Storage Services

Robert Illingworth

FCRSG

2 June 2022

Scope

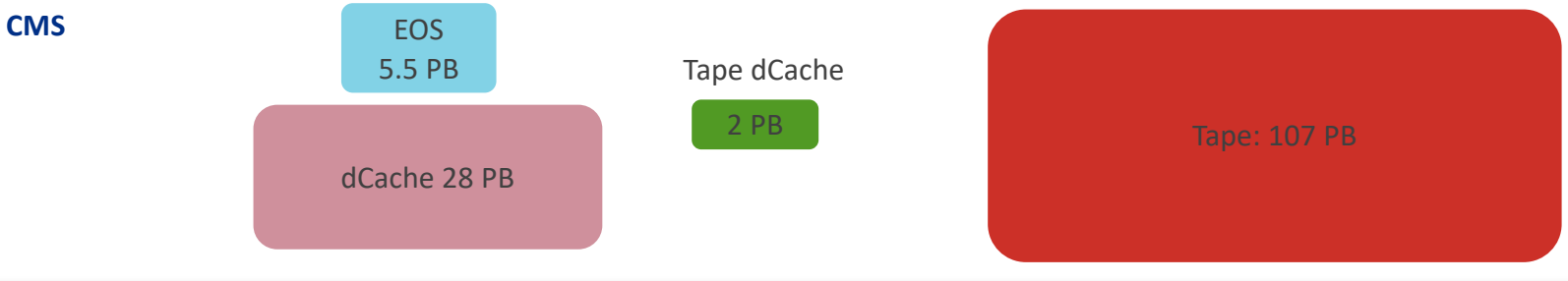
- Scientific Data Services department
  - Covers storage, data management, and scientific database applications

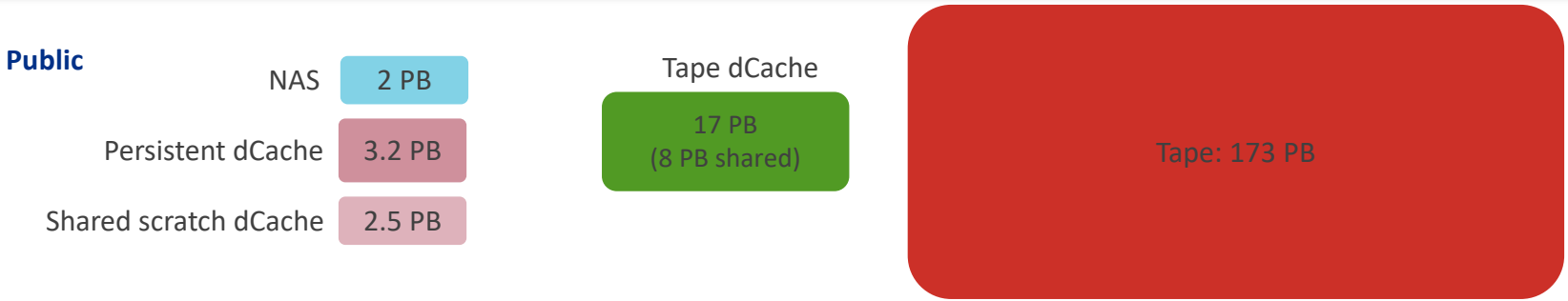
- Storage service
  - Bulk disk
    - dCache, EOS (CMS only)
  - Tape/archival storage

- Data management service
  - Newer experiments moving to Rucio
  - Maintain legacy DM support for ongoing Fermilab experiments

Resources – disk [update]

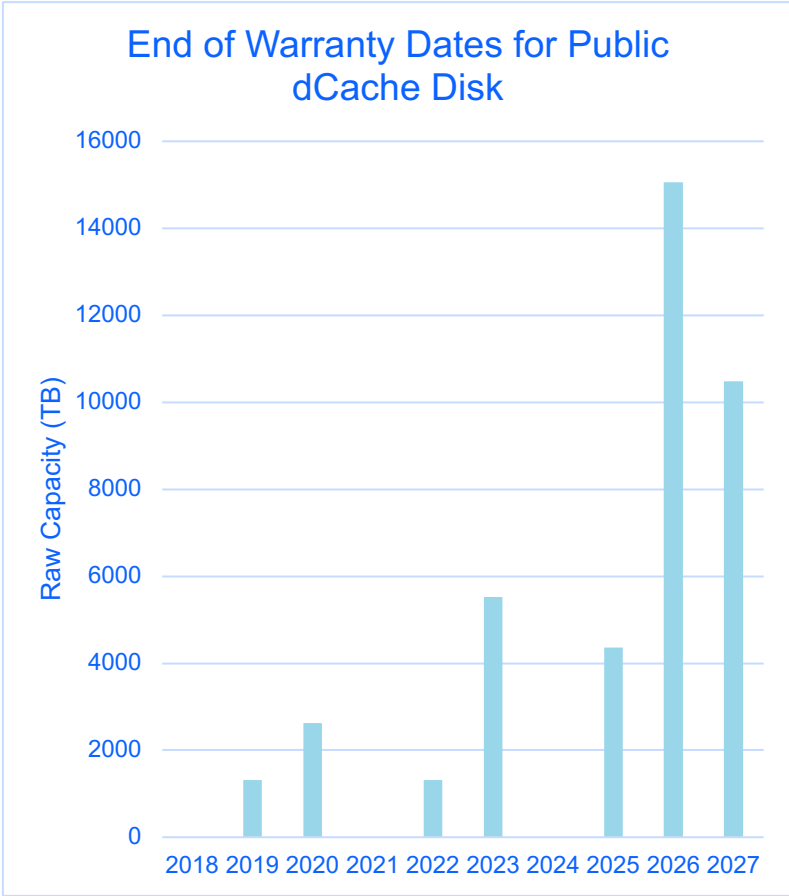
- FNAL dCache (disk) and Enstore (tape) systems are split into two pieces – CMS and “Public” (everything else)
  - I will not be discussing CMS in this presentation.





Resources – disk

- Public dCache disk
  - 28 PB usable available + 9.4 PB just added (FY21 purchase, long procurement)
  - ~5 PB (usable) out of warranty
  - Purchases have been erratic and primary based on available funding rather than need
  - ~~Currently not expecting to add any more disk this year~~
- Transferred some old disk to R&D purposes
  - Investigate non-dCache options (Ceph)
  - Will need some of the in-warranty disks to make further progress



Fermilab

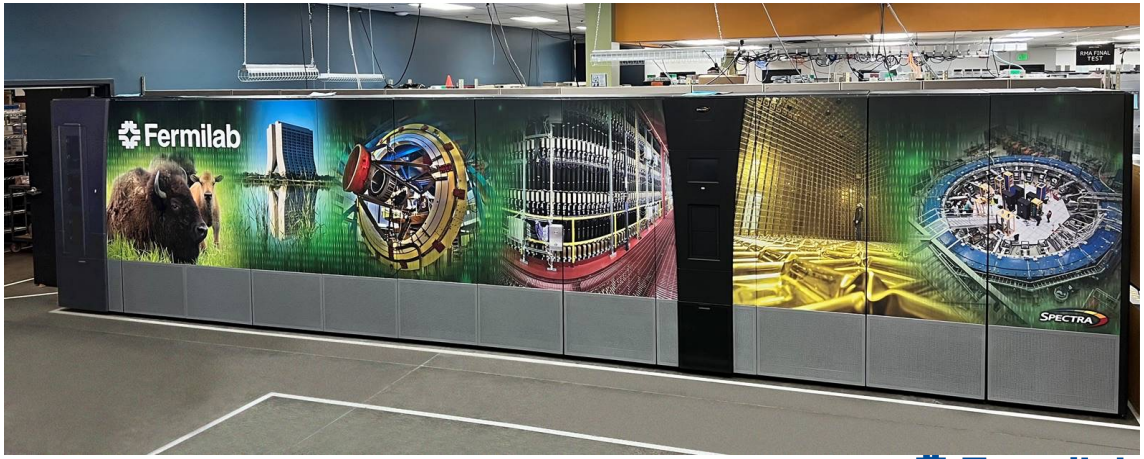
Resources – disk

- Main cache is backed by tape; data staged on access
  - 7.7 PB – aim to maintain 30 day lifetime
    - Shrunk since last year because some was moved to experiment specific pools
  - Past years asking experiments to estimate usage of this didn't prove very useful
- Scratch is another shared resource; LRU file removal, but not tape backed
  - 2.5 PB – similarly aim for 30 day lifetime
- Dedicated tape areas are primarily for raw & production data
  - 9.3 PB
- Persistent space management is permanently resident under experiment control
  - 3.2 PB
- Outside FCRSG scope (tape migration, small experiments, external customers)
  - 1.7 PB

Resources - tape

- Tape complex changes since last year
  - Last year purchased Spectra Logic TFinity library for CMS (12500 slots)
  - Retired and removed 2 more Oracle SL8500
  - Installing new TFinity for Public next week (12500 slots)
  - Now (almost)
    - 3 x TS4500 (120 PB capacity w/ LTO8) – 2 public; 1 CMS
    - 2 x TFinity (150 PB capacity w/ LTO8, 225 PB w/ LTO9) – 1 public; 1 CMS
    - 1 x SL8500 – 1 public; retire when data is migrated to newer libraries
  - Public LTO8 tape drives
    - FCC IBM library (almost full) – 38 LTO8 drives
    - GCC IBM library (most writes now go here) – LTO8 36 drives
    - FCC TFinity library (upcoming) – 40 LTO9 drives
    - Shared drives make it hard to guarantee drive allocation for a specific experiment

In pictures…



Clockwise from above

- New CMS library
- Removing the old library
- New Public library (at Spectra)

Tape management software evolution

- Decision made to move from Enstore to CTA for tape management
  - Need to add reading Enstore formatted tapes to CTA to allow metadata only migration
    - Working with DESY to do this – they are also moving to CTA and their tape layout is very similar to Enstore's
    - DB schema change made in core CTA to track imported tape layout
  - Do not plan to write Enstore format with CTA which simplifies things

- Developing plan/schedule for CMS migration
  - Some policy decisions to make – retain dCache as buffer or use EOS
    - Do we need SSDs for the buffer?

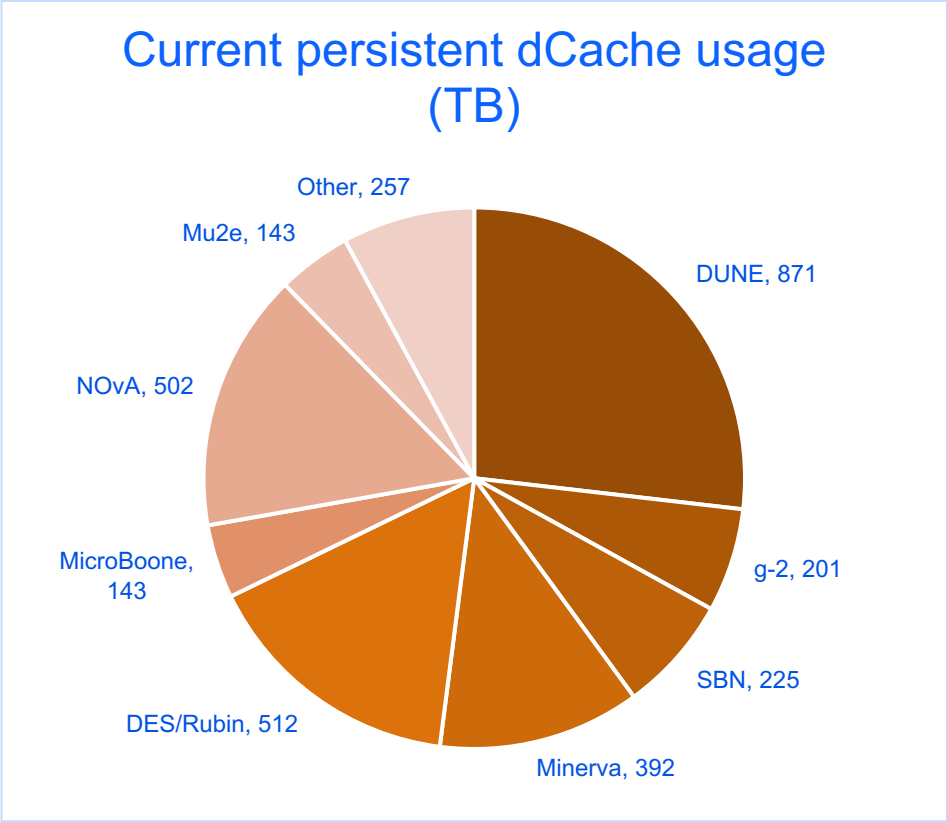
- Public migration is more complex because of Small File Aggregation data
    - Read-only solution looks possible within dCache (no change to CTA needed)
    - More difficult if we need to write – but CERN claim it's not necessary

Requests from experiments – persistent dCache

Experiment	2022 request	2023 request	2024 request
DES	512	512	
DUNE	900	2000	2000
MicroBooNE	151	151	151
Mu2e	148	148	148
g-2	201	200	200
NOvA	500	500	500
SBN	250	300	300
MINERVA	350	350	350
Other	90	90	90
Total	3101	4451	4039

Current allocation is 3246 TB

“Other” only includes FCRSG requests. There are other users beyond these.



Fermilab

Dedicated dCache

- Most tape-backed dCache disk is a common pool; LRU eviction
  - Some gets allocated to a particular experiment or purpose
  - The primary use for this is for raw data uploads
    - Experiments taking data have a dedicated pool that allows flushing to tape more reliably than if they were mixed in with the general pool

In some cases, a larger dedicated allocation has been provided to an experiment to give more reliable access to existing data

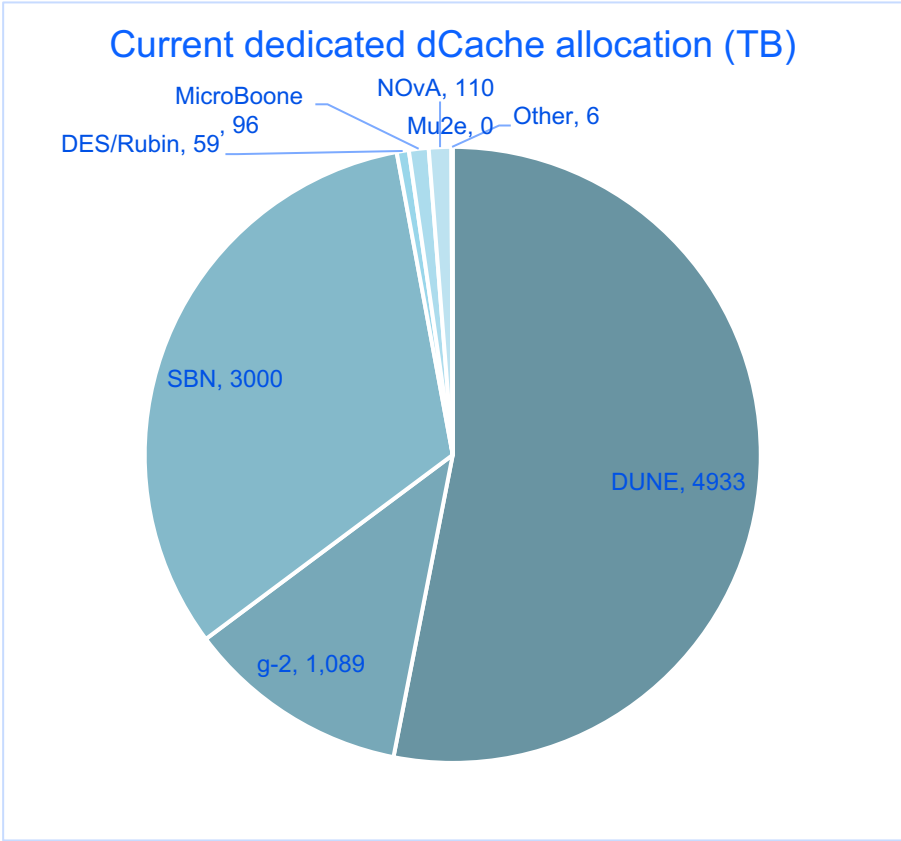
- This has often been done on an ad-hoc basis when experiments complain enough (g-2!)

- Some of these issues are because we are hitting scalability/queuing limits with dCache and Enstore.
  - CTA should alleviate this in the longer term and there are potential near term improvements to the dCache/enstore interface

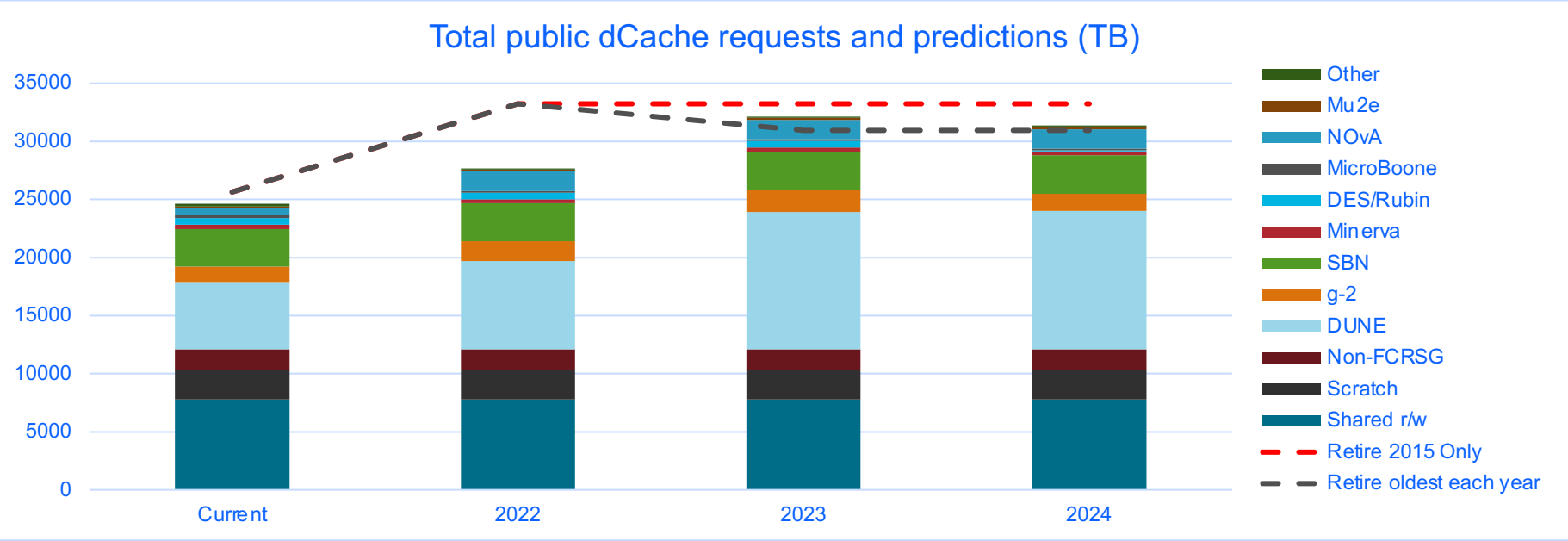
Requests from experiments – dedicated dCache

Experiment	2022 request	2023 request	2024 request
DUNE	6700	9800	9900
SBN	3000	3000	3000
g-2	1500	1500	1000
NOvA	1191	1191	1191
MicroBooNE	0	0	0
Mu2e	20	40	80
Other	6	6	6
Total	12476	15596	15236

- Current allocated total 9300 TB



Total dCache requests

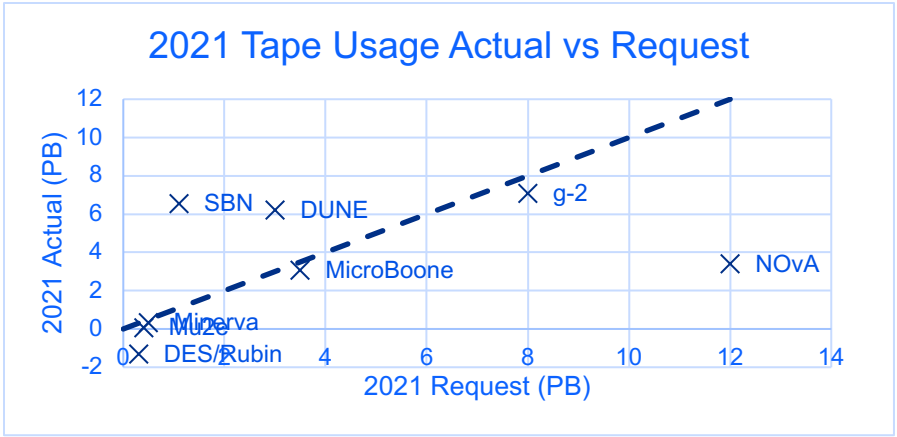


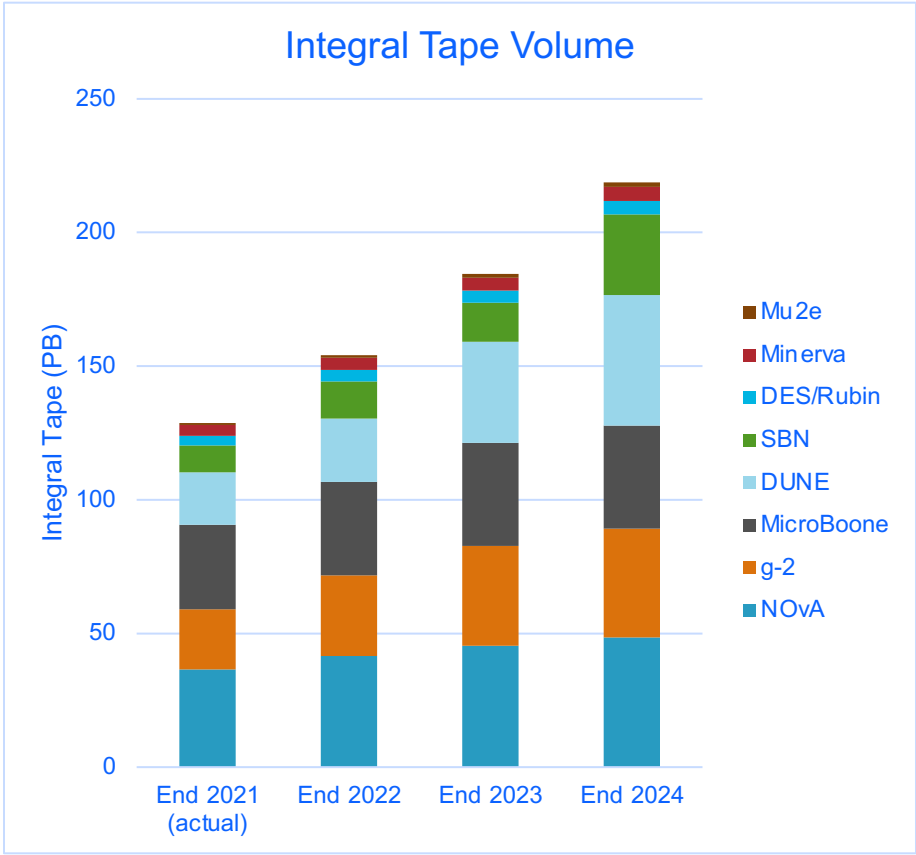
- Assumes no increase in scratch or shared space
- Dashed lines show capacity (usable, no replication).
  - Assumes no additional purchases before 2023
- Red line – retire/repurpose 2015 disks only
- Blue line – retire/repurpose oldest each year

Fermilab

Requests from experiments - tape

- Last year's requests were not a very good guide to actual usage
  - But better than some years
- Most experiments are not considering significant deletion of data on tape
  - Exceptions are SBN and Mu2e





Fermilab

Disk R&D

- Investigating Ceph as additional service for disk storage
- Currently two main foci
  - Object store
    - CMS funded – vague expressions of interest from DUNE, but no effort
  - Cephfs filesystem for interactive/analysis facility usage – **not Grid**
    - Experimenters much prefer POSIX for this; dCache isn't really suitable

- At this point we plan to keep dCache for bulk data storage
  - But long-term going to need an alternative to RAID (industry trends)

- If we pursue Ceph for this will have to trade off dCache space
  - But unlikely 1:1; convenience & performance come at a cost (still cheaper than NAS)

Resources

- In order of effort
  - Tape operations
  - Disk operations
  - Storage development (ramping up)
  - Data Management (mainly Rucio for DUNE)

- Operations is effort constrained
  - Deployed a new more automated migration method over the last year that requires significantly less effort to run
  - For DM, primary goal is full Rucio deployment for DUNE; lack effort for other work (e.g. Icarus)
    - Rubin is providing additional funding for DM, but obviously they want it to go on their needs
      - But work on Rucio benefits everybody

Personnel

- Including IF, CMS, and Rubin funding
  - Operations
    - Storage 8 FTE + 1 pending hire
    - Data management 0.75 FTE
  - Development
    - Storage 5 FTE + 1 pending hire
    - Data management ~2.5 FTE

- Retirements/departures
  - Primary Enstore developer retired early this year
  - Expert tape operator retired early this year

- Hiring & replacements
  - Filled 2 new storage R&D positions, 1 more starting August
  - Filled 0.5 FTE new developer for data management
  - Final stages of filling operations position
  - Internal transfer transitioning to tape development; will fully transfer when backfilled

Conclusions

- Requested tape usage growth still significant
- Disk purchases are still based on available funds rather than needs, but are managing to keep pace with requests
- Still a number of places where resources could be used more efficiently
  - Both on service side
    - Poor tape request scheduling between enstore & dcache
    - Better control of tape staging (need QoS management in Rucio)
  - And experiments
    - Dataset lifecycles and deleting unneeded data
    - Transition to new technologies (e.g. Rucio)
- Added effort to storage R&D
  - CTA transition
  - Potential Cephfs replacement for NAS
- But both of these will cause more operations load