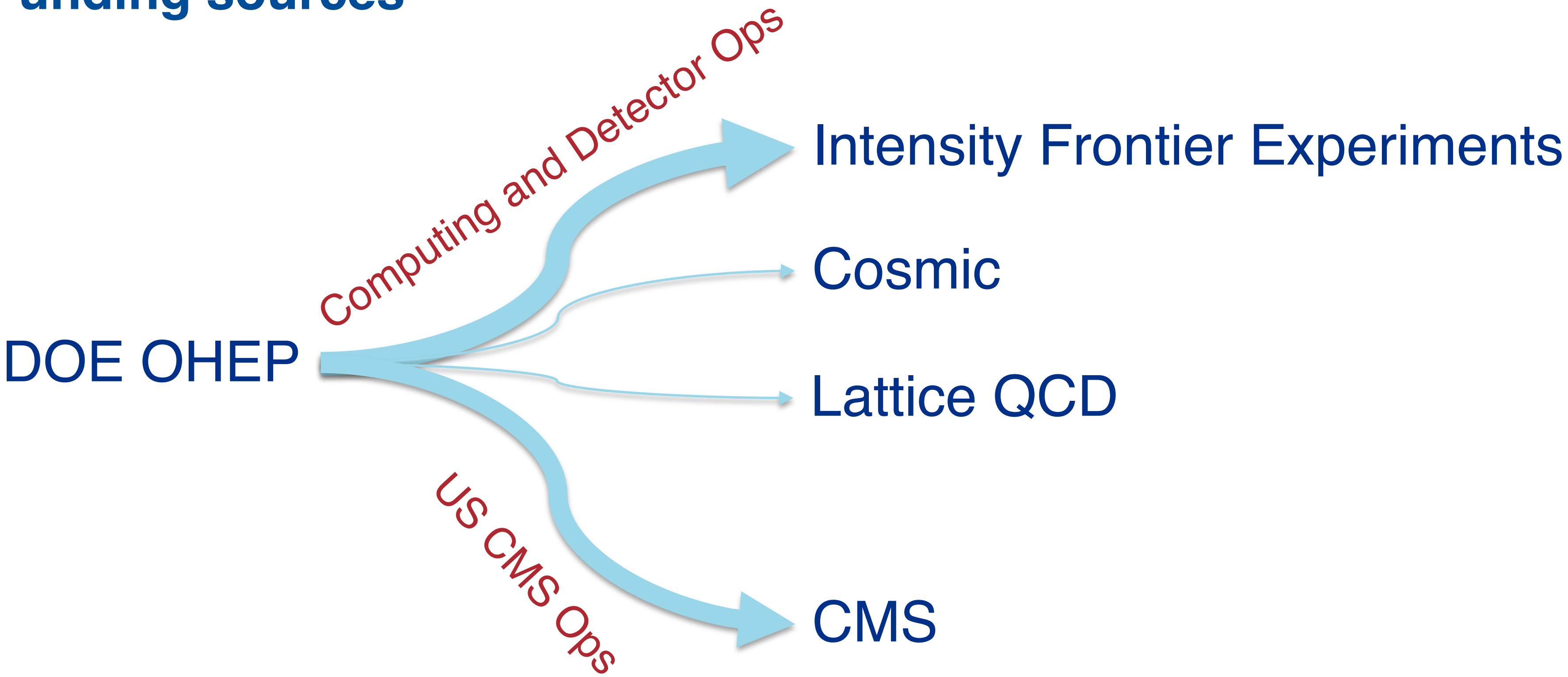
Fermilab's Process for Allocating Computing Resources

James Amundson
SBN Oversight Board Meeting
June 11, 2021

Funding sources



DOE OHEP

Computing and Detector Ops

Intensity Frontier Experiments

Cosmic

Lattice QCD

US CMS Ops

CMS

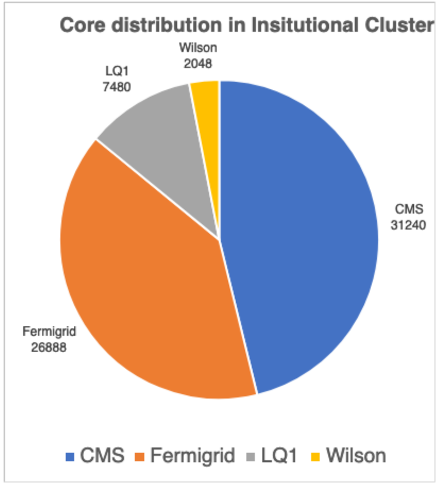
Allocation Process

- Resource allocation provided by Fermilab Computing Resource Scrutiny Group (FCRSG)
  - Committee membership comes from both within Fermilab and outside institutions
- Annual review in spring
  - March 29-30, 2021: https://indico.fnal.gov/event/47845/
  - Experiments present computing models
    - New this year
    - Large experiments with future runs
  - Experiments present resource requests
  - Greatest scrutiny given to incremental costs
  - Scientific Computing Division presents facility status and resource history
  - Committee writes report
  - SCD sets Intensity Frontier allocations for the year
    - Other frontiers do not require subdivision

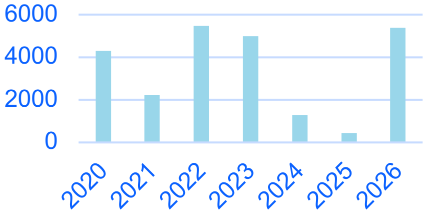
Compute Resources

Resources – Institutional Cluster

- 210M core hours available in Fermigrid, (another 28M for Rubin) 18M in Wilson
- Recently added older nodes to FermiGrid. They are DNR.
- Wilson has high speed interconnects for parallel processing and is a steppingstone to large HPC resources
- 4 x 2 NVidia Tesla V100 GPUs
- 27 x 4 NVidia Tesla K40m GPUs
- One power9 + 4 volta GPUs (Oakridge Summit)
- One KNL (NERSC Cori/ALCF Theta)
- To maintain 24000 cores, we'd need about 300k per year to buy 5000 cores

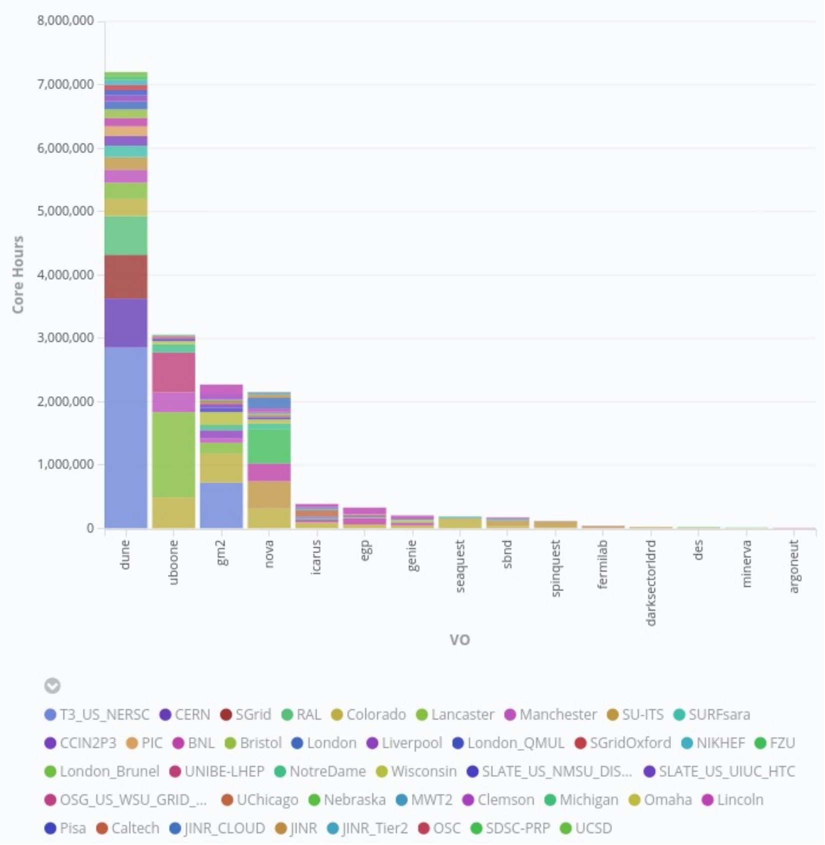


FermiGrid Cores Falling Off Warranty



Resources – Outside our walls

- HPC sites (allocations)
- OSG (opportunistic)
- GCE, AWS (paid)
- If experiments have special agreements with collaborating sites, we can enable access to their individual allocations
- Containers should limit issues at remote sites
- Not everyone submitting everywhere

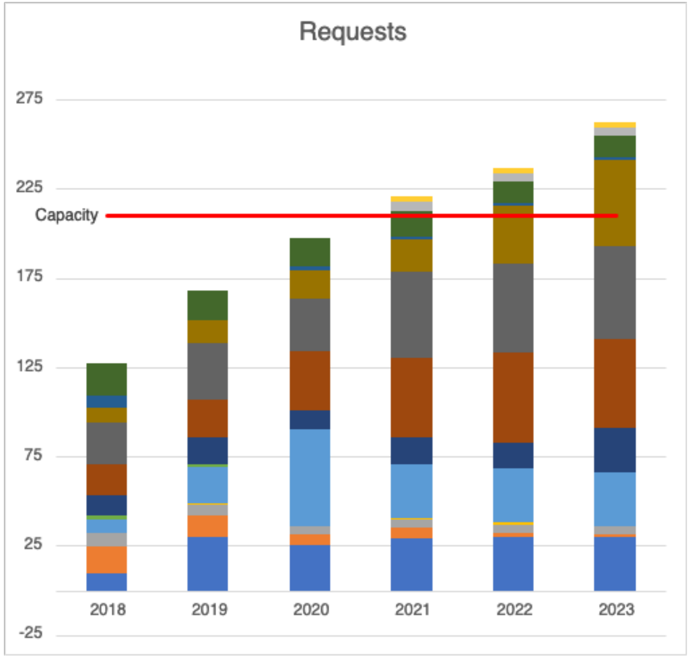


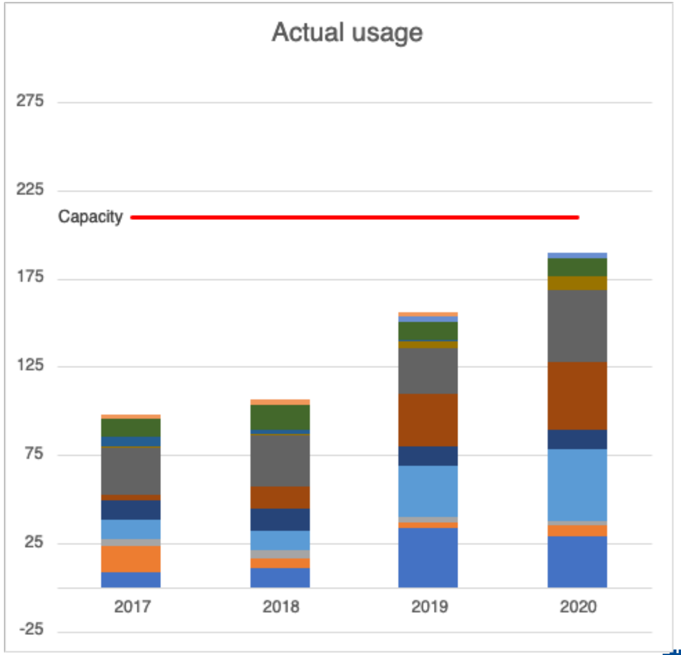
Resources – Outside our walls

- HPC Center allocations
  - NERSC CMS: 59.5M of 105M (57%) used since 20/1/2021
  - NERSC FIFE: 8.8M of 75M (12%) used since 20/1/2021
  - TACC Frontera CMS: 273/500K (54%) used
  - TACC Stampede2: 65/100K(65%) used
  - SDSC Expanse: 4.2M / 4.4 M (95%) used
  - PSC Bridges: 1.1E7 hours before decommission
  - PSC Bridges2: 1.3M of 5.6M used (25%)
  - ANL Theta being tested by CMS and mu2e now
- Last year we ran 236.5M hours of compute at NERSC via HEPCloud

Summary of requests from experiments

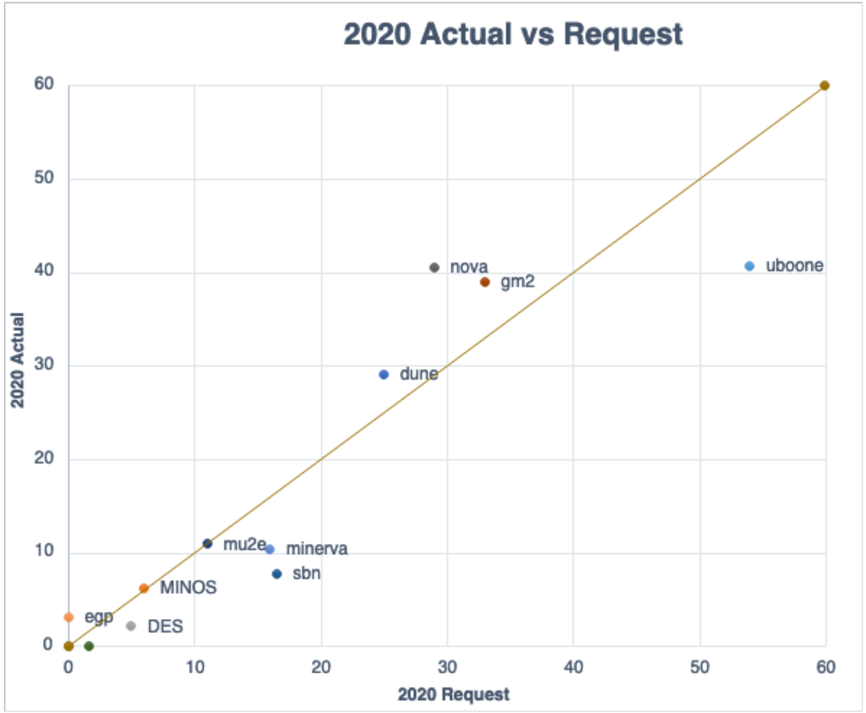
- Requests continue to climb. There may be contention for onsite resources this year. Experiments should be encouraged to submit everywhere.





Summary of requests from experiments

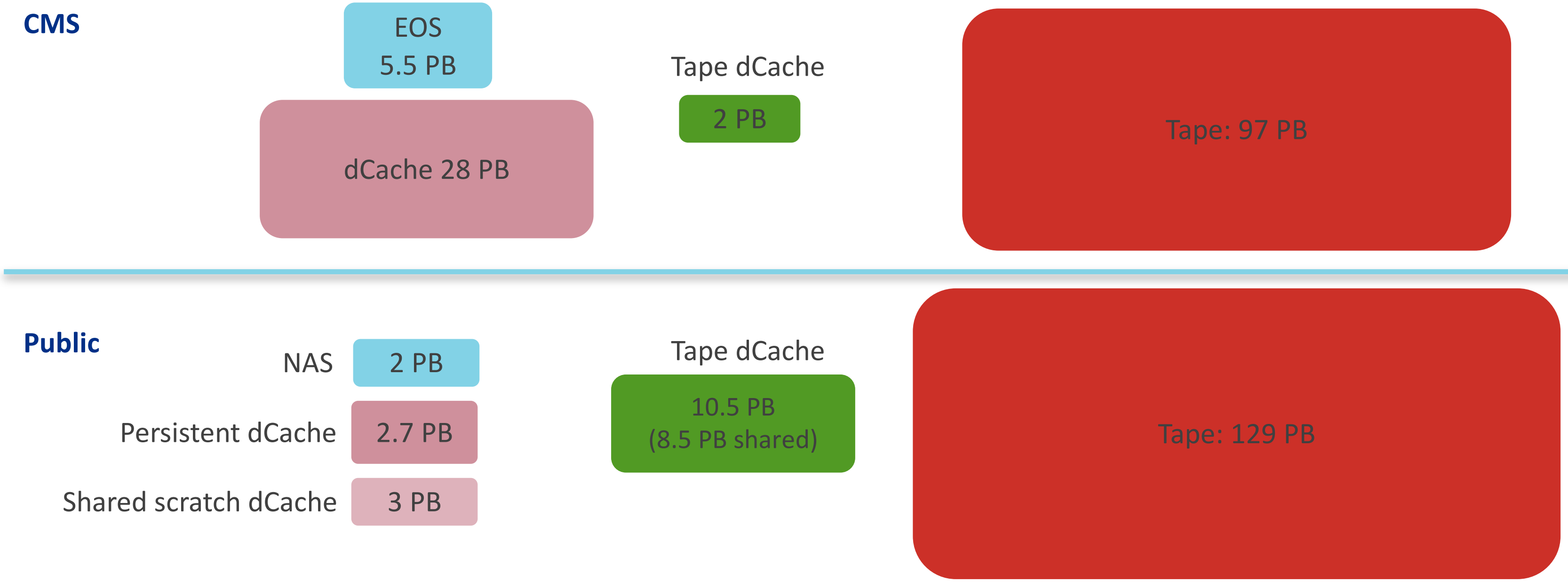
- Comparing 2020 request made in 2019 to actual 2020 usage shows reasonable predictions
- Some experiments were above request and some below, so it averages out overall.



Storage Resources

Resources - disk

- 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.



CMS

EOS
5.5 PB

dCache 28 PB

Tape dCache

2 PB

Tape: 97 PB

Public

NAS 2 PB

Persistent dCache 2.7 PB

Shared scratch dCache 3 PB

Tape dCache

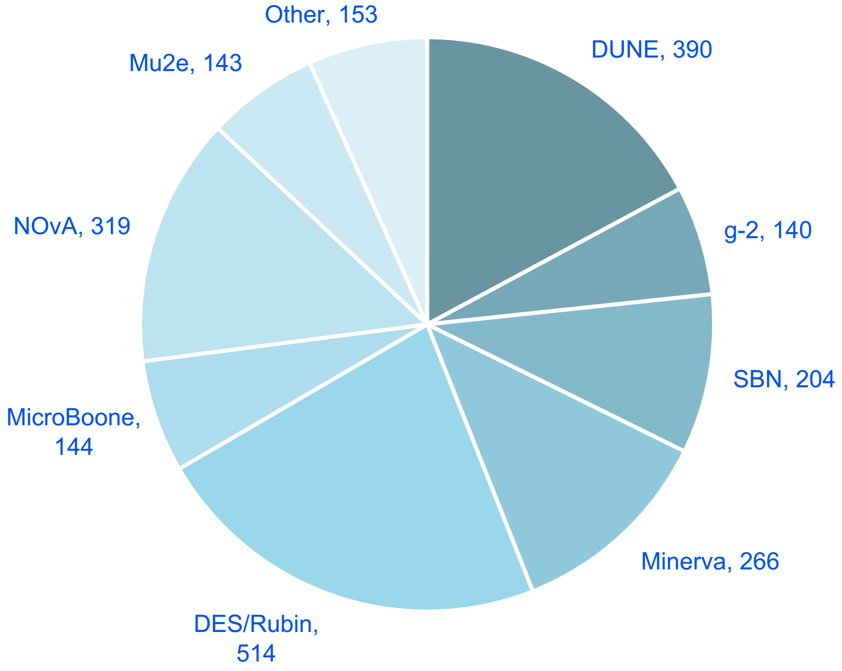
10.5 PB
(8.5 PB shared)

Tape: 129 PB

Requests from experiments – persistent dCache

Experiment	2021 request	2022 request	2023 request
DES	538	538	538
DUNE	600	800	800
MicroBooNE	151	151	151
Mu2e	100	100	150
g-2	150	200	400
NOvA	320	345	375
SBN	250	250	300
MINERVA	250	250	250
Other	174	175	175
Total	2533	2809	3139

Current persistent dCache usage (TB)



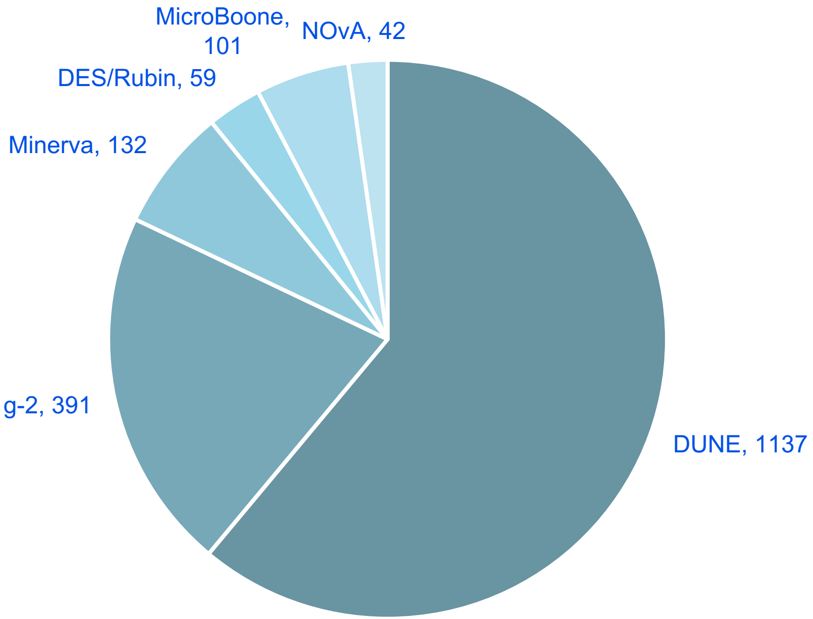
Note – "other" only includes FCRSG requests. There are other users beyond these.

Requests from experiments – dedicated dCache

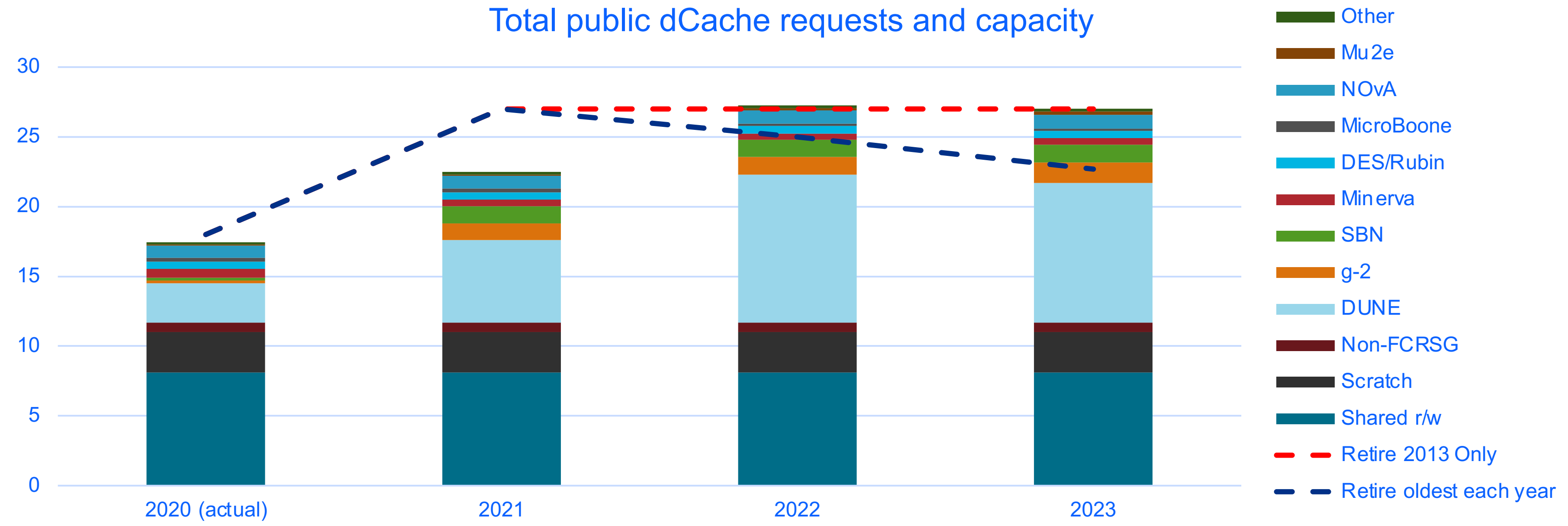
Experiment	2021 request	2022 request	2023 request
DUNE	5300	9800	9200
SBN	1000	1000	1000
g-2	54+1000	60+1000	60+1000
NOvA	610	610	610
MINERVA	200	200	200
MicroBooNE	100	0	0
Mu2e	0	50	100
Other	24	24	24
Total	8288	12744	12194

- Current total 3600 TB
- DUNE request is for ProtoDUNE II
- SBN is requesting a large increase for data taking

Current dedicated dCache allocation (TB)



Total dCache requests



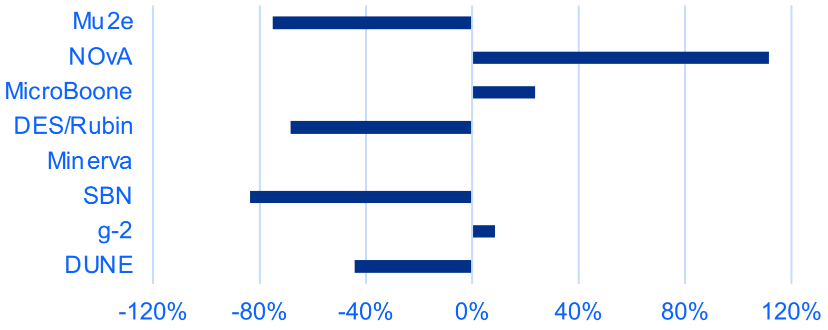
- Assumes no increase in scratch or shared space
- Dashed lines show capacity (usable, no replication).
  - Assumes no additional purchases before 2023
- Red line – retire/repurpose 2013 disks only
- Blue line – retire/repurpose oldest each year

Requests from experiments - tape

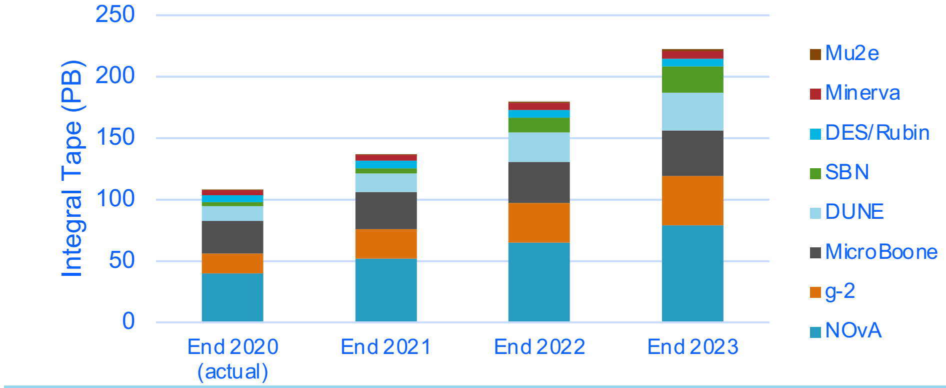
- Last year’s requests were not a very good guide to actual usage
  - SBN used much less and have significantly reduced projected future usage

- Most experiments are not considering significant deletion of data on tape
  - Exceptions are SBN and Mu2e

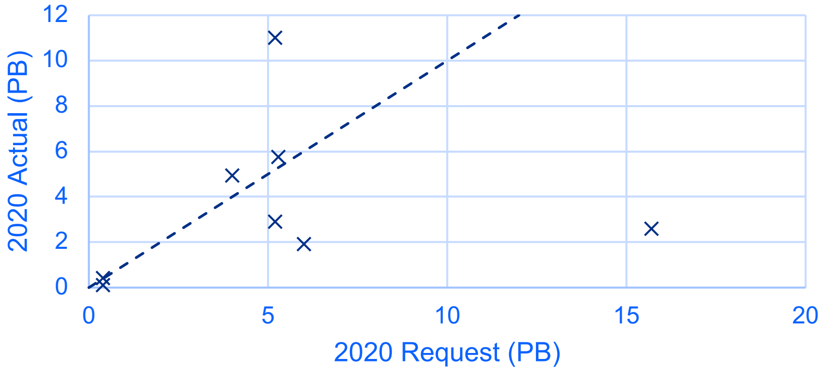
2020 Tape Actual Use Difference from Request



Integral Tape Volume



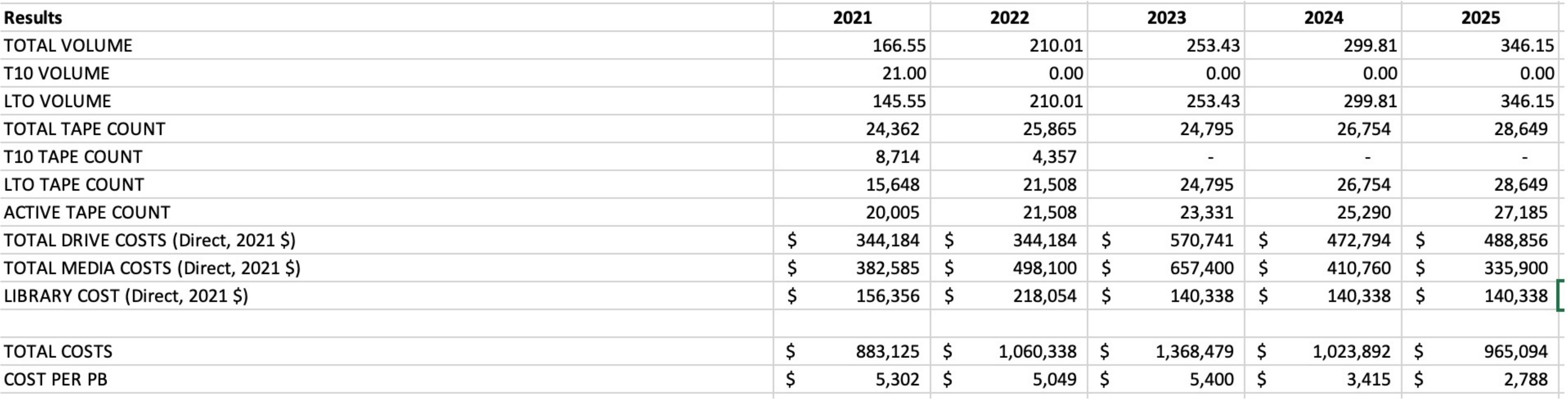
2020 Tape Usage Actual vs Request



Tape Storage Costs

- Tape storage has many components
  - Tape media
  - Libraries
  - Drives
    - Data rates are limited by the number of drives
      - Data rates are becoming more of a problem than data volume
    - Drives are expensive
  - Maintenance
  - Effort
- Tape costs are ongoing
  - Media continually needs to be migrated to the current tape storage technology

Tape Cost Model



Results	2021	2022	2023	2024	2025
TOTAL VOLUME	166.55	210.01	253.43	299.81	346.15
T10 VOLUME	21.00	0.00	0.00	0.00	0.00
LTO VOLUME	145.55	210.01	253.43	299.81	346.15
TOTAL TAPE COUNT	24,362	25,865	24,795	26,754	28,649
T10 TAPE COUNT	8,714	4,357	-	-	-
LTO TAPE COUNT	15,648	21,508	24,795	26,754	28,649
ACTIVE TAPE COUNT	20,005	21,508	23,331	25,290	27,185
TOTAL DRIVE COSTS (Direct, 2021 $)	$ 344,184	$ 344,184	$ 570,741	$ 472,794	$ 488,856
TOTAL MEDIA COSTS (Direct, 2021 $)	$ 382,585	$ 498,100	$ 657,400	$ 410,760	$ 335,900
LIBRARY COST (Direct, 2021 $)	$ 156,356	$ 218,054	$ 140,338	$ 140,338	$ 140,338
TOTAL COSTS	$ 883,125	$ 1,060,338	$ 1,368,479	$ 1,023,892	$ 965,094
COST PER PB	$ 5,302	$ 5,049	$ 5,400	$ 3,415	$ 2,788