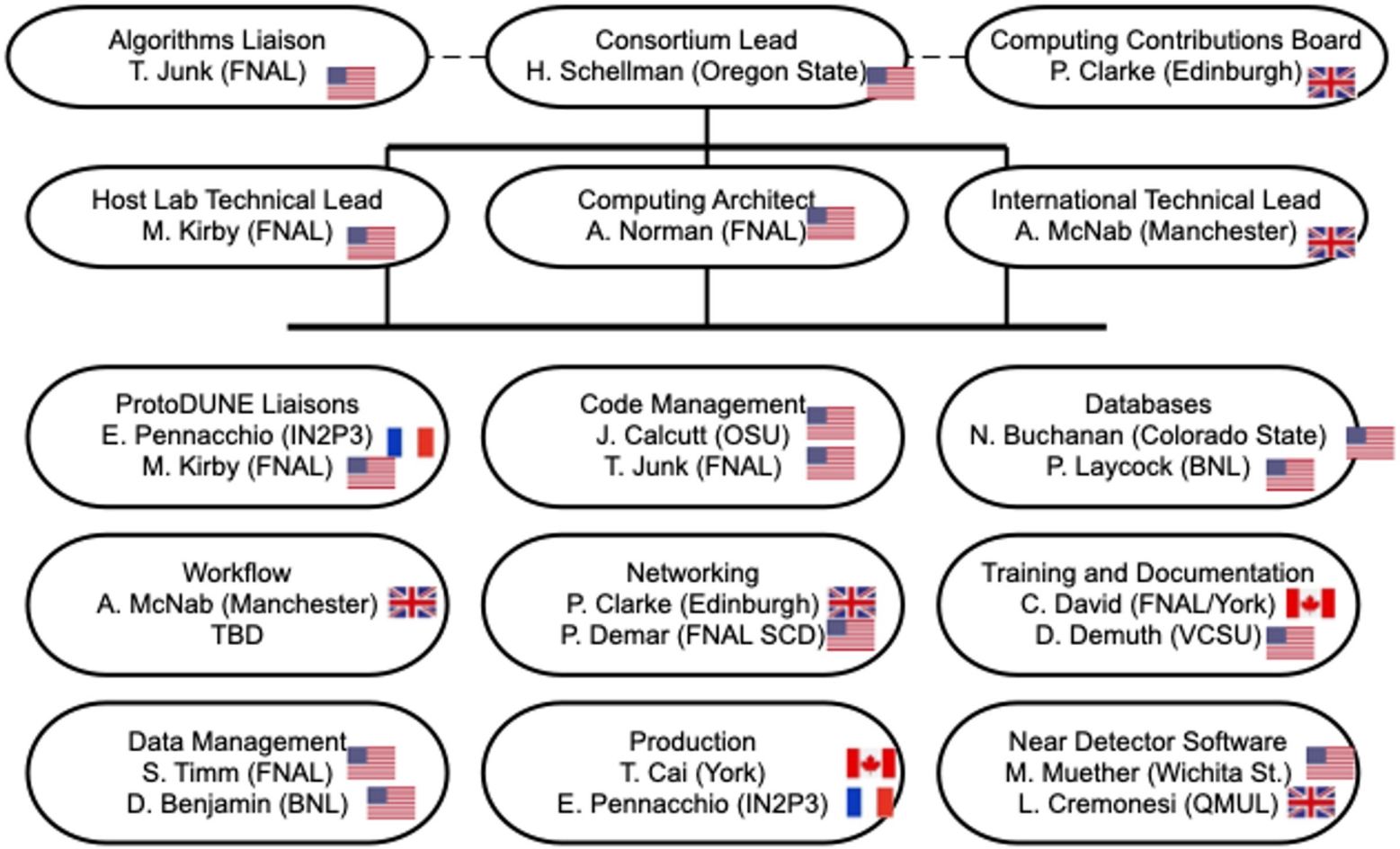
FCRSG FY23 DUNE

H. Schellman, M. Kirby, S. Timm, S. Fuess

15 Feb 2023

DUNE Organization Chart for Offline Computing



Liaison: Steve Timm

Fermilab

Notes on DUNE computing in 2022

This has been an off year for computing use as we are only running small test modules (cold boxes) at CERN

A very large fraction of our CPU resources come from offsite, mainly UK, CERN, EU
Current resource model assumes 40/10/50 CPU and Disk from FNAL/CERN/Others

Few production runs were made.

We are using NERSC HPC resources, including GPU, for near detector simulation and, in future, reconstruction. We used our normal DUNE allocation and GPU work was done using a special NERSC allocation.

Full resource model description is at https://docs.dunescience.org/cgi-bin/sso/RetrieveFile?docid=27487&filename=FCSRG-Report-2023-v02.pdf and in the CDR (arXiv 2210.15665)

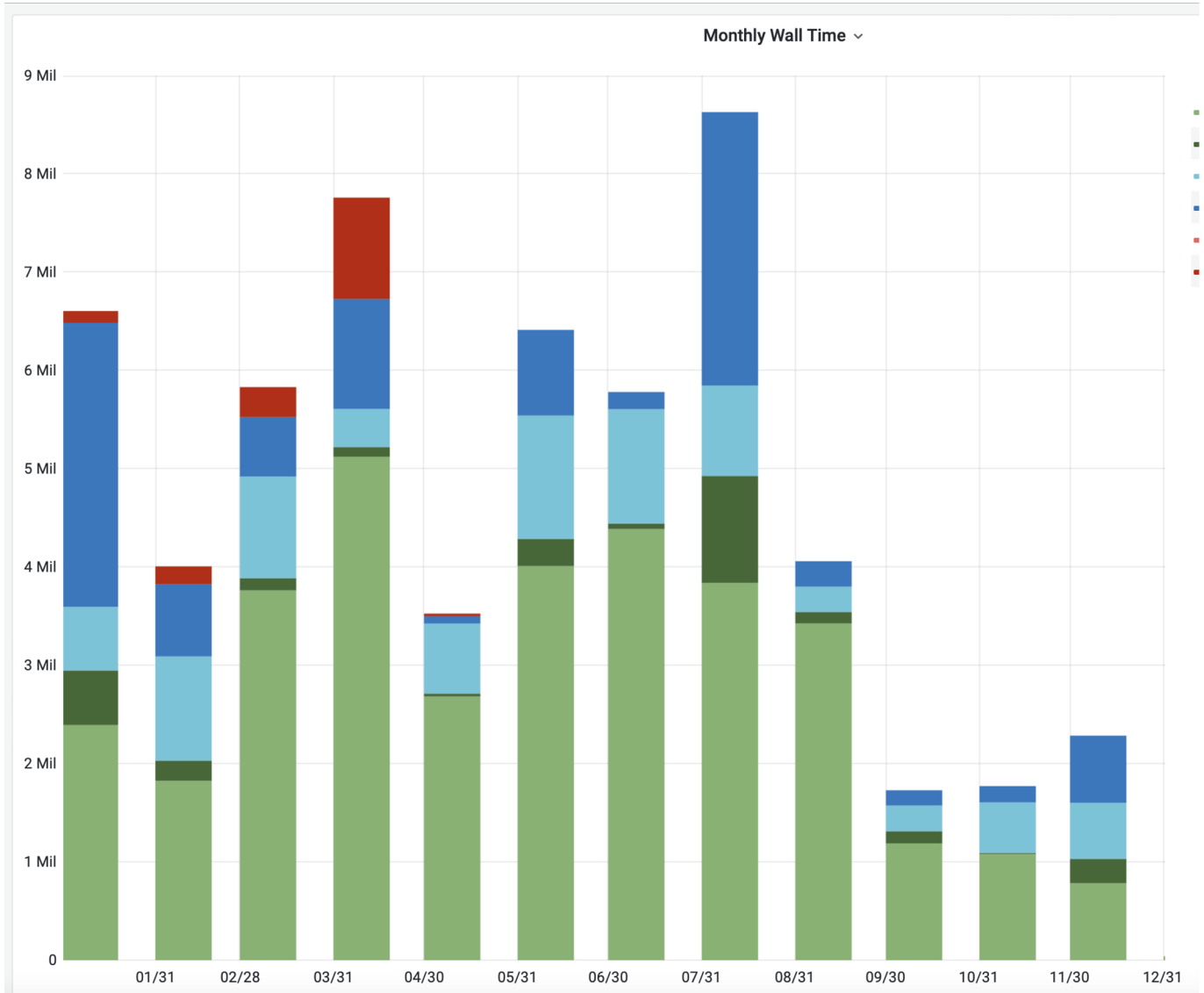
2023 and on

We expect to run a small prototype of Liquid Argon Near Detector (Argon Cube 2x2) in the NuMI hall this spring/summer and the ProtoDUNEs at CERN in the Fall. This means several PB of raw data.

Longer term strategy with NERSC

- LBNL collaborators have their own NERSC allocation for R+D.
- The main CPU and disk allocations at NERSC are currently shared with other FNAL IF experiments.
- DUNE needs more disk space at NERSC- we requested 200 TB for DUNE and all of FNAL got 100 TB. We are working with NERSC/LBNL colleagues to integrate into Rucio

Experiment CPU Usage over the past year



14Mil of “FermiGrid Analysis” is MARS/LBNF

OSG resource numbers include US and International sites

	avg	total
FermiGrid Analysis	2.71 Mil	35.3 Mil
FermiGrid Production	193 K	2.51 Mil
OSG Analysis	743 K	9.66 Mil
OSG Production	626 K	8.13 Mil
NERSC Analysis	0	0
NERSC Production	128 K	1.66 Mil

Fermilab

Comments on memory

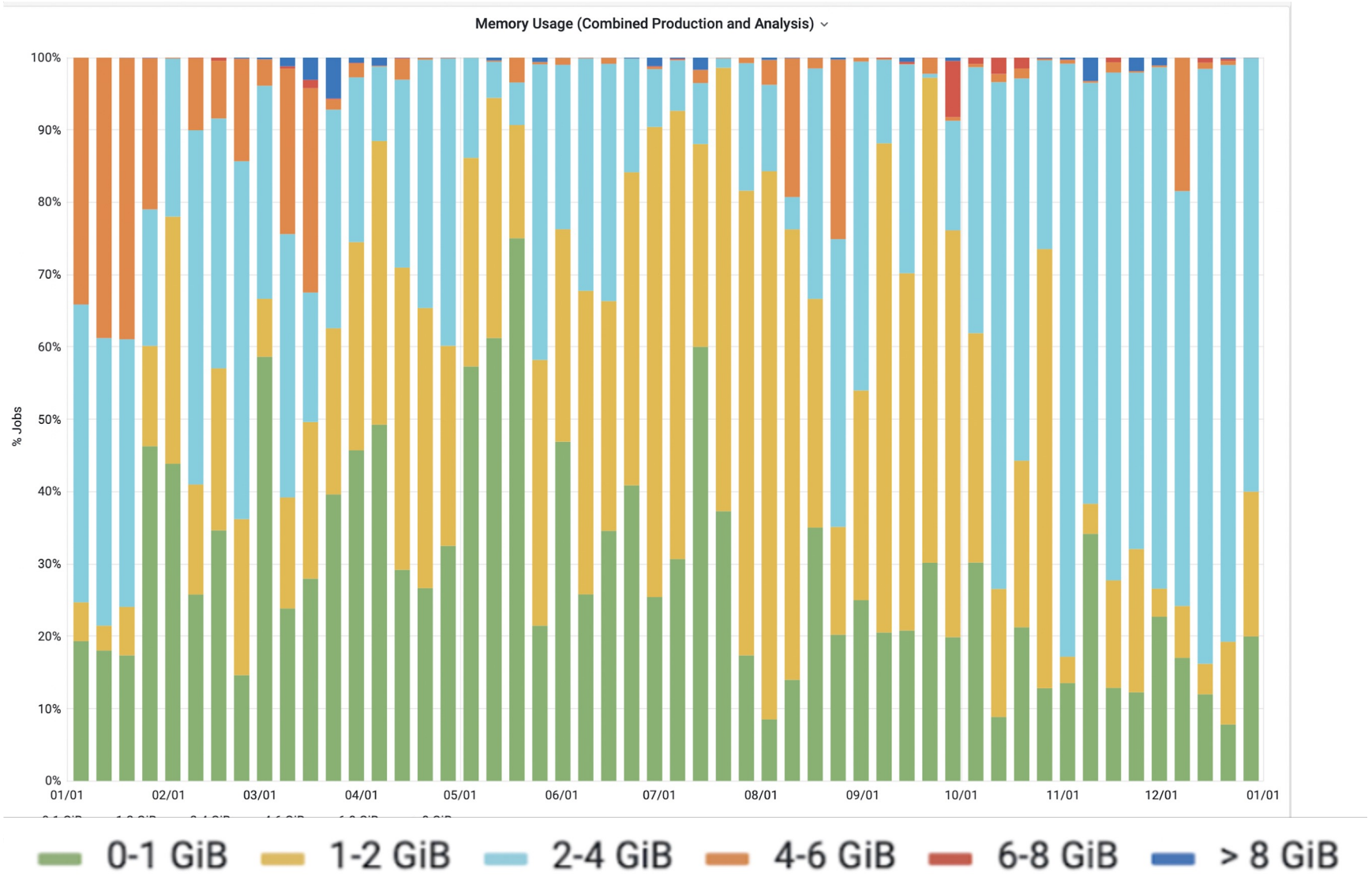
Our full sim/reconstruction code uses 4-6 GB of memory because multiple large objects are stuck in memory at once.

FNAL and Argonne people are working on this, funded by LDRD, HEP-CCE and FOA funds.

We anticipate a decrease (but probably not to 2GB) in memory footprint once this work is done.

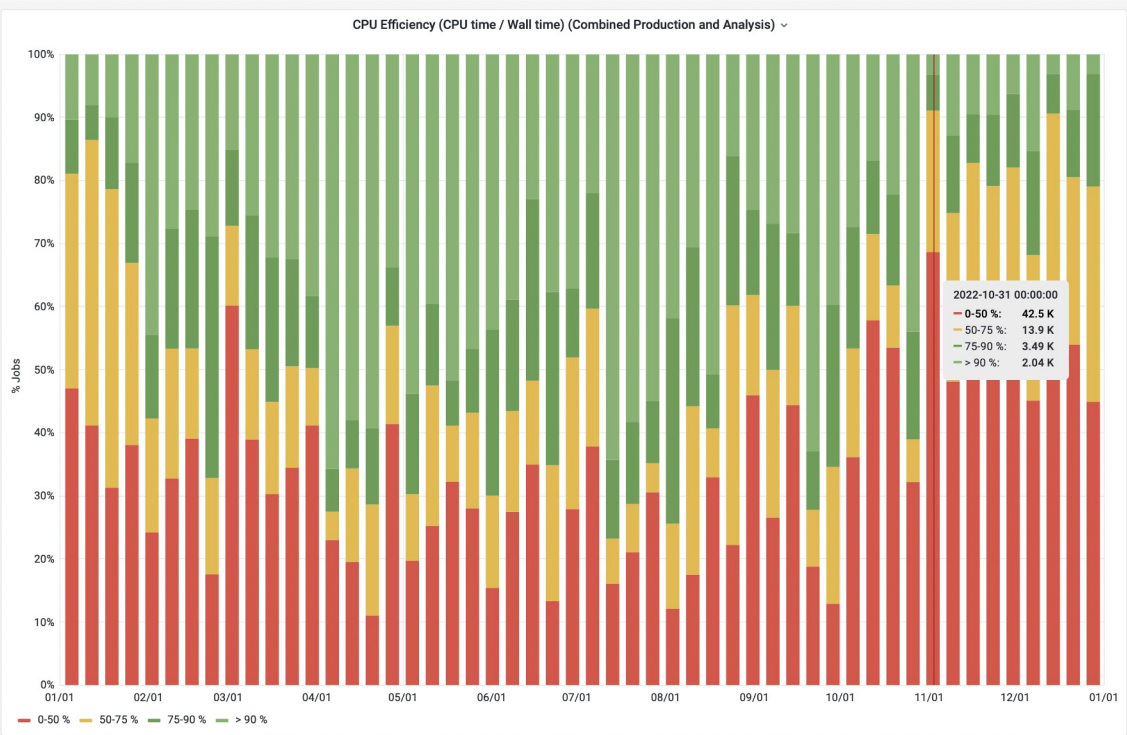
We have access to machines with large memory/core in the UK (4GB), India (6GB), NL (7.5) where we preferentially route the large memory jobs.

Memory footprint over the past year



Fermilab

CPU and memory efficiency over the past year



0-50 % 50-75 % 75-90 % > 90 %

CPU efficiency was down at the end of the year due to workflow testing which is dominated by startup effects. Usage was also down so wasted hrs did not increase. Memory efficiency may largely be users using 2000 MB for 500 MB jobs

Fermilab

What do you want to achieve in computing over the next 5 years?

Goals	Where does the experiment need to contribute	Where does SCD need to contribute
Successful ProtoDUNE Beam runs data collection 2023-2024	Port of experiment specific code.	Has already made significant contribution with new Metadata catalog, new Rucio deployment. tape and disk storage.
Successful Near Detector simulations (having heterogeneous GPU/CPU workflow)	Organizing SW in more logical way (integration of non-art-based software into a coherent framework)	Data movement, workflow deployment
Framework Extensions/ Development	Requirements for new features and use cases that don't match current art features, integration effort	effort with expertise in framework construction and development
Full transition to Token auth	coordination with CE and SE integration into workflows	potentially development in workload software and data management centralized security coordination
Multiple Offline processing of ProtoDUNE Beam runs	workflow development, production coordination, and expansion of infrastructure	support of DUNE global pool database services

What do you want to achieve in computing over the next 5 years?

Goals	Where does the experiment need to contribute	Where does SCD need to contribute
Transition to US DUNE Computing Operations funding stream	submit a successful FWP	provide assistance in estimates for cost and effort and resources
Plan for long term computing model going out to 2040	better understand offline reconstruction model, computing model, and data model	assistance with building parameterized resource model
Start commissioning the DAQ in South Dakota	DAQ group	Commission data flow, high speed storage, networks.
Analysis using GPU's	Organizing code base into unified frameworks	Ways to deal with heterogeneous workflow with GPU and CPU stages
Better collaboration and communication tools (including publications/talks database and common documentation)	Clear requirements	Staff support, ideas for tools

Campaign Schedules

	2023	2024	2025	2026	2027
Processing campaigns (start month-end month if known).	Far Detector Horizontal Drift simulation 2Q 2023 ND 2x2 simulation Feb 2023 Full ND simulation 2Q 2023 ProtoDUNE HD Keep up processing (Nov 2023)	Protodune HD keep up processing Protodune VD keep up processing Simulation campaigns	Data reprocessing pass (month TBD) Simulation campaigns	Data reprocessing pass (month TBD) Simulation campaigns	Data reprocessing pass (month TBD) Simulation campaigns
Storage + CPU estimates (call out any special resource needs if known, e.g. HPC or GPU).	FD HD: 2M hours, 600 TB ND 2x2: 20k NERSC node hours, 150 TB (needs HPC GPU) Aggregate usage of all campaigns expected to be 11PB Expect 2PB of prestages in 2Q 2023 for MC and data distribution to remote sites	In general each large simulation campaign generates 0.5 - 2 PB. Size of reco output is less than raw data.			
Conference or result targets (month if known)	2x2 Demonstrator input into ND-LAr TDR	Neutrino 2024	ProtoDUNE publications	Neutrino 2026	Commissioning

CPU @ Fermilab Prediction Going Forward and Accuracy of Your Predictions [units of Million (1 CPU, 2GB) wall hours per CY]

	2019	2020	2021	2022	2023	2024	2025	2026	2027
Requested (could have multiple values for different MWC combinations)		25 M (FNAL) 36 M (Total)	29 M (FNAL) 40 M (Total)	49 M (FNAL) 98 M (Total)	56 M (FNAL) 139 M (Total)	61 M (FNAL) 153 M (Total)	70 M (FNAL) 175 M (Total)	55 M (FNAL) 138 M (Total)	33 M (FNAL) 83 M (Total)
Actual Used		29 M (FNAL) 40 M (Total)	36 M (FNAL) 56 M (Total)	38 M (FNAL) 58 M (Total)	N/A	N/A	N/A	N/A	N/A
Efficiency	%	%	80%	64.5% 44.3% (mem)	N/A	N/A	N/A	N/A	N/A

Need to convert to HEPScore-based units going forward for better compatibility with international pledge requests.

N HEPScore x 2GB
N HEPScore x 4GB
N HEPScore x 6GB
N HEPScore x 8GB

Almost all requests in the table are 4-6 GB/core

Does not include the 15 (15*2GB) MWC/year at FNAL used for MARS

CPU – non-FNAL HTC Resources Going Forward and Accuracy of Your Predictions [units of Million (1 CPU, 2GB) wall hours per CY]

	2019	2020	2021	2022	2023	2024	2025	2026	2027
Requested (could have multiple values for different MWC combinations)					84 M	92 M	105 M	83 M	50 M
Actual Used					N/A	N/A	N/A	N/A	N/A
Efficiency	%	%	%	%	N/A	N/A	N/A	N/A	N/A

Looking for five-year projections this cycle

HPC-CPU Resources Going Forward and Accuracy of Your Predictions [units of Million (1 CPU, 2GB) wall hours per CY]

	2019	2020	2021	2022	2023	2024	2025	2026	2027
Requested (could have multiple values for different MWC combinations)					3.6 M for PD sim 1 M for 2x2Prod	5.0 M for PD+FD sim 1 M for 2x2 prod	6.5 M for PD +FD sim	???	???
Actual Used					N/A	N/A	N/A	N/A	N/A
Efficiency	%	%	%	%	N/A	N/A	N/A	N/A	N/A

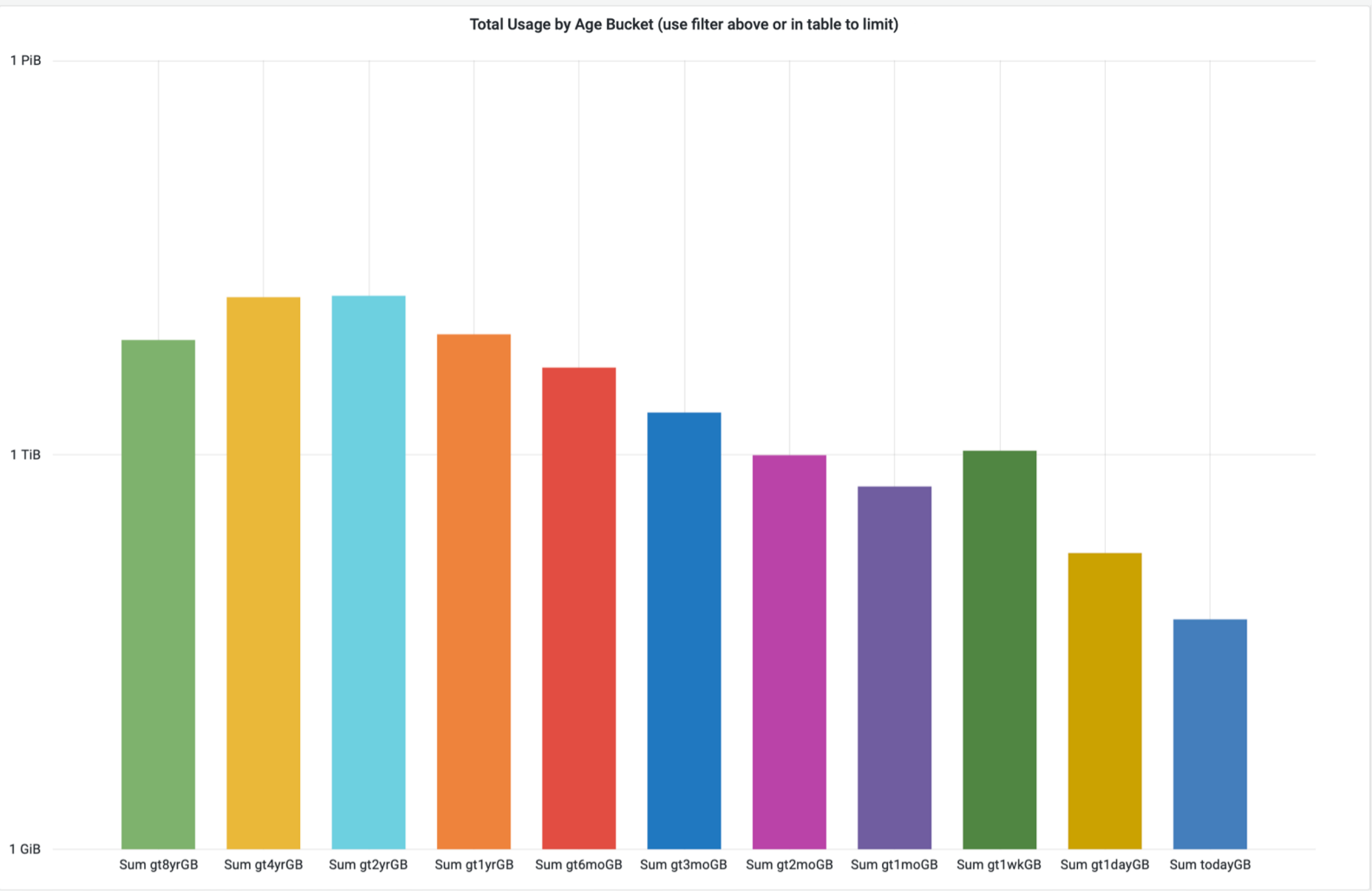
Looking for five-year projections this cycle

HPC-GPU Resources Going Forward and Accuracy of Your Predictions [units of Million (1 CPU, 2GB) wall hours per CY]

	2019	2020	2021	2022	2023	2024	2025	2026	2027
Requested (could have multiple values for different MWC combinations)					8k N-hrs PD+FD 20k N-hrs 2x2 Sim	12k N-hrs PD+FD 20k N-hrs 2x2 Sim	16k N-hrs PD+FD	???	???
Actual Used					N/A	N/A	N/A	N/A	N/A
Efficiency	%	%	%	%	N/A	N/A	N/A	N/A	N/A

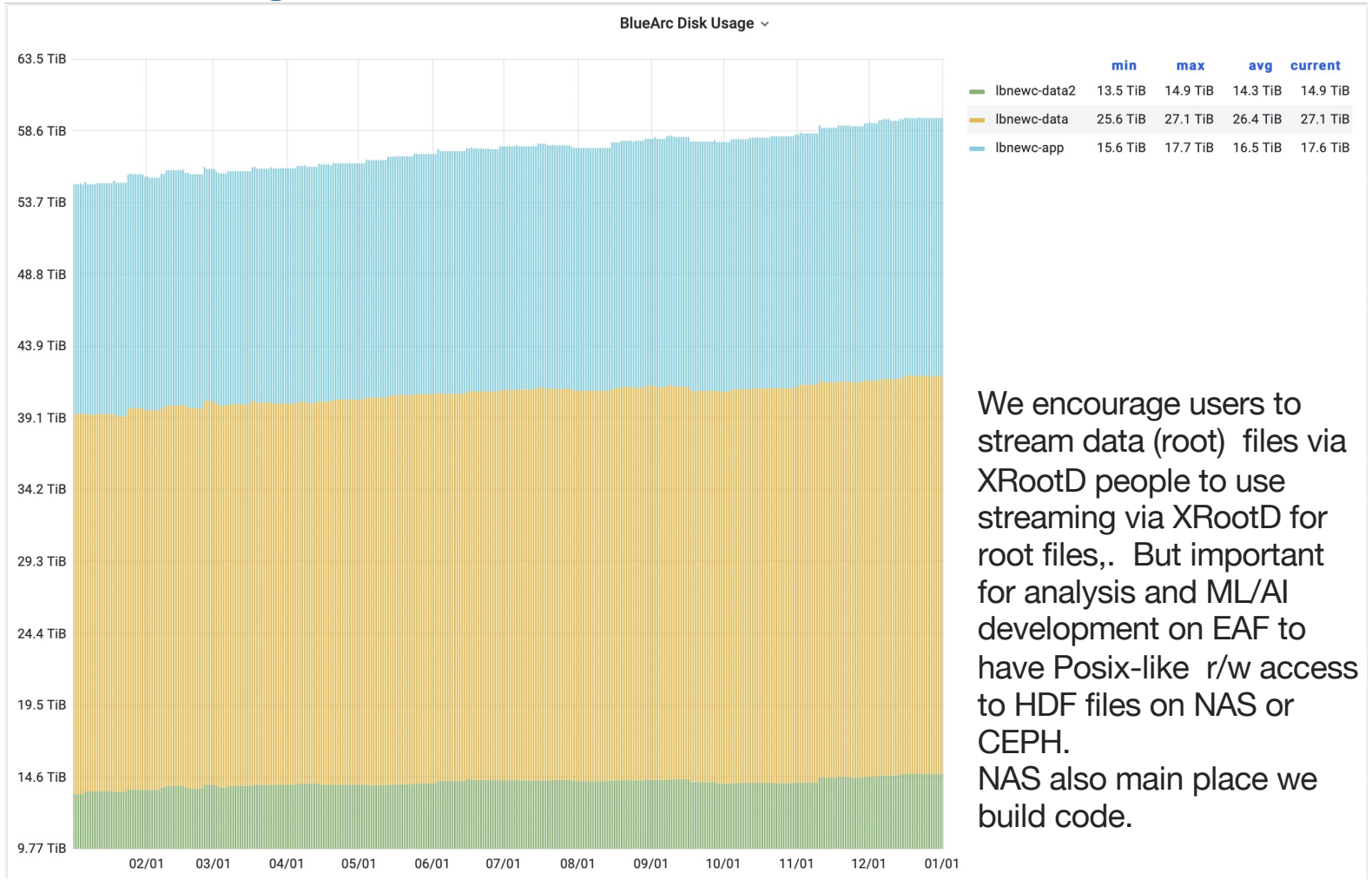
1 N-hr GPU = 4 GPU cores + 64 CPU cores

Age of files in NAS



Fermilab

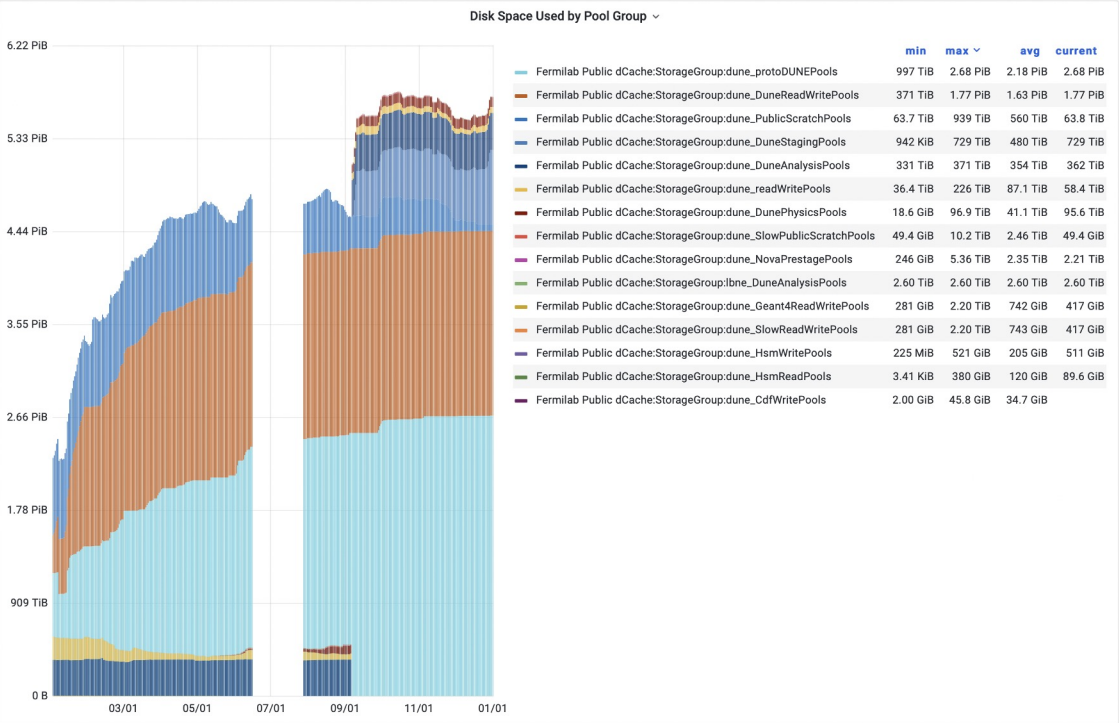
NAS Usage and Projections



We encourage users to stream data (root) files via XRootD people to use streaming via XRootD for root files,. But important for analysis and ML/AI development on EAF to have Posix-like r/w access to HDF files on NAS or CEPH.
NAS also main place we build code.

	App	Data
2022	17	62
2023	19	70
2024	21	80
2025	23	90
2026	25	100
2027	27	100

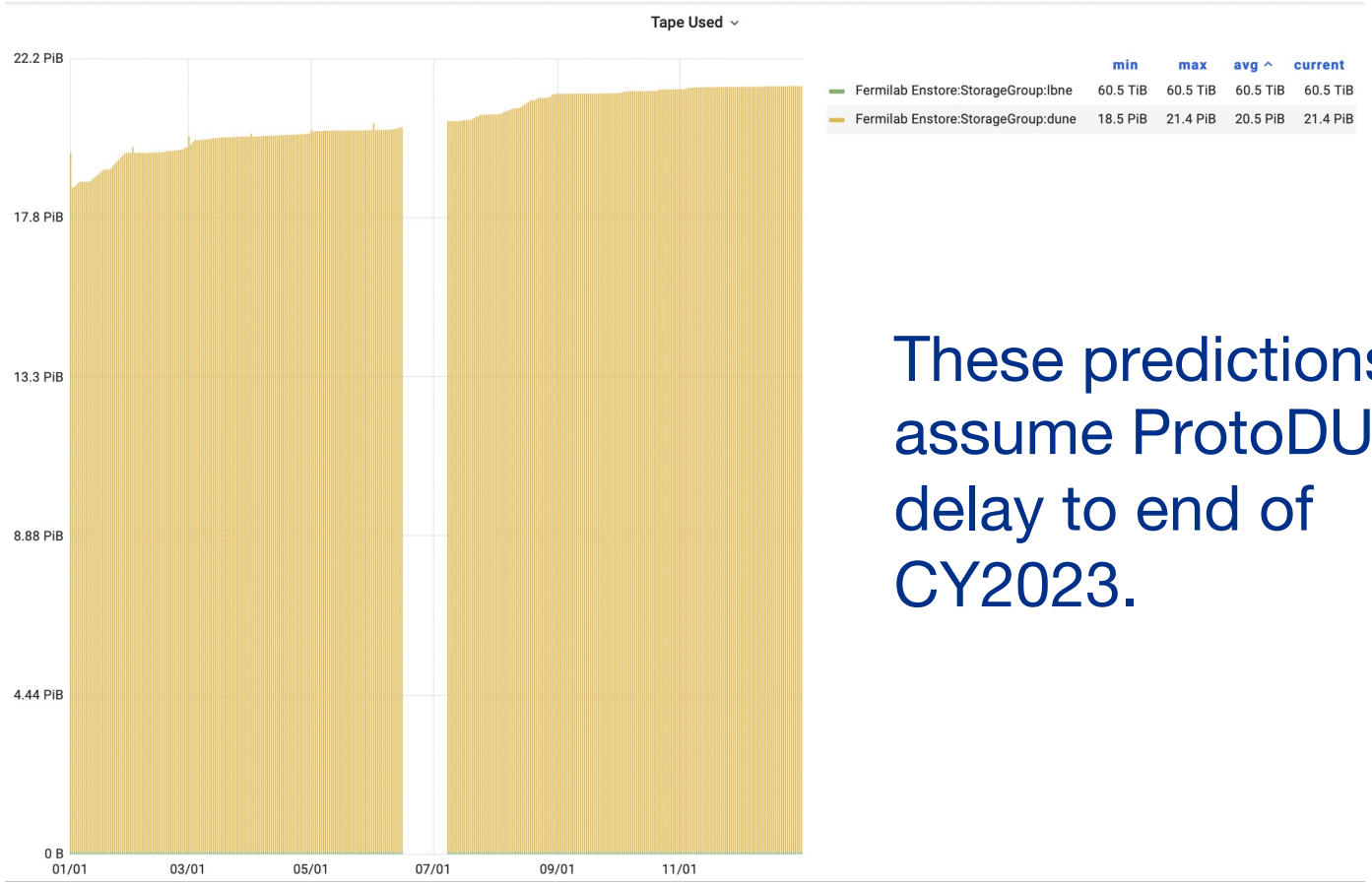
dCache Usage and Predictions (in TB)



Total dedicated available r/w (tape backed): 7000 TB
Total persistent: 950 TB
Total dedicated other (staging pools, etc.): 1000 TB

000	Analysis (Persistent)	Other Dedicated (Write)	Offsite dedicated disk
Current	371/396 TB (user) 96/550TB (physics group)	4450/7000 (RW) 730/1000 (staging)	11000TB
2022 (pledged)	950 TB	9800 TB	11000TB
2023	2000 TB	9800 TB	15200 TB
2024	2000 TB	9800 TB	18100 TB
2025	3500 TB	9800 TB	19000 TB
2026	3500 TB	7000 TB	15700 TB
2027	3500 TB	5000 TB	11000 TB

Tape usage and predictions (in TB)



These predictions assume ProtoDUNE delay to end of CY2023.

	Total Added By End of Year	Offsite
End 2021	Total 17.8	
End 2022	Actual 21.4 (Predicted 24.3)	5.1
2023	+12.1	+4.3
2024	+12.1	+3.6
2025	+12.6	+2.5
2026	+6.0	-0.3
2027	+4.5	+1.3

Data Lifetimes

DUNE is about to delete ~1PB of older Monte Carlo samples that are no longer used from tape.
New MetaCat data catalog features a "data retention type". Known types are "test" (2 years), "study" (5 years) "raw data" (indefinite).
Only latest 2 MC samples and reco samples will be available on disk.
"data retention types" in MetaCat will be accompanied with expiration times in the Rucio replica catalog.

Analysis Facility Use

- A small number of DUNE users already using the Elastic Analysis Facility
- Expect to grow to 10-20 users over this year growing to ~100 in the long term
- Use case currently is fast turnaround from freshly taken data, much higher IO bytes per CPU instruction than analyzing reconstructed data.
- Stream raw data from dCache frequently. xCache may be solution here, beginning to investigate that
- Expect GPU demand to escalate as more users are working on developing ML-based reconstruction
- Expect more demand for POSIX-like disk as EAF grows.