Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Fermilab Storage Strategy

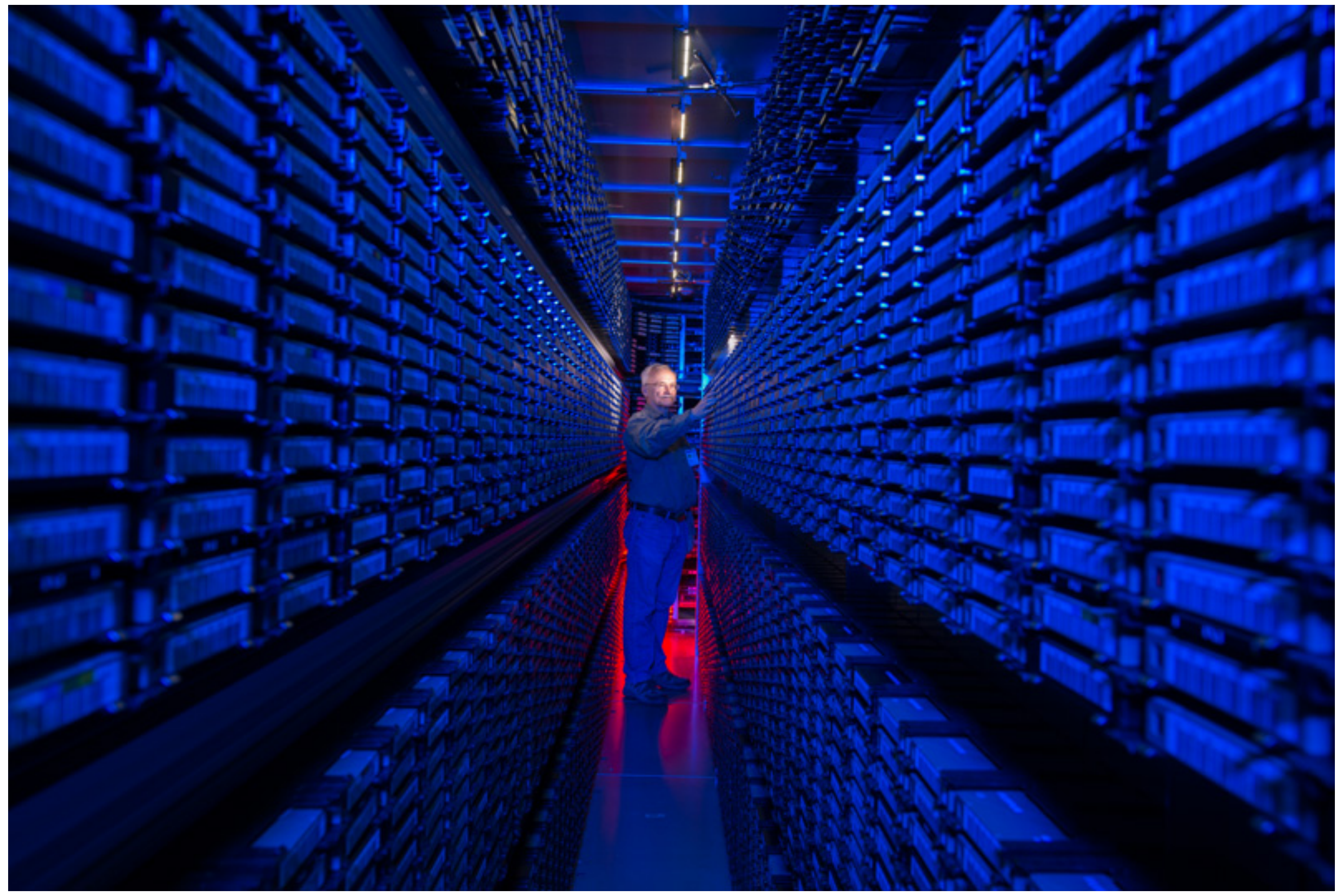
Bo Jayatilaka (Fermilab SCD)
2nd International Computing Advisory Committee Meeting
15 October 2019

Current storage landscape

- Custodial and active storage for all Fermilab experiments’ scientific data
  - This includes considerable storage for “external” experiments/projects (e.g. CMS and DES)

- Utilizing a tape+(spinning) disk HSM
  - Tape managed by **Enstore** (Fermilab)
  - Disk managed by **dCache** (DESY+Fermilab+NDGF)

- **198 PB** of tape in use (178 PB active) as of 10/1



Storage infrastructure

- Two major storage "instances"
  - Public: experiments/projects on detector ops funding, DES, LQCD, AAF
  - CMS: dedicated for CMS Tier 1 storage
    - Also managed: analysis-only EOS pool

- Dedicated hardware for each instance
  - Tape library complexes
  - Multiple dCache pools

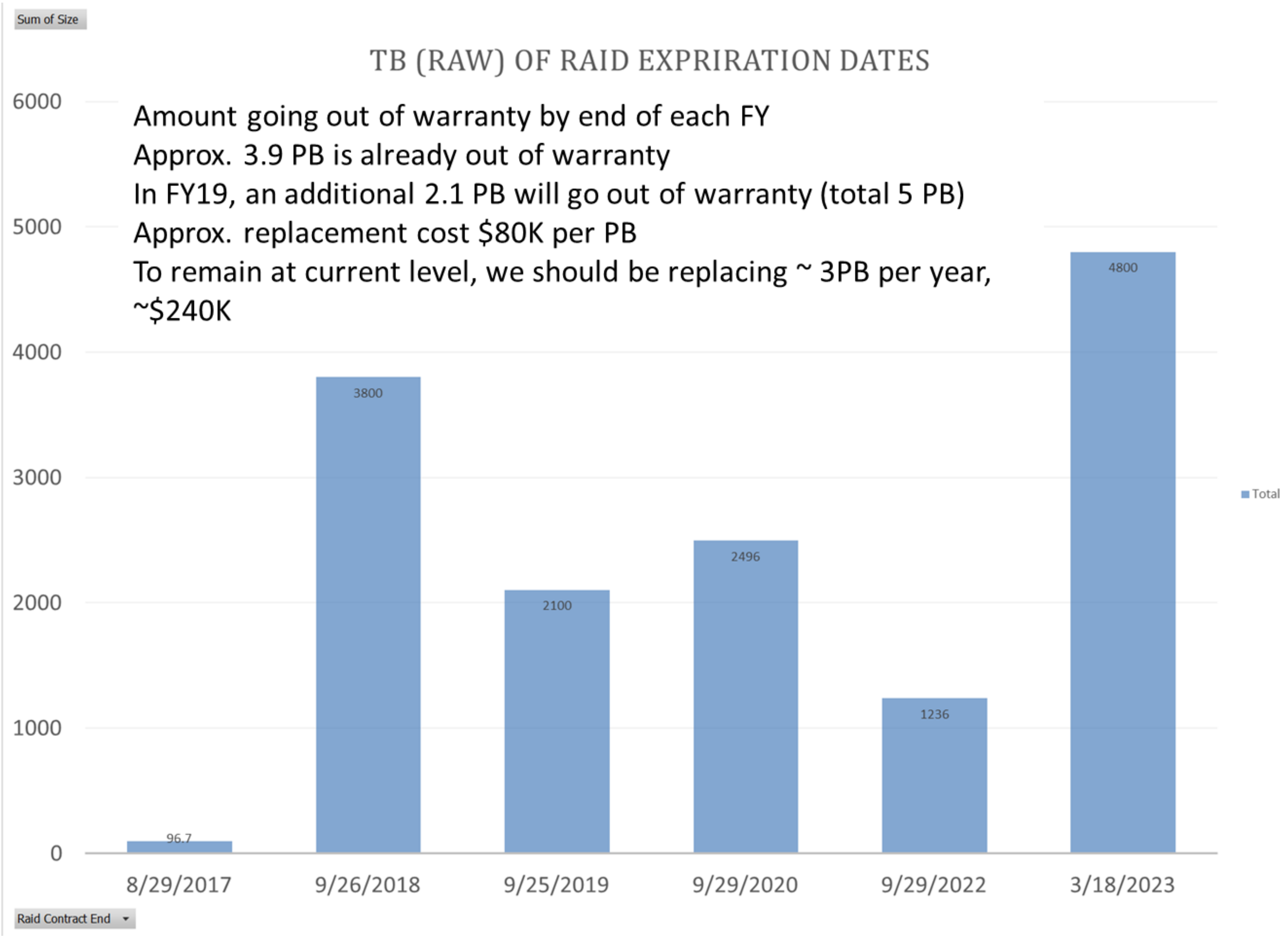
- **Resource allocation** and **use cases** differ considerably between the two

Tape hardware: current state

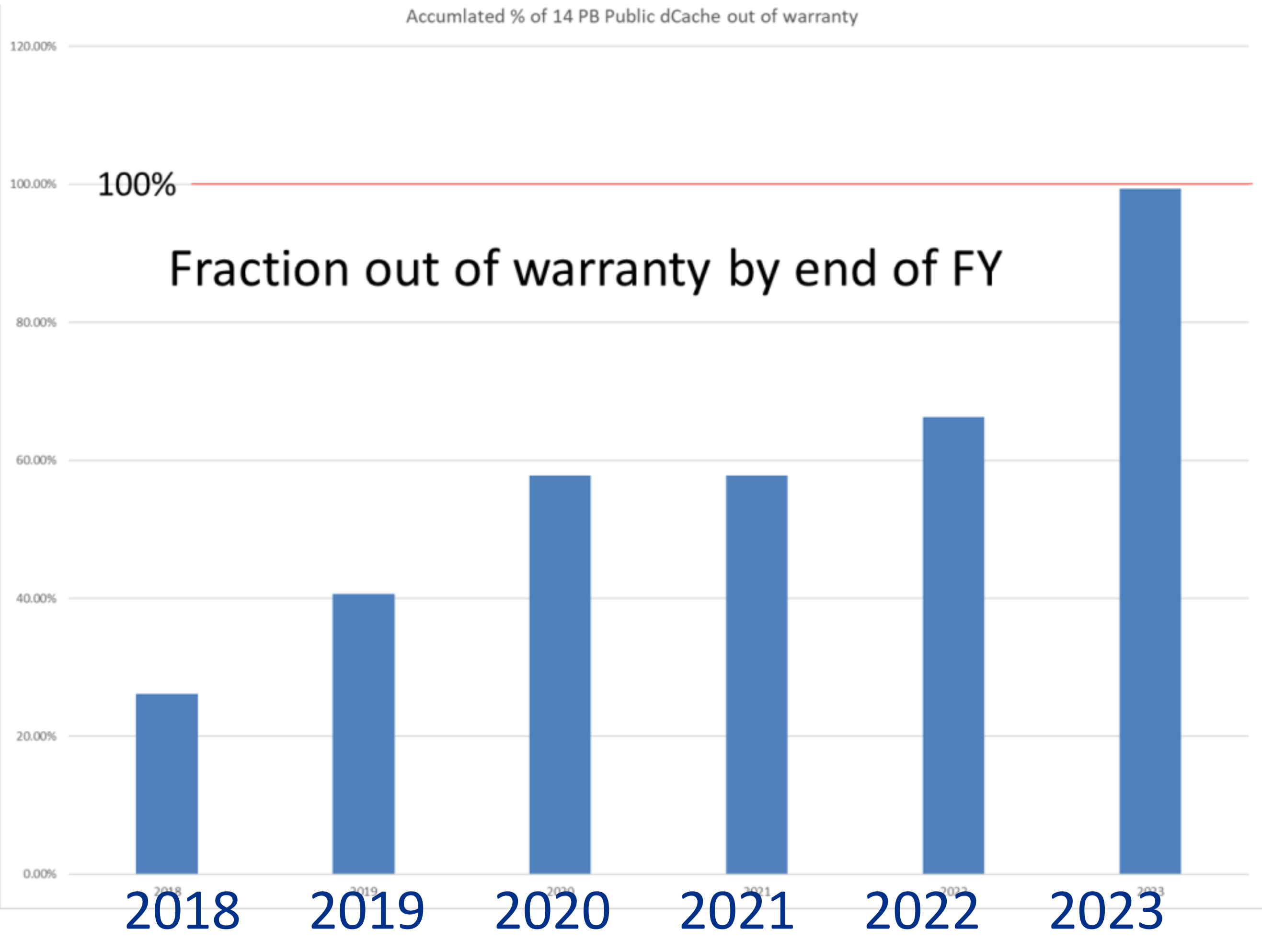
- Prior to June 2018, all tape libraries were Oracle/StorageTek SL8500 (7*10k-slot)
  - Combination of T10KC, T10KD, and LTO4 drives
  - 125 PB active storage
  - Plan was to acquire T10KE (~15TB) drives when available
    - Oracle informed us in mid-2017 that their enterprise tape line would end
- RFP process in early 2018: IBM TS4500 libraries with LTO8 drives chosen
  - Three libraries (two in production)
    - One for Public (56 LTO8 drives) and one for CMS (36 LTO8 drives)
    - Apart from initial batch of 100 cartridges, all new media has been "M8" (LTO7 formatted to 9TB capacity)
  - *Considerable effort in development (Enstore) and operational integration*

Disk hardware: current state

- Commodity SAN configuration (storage servers+disk arrays)
  - Identical configurations purchased, when possible, for Public and CMS
  - Most recent purchases result in ~70TB usable storage per array



Amount going out of warranty by end of each FY
Approx. 3.9 PB is already out of warranty
In FY19, an additional 2.1 PB will go out of warranty (total 5 PB)
Approx. replacement cost $80K per PB
To remain at current level, we should be replacing ~ 3PB per year, ~$240K

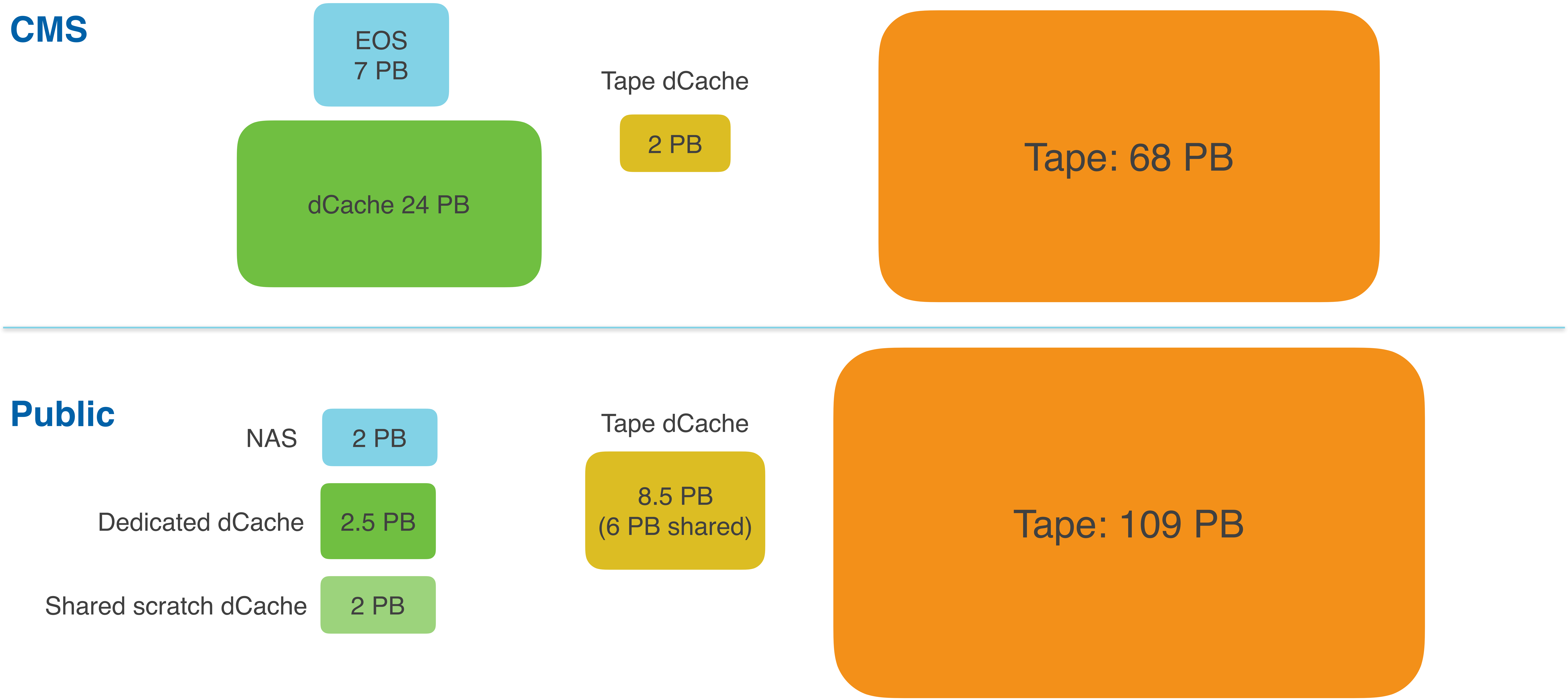


Fraction out of warranty by end of FY

Bottom line:
Funding constraints unlikely to allow little expansion of Public disk

S. Fuess, 1st ICAC meeting

Storage infrastructure in a nutshell



CMS

EOS
7 PB

dCache 24 PB

Tape dCache

2 PB

Tape: 68 PB

Public

NAS 2 PB

Dedicated dCache 2.5 PB

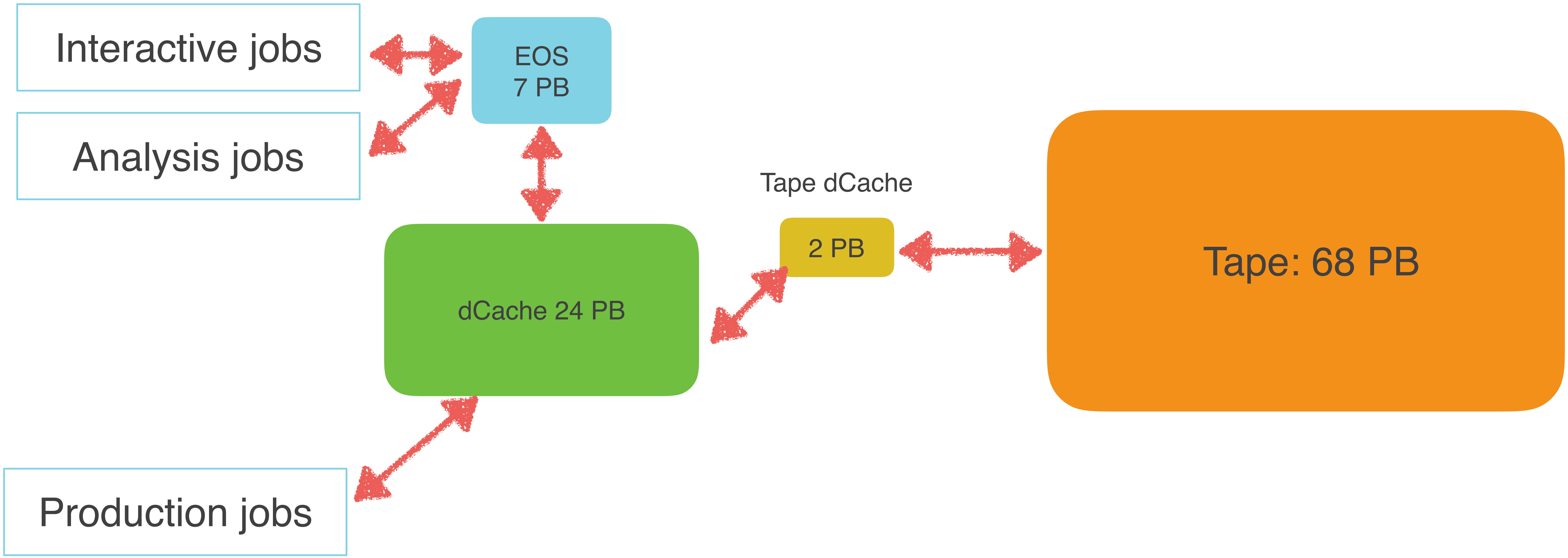
Shared scratch dCache 2 PB

Tape dCache

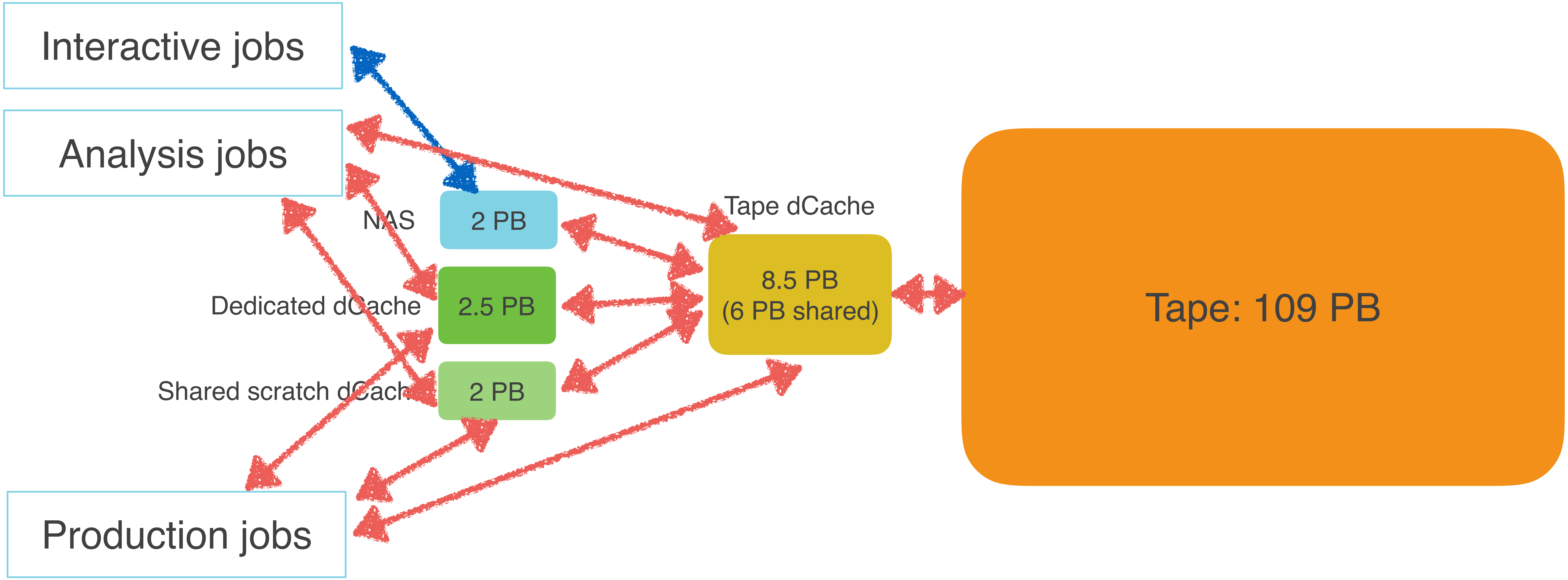
8.5 PB
(6 PB shared)

Tape: 109 PB

What goes where: CMS



What goes where: Public



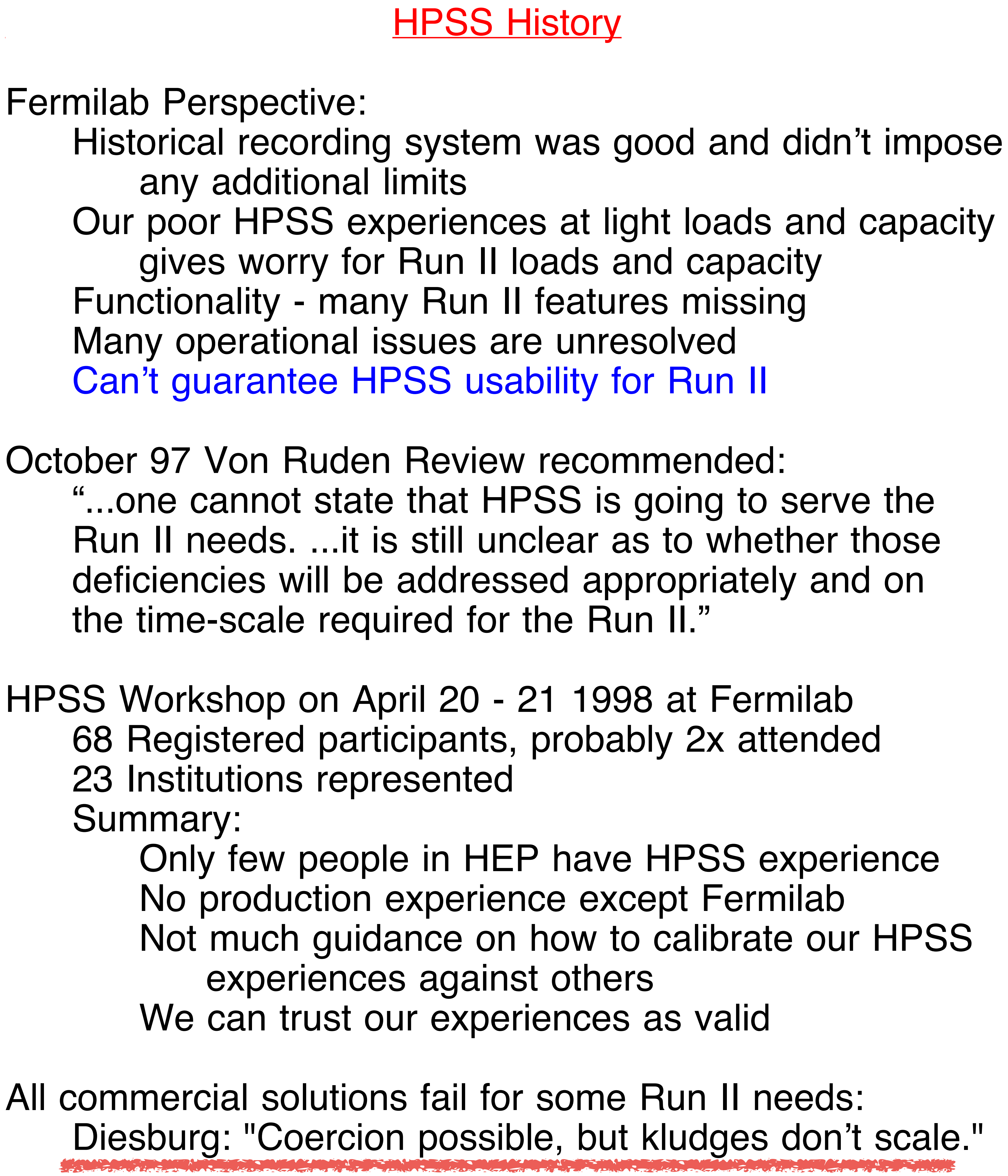
Differences between Public and CMS

- **Disk:Tape ratio** is considerably higher in CMS
  - CMS, globally, has approximately 1:1 tape:disk
- **Tape-cache disk and general-use disk separation** in CMS
  - Shared Public dCache pools are simultaneously used for tape recall and grid-job input
- **Dedicated analysis disk** available for CMS
  - Access patterns between the two can differ considerably
  - CMS has a dedicated EOS instance for analysis use
- **User data** is not tape-backed in CMS
  - User data is a large contribution to the proliferation of files (>1B) on Public
- **Multiple VOs** contend for shared tape and disk resources in Public
  - Some larger experiments have dedicated dCache pools
  - Tape drive contention cannot be similarly avoided

Storage model evolution: near term

- Implementing some features of the CMS model for Public is low-hanging fruit
  - e.g. disk/tape separation, analysis disk separation
- Introducing additional storage layers for high-IOPS use cases
  - e.g., NVMe
  - Currently dealt with “resilient” dCache (replicated pools)
- ~~Sociological conditioning~~ user/experiment training
  - Plan for large campaigns ahead of time (e.g., allow for pre-staging)
  - Optimizing workflows/code to available IO
- Operations streamlining
  - Understand if hardware running services is currently optimal
  - Efficiency improvements in service software
    - e.g., is queue logic suitable for future tape capacities and use cases?

Enstore: some history



HPSS History

Fermilab Perspective:

- Historical recording system was good and didn't impose any additional limits
- Our poor HPSS experiences at light loads and capacity gives worry for Run II loads and capacity
- Functionality - many Run II features missing
- Many operational issues are unresolved
- Can't guarantee HPSS usability for Run II

October 97 Von Ruden Review recommended:

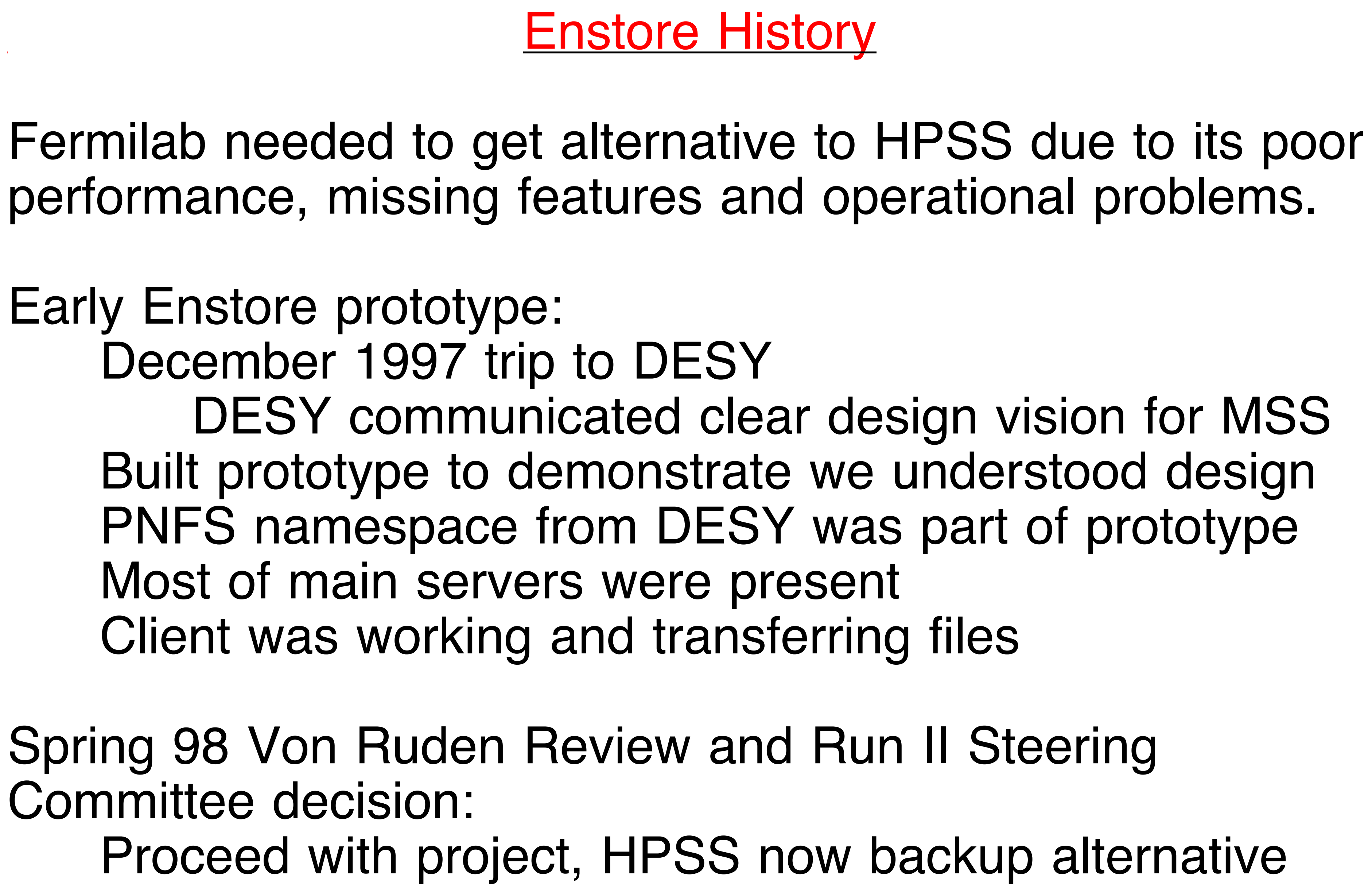
> "...one cannot state that HPSS is going to serve the Run II needs. ...it is still unclear as to whether those deficiencies will be addressed appropriately and on the time-scale required for the Run II."

HPSS Workshop on April 20 - 21 1998 at Fermilab

- 68 Registered participants, probably 2x attended
- 23 Institutions represented
- Summary:
  - Only few people in HEP have HPSS experience
  - No production experience except Fermilab
  - Not much guidance on how to calibrate our HPSS experiences against others
  - We can trust our experiences as valid

All commercial solutions fail for some Run II needs:

- Diesburg: "Coercion possible, but kludges don't scale."



Enstore History

Fermilab needed to get alternative to HPSS due to its poor performance, missing features and operational problems.

Early Enstore prototype:

- December 1997 trip to DESY
  - DESY communicated clear design vision for MSS
- Built prototype to demonstrate we understood design
- PNFS namespace from DESY was part of prototype
- Most of main servers were present
- Client was working and transferring files

Spring 98 Von Ruden Review and Run II Steering Committee decision:

- Proceed with project, HPSS now backup alternative

J. Bakken

D0 Workshop

7/1/1999

Enstore: today

- Provides access and control to a data volume ~10 times that of Run II
  - Sustained effort of ~4 FTE operations and ~1 FTE development
  - Operates on ~6 dedicated servers

- Most development efforts have been operations-driven
  - Primarily to implement functionality on new tape hardware
  - Last major feature development was small-files aggregation

- Expect ~350 PB of data on tape by the end of 2022
  - Current system is expected to scale to those levels

- Scaling Enstore to the HL-LHC/DUNE era is **not** a given

Tape: evolution

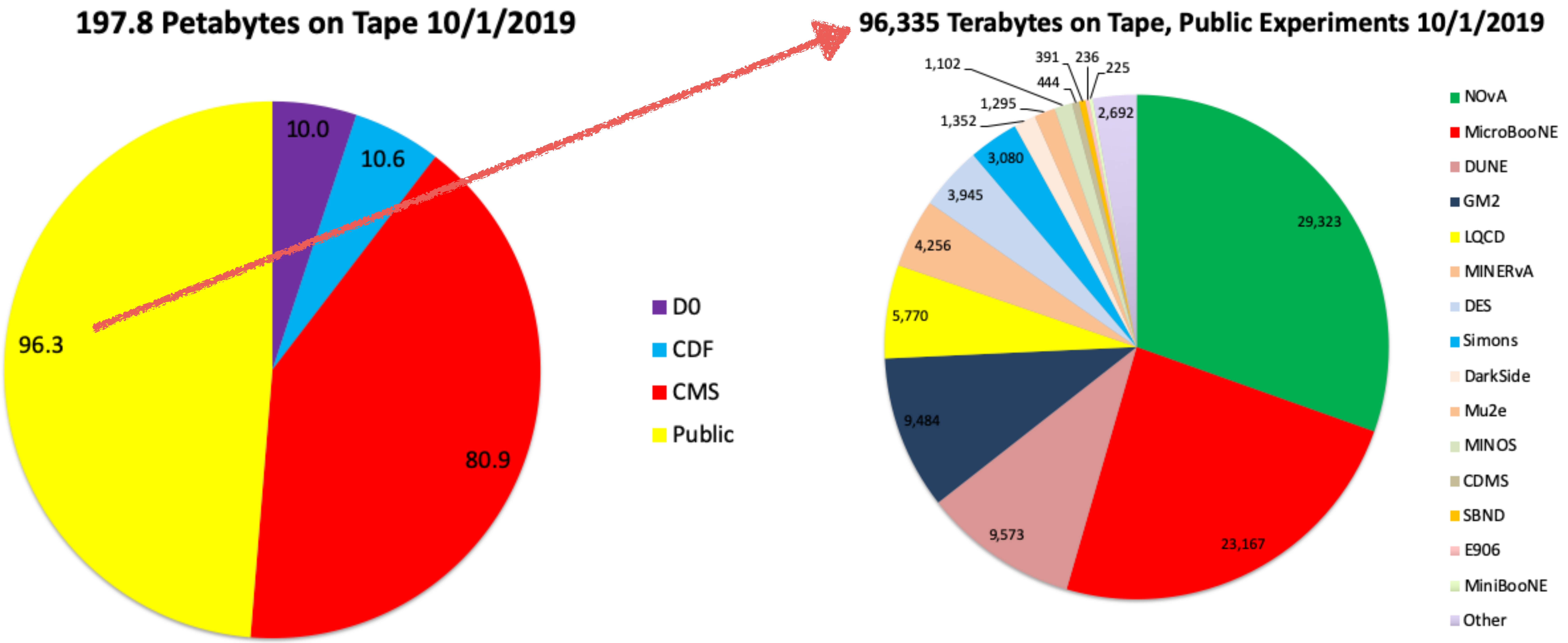
- Not many options for large-scale tape control software
  - CERN developing CTA to replace CASTOR
    - No current plans to make it a “customer” product
    - Can be evaluated when it does
  - HPSS has come a long way since 1999
    - Still largely geared to “backup” tape customers
  - Chosen solution must be **community supported**
  - Taking part in **DoE Distributed Archive Storage System** (DASS) RFP development
    - Goal 10-30 EB archive/active storage system across national labs
    - Planning an RFP for early 2020
- **Whither tape?**
  - No serious, **cost-effective** alternative to tape as archival storage is foreseen
  - Industry trends can shift fast, however

Longer term considerations and questions

- Object storage or other non-file paradigms?
  - Event-based organization continues to be optimal for reconstruction/production of collider data
    - Not necessarily the best for DUNE-like experiments or physics analysis of any kind
  - Changes in storage paradigms might necessitate development effort in Rucio
- Data lifetimes
  - Should all archival data have a set lifetime?
  - Migration across tape formats threatens to be a continuous process otherwise
- Community alignment
  - How do storage/DOMA efforts across the field stay in sync?
  - Should Fermilab be more involved in ongoing community DOMA efforts
    - How do we accomplish this?

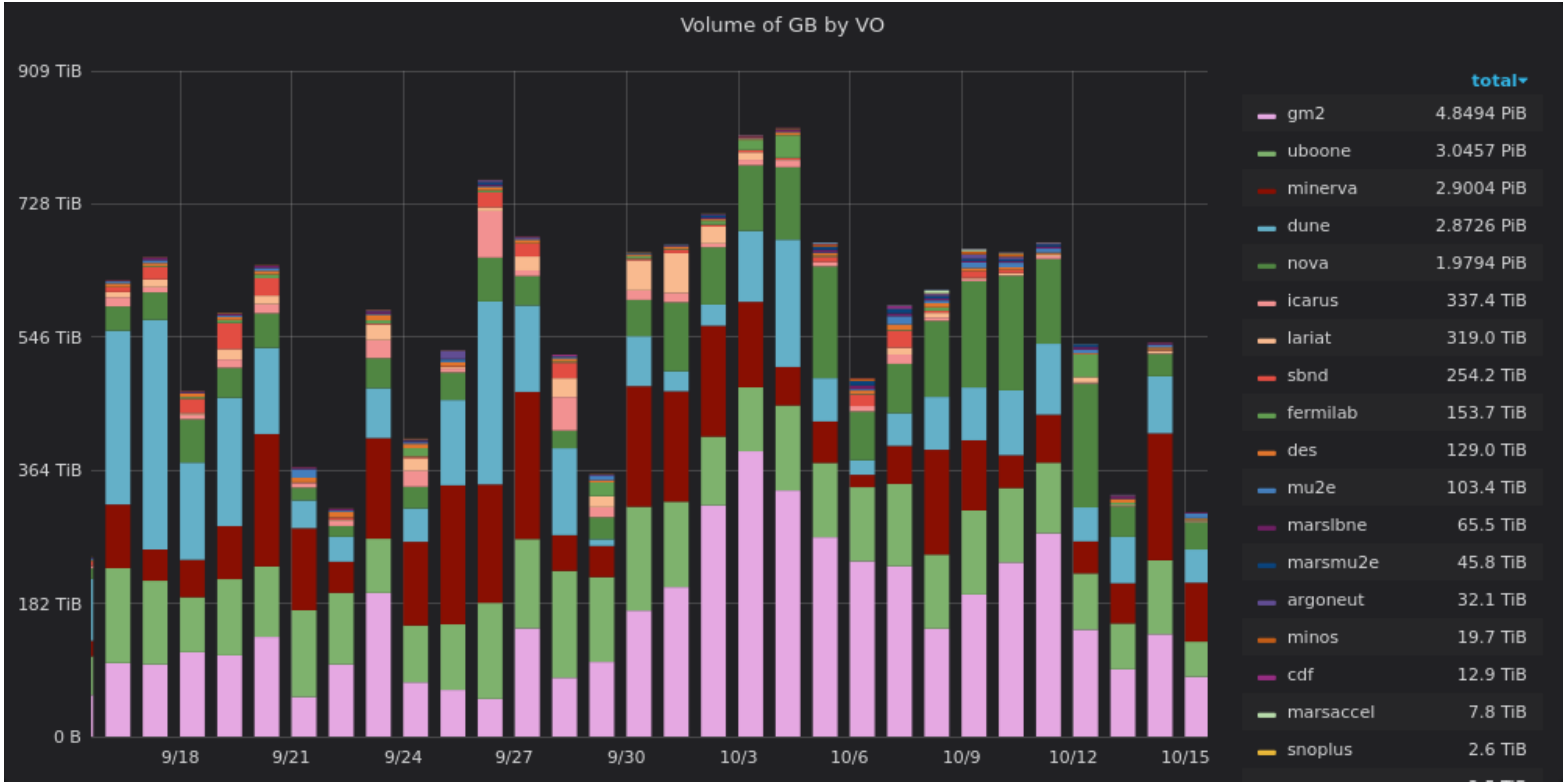
Backup

Tape storage use (as of 10/1/19)



Includes data marked as deleted and some copies

Public dCache transfers (past 30 days)

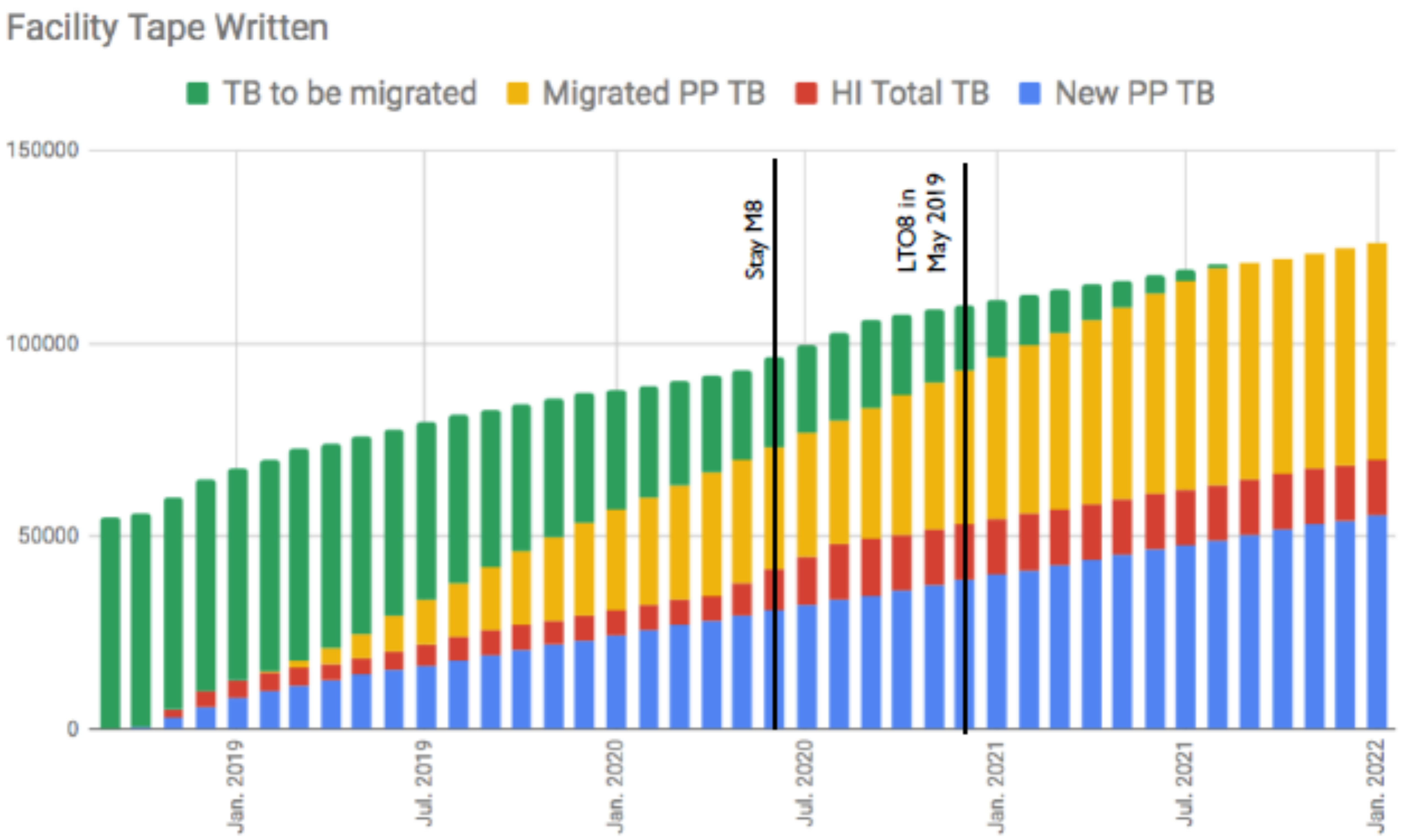


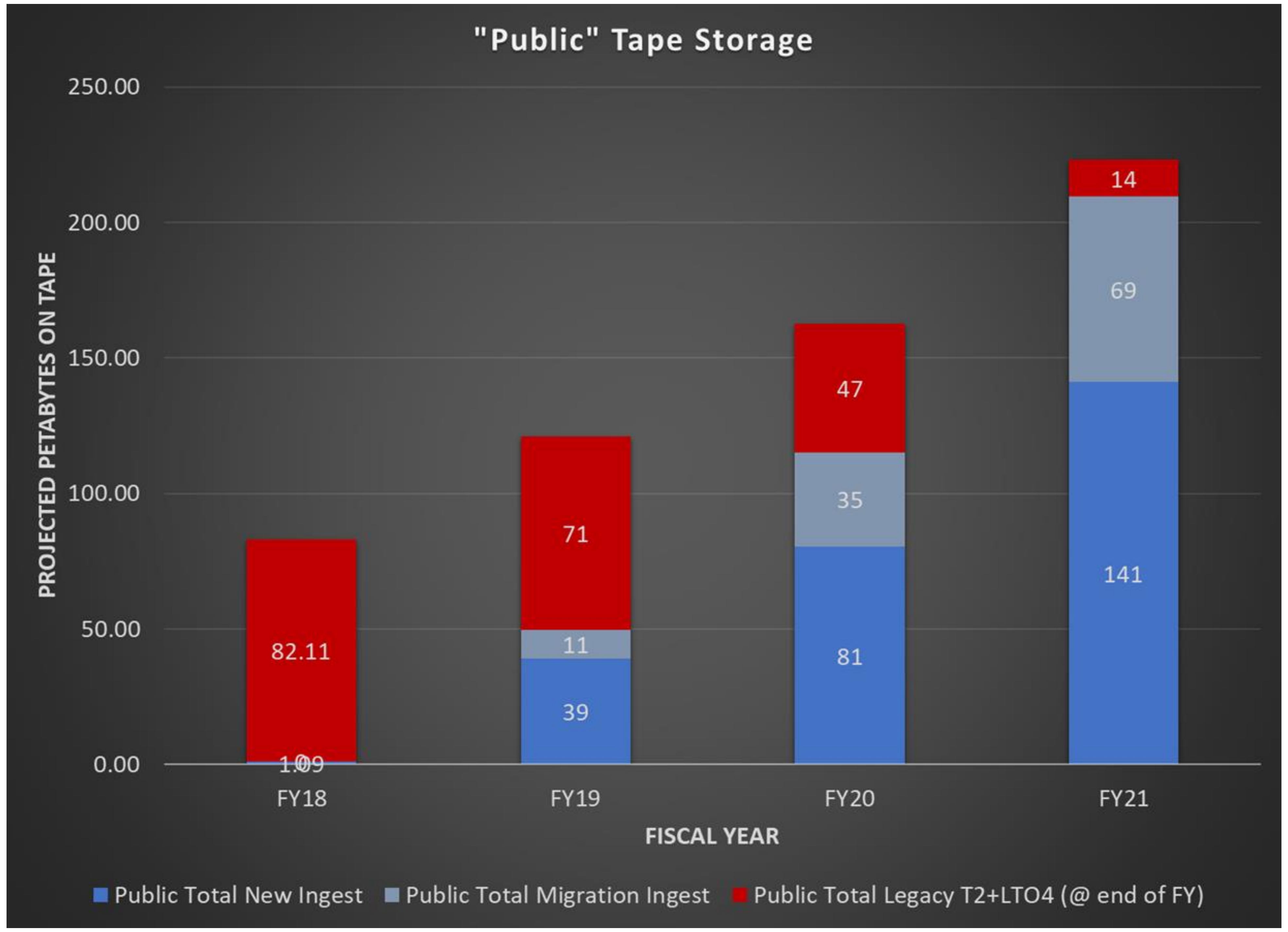
	total▾
gm2	4.8494 PiB
uboone	3.0457 PiB
minerva	2.9004 PiB
dune	2.8726 PiB
nova	1.9794 PiB
icarus	337.4 TiB
lariat	319.0 TiB
sbnd	254.2 TiB
fermilab	153.7 TiB
des	129.0 TiB
mu2e	103.4 TiB
marslbne	65.5 TiB
marsmu2e	45.8 TiB
argoneut	32.1 TiB
minos	19.7 TiB
cdf	12.9 TiB
marsaccel	7.8 TiB
snoplus	2.6 TiB

Tape storage projections

CMS (125PB by 2022)

Public (225PB by 2022)


Facility Tape Written
🟩 TB to be migrated
🟨 Migrated PP TB
🟥 HI Total TB
🟦 New PP TB
Stay M8
LTO8 in May 2019
Jan. 2019
Jul. 2019
Jan. 2020
Jul. 2020
Jan. 2021
Jul. 2021
Jan. 2022


"Public" Tape Storage
PROJECTED PETABYTES ON TAPE
FY18
FY19
FY20
FY21
FISCAL YEAR
Public Total New Ingest
Public Total Migration Ingest
Public Total Legacy T2+LTO4 (@ end of FY)