Storage Services

Robert Illingworth

FCRSG

16 Feb 2023

Scope

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

- Storage service
  - Bulk disk
    - dCache, EOS (CMS only), Lustre (Wilson Cluster/LQCD), Ceph (not yet in production)
  - Tape/archival storage
    - Enstore, CTA (not yet in production)

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

Personnel

- Including IF, CMS, and Rubin funding
  - Operations
    - Storage 9 FTE
    - Data management 0.75 FTE
  - Development
    - Storage 6 FTE
    - Data management ~2.5 FTE

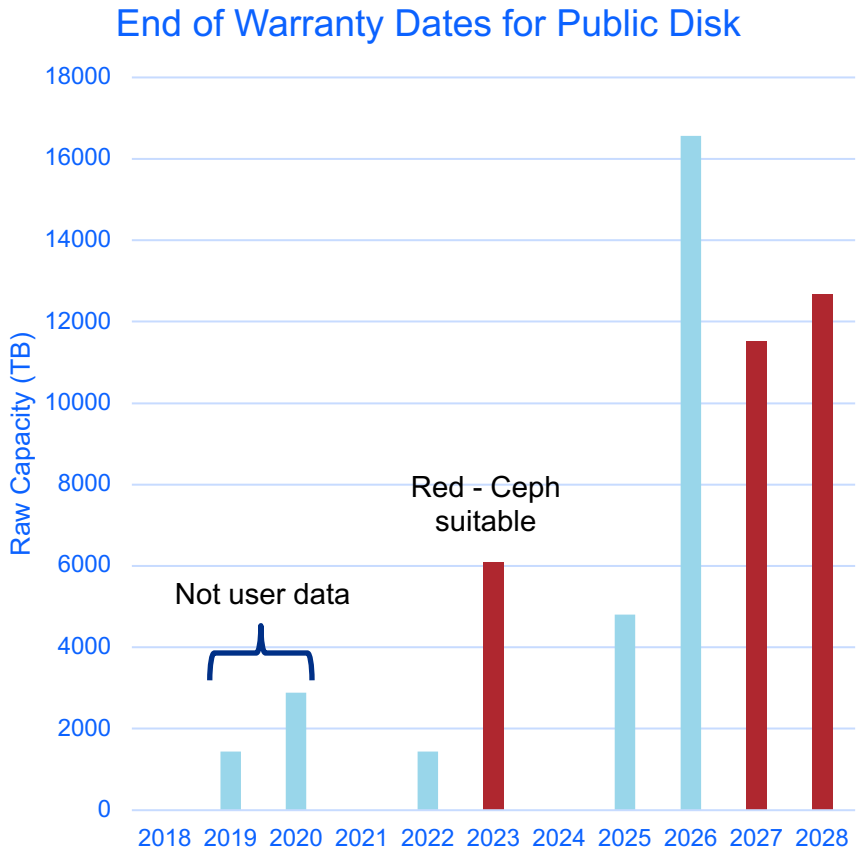
Added more developer and R&D effort recently, but long term Enstore developer retired last year, which has caused some disruption

Resources – public disk

- "Public" basically means non-CMS
- 36 PB raw HDD space available for user data + 13 PB recently delivered
  - Additional ~5 PB raw OOW used for tape media migration and R&D tasks
- ~1 PB NVMe on order
  - Could be used for data storage or as tape buffer

- Vast majority of HDD are configured for dCache as RAID
  - Plan to transfer all 6 PB (raw) of the very nearly OOW disk to Ceph (see later slide)

- Past purchasing history means that we will have large amounts of disk leaving warranty in 2025 onwards



dCache

- dCache provides grid accessible bulk storage and tape interface
- Main cache is backed by tape; data staged on access
  - 8.7 PB – aim to maintain 30 day lifetime; +1 PB since last year
- Scratch is another shared resource; LRU file removal, but not tape backed
  - 2.5 PB – similarly aim for 30 day lifetime; no change since last year
- Dedicated tape-backed areas are allocated to specific experiments primarily for raw & production data
  - 9.3 PB; +1.3 PB since last year
- Persistent space is permanently resident on disk (not tape cache) under experiment control
  - 7.9 PB; +4 PB since last year: a move towards more experiment managed tape recall
- Outside FCRSG scope (small experiments, external customers, some unallocated space)
  - 1.4 PB

Ceph

- dCache is less effective for interactive/POSIX filesystem style use
  - dCache NFS is a significant source of support load
  - Anticipate EAF creating more demand for this type of interactive filesystem usage

- The scientific disk area of the Central NAS (ITD managed) is reaching end-of-life and a direct replacement would be very expensive

- In the process of deploying Ceph cluster to provide filesystem space to replace the NAS
  - Level of demand is currently not clear
  - Plan to move 6 PB raw of JBOD capable disk (but OOW) from dCache to this.
    - Much of the existing disk is HW RAID unsuitable for Ceph
    - Question: how much new, JBOD capable, disk should be allocated? The responses from experiments didn't help much, likely due to lack of any experience with this system
    - Size of individual servers and need for redundancy means that the minimum allocation is a lot more than our current perceived requirements for the system

Ceph cont.

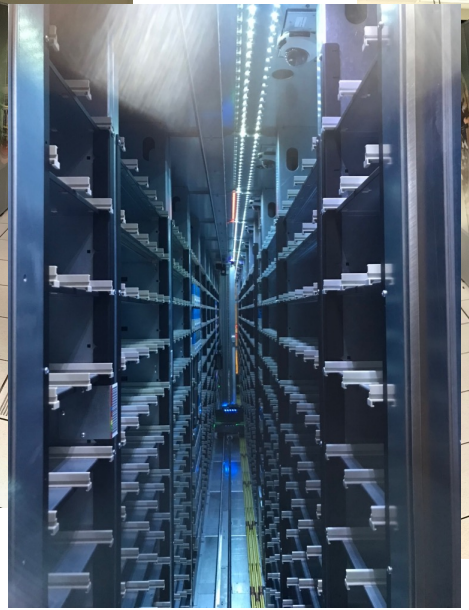
- Would also consider SSD space if there are use cases
  - Resource requests generally focus on capacity; not performance
  - Working space for EAF, but beyond that…?

- Additionally, continuing investigations on using the object store for physics data
  - Effort coming from CMS; some vague interest from DUNE, but no people
  - No immediate plans to put S3 object store on Public cluster, but it's certainly something we would pursue if the requirement is there

Resources - tape

- Added new Spectra Logic TFinity tape library for Public
- Tape complex is now
  - 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 later this year
  - Public LTO tape drives
    - FCC IBM library (almost full) – 38 LTO8 drives
    - GCC IBM library (almost full, currently most reads) – LTO8 36 drives
    - FCC TFinity library (most writes now go here) – 40 LTO9 drives
    - Shared drives make it hard to guarantee drive allocation for a specific experiment
    - Unbalanced usage between libraries is an issue
      - Moved g-2 tapes between GCC and FCC last year to maximize drive utilization
      - We need to consider further efforts to spread the load

New tape library



Fermilab

Tape concerns

- LTO8 - delayed by patent dispute
- LTO9 - a year later than expected and undersized (18 TB vs 24 TB)
- LTO10 - ?

- Drive r/w rates are increasing much slower than capacity (even taking into account smaller than expected LTO9)

- Industry trends
  - Judging by vendors' latest offerings the hyperscalers are driving development, and their requirements are not well aligned with ours
  - Single supplier of tape drives; single manufacturer of r/w heads; two manufacturers of tape

- Requires more organized recall effort from experiments and better tools to enable this
  - Things like the method we are using to stage g-2 data recently are not sustainable long term
  - Choice of layout on tape may come back to bite you 3-4 years later
  - Efficient staging of datasets may require knowledge of tape locations (whole tape vs individual files)

Tape management software evolution

- Proceeding with plan to move from Enstore to CTA for tape management
  - We have demonstrated (small-scale) migration of Enstore metadata to CTA
  - Successfully read back entire Enstore written tapes with CTA
  - Demonstrates key requirements to perform the migration for CMS
  - But, of course, many details still to work out, and larger scale testing

- For public data, main problem is how to handle Small File Aggregation data
  - CTA has no support for anything like like that
  - Conceptual ideas about how to read existing data, but needs to be turned into a concrete design
  - Unclear if write is still needed; it will be much easier if we don't need to implement this

Data management

- Continuing push to move experiments off SAM to Rucio/Metacat/etc
  - Primary focus has been DUNE (for ProtoDUNE II), plus some ICARUS
  - Need to move forward with Mu2e
  - Added extra 0.5 FTE to IF efforts this year (to 1 FTE)
  - Path for legacy experiments (particularly NOvA) is still unclear
    - Aside from token authentication and some changes to address issues with tape staging, essentially no SAM development recently (or envisioned)

- A request from multiple experiments is for easier ways to move data to/from HPC sites
  - Starting to experiment with bridge from dCache to Globus Online (via Cephfs space)
  - Automation should be possible via Rucio

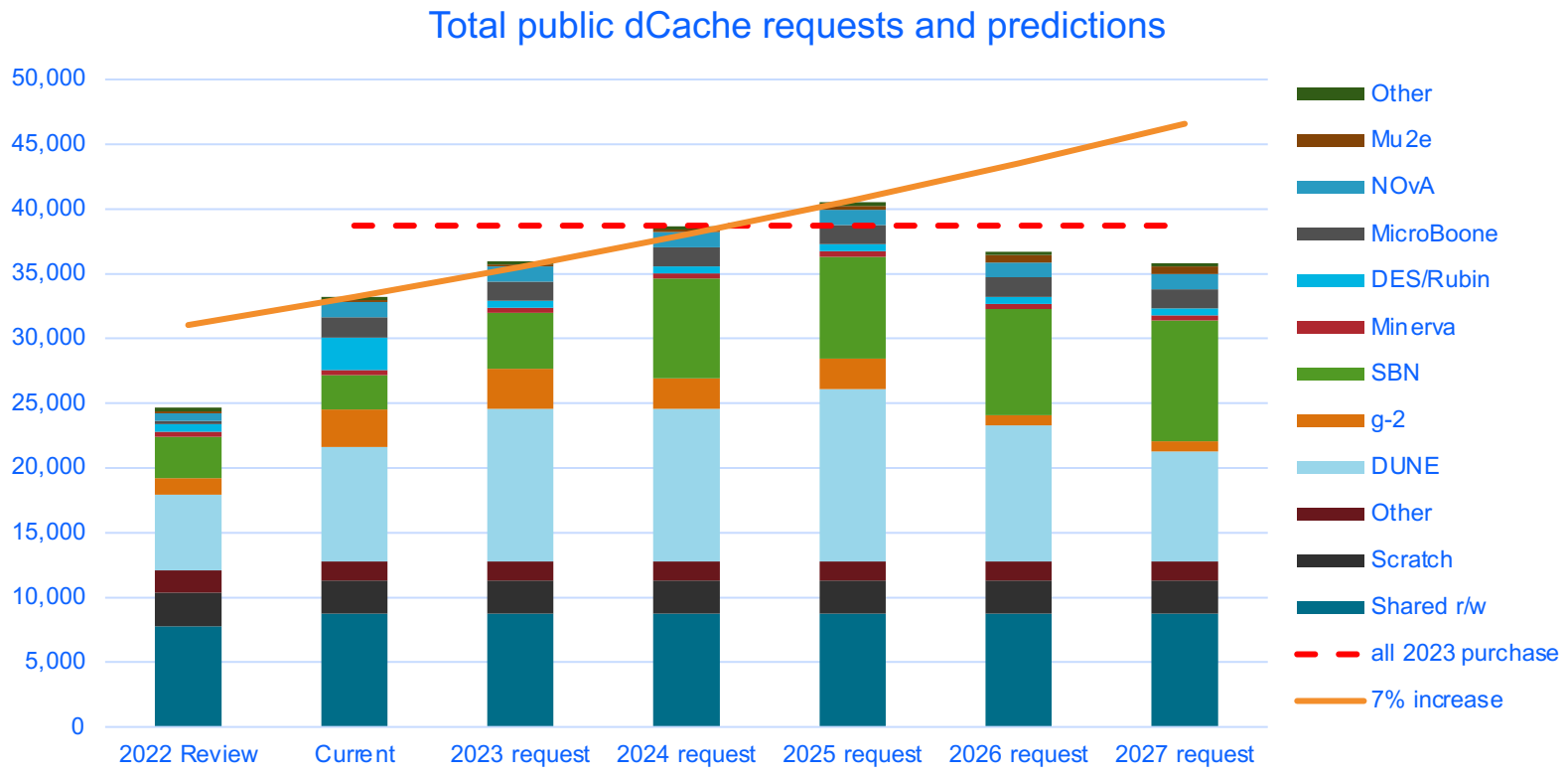
dCache disk requests

Allocated TB	2022 Review	Current	2023 request	2024 request	2025 request	2026 request	2027 request
DUNE	5,804	8,860	11,800	11,800	13,300	10,500	8,500
g-2	1,290	2,886	3,090	2,350	2,350	800	800
SBN	3,225	2,655	4,300	7,700	7,900	8,200	9,300
Minerva	392	398	400	400	400	400	400
DES/Rubin	571	2,495	540	540	540	540	540
MicroBoone	239	1,572	1,500	1,500	1,500	1,500	1,500
NOvA	612	1,172	1,171	1,171	1,171	1,171	1,171
Mu2e	143	143	148	168	320	560	560
Other	263	259	258	258	258	258	258
Shared r/w	7,785	8,732	8,732	8,732	8,732	8,732	8,732
Scratch	2,574	2,576	2,576	2,576	2,576	2,576	2,576
Other	1,750	1,464	1,464	1,464	1,464	1,464	1,464
Total	24,648	33,210	35,979	38,659	40,511	36,701	35,801

Notes:

Combined “persistent” and “dedicated” requests for experiments
”Other” includes unallocated space
Assumes no change in shared or Non-FCRSG totals
Old disks used for migration are excluded
DES/Rubin current include 2 PB temporary “lent” capacity they are in the process of removing

dCache disk requests



Fermilab

Tape requests – annual delta

Delta TB Experiment	2022 (actual)	2023	2024	2025	2026	2027
DUNE	3,170	12,100	12,100	12,600	6,000	4,500
g-2	10,771	7,800	1,500	500	0	0
SBN	6,659	3,800	12,300	14,400	8,600	6,200
DES/Rubin	1,035	0	1,000	0	0	0
MicroBoone	-4,772	2,500	1,500	1,500	0	0
NOvA	6,242	4,000	3,000	3,000	2,000	2,000
Mu2e	-5	500	1,000	2,000	15,000	5,000
Subtotal	23,100	30,700	32,400	34,000	31,600	17,700
Other	2,560	2,560	2,560	2,560	2,560	2,560
Grand total (excl CMS)	25,660	33,260	34,960	36,560	34,160	20,260

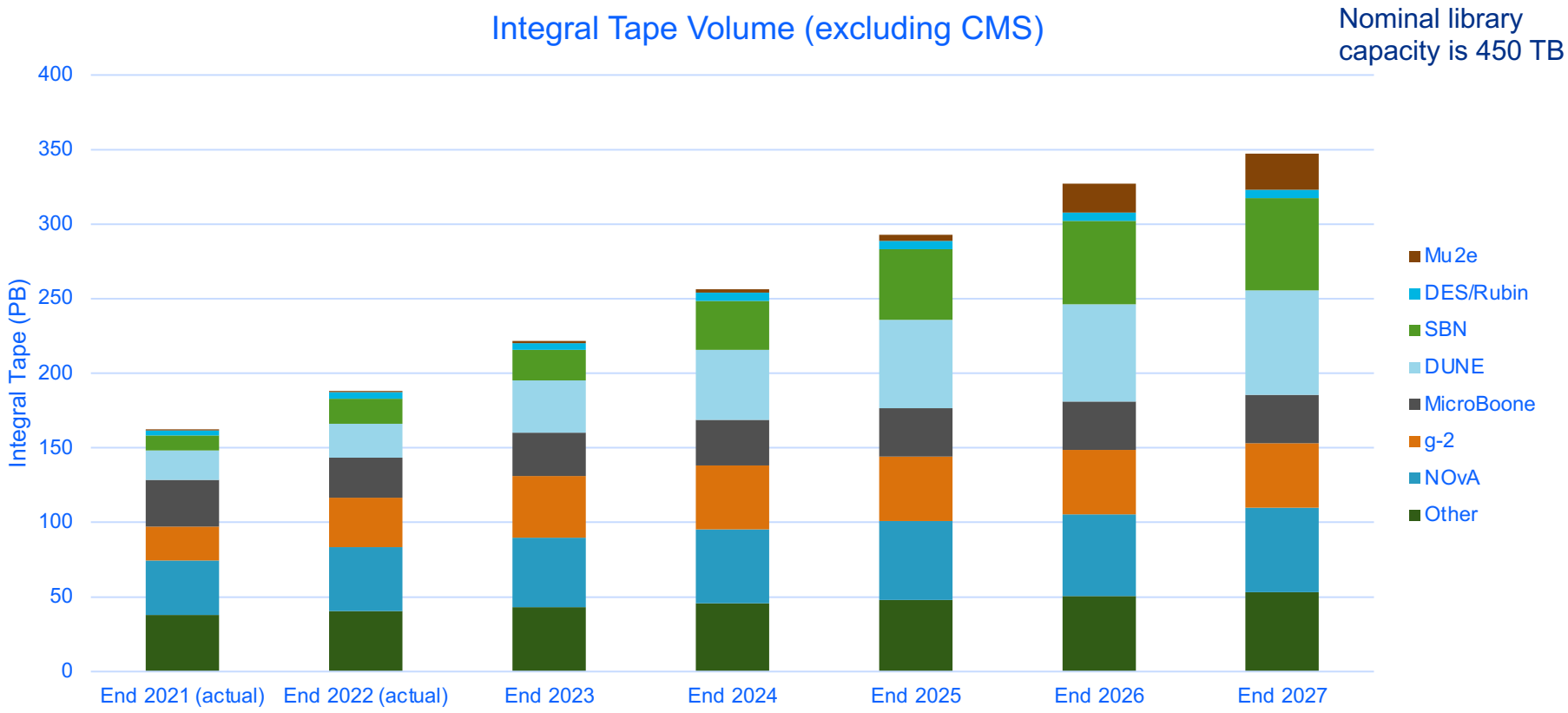
MicroBoone deleted over 6 PB this year!

Tape requests – integrals

Integral TB	End 2021 (actual)	End 2022 (actual)	End 2023	End 2024	End 2025	End 2026	End 2027
DUNE	19,504	22,674	34,774	46,874	59,474	65,474	69,974
g-2	22,655	33,426	41,226	42,726	43,226	43,226	43,226
SBN	10,032	16,691	20,491	32,791	47,191	55,791	61,991
DES/Rubin	3,667	4,702	4,702	5,702	5,702	5,702	5,702
MicroBoone	31,552	26,780	29,280	30,780	32,280	32,280	32,280
NOvA	36,517	42,759	46,759	49,759	52,759	54,759	56,759
Mu2e	758	753	1,253	2,253	4,253	19,253	24,253
Subtotal	124,685	147,785	178,485	210,885	244,885	276,485	294,185
Other	37,909	40,469	43,029	45,589	48,149	50,709	53,269
Grand total (excl CMS)	162,594	188,254	221,514	256,474	293,034	327,194	347,454

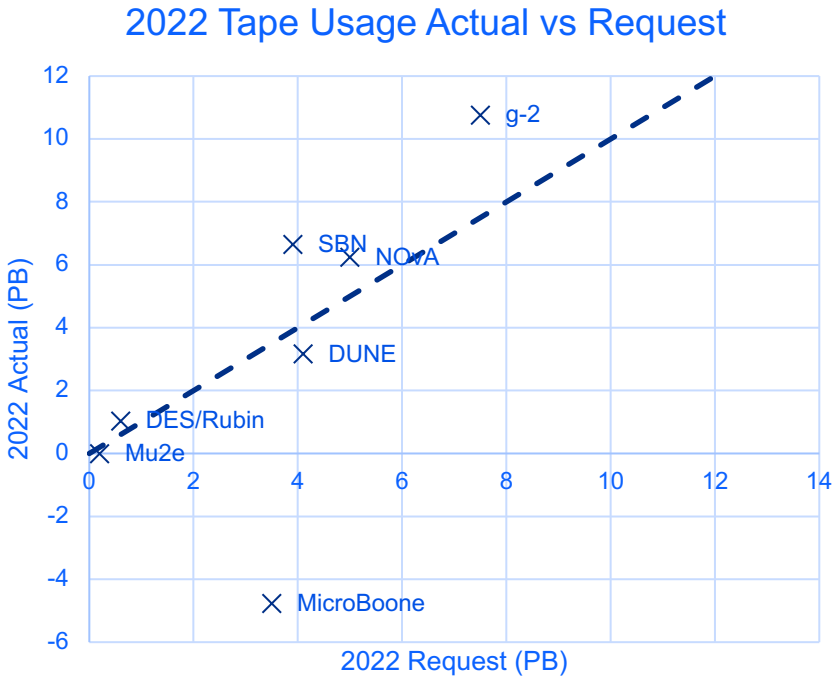
“Other” includes ~20 PB of Tevatron Run II data

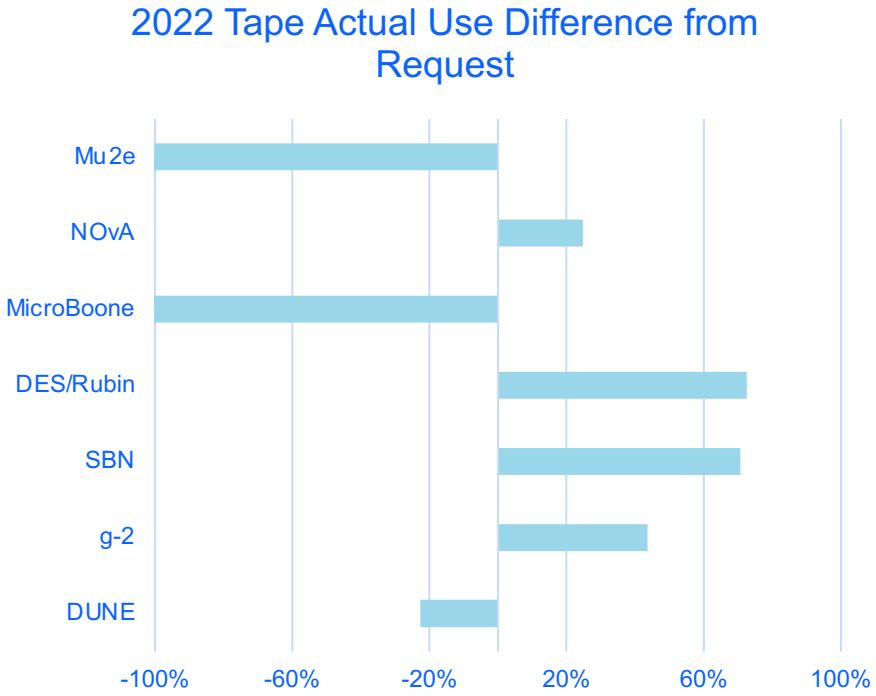
Tape requests - totals



Fermilab

Last year requests compared to actual





Fermilab