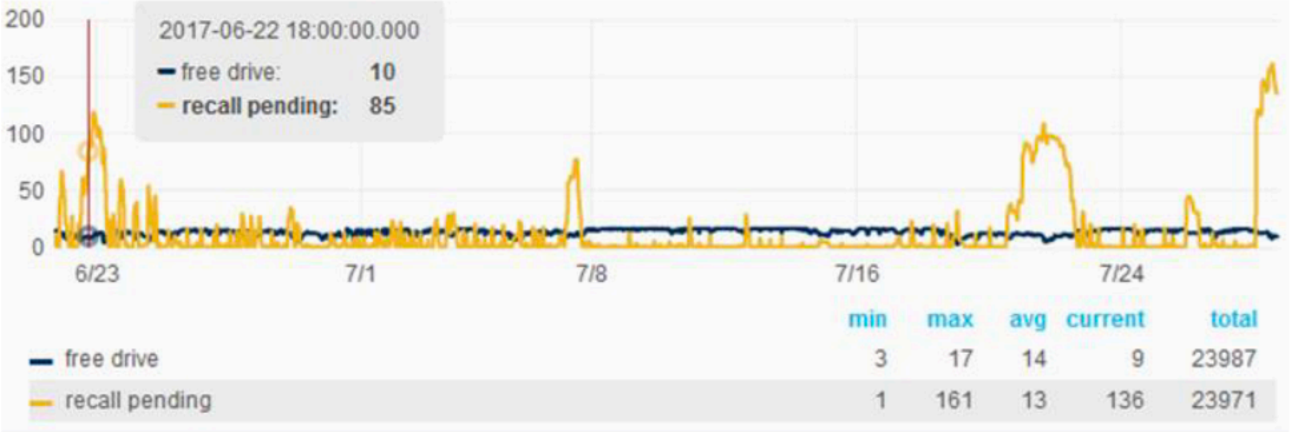
~20 tapes/day

Allocate more drives for repacking

~25 tapes/day

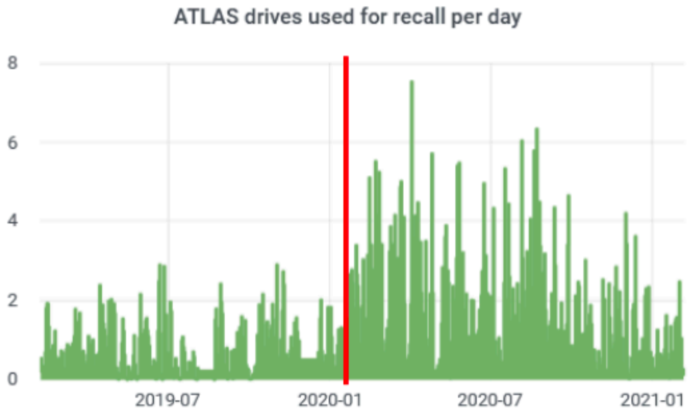
CNAF - Dynamic resources allocation

- CNAF noticed that some tape drives were unused even when large scale activity was ongoing.
  - Due to static partitioning of drives between VOs and activities.
- Monitor tape drive usage in InfluxDB and use that information to allocate drives where needed automatically.
  - Increased average throughput by 85%.



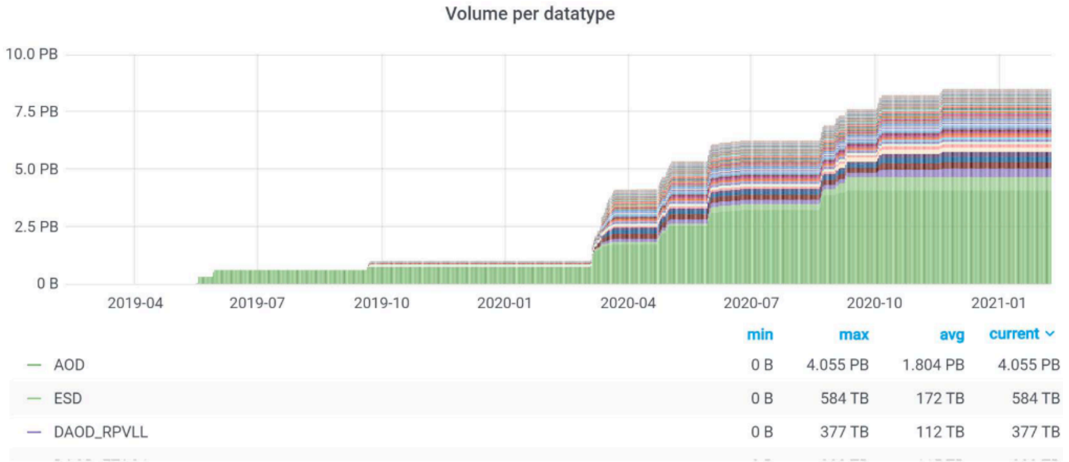
2017-06-22 18:00:00.000
free drive: 10
recall pending: 85

	min	max	avg	current	total
free drive	3	17	14	9	23987
recall pending	1	161	13	136	23971



BNL - Tape Recalls and MAS

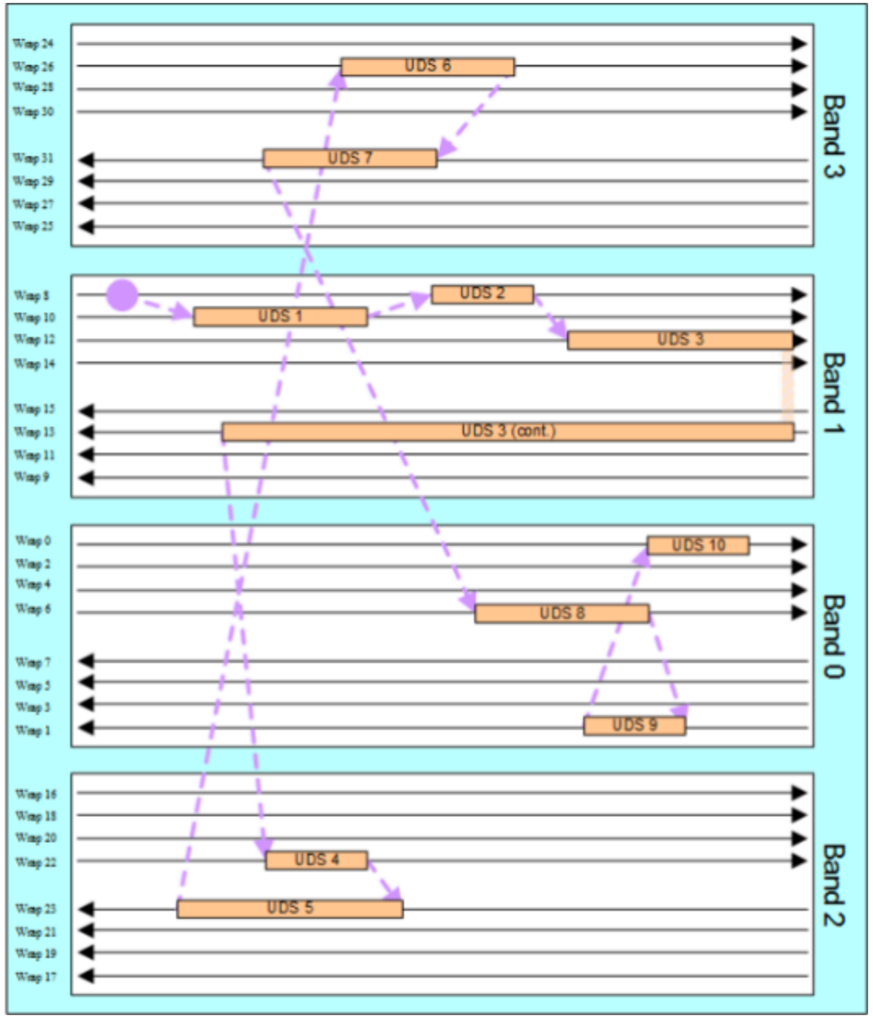
- Hiro gave a very detailed talk about directory based sorted writes to tape to maximize recall speed.
- Conclusions were, to improve efficiency:
  - Larger files
  - Fewer gaps between files on Tape
  - Read requests come in bulk to HPSS.
- MAS archives to tape unused files.
- Created 8PB additional space on disk.



	min	max	avg	current ⌄
— AOD	0 B	4.055 PB	1.804 PB	4.055 PB
— ESD	0 B	584 TB	172 TB	584 TB
— DAOD_RPVLL	0 B	377 TB	112 TB	377 TB

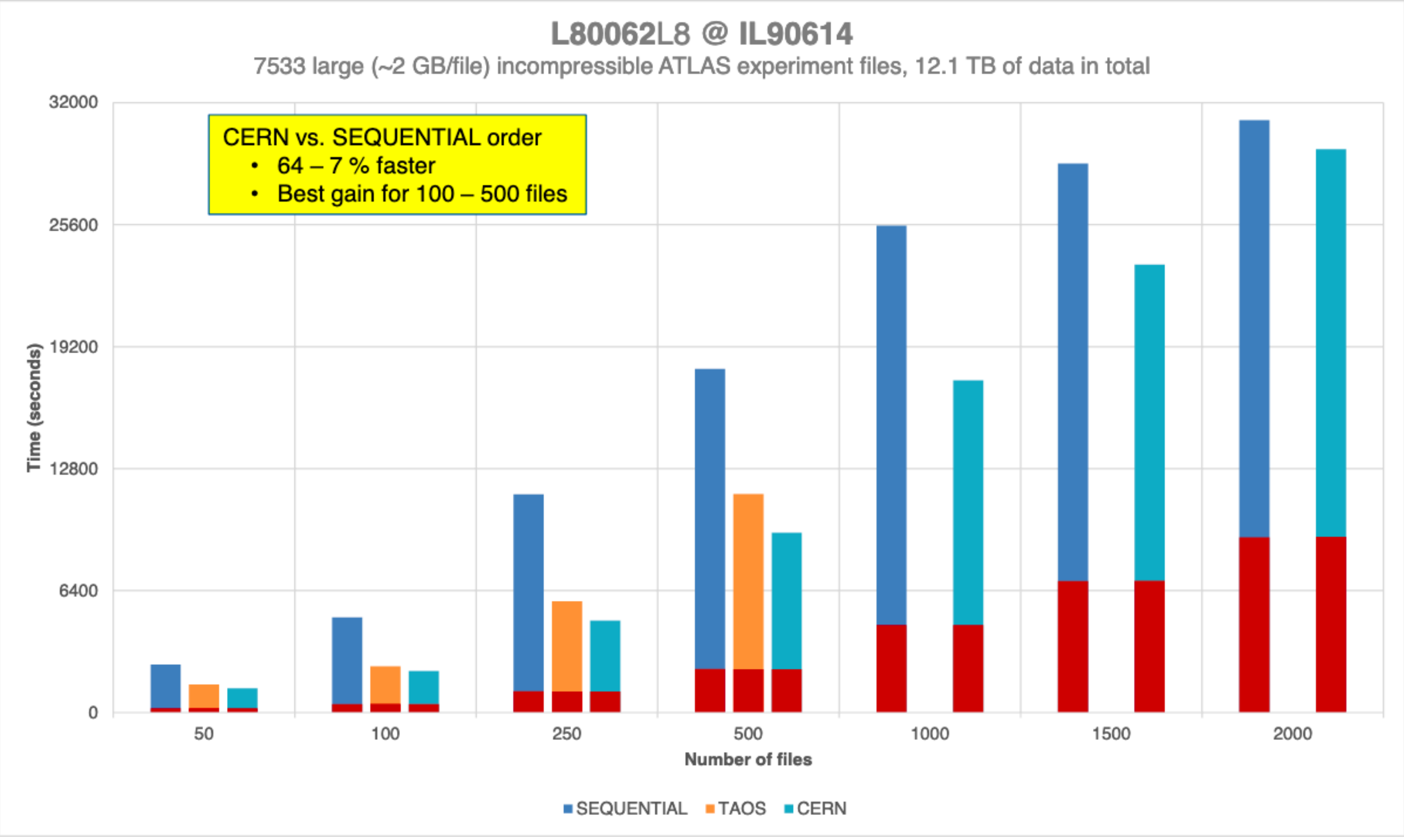
RAO (1)

- RAO stands for Recommended Access Ordering.
  - It orders the way data is read from tape and can make a significant difference in performance.
- A Tape is:
  - 960m long
  - Is made up of 4 bands, only one of which can be read at a time.
  - Each band if made up of 52 wraps, which the read head needs to read in the correct direction.
  - Each wrap holds 32 tracks which actually contain the 1 and 0s.



RAO (2)

- Enterprise drives have RAO.
- CERN have implemented RAO for LTO drives.
- Produces up to 64% improvement against sequential reads.
- Lower gains when there are many files (or very few) because sequential works ok.



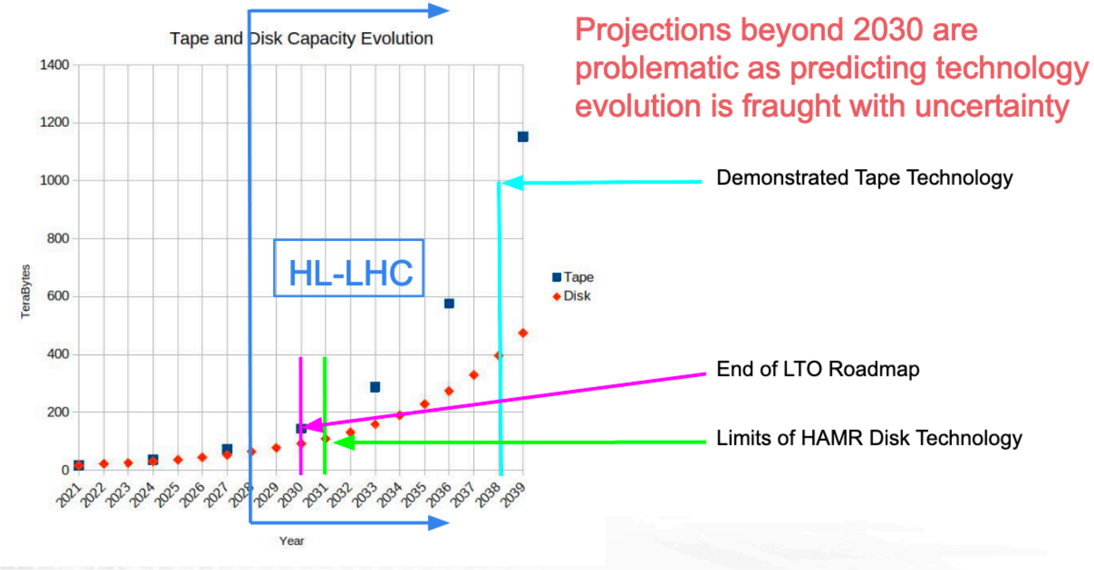
L80062L8 @ IL90614

7533 large (~2 GB/file) incompressible ATLAS experiment files, 12.1 TB of data in total

CERN vs. SEQUENTIAL order
- 64 – 7 % faster
- Best gain for 100 – 500 files

Cost modelling

- Shigeki Misawa gave a detailed presentation on long term cost modelling.
  - This was a requested follow up from the WLCG Workshop.
- Inefficient use of tape hardware drives up costs!



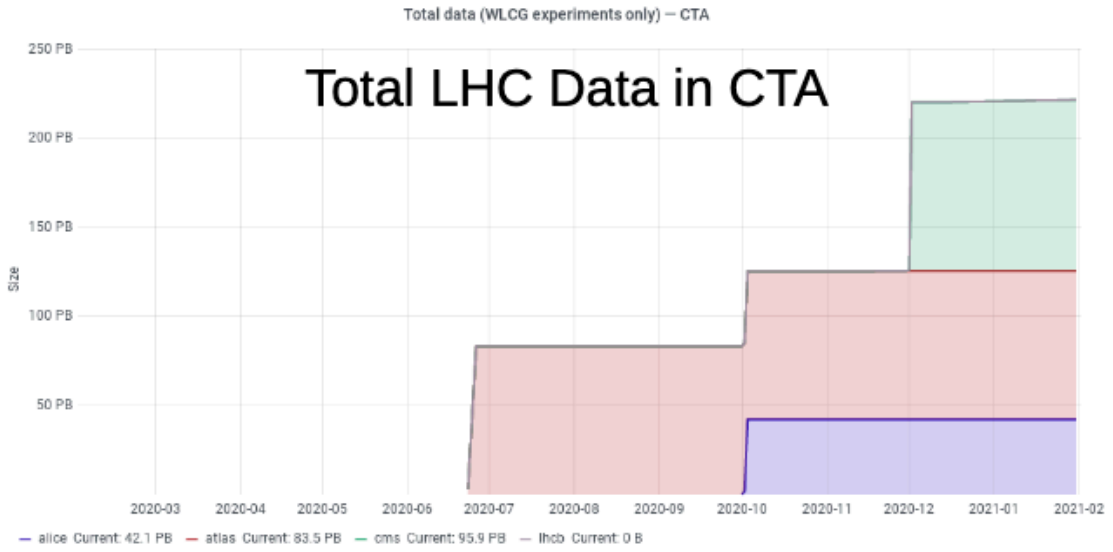
Projections beyond 2030 are problematic as predicting technology evolution is fraught with uncertainty

Demonstrated Tape Technology

End of LTO Roadmap

Limits of HAMR Disk Technology

CTA Outlook


Total LHC Data in CTA

- CTA has deployed production instances for Alice, Atlas and CMS.
- Lots still to do. Focusing on ensuring critical features are ready for start of Run 3.
  - CTA is feature-frozen in preparation for Run 3.
- CTA is designed to exploit tape hardware to the max.
  - SSD buffers are needed for throughput.
- CTA is in the early stages of establishing a community.
  - Welcomes contributions / collaborators

dCache / DESY Plans

- Tigran gave an update of DESY tape setup and use cases.
- Also gave an update of some of the new dCache developments.
- DESY are evaluating HSM (Tape) Software to replace OSM.
  - Maximize Tape hardware efficiency
- Possible candidates:
  - Plan A (Open Source): CTA or Enstore
  - Plan B (Proprietary): HPSS or TSM
- DESY are looking to make a decision in the next 6 months.
- It should be noted that FNAL are also evaluating the long term status of Enstore.

Science and
Technology
Facilities Council

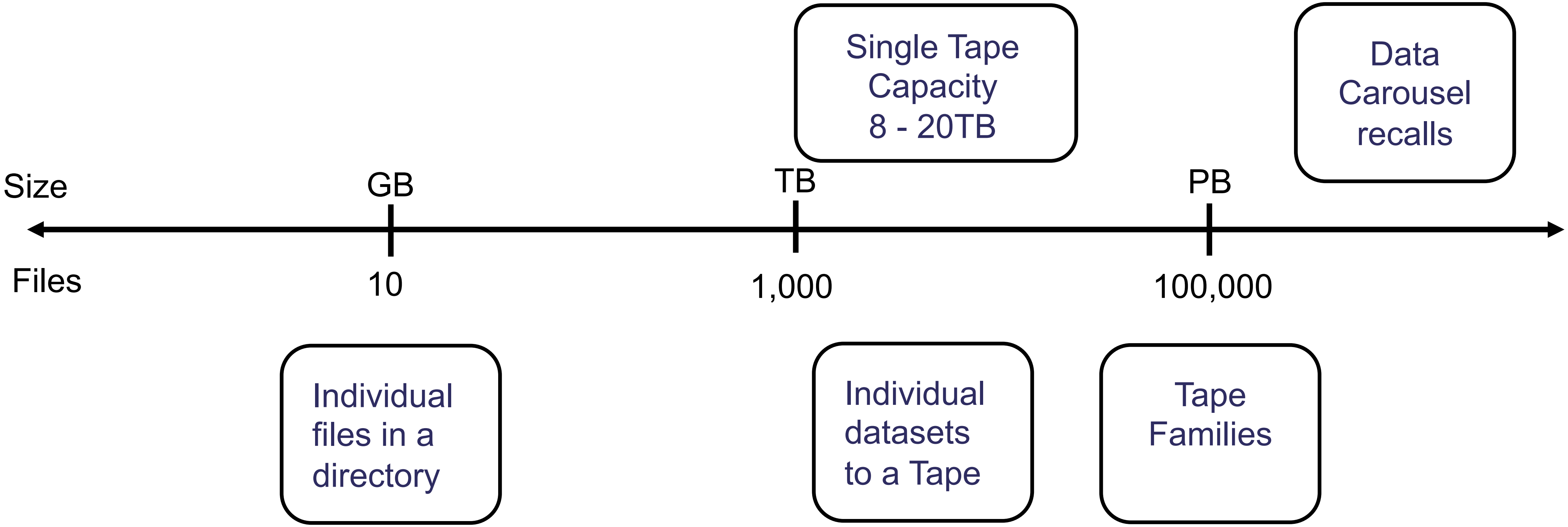
Alastair’s Opinions

Motivation

- I wanted to organize the meeting for the following reasons:
  - Other sites have been interested in RAL's CTA evaluation and the progress we are making with deploying it.
  - I am aware that other sites have been making big improvements in their tape performance and 2020 was not the easiest year to collaborate and share ideas.
- I think Tape can play an incredibly important role in solving HL-LHC data challenges because it is so much cheaper than Disk.
  - Tape systems are complicated and I feel some opportunities are not capitalized on.

Tape Recalls

- There is a lot of discussion about how to optimize writes.
- Use cases vary, but they are all probably good enough.


Single Tape Capacity 8 - 20TB
Data Carousel recalls
Size
GB
TB
PB
Files
10
1,000
100,000
Individual files in a directory
Individual datasets to a Tape
Tape Families

Accessing Tape Endpoints

ALICE

ATLAS EXPERIMENT

CMS

LHCb

Communication between VOs and Tape frontends is done really well by various DOMA groups.

CASTOR
CERN Advanced STORage manager

RAL

CERN
Tape Archive

CERN

dCache

BNL, CNAF, FNAL, JINR, KISTI, KIT, IN2P3, NDGF, PIC, SARA, Triumf

StoRM
Storage Resource Manager

CNAF

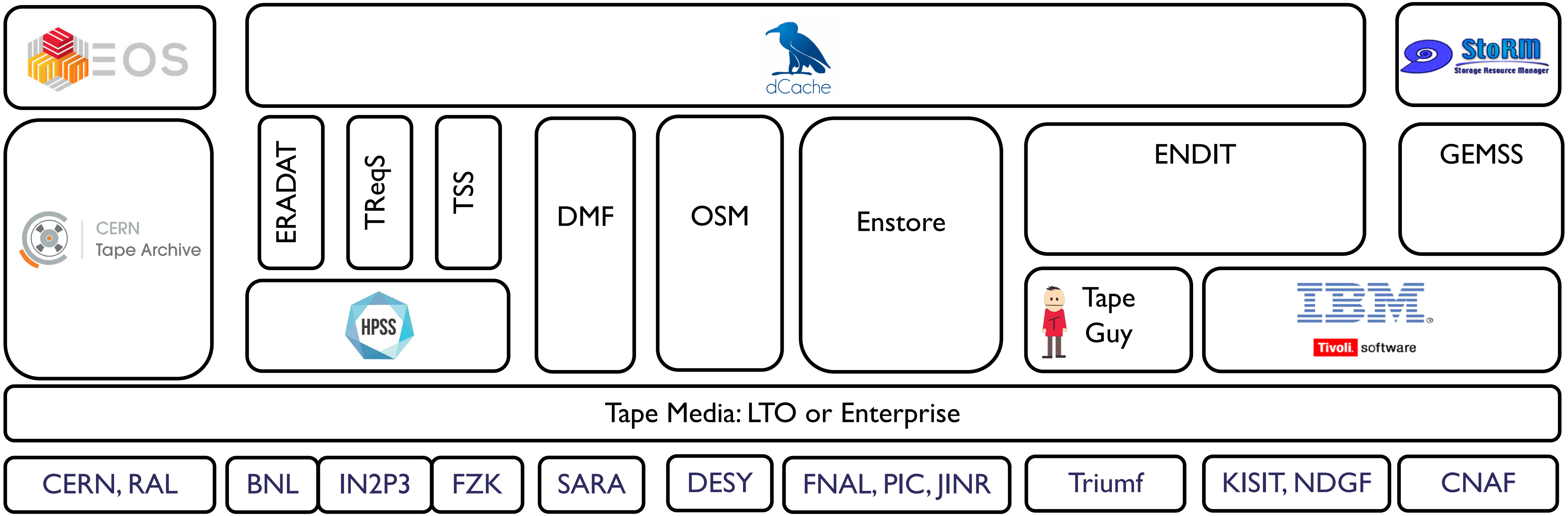
Optimizing Tape Endpoints

ALICE

ATLAS EXPERIMENT

CMS

LHCb



EOS

dCache

StoRM Storage Resource Manager

CERN Tape Archive

ERADAT

TReqS

TSS

HPSS

DMF

OSM

Enstore

ENDIT

GEMSS

Tape Guy

IBM Tivoli software

Tape Media: LTO or Enterprise

CERN, RAL

BNL

IN2P3

FZK

SARA

DESY

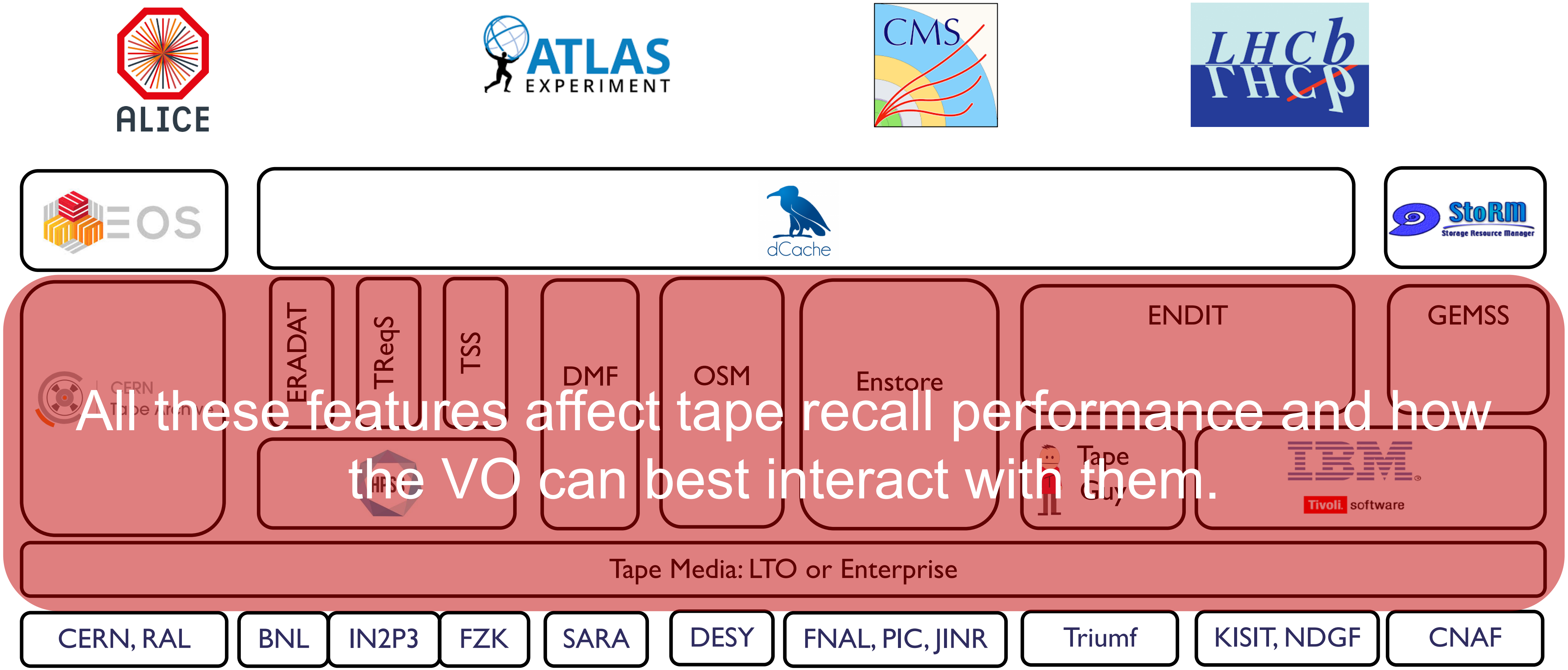
FNAL, PIC, JINR

Triumf

KISIT, NDGF

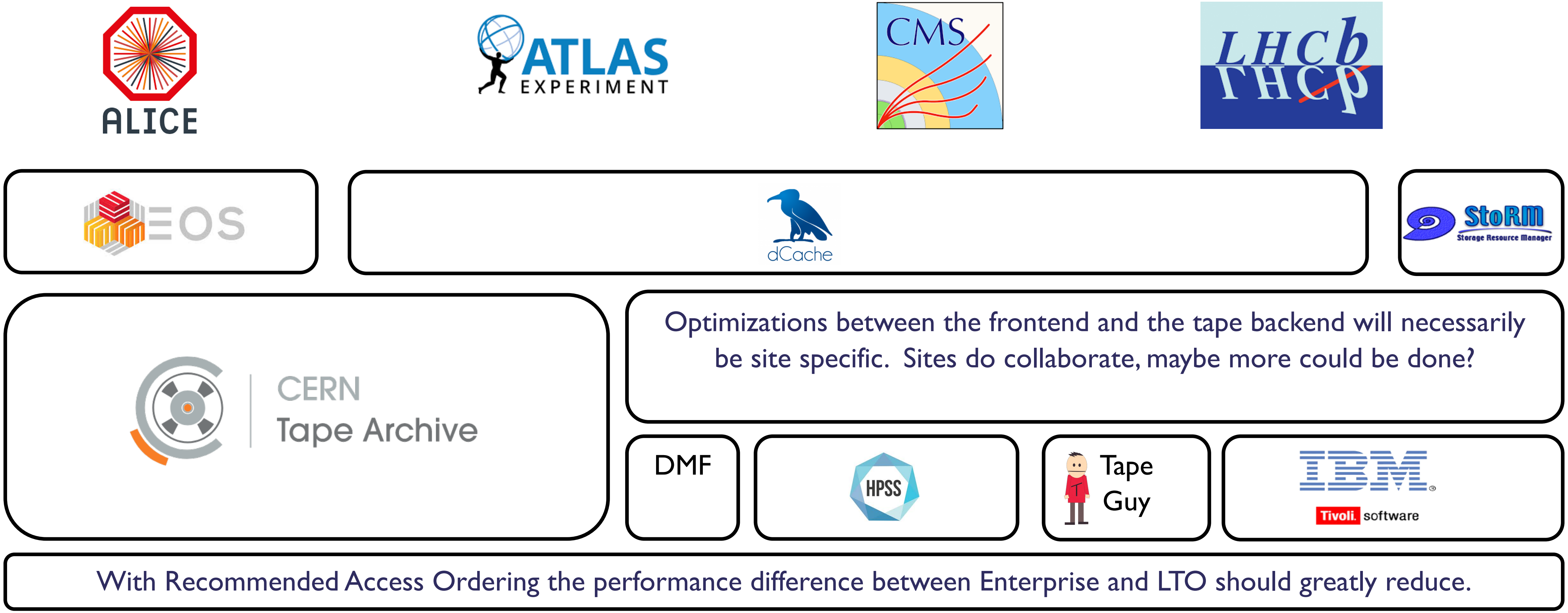
CNAF

Optimizing Tape Endpoints



All these features affect tape recall performance and how the VO can best interact with them.

A more consolidated future?



Optimizations between the frontend and the tape backend will necessarily be site specific. Sites do collaborate, maybe more could be done?

With Recommended Access Ordering the performance difference between Enterprise and LTO should greatly reduce.

Conclusions

- An awful lot of work has gone into improving and optimizing tape systems in the last few years.
- Tape drives are expensive, therefore a lot of optimization focus on making the best use of this.
- There is currently a fantastic opportunity for greater collaboration between Tape providers.
- I am sure there will be many smaller discussions but I would like to arrange another pre-GDB (or similar) in ~9 months.

Science and
Technology
Facilities Council

Questions?