# Long-term Fermilab Storage Facility Planning

James Amundson

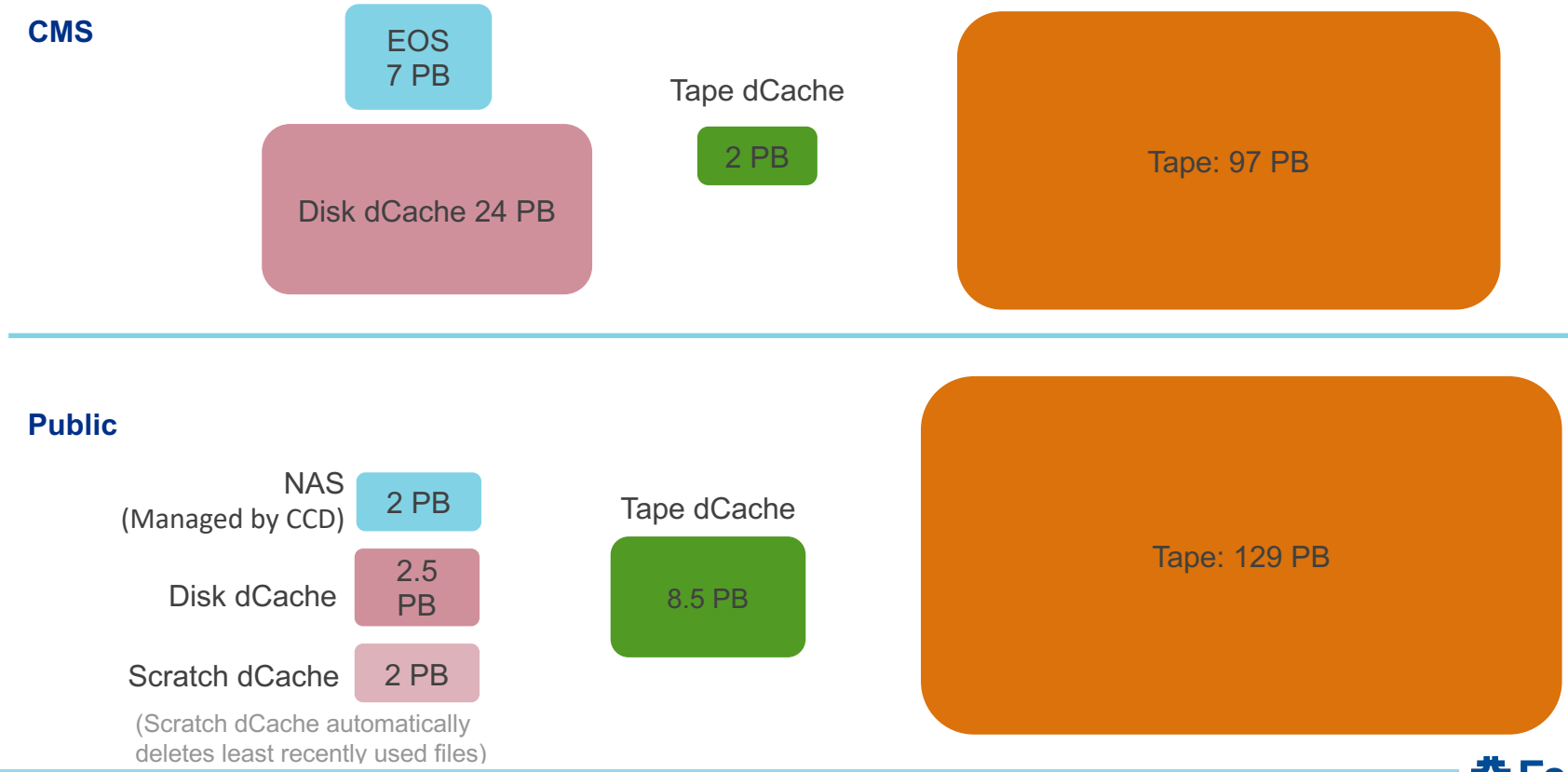
4th Meeting of the International Computing Advisory Committee

February 9, 2022

# Overview

- Current Fermilab mass storage configuration
- Projected needs for the next decade
- Current and projected compute facility
- Limitations of current configuration
- Planning to meet the needs of the next decade

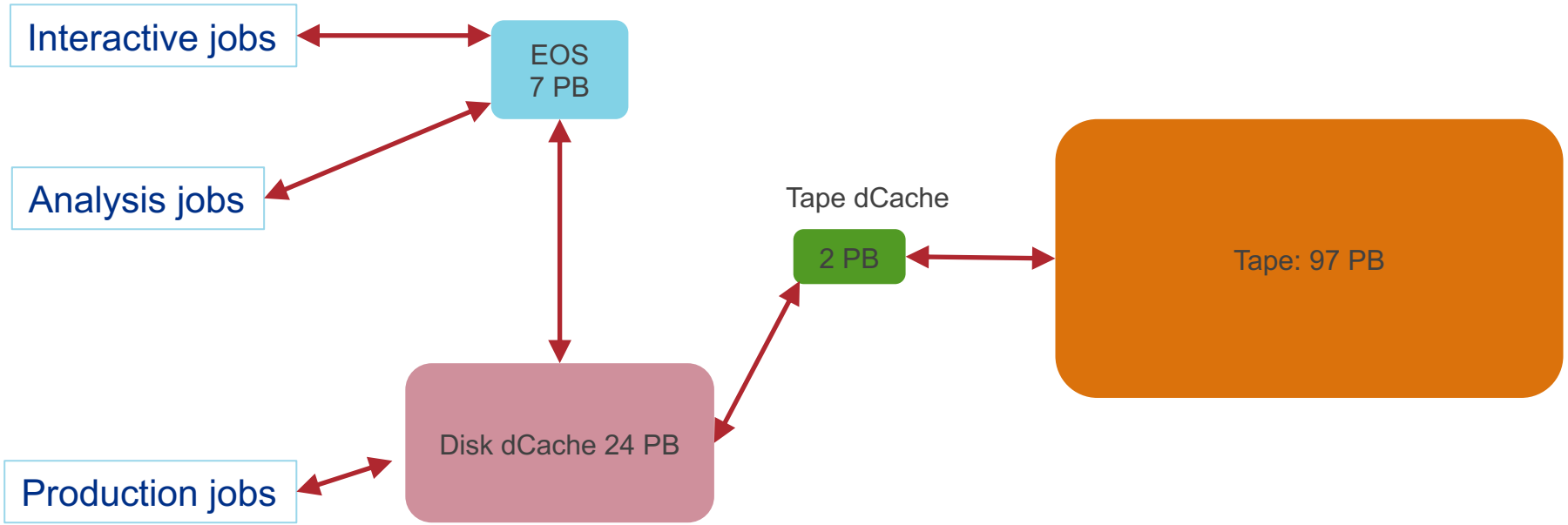
# Public and CMS overview



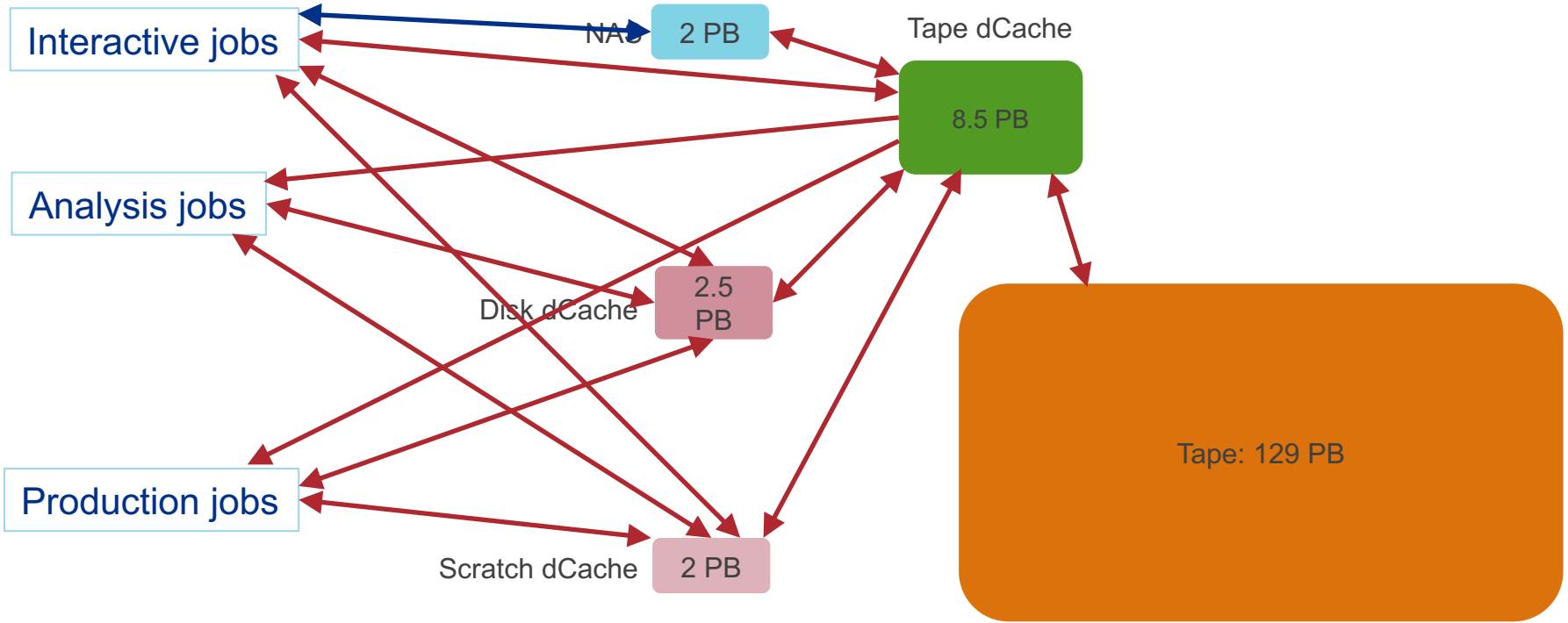



Fermilab

# What goes where: CMS

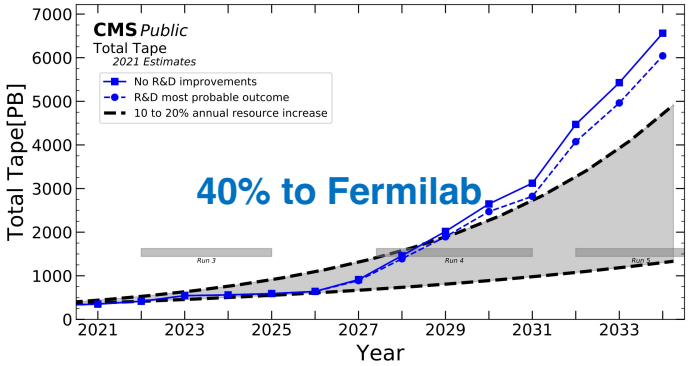


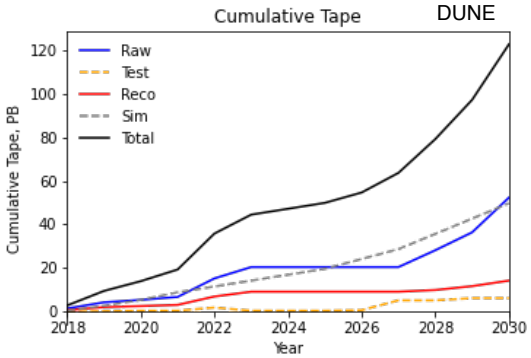
# What goes where: Public



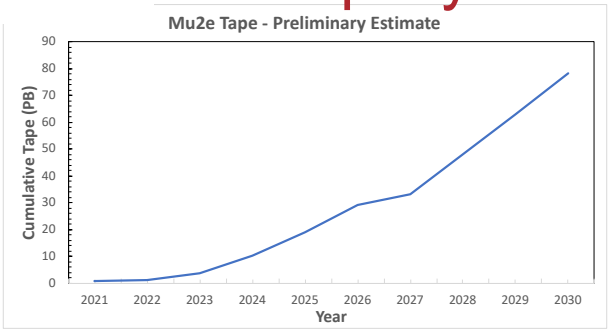
# Projected needs for the next decade

- Fermilab's data volume on tape today is 269 TB (225 TB active)
  - Two main categories
    - CMS
    - Public
- CMS during HL-LHC
  - ~2.4 EB by 2035
- DUNE
  - ~120 PB by 2030
- SBN
  - ~120 PB by 2030
- Mu2e
  - ~80 PB by 2030
- Small experiments
  - Lacking detailed plans, but small compared to above experiments
  - Adds to support load and complexity
    - complexity grows *more slowly than linearly with data volume*
    - complexity grows *faster than linearly with number of experiments*
- Legacy experiments
  - The support load of these experiments is easy to underestimate




Dates and estimates are rapidly changing!



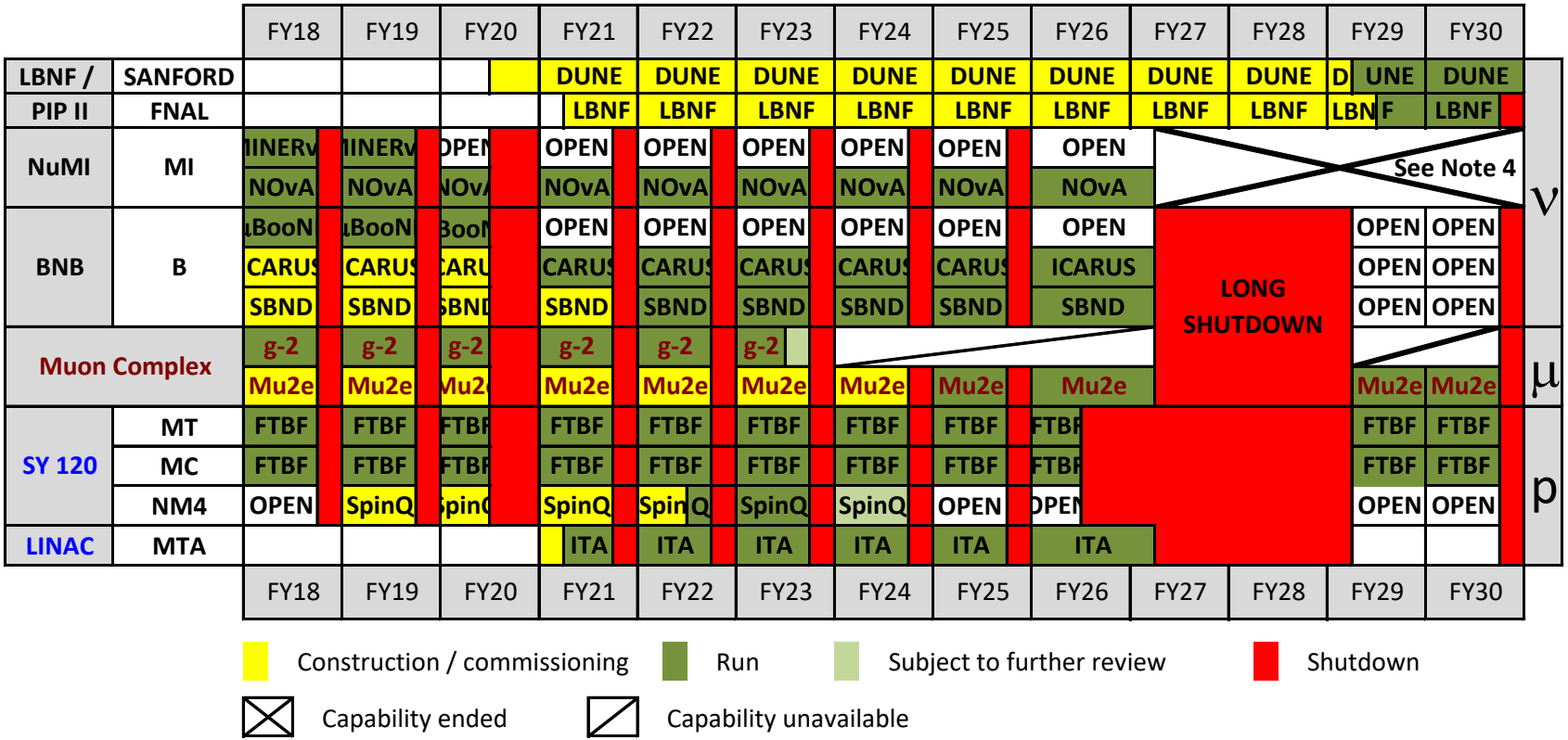



Fermilab

# Fermilab 10-year plan

Office of the CRO January 2022

**DRAFT** LONG-RANGE PLAN






Fermilab

# 10-year plan notes

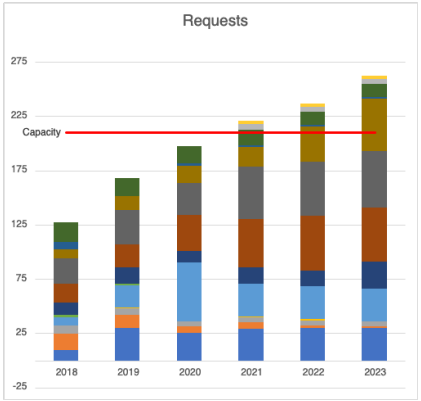
**NOTES**

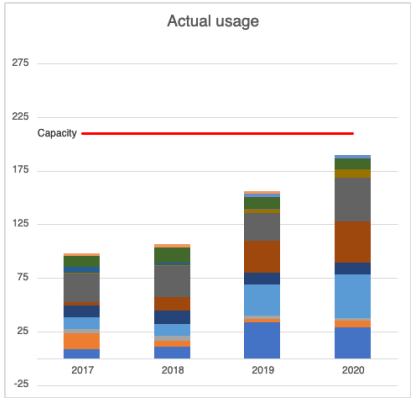
1. This draft long-range plan is updated bi-annually, typically following PAC meetings.
2. The timing and length of the Long Shutdown associated with the major construction activities at the lab will become clearer as the projects are baselined. Optimized commissioning and physics startup plans will be developed. Summer shutdowns will typically last about 4 months during the construction of LBNF/DUNE and PIP-II.
3. There will be no SY120 running from 6/2026 through the end of the long shutdown.
4. NOvA will run at least until the beginning of the Long Shutdown. A decision on whether to run after the Long Shutdown using PIP-II will be made before the Long Shutdown begins. The NOvA experiment will continue to alternate between neutrino and anti-neutrino running.
5. SpinQuest is expected to finish commissioning and start running late in FY22. Running beyond FY23 is subject to further review.
6. The MTA beamline and the Irradiation Test Area (ITA) began operations in FY21. It will not return in FY29.
7. The optimal timing of the Muon Complex switch from Muon g-2 to Mu2e commisioning and data running will continue to be monitored as Mu2e construction and g-2 data collection progress.

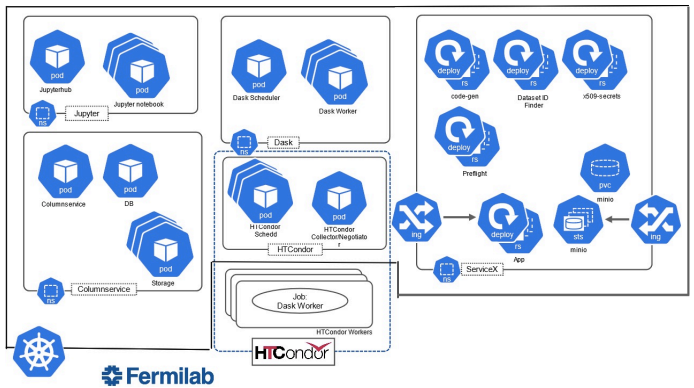
# Current and projected compute facility

On prem

- Fermigrid
  - Gradual evolution – assuming flat funding
  - More GPUs – give and take with experiments' needs
- Analysis facility
  - Take advantage of industry big data tools
  - Fast data access is a key ingredient
  - Goal is to minimize time to scientific insight



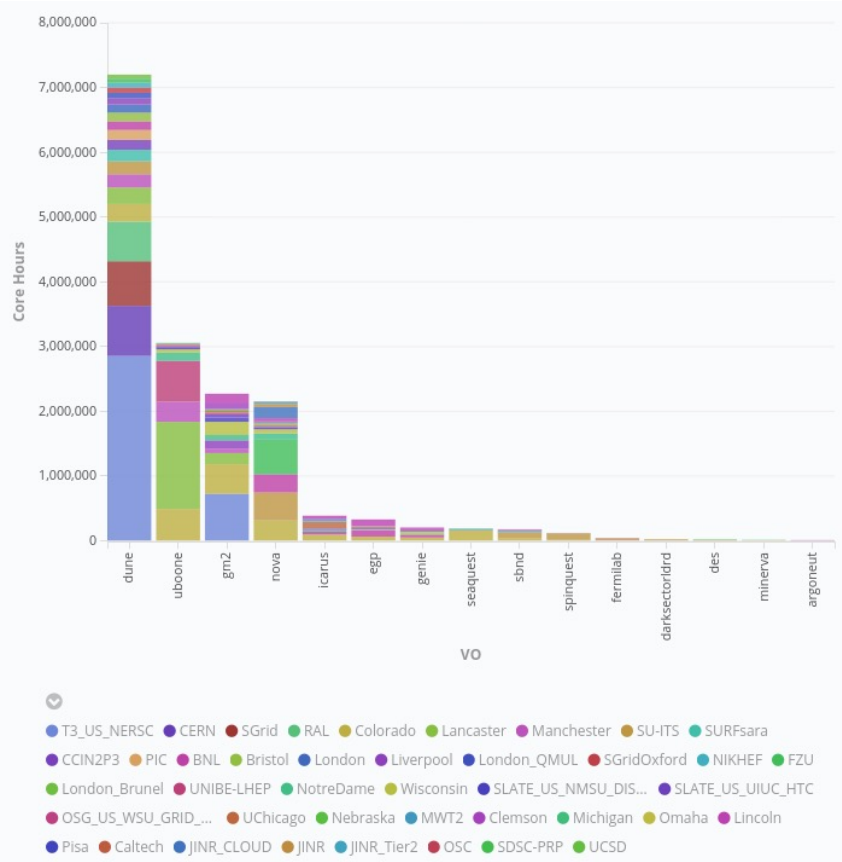



Fermilab

# Current and projected compute facility

Off prem/in field

- Open Science Grid
  - Also expect gradual evolution






Fermilab

# Current and projected compute facility



Off prem

- HPC
  - Expect expanded use
  - No dedicated storage component – Fermilab storage will be required to feed HPC site
  - Working with OHEP on allocation strategies
  - Networking
- Commercial cloud
  - Primarily peak use
  - Exotic hardware access? R&D
  - Storage not expected to be cost effective

- HPC Issues
  - HEPCloud working well with HPC
  - Experiments doing well with pre-Exascale machine
    - Generally able to utilize all allocated time for current machines
  - Ongoing experiment development efforts to utilize GPUs (required for Exascale)
    - Remains to be see how much of available GPU capacity can be utilized
    - Progress being made
  - Exascale machines are late

# Funding issues

- All ”public” computing hardware covered by a single large funding source (Detector Operations B&R)
  - Single budget for salary + hardware
  - Generally flat funding
  - Fluctuations in salary budget are extremely difficult to accommodate
    - Hardware was routinely squeezed out of budget in previous years
    - Typically spend remaining budget on hardware at the end of the fiscal year
- We are operating under a continuing resolution (no approved FY22 budget)
  - FY22 funding is still significantly unclear four months into FY
- In the last year the DUNE computing project has received limited direct funding
  - Roughly $1M spread over multiple labs and universities
  - All salary
  - Hoped to be first step towards a significantly funded project
- Public spending for FY21 included significant disk purchase
  - 11.5 PB raw, 8-9 PB usable
  - Still waiting (May?)
- FY22 hardware funds will prioritize the purchase of a new public tape library
  - Not waiting until the end of fiscal year
  - Effect of supply chain issues difficult to predict

# Current issues

- Capacity
- Throughput
  - Inefficiency of treating tape as a random-access system
- Technologies
  - Tape
    - Enstore
  - Disk
    - reliance on hardware raid
      - Already moving away, but slowly
        - New purchases not Nexsan
        - Still hardware raid
        - Can be used in JBOD mode
    - EOS
    - dCache
    - Inappropriate use of NAS
      - Instituted per-experiment quotas in data areas (100 TB)
      - Moved data to DNR portion of array
      - NAS usage for builds, etc., still necessary
- Migration
  - Substantial progress in last year
    - Minimize effort

# Current issues

- Capacity – buy more of everything, build a new data center (out of scope)
- Throughput
  - Inefficiency of treating tape as a random-access system
    - Tape/disk global architecture
    - Experiment usage
      - Tape carousels, etc.
- Technologies
  - Tape
    - Enstore ➔ CTA
      - **Promoted CTA from "investigation" to "decision"**
        - Takes into account working relationship built with CERN so far
    - Tape/disk global architecture
  - Disk
    - JBOD
    - Reevaluate EOS/dCache mix
    - Investigate Ceph for object stores
    - Investigate Ceph for NAS use
    - Implemented 100 Tb/experiment quotas on NAS
      - Migrated old data to out-of-warranty disk
- Migration
  - Integral part of new tape implementation