SBN Presentation for FCRSG 2023

Giuseppe Cerati, Steven Gardiner, Wes Ketchum

15 February 2023

Organization Chart for Offline Computing

ICARUS Simulation and Reconstruction Software
Tracy Usher, Daniele Gibin

Release Management
Sergey Martynenko, Matthew Rosenberg

Production Management
Ivan Caro Terrazas

SBN Computing Sector Liaison
Vito Di Benedetto

SBN Analysis Infrastructure
Giuseppe Cerati, Steven Gardiner

Analysis Framework
Gianluca Petrillo, Fernanda Psihas

Production and Data Management
Mateus Carneiro, Ivan Caro Terrazas

Release Management
Miquel Nebot-Guinot

Simulation Infrastructure
Marco Del Tutto, Alessandro Menegolli

Validation Framework
Chris Hilgenberg

Workflow Management
Sophie Berkman, Herb Greenlee

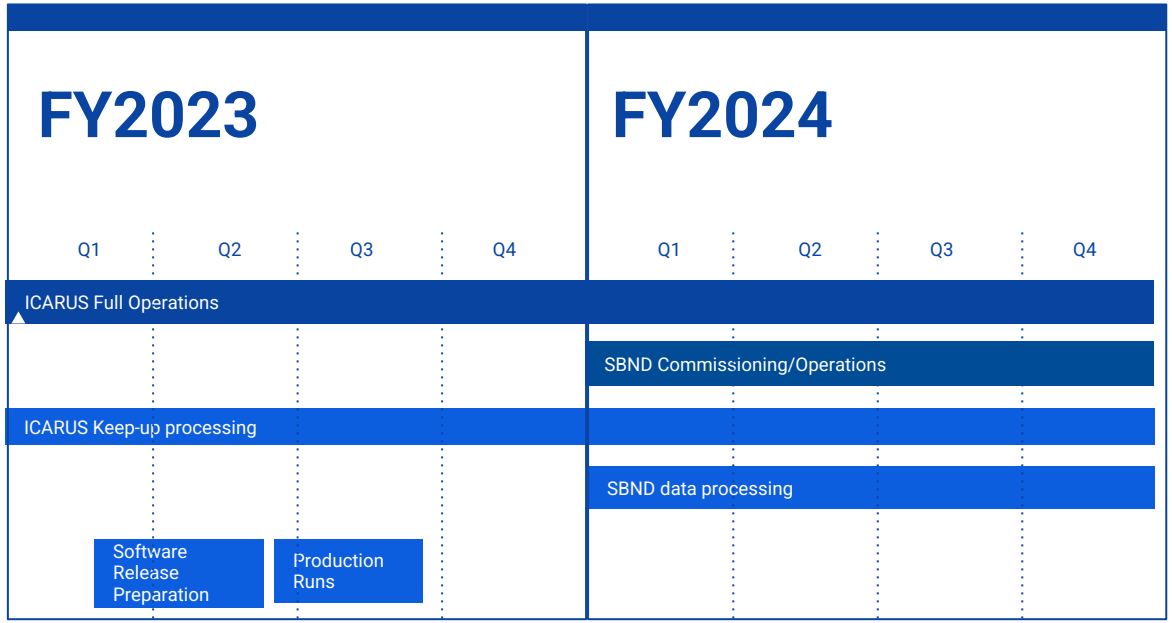
SBN Analysis
Daniele Gibin, Georgia Karagiorgi

SBND Physics and Analysis Tools
Costas Andreopoulos, Bill Louis, Andrzej Szelc

Release Management
Francisco Javier Nicolás Arnaldos

Production Management
Mateus Carneiro

Overview Timeline

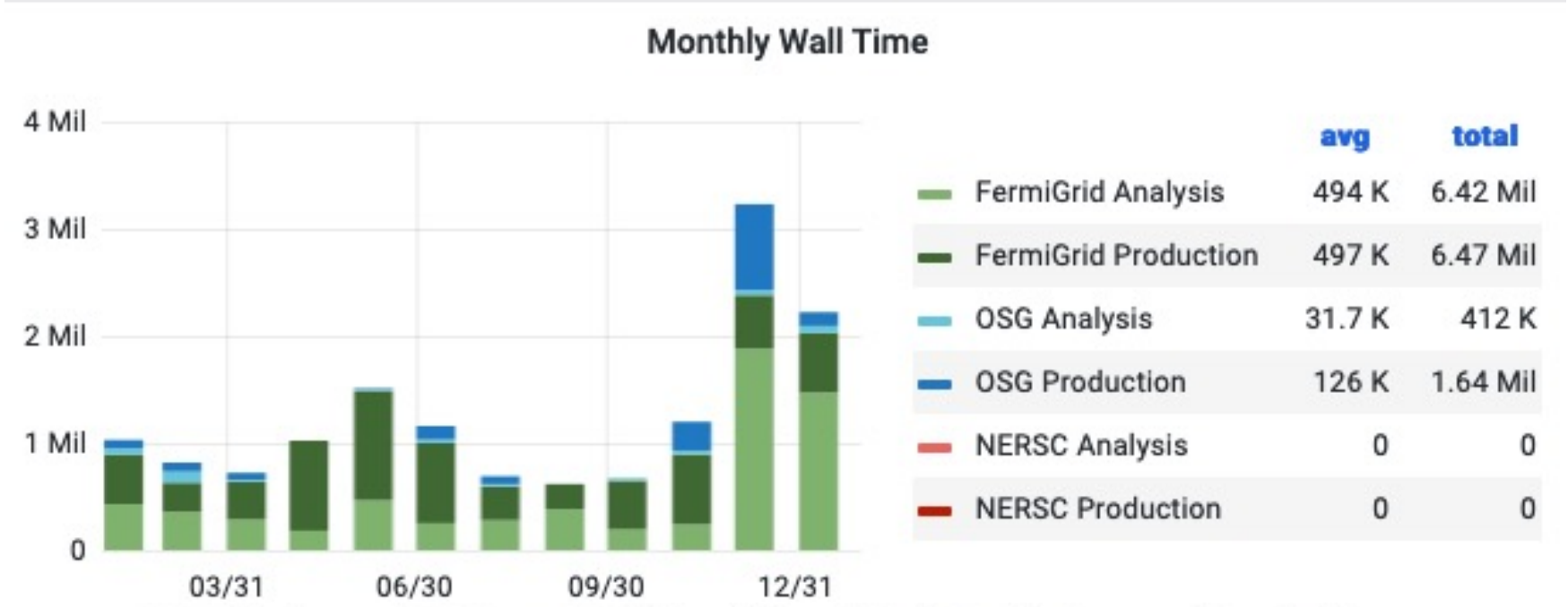


- ICARUS started physics data taking at the end of calendar year (CY) 2022
  - First physics results as early as FY2023
- SBND to enter commissioning in FY2024 and move (hopefully rapidly) to full operations

Computing model updates

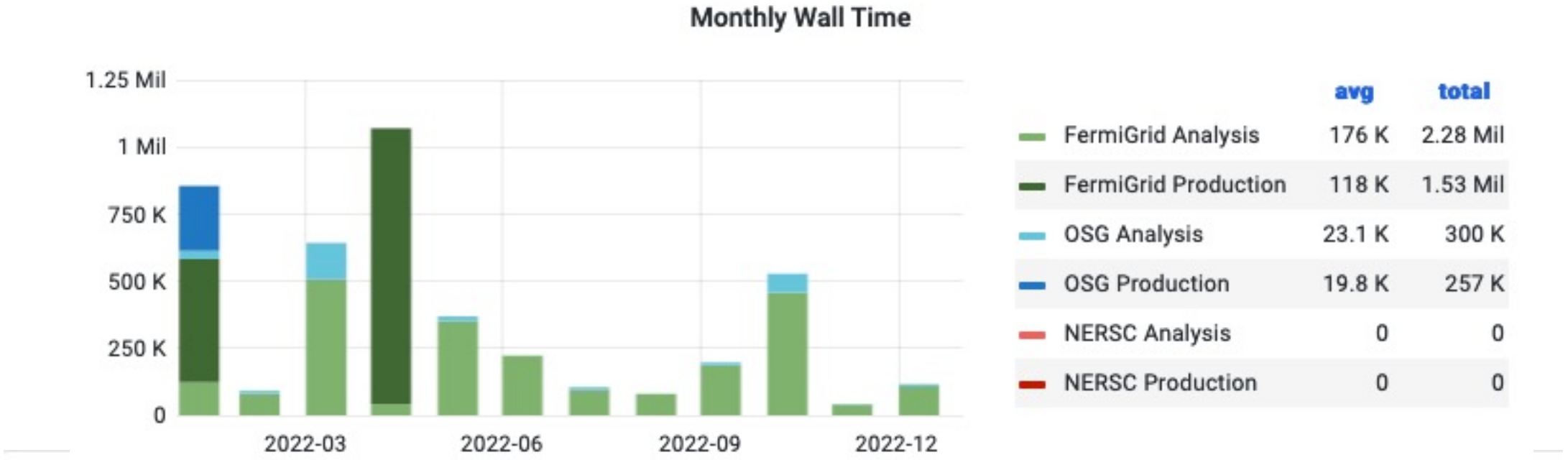
- Run keep-up processing and save output of first stage reconstruction on disk
  - size of stage0 output is 6.5x smaller than compressed RAW
  - Allows for re-processing of data without reading back from tape
- Run MC campaigns from scratch at least once per year
  - Avoid tape bottleneck
  - Allows for updates to interaction and simulation model
  - Drop most artroot files and keep analysis file format on disk
  - Store first pass reco on tape in case of reprocessing
  - Keep 0.1-1M events on disk after full reco for development
- Main use of tape is for perennial archival data
  - and for storing other samples for limited time before deletion

ICARUS CPU Usage over the past year



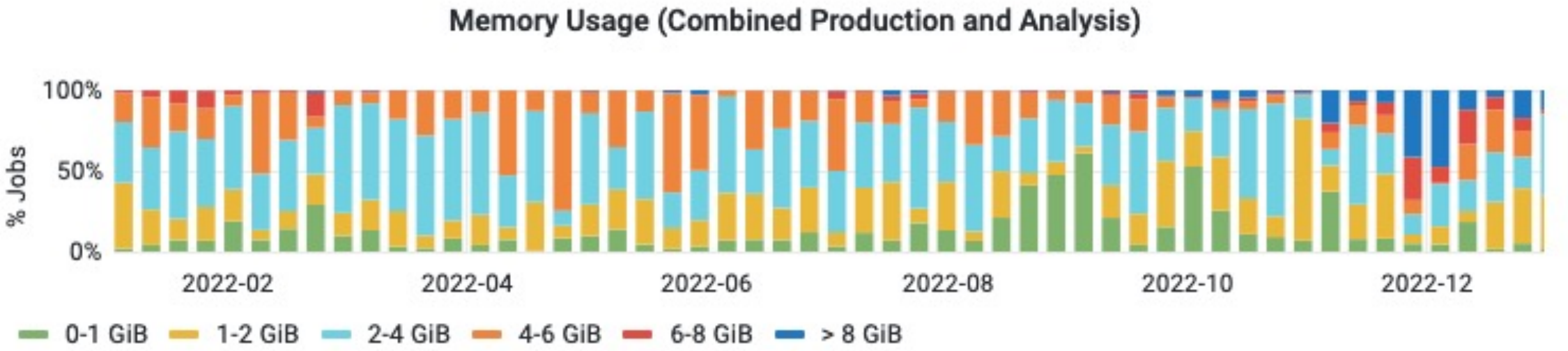
Fermilab

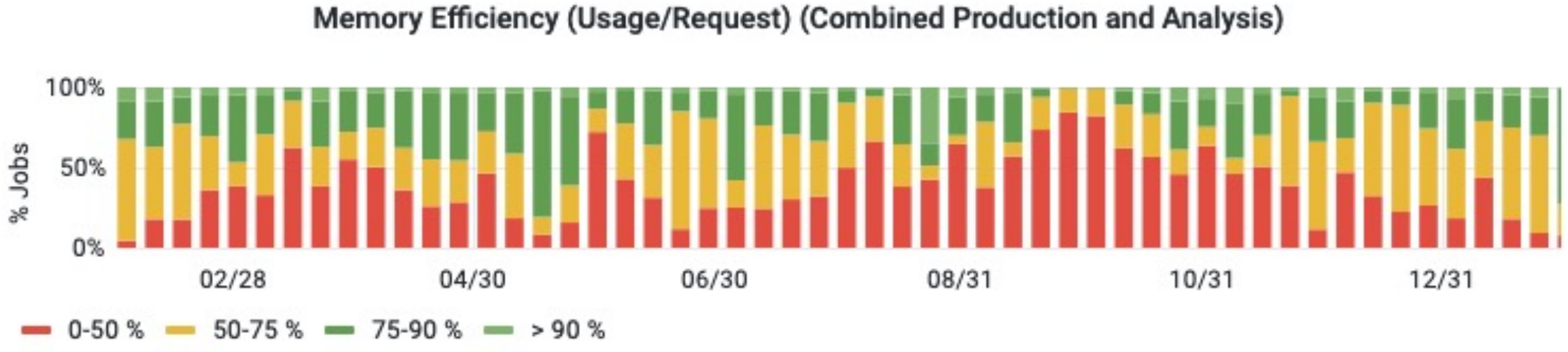
SBND CPU Usage over the past year



Fermilab

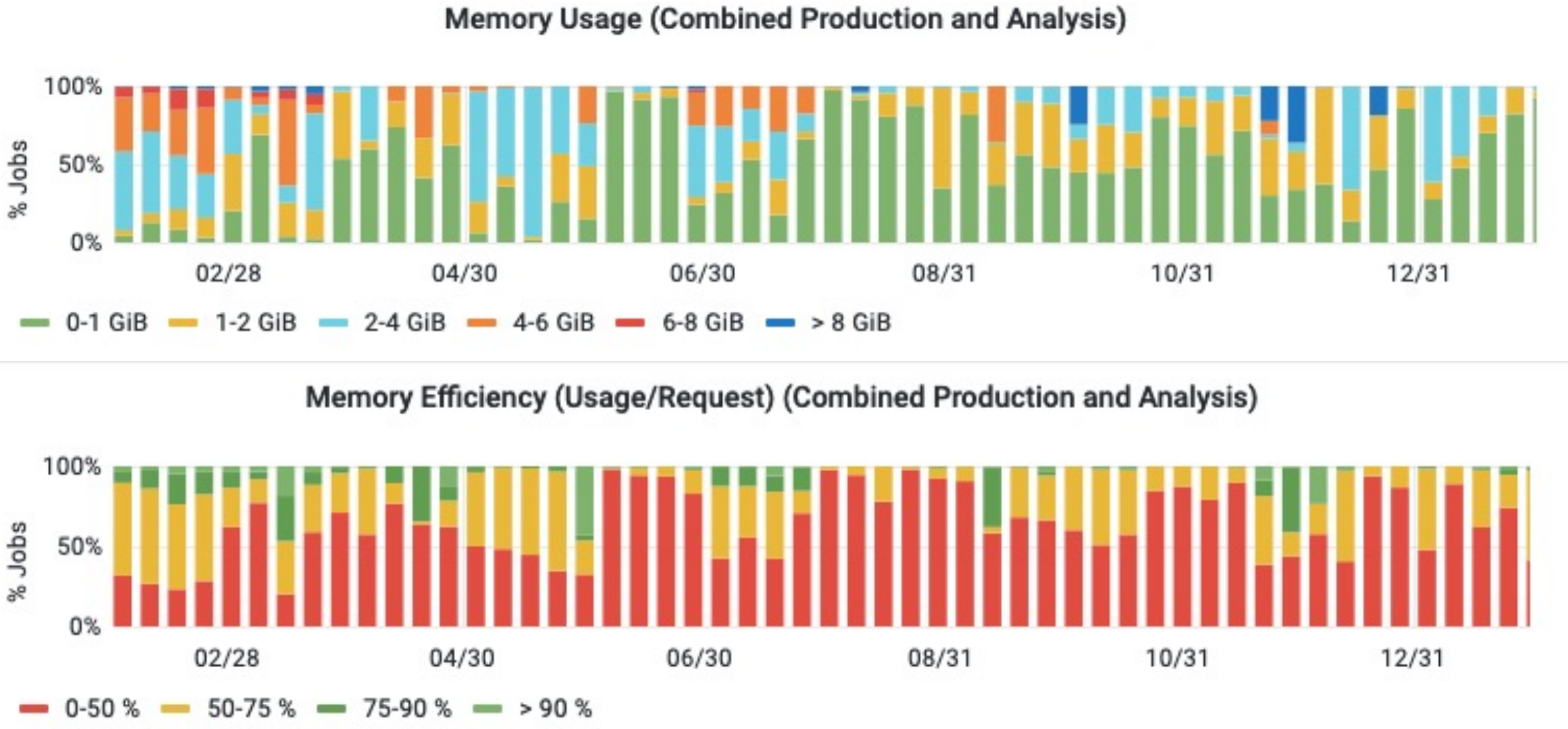
ICARUS memory usage over the past year





Fermilab

SBND memory usage over the past year



Fermilab

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

Goals	Where does the experiment need to contribute	Where does CSAID needs to contribute
Deploy PMT-CRT matching as software filter for early cosmic rejection	Experiment goal	N/A
Enable data overlay simulation	Experiment goal	May need support for optimization of workflow (e.g. if cosmic data needs to be read from tape)
Improve resource usage in production with multithreading and/or reduced memory footprint	Joint effort	Joint effort
Integration of external resources (CNAF) through RUCIO	Integration of non-FNAL resources in prod. workflows	Need to enable data transfers to FNAL, and integrated data management
Better integration of AI inference in production workflows	Joint effort	Centrally supported mechanism for running inference in production
Harmonize infrastructure between ICARUS and SBND (simulation, AI, etc.)	Experiment goal	N/A
Usage of HPC resources for production jobs	Need to transition from R&D to integrated resource	Need to transition from R&D to integrated resource
Enable usage of Elastic Analysis Facility for analysis and AI training	Define use cases, advertise in collaborations	User support (docs, tutorials, support of experiment resources)

ICARUS Campaign Schedules

	2023	2024	2025	2026	2027
Processing campaigns (start month-end month if known). **Include when you expect to be prestaging**	Data: All Year keepup process MC Campaign 1: Mar-May MC Campaign 2: Aug-Oct **No significant prestaging**	Data: All Year keepup process MC Campaign 1: Jan-Mar MC Campaign 2: Aug-Oct **Prestaging of MC1 for MC2**	Data: All Year keepup process Overlay Camp. 1: Jan-Mar Overlay Camp. 2: Aug-Oct **Prestage OffBeam data** **Prestaging of MC1 for MC2**	Data: All Year keepup process Overlay Camp. 1: Jan-Mar Overlay Camp. 2: Aug-Oct **Prestage OffBeam data** **Prestaging of MC1 for MC2**	Data: All Year keepup process Overlay Camp. 1: Jan-Mar Overlay Camp. 2: Aug-Oct **Prestage OffBeam data** **Prestaging of MC1 for MC2**
Storage + CPU estimates (call out any special resource needs if known, e.g. HPC or GPU). **Include amount(s) to be prestaged and file families, in addition to space needed for new outputs.** **Note: currently not including deletions from disk nor tape (see later slide)**	Data: CPU 4.5M hrs Write 3.7 PB to tape Write 1.9 PB to disk MC1: CPU 5.7M hrs No significant prestaging No write to tape Write 50 TB to disk MC2: CPU 5.7M hrs No significant prestaging Write 50 TB to tape Write 50TB to disk	Data: CPU 5.7M hrs Write 3.7 PB to tape Write 0.9 PB to disk MC1: CPU 10.4M hrs No significant prestaging Write 4.4 PB to tape Write 80 TB to disk MC2: CPU 5.5M hrs Prestage 4.4 PB (SIM) Write 80 TB to tape Write 80 TB to disk	Data: CPU 4.3M hrs Write 3.7 PB to tape Write 0.9 PB to disk MC1: CPU 9.8M hrs Prestage 4.9 PB (rawoffbeam) Write 6.7 PB to tape Write 100 TB to disk MC2: CPU 2.8M hrs Prestage 6.7 PB (SIM) Write 100 TB to tape Write 100 TB to disk	Data: CPU 4.9M hrs Write 3.7 PB to tape Write 0.9 PB to disk MC1: CPU 12.8M hrs Prestage 6.5 PB (rawoffbeam) Write 8.8 PB to tape Write 130 TB to disk MC2: CPU 3.7M hrs Prestage 8.8 PB (SIM) Write 130 TB to tape Write 130 TB to disk	Data: CPU 5.5M hrs Write 3.7 PB to tape Write 0.9 PB to disk MC1: CPU 15.7M hrs Prestage 8.1 PB (rawoffbeam) Write 11 PB to tape Write 150 TB to disk MC2: CPU 4.5M hrs Prestage 11 PB (SIM) Write 150 TB to tape Write 150 TB to disk
Conference or result targets (month if known)	First ICARUS results	Neutrino 2024			

SBND Campaign Schedules

	2023	2024	2025	2026	2027
Processing campaigns (start month-end month if known). **Include when you expect to be prestaging**	No Data MC Campaign 1: Mar-May (or later) **No significant prestaging**	Data: All Year keepup process MC Campaign 1: Jan-Mar MC Campaign 2: Aug-Oct **No significant prestaging**	Data: All Year keepup process Overlay Camp. 1: Jan-Mar Overlay Camp. 2: Aug-Oct **No significant prestaging**	Data: All Year keepup process Overlay Camp. 1: Jan-Mar Overlay Camp. 2: Aug-Oct **Prestage OffBeam data** **Prestaging of MC1 for MC2**	Data: All Year keepup process Overlay Camp. 1: Jan-Mar Overlay Camp. 2: Aug-Oct **Prestage OffBeam data** **Prestaging of MC1 for MC2**
Storage + CPU estimates (call out any special resource needs if known, e.g. HPC or GPU). **Include amount(s) to be prestaged and file families, in addition to space needed for new outputs.** **Note: currently not including deletions from disk nor tape (see later slide)**	MC1: CPU 6.2M hrs No significant prestaging No write to tape Write 50TB to disk	Data: CPU 10M hrs Write 4.2 PB to tape Write 2.1 PB to disk MC1: CPU 6.2M hrs No significant prestaging Write 50 TB to tape Write 50 TB to disk MC2: CPU 6.2M hrs No significant prestaging Write 50 TB to tape Write 50 TB to disk	Data: CPU 12M hrs Write 3.2 PB to tape Write 1.6 PB to disk MC1: CPU 7.7M hrs No significant prestaging Write 50 TB to tape Write 60 TB to disk MC2: CPU 7.7M hrs No significant prestaging Write 60 TB to tape Write 60 TB to disk	Data: CPU 15M hrs Write 3.2 PB to tape Write 0.8 PB to disk MC1: CPU 10.5M hrs Prestage 5.9 PB (rawoffbeam) Write 1.3 PB to tape Write 100 TB to disk MC2: CPU 5.8M hrs Prestage 1.3 PB (SIM) Write 100 TB to tape Write 100 TB to disk	Data: CPU 18M hrs Write 3.2 PB to tape Write 0.8 PB to disk MC1: CPU 15.0M hrs Prestage 8.5 PB (rawoffbeam) Write 1.8 PB to tape Write 130 TB to disk MC2: CPU 8.3M hrs Prestage 1.8 PB (SIM) Write 130 TB to tape Write 130 TB to disk
Conference or result targets (month if known)	MC studies for summer/fall conferences	Neutrino 2024			

SBN CPU @ Fermilab Prediction Going Forward and Accuracy of the 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)	13.2	16.5	19.5	27	**29**	**53±15**	**48±15**	**58±15**	**74±15**
Actual Used	4.46	7.7	11.4	15	N/A	N/A	N/A	N/A	N/A
Efficiency	33.8%	46.7%	58.5%	55.6%	N/A	N/A	N/A	N/A	N/A

Assumed that efficiency = requested / actual

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

We plan to use OSG in opportunistic mode.
We've been using it for up to 20% for production jobs.

ICARUS has allocation at CNAF (Italy), re-negotiated each year.
Currently 3M CPU hours (1 PB disk, 2 PB tape), but needs deployment of RUCIO transfers from CNAF to FNAL to be fully integrated in workflows.
Short term plan is to use this resource for reprocessing of data from runs before summer 2022.

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

HPC resources used by individual groups, not by SBN experiments themselves (no centralized accounting at present). Current R&D efforts may lead to centralized requests in future planning. Notable examples:

ICARUS ML workflow: plan to run end-to-end ML reconstruction at HPC centers (GPU) using allocation from a collaborator's grant. Dedicated input files need to be transferred from FNAL to offsite storage during or shortly after production. Analysis file format eventually needs to be accessible to the full collaboration.

SBNfit: Run Feldman-Cousins fits for sensitivity studies of SBN oscillation analyses at NERSC, using development version of the code specific for HPC. Leverages SciDAC4 allocation.

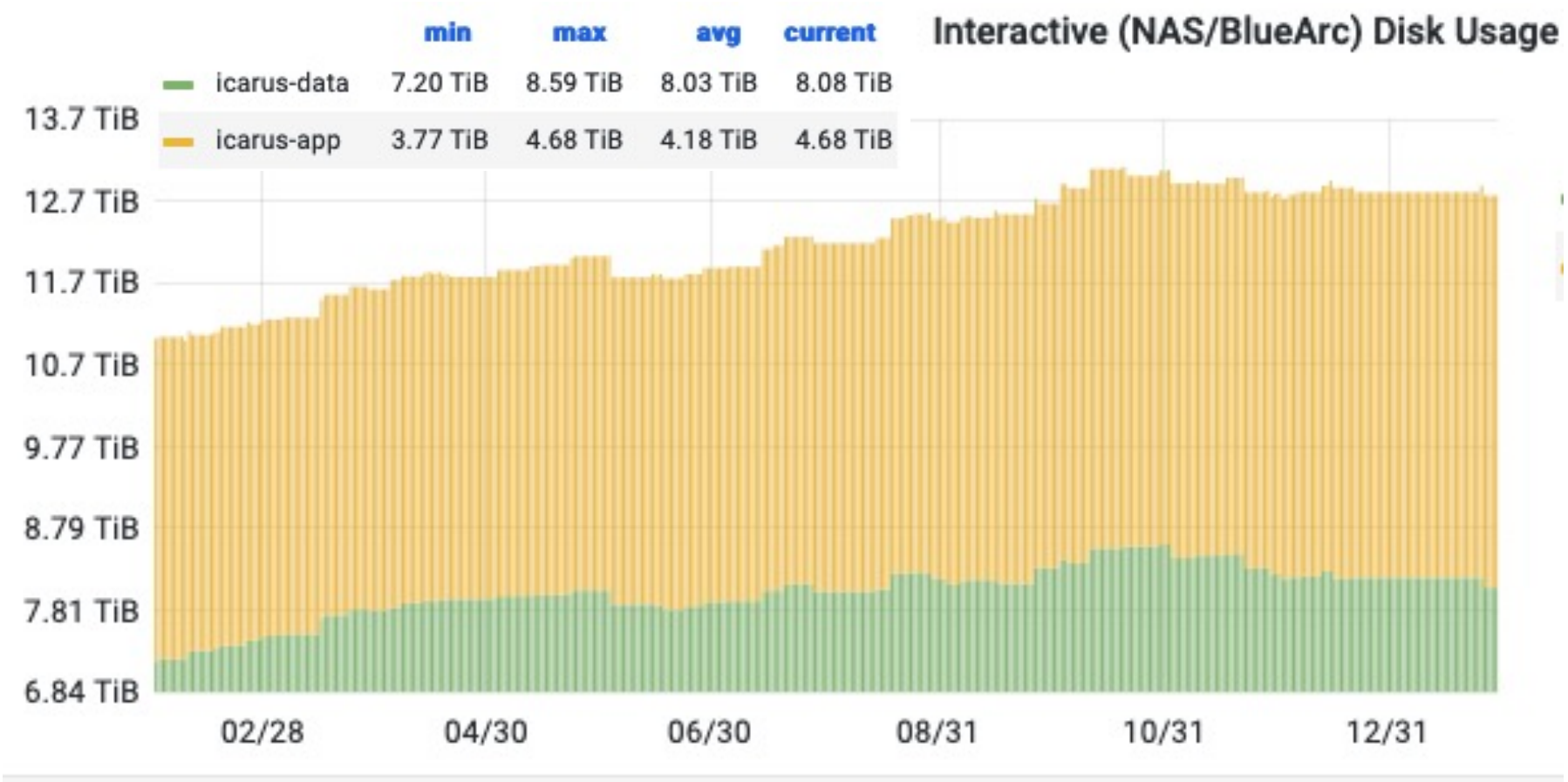
HPC LArSoft Workflow: development of HPC processing workflow based on Spack built central LArSoft code is under test on Theta@ALCF. Part of SciDAC projects as well.

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

No current roadmap to evolving central SBN software to run on GPUs. Support of GPU offloading for common tasks in LArSoft (e.g. Geant4?) can be a game changer.

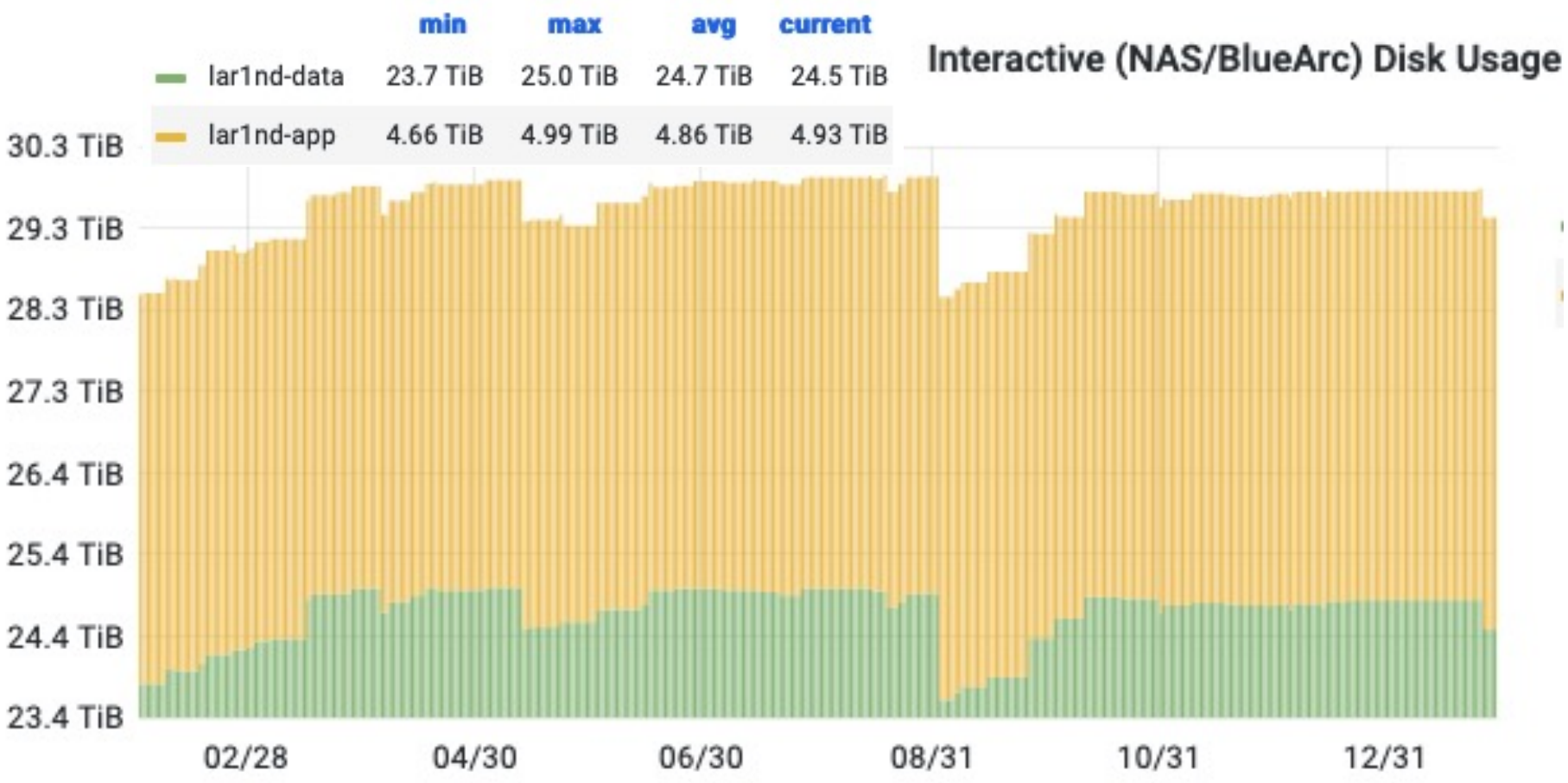
GPUs are used for ML training and inference, both onsite (see also later slide on Elastic Analysis Facility) and offsite.

ICARUS NAS+Ceph Usage and Projections



	App [TB]	Data [TB]
2022	5	9
2023	5	25
2024	6	30
2025	7	35
2026	8	40
2027	8	40

SBND NAS+Ceph Usage and Projections

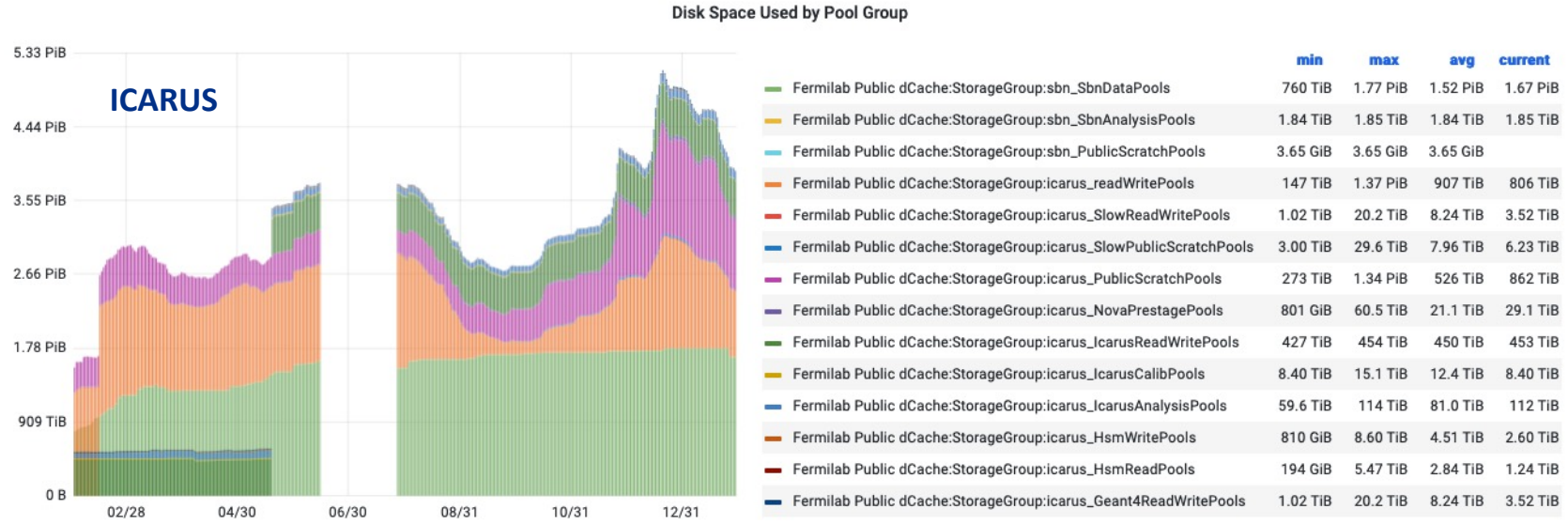


	App [TB]	Data [TB]
2022	5	25
2023	7	30
2024	8	35
2025	8	40
2026	9	45
2027	9	45

Fermilab

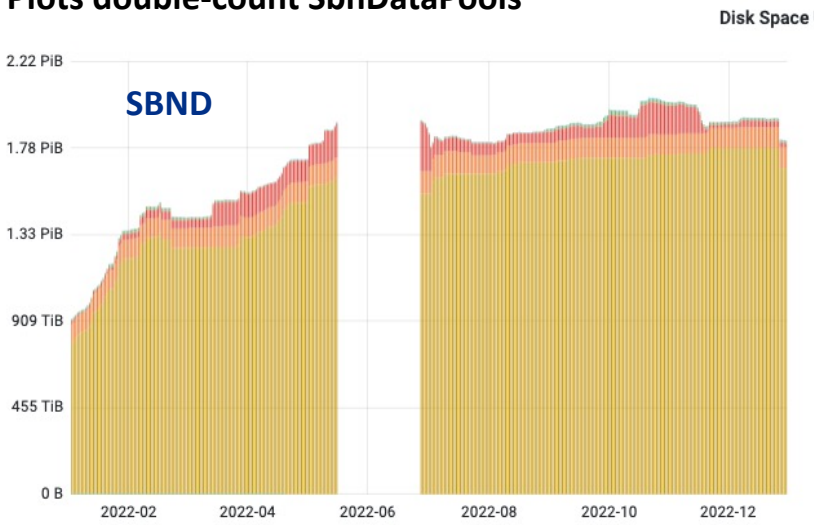
ICARUS + SBND dCache Usage and Predictions

ICARUS has 1 PB allocated at CNAF (currently 60% used)



	min	max	avg	current
Fermilab Public dCache:StorageGroup:sbn_SbnDataPools	760 TiB	1.77 PiB	1.52 PiB	1.67 PiB
Fermilab Public dCache:StorageGroup:sbn_SbnAnalysisPools	1.84 TiB	1.85 TiB	1.84 TiB	1.85 TiB
Fermilab Public dCache:StorageGroup:sbn_PublicScratchPools	3.65 GiB	3.65 GiB	3.65 GiB	
Fermilab Public dCache:StorageGroup:icarus_readWritePools	147 TiB	1.37 PiB	907 TiB	806 TiB
Fermilab Public dCache:StorageGroup:icarus_SlowReadWritePools	1.02 TiB	20.2 TiB	8.24 TiB	3.52 TiB
Fermilab Public dCache:StorageGroup:icarus_SlowPublicScratchPools	3.00 TiB	29.6 TiB	7.96 TiB	6.23 TiB
Fermilab Public dCache:StorageGroup:icarus_PublicScratchPools	273 TiB	1.34 PiB	526 TiB	862 TiB
Fermilab Public dCache:StorageGroup:icarus_NovaPrestagePools	801 GiB	60.5 TiB	21.1 TiB	29.1 TiB
Fermilab Public dCache:StorageGroup:icarus_IcarusReadWritePools	427 TiB	454 TiB	450 TiB	453 TiB
Fermilab Public dCache:StorageGroup:icarus_IcarusCalibPools	8.40 TiB	15.1 TiB	12.4 TiB	8.40 TiB
Fermilab Public dCache:StorageGroup:icarus_IcarusAnalysisPools	59.6 TiB	114 TiB	81.0 TiB	112 TiB
Fermilab Public dCache:StorageGroup:icarus_HsmWritePools	810 GiB	8.60 TiB	4.51 TiB	2.60 TiB
Fermilab Public dCache:StorageGroup:icarus_HsmReadPools	194 GiB	5.47 TiB	2.84 TiB	1.24 TiB
Fermilab Public dCache:StorageGroup:icarus_Geant4ReadWritePools	1.02 TiB	20.2 TiB	8.24 TiB	3.52 TiB

Plots double-count SbnDataPools



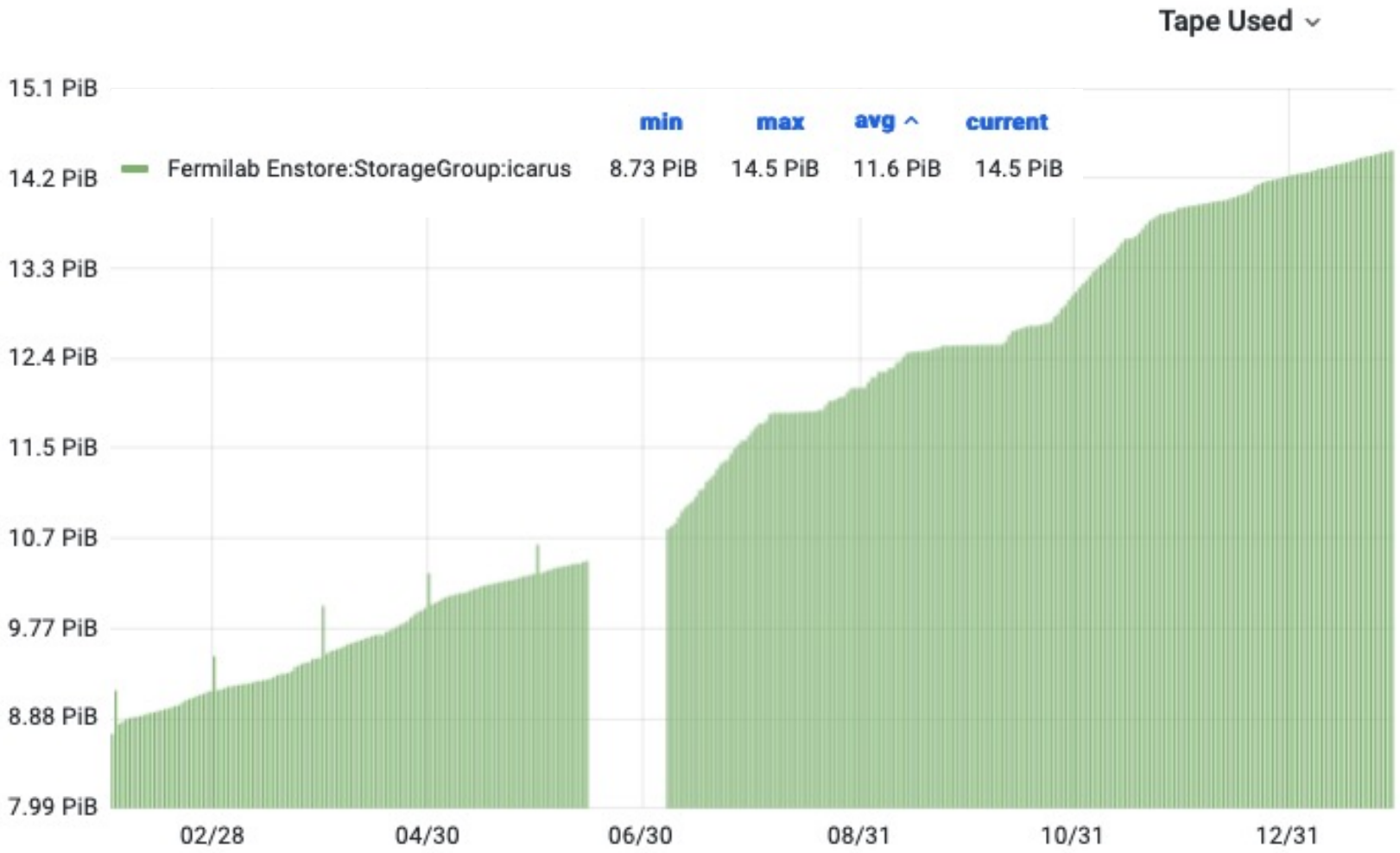
	min	max	avg	current
Fermilab Public dCache:StorageGroup:sbn_SbnDataPools	760 TiB	1.77 PiB	1.52 PiB	1.67 PiB
Fermilab Public dCache:StorageGroup:sbnd_SbndAnalysisPools	95.3 TiB	118 TiB	105 TiB	110 TiB
Fermilab Public dCache:StorageGroup:sbnd_PublicScratchPools	5.87 TiB	261 TiB	77.5 TiB	19.8 TiB
Fermilab Public dCache:StorageGroup:sbnd_readWritePools	2.46 TiB	22.8 TiB	9.87 TiB	11.2 TiB
Fermilab Public dCache:StorageGroup:sbn_SbnAnalysisPools	1.84 TiB	1.85 TiB	1.84 TiB	1.85 TiB
Fermilab Public dCache:StorageGroup:sbnd_NovaPrestagePools	20.3 GiB	771 GiB	429 GiB	713 GiB
Fermilab Public dCache:StorageGroup:sbnd_SlowPublicScratchPools	518 MiB	4.99 TiB	2.37 TiB	246 GiB
Fermilab Public dCache:StorageGroup:sbnd_Geant4ReadWritePools	23.9 GiB	159 GiB	69.5 GiB	153 GiB
Fermilab Public dCache:StorageGroup:sbnd_SlowReadWritePools	23.9 GiB	159 GiB	69.5 GiB	153 GiB
Fermilab Public dCache:StorageGroup:sbnd_HsmWritePools	264 MiB	36.5 GiB	23.4 GiB	30.9 GiB
Fermilab Public dCache:StorageGroup:sbnd_HsmReadPools	254 MiB	247 GiB	51.2 GiB	9.79 GiB
Fermilab Public dCache:StorageGroup:sbn_PublicScratchPools	3.65 GiB	3.65 GiB	3.65 GiB	

	Analysis (Persistent)	Other Dedicated (Write)
Current	1.9 PB (actual)	0.8 PB (actual)
2023	3.3 PB	1 PB
2024	5.7±1 PB	2 PB
2025	5.9±1 PB	2 PB
2026	6.2±1 PB	2 PB
2027	7.3±1 PB	2 PB

Analysis = IcarusAnalysis + SbndAnalysis + SbnData pool
Other = readWrite pool

Fermilab

ICARUS Tape usage and predictions



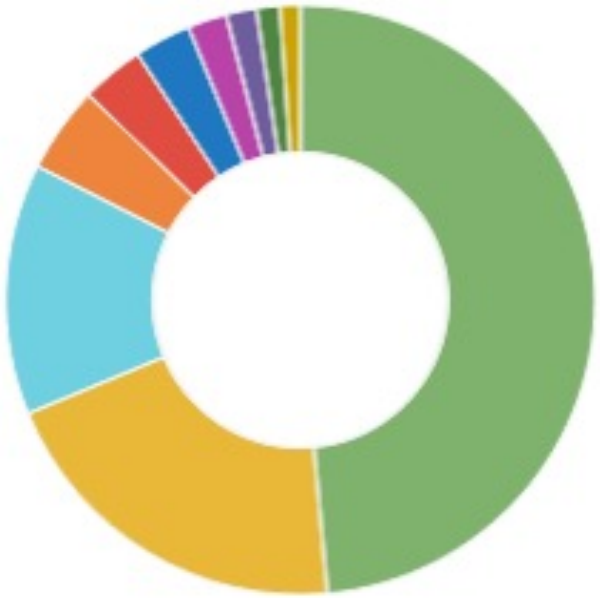
	Total Added By End of Year
At end 2021	8.7 PB
2022	+3.9 PB
2023	+3.8 PB
2024	+8±2 PB
2025	+11±5 PB
2026	+4±5 PB
2027	+1±5 PB

ICARUS Tape usage and predictions

File families organized by data stream.

Half of volume is in generic family

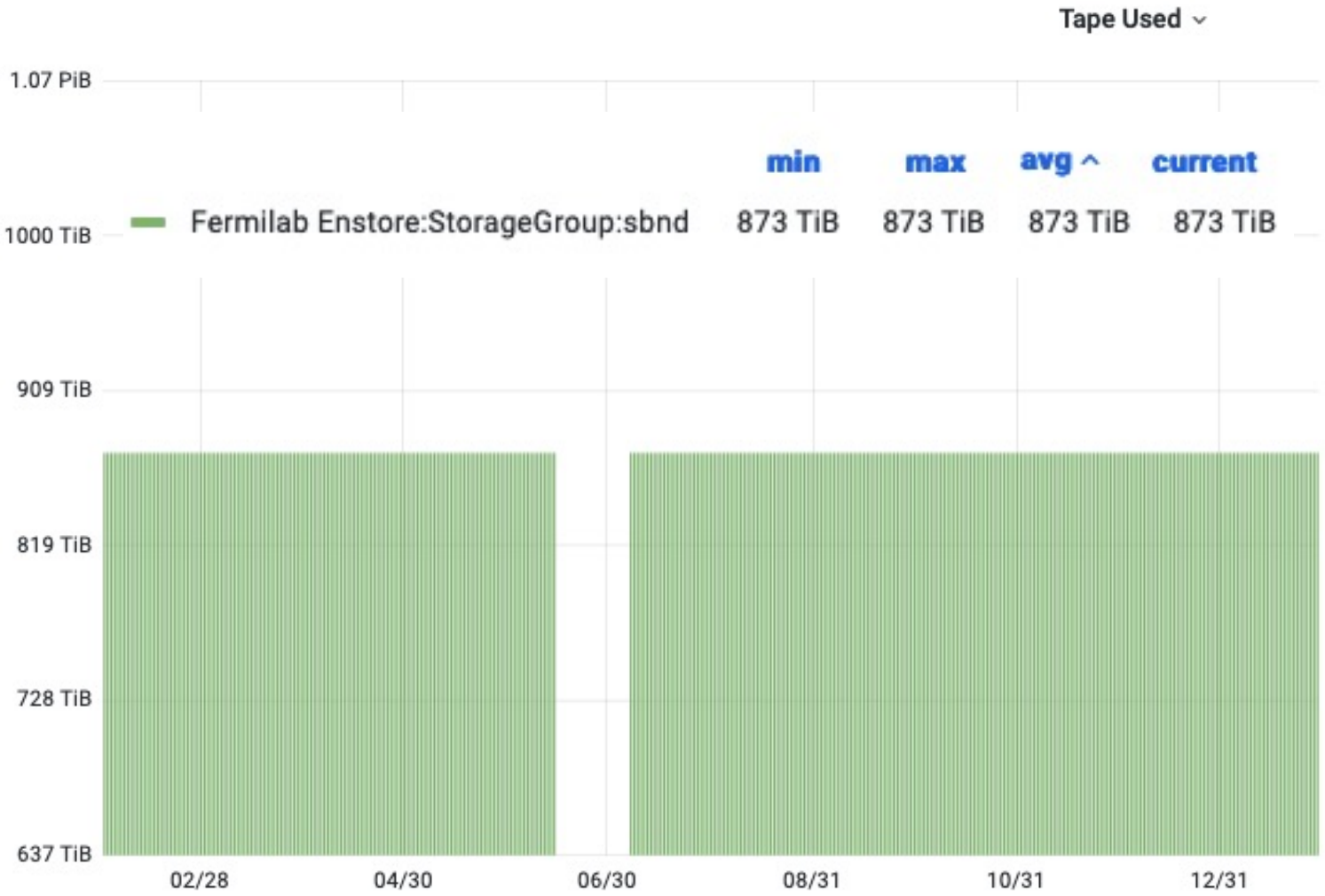
We need more granular definitions, e.g.:
- split by run?
- artroot MC?
- analysis files?



Data Volume by File Family

	current	percentage
icarus	7 PB	49%
data_artroot_raw_offbeambnbminbias	3 PB	20%
raw	2 PB	14%
data_artroot_reconstructed_bnb	729 TB	5%
data_artroot_raw_bnb	527 TB	3%
data_artroot_raw_unknown	477 TB	3%
data_artroot_raw_offbeambnb	329 TB	2%
data_artroot_raw_numi	254 TB	2%
data_artroot_raw_offbeamnumi	195 TB	1%
icarusdata_artroot_reconstructed_numi	171 TB	1%

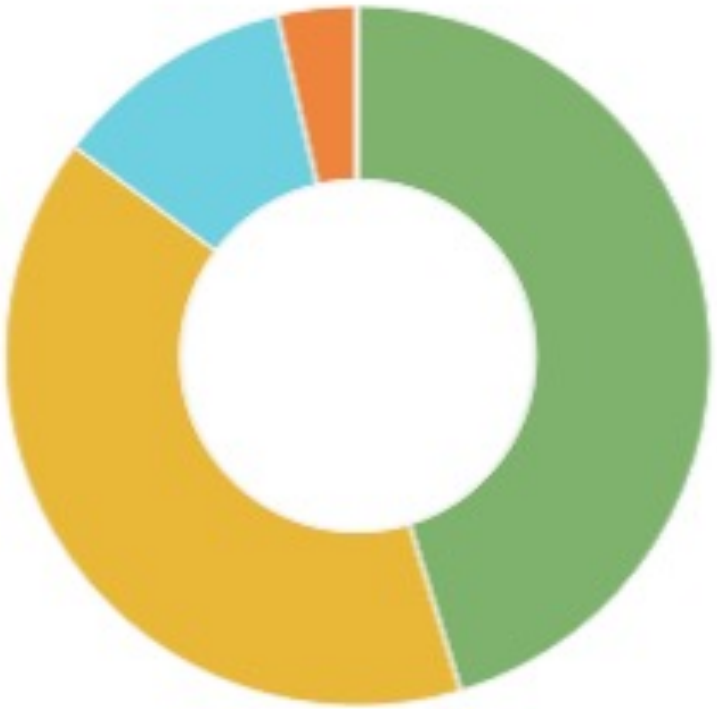
SBND Tape usage and predictions



	Total Added By End of Year
At end 2021	+0.9 PB
2022	+0 PB
2023	+0±1 PB
2024	+4.3±2 PB
2025	+3.4±2 PB
2026	+4.6±2 PB
2027	+5.2±2 PB

Fermilab

SBND Tape usage and predictions



	current	percentage
sbnd	400 TB	45%
mc_reco	353 TB	40%
mc_detsim	99 TB	11%
mc_g4	32 TB	4%
mc_anatree	318 GB	0%
mc_gen	16 GB	0%
mc_root_small	4 GB	0%

Fermilab

Data Lifetimes

Summary of plan for data lifetimes:

- Raw (detector) data: permanent on tape
  - At the beginning of a new run, keep a few months of RAW data on disk to allow fast reprocessing in case calibrations are not ready
- First pass reco data: keep permanently on disk for reprocessing
- Other derived data: 1-2 year lifetime
- Simulation: 1-2 year lifetime

Analysis Facility Use

- A few SBN users have started using the facility for ML training
  - In the longer term we envision a combination of analysis work + ML training
- Mounting /pnfs would be great
  - xrootd already works, but not sure if there are bandwidth concerns for massive usage
  - Our analysis files take O(100) TB per campaign
- Likely combination of NVMe (solid state) + HDD storage would be optimal
  - Just a guess at this point
- Mixed CPUs + GPUs needed given the different applications

Backup