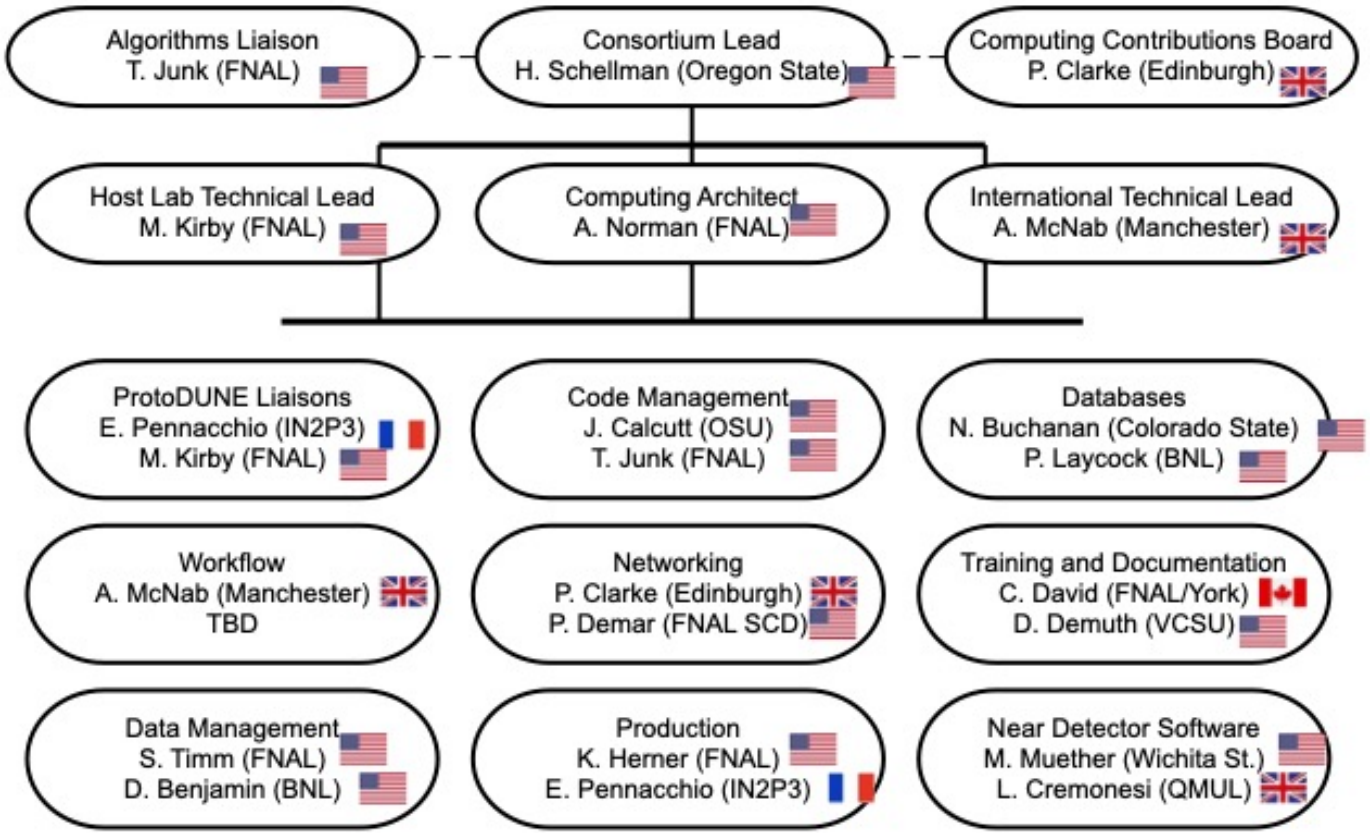
Fermilab U.S. DEPARTMENT OF ENERGY | Office of Science

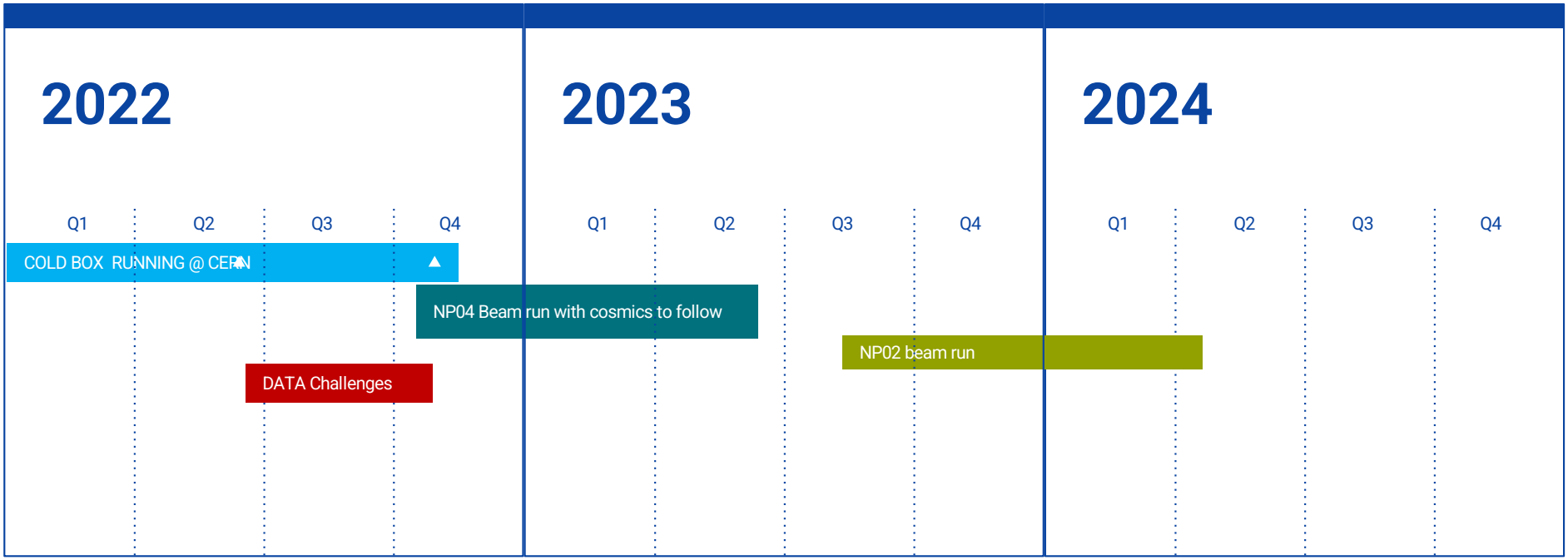
DUNE FCRSG 2022

Heidi Schellman for the Computing Consortium

Experiment Organization Chart for Offline Computing



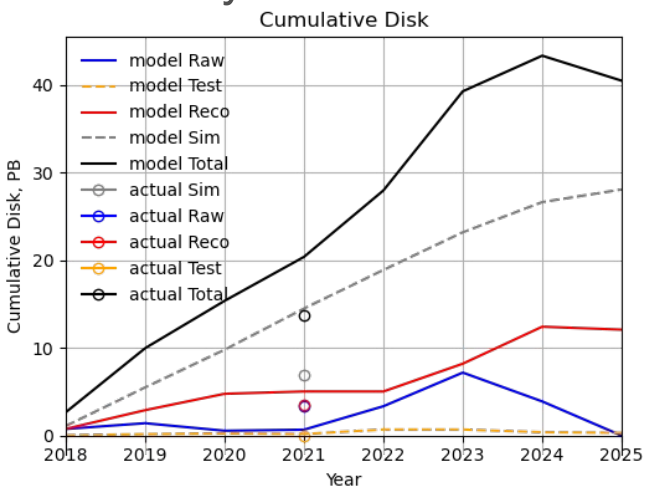
Important Dates to Remember

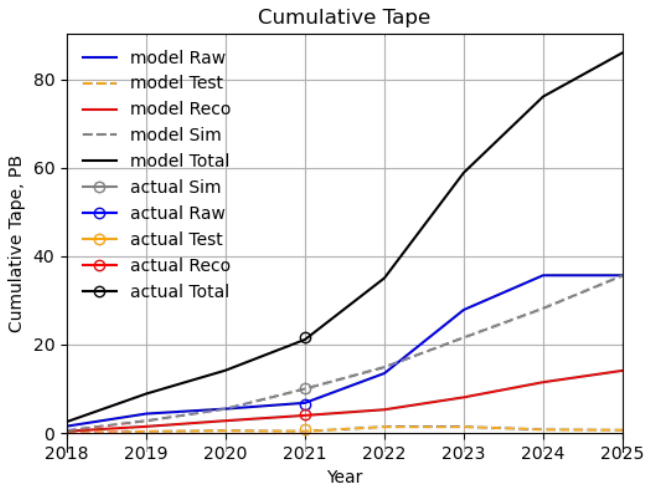


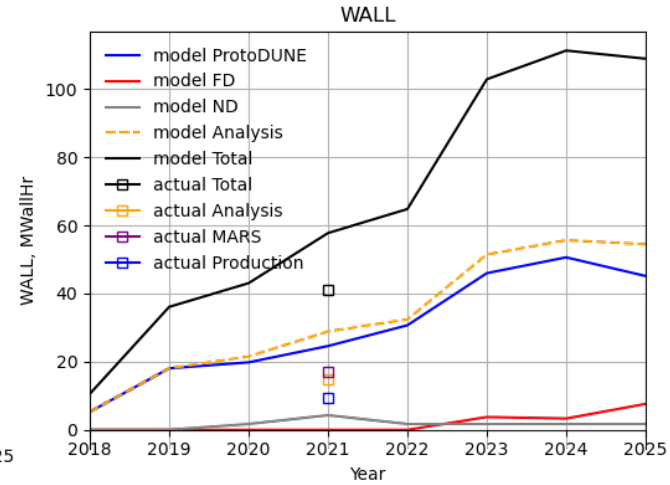
Big DUNE DAQ commissioning starts in 2027, first “light” in 2029

Data Model Calculations

- Dune has a python based resource prediction model that is updated 1-2 times/year.







This shows total needs. Collaboration contributions are negotiated by the Computing Contributions Board each year

Important notes on DUNE CPU usage

- 17 **MHrs of wall time/year is used for MARS which was not in our prediction models**
    - MARS is pretty CPU and memory efficient
- **Analysis and production ran at about 80% efficiency at FNAL and used 27 MHrs of FNAL slot-weighted wall time according to Kibana – but typical jobs use 4-6 GB of memory because LAr data is BIG!!!!!**
- **So average memory and CPU for DUNE are looking at 2 very different populations.**
- We now phrase our request as Model Wall time x 2 slots + MARS

Making use of external resources

Goal is 25% FNAL 75% Collab
Except for raw tape which is 50% FNAL

- International Pledges as of January 2022 CPU is not memory slot corrected.

DUNE Pledges as of Jan 2022		2021			2022		
		CPU Cores)	Disk (PB)	Tape (PB)	CPU Cores)	Disk (PB) [2]	Tape (PB)
Need		6594	20.4	24.2	**7780**	**27.3**	**33.4**
FNAL		3310	2.2	24.2	**1945**	**7.6**	**21.8**
CERN		3310	2.2	24.2	**950**	**3**	**10**
BNL	BNL	100	0.5		100	0.5	
USA - other	(OSG opportunistic)	1150			1150	0	0
UK	IRIS (UK)	1000	4	3	1000	4	3.1
FR	CC-IN2P3	310	0.5	2	250	0.5	2
ES	PIC Tier-1	500	0.5		512	0.72	
NL	NL/LHC Tier-1	696	1.9		788	1.8	
CZ	CZ-Prague-T2	1560	0.3		2400	1	
IT							
CH		200	0.2		200	0.2	
BR	CBPF	100	0				
IN	Tata	450	0.75		450	0.75	
RU	JINR				1000	0.5	
Total pledge		**12686**	**13.05**	**53.4**	**10745**	**20.57**	**36.9**
Shortfall		**6092.00**	**(7.35)**	**29.20**	**2965.00**	**(6.73)**	**3.50**

Model based requests – updates May 2022

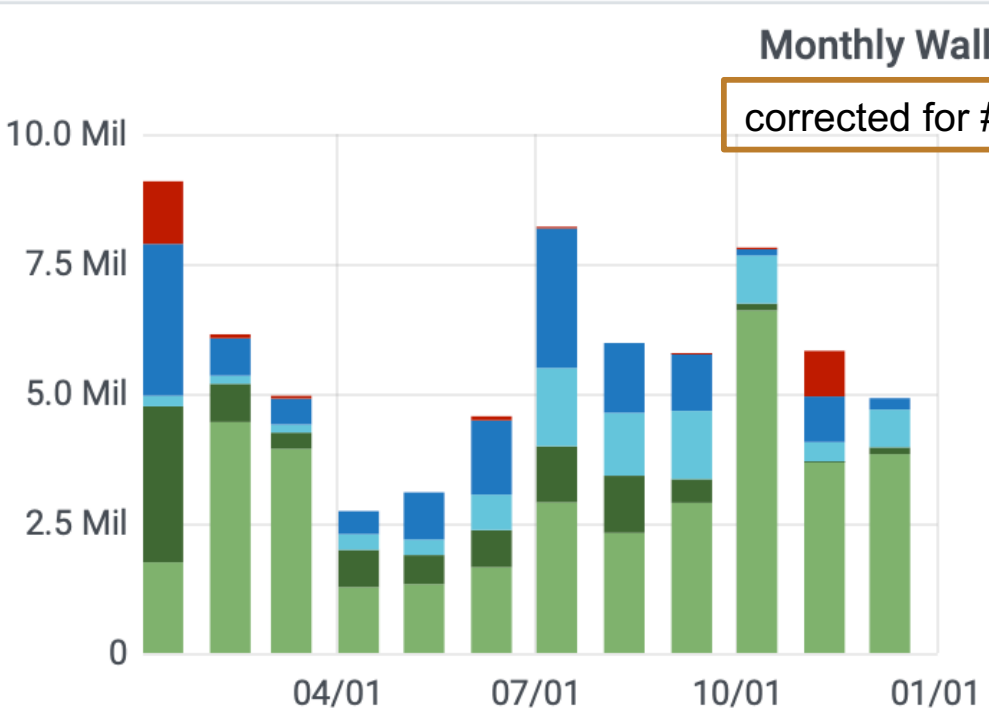
Change from Contributions board numbers is increased data size from PD
Wall hours are not corrected to slot Hrs in this table as many sites have > 2 GB/core

F means Fermilab, C means CERN, Collab means Collaboration

Years	CPU (Mhrs)	Wall (Wall-Mhrs)	Wall F+MARS/Collab (Wall-Mhrs)	cores	Tape Total(PB)	Tape F/C/Collab	Disk Total(PB)	Disk F/C/Collab
2021	40	58	(14+ 0)/ 43	6594	21.1	14.1/ 3.6/ 3.5	20.4	5.3/ 0.4/ 14.7
2022	45	65	(16+ 0)/ 49	7399	35.1	22.6/ 7.4/ 5.0	28.0	8.0/ 2.0/ 17.9
2023	72	103	(26+ 0)/ 77	11747	58.9	36.9/ 14.6/ 7.4	39.3	11.8/ 3.9/ 23.6
2024	78	111	(28+ 0)/ 84	12710	76.2	48.0/ 18.2/ 9.9	43.3	11.9/ 2.1/ 29.3
2025	76	109	(27+ 0)/ 82	12438	86.1	55.5/ 18.2/ 12.4	40.5	10.2/ 0.2/ 30.1

- FNAL request for 2022 – 16 MHrs x2 mem + 17 Mhrs of Mars = *49 2GB Slot-hrs.*
  Increase disk to 8.0 PB in 2022, 11.8 in 2023

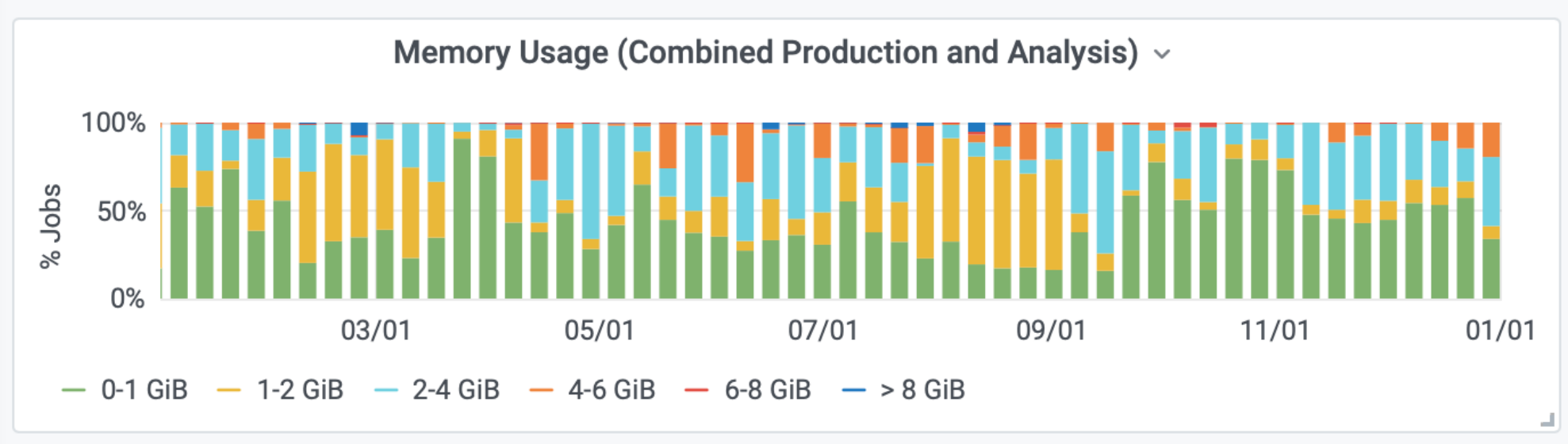
CPU - Experiment Usage Calendar 2021



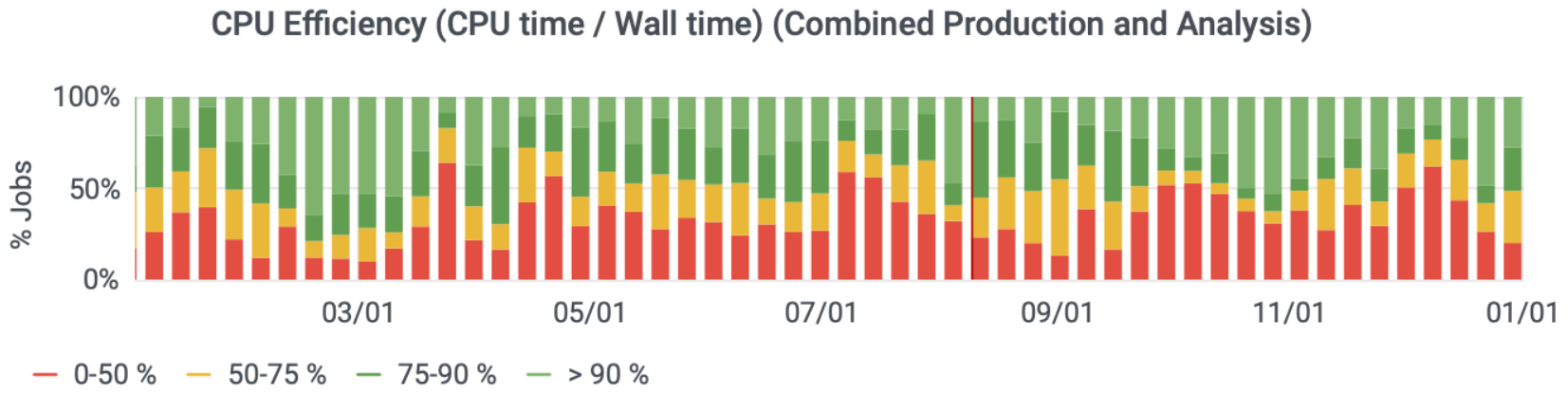
	avg	total
FermiGrid Analysis	3.05 Mil	36.59 Mil
FermiGrid Production	747 K	8.96 Mil
OSG Analysis	655 K	7.87 Mil
OSG Production	1.11 Mil	13.30 Mil
NERSC Analysis	0	0
NERSC Production	197 K	2.36 Mil

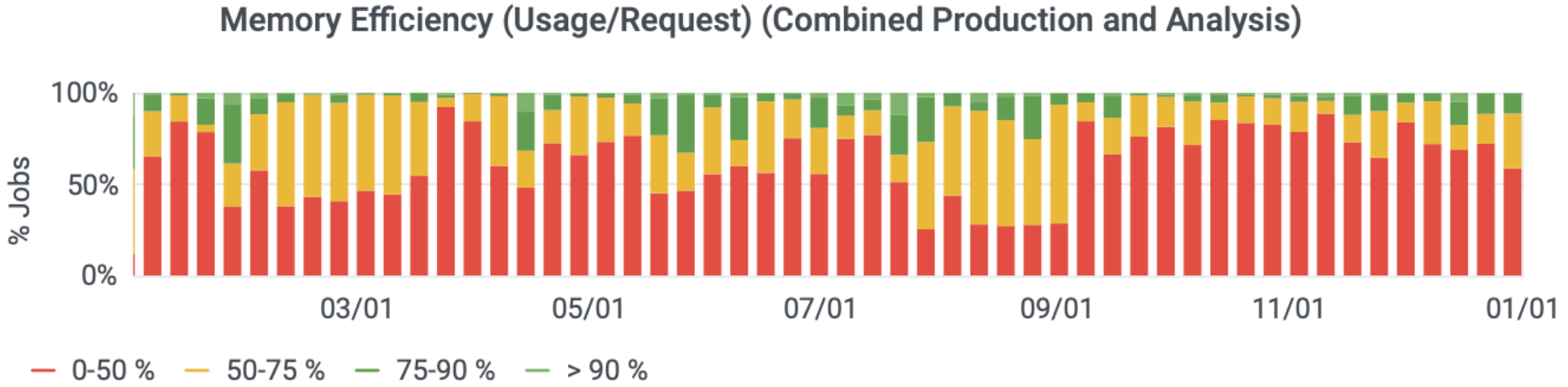
Includes 17 MHrs of MARS

Memory Footprint (Combined)

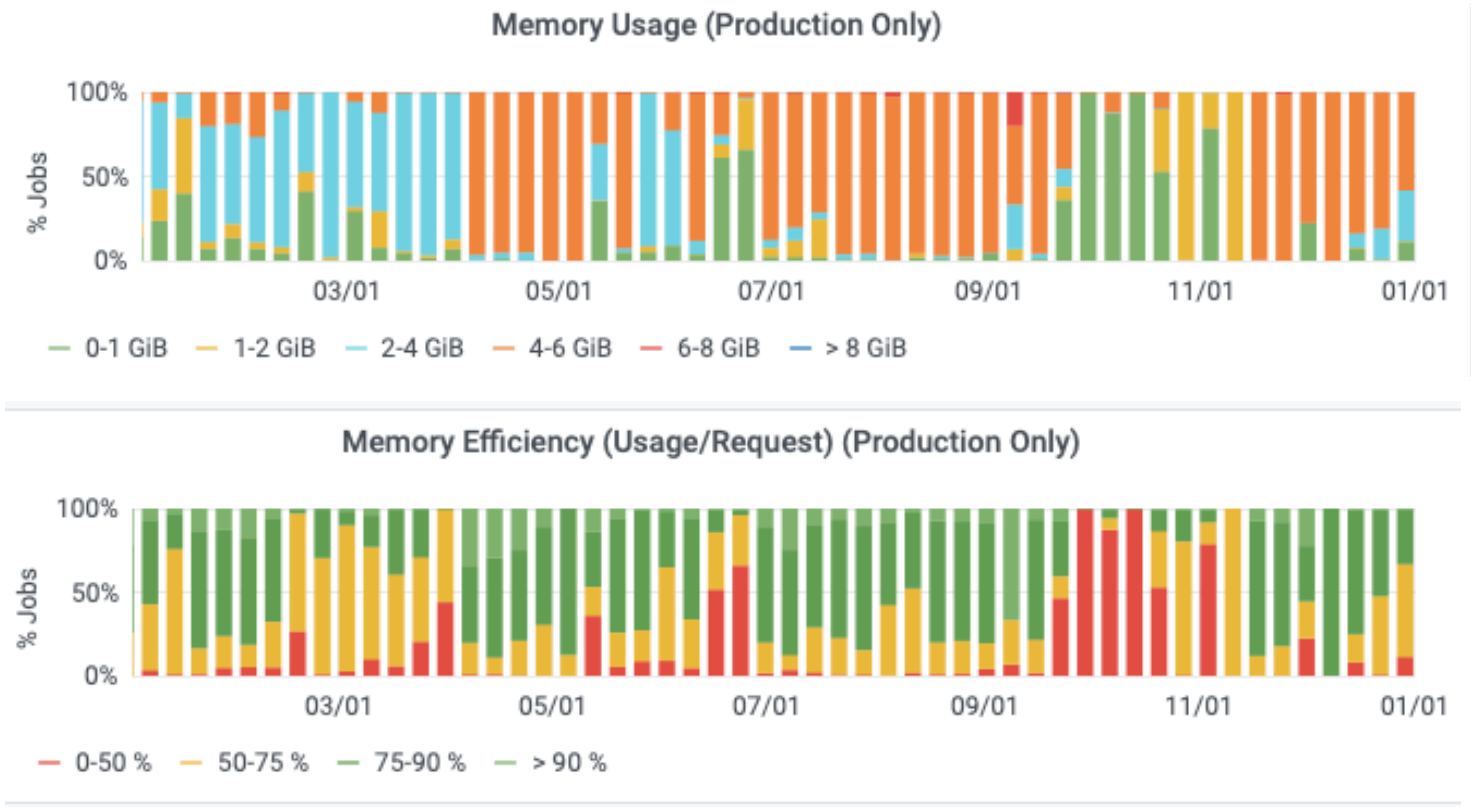


CPU and Memory Efficiency Calendar 2021 (Combined)

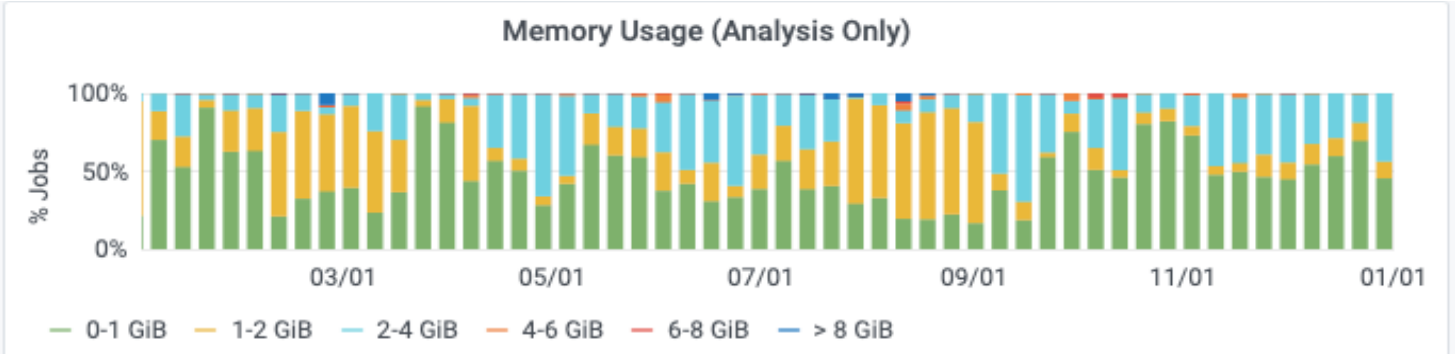


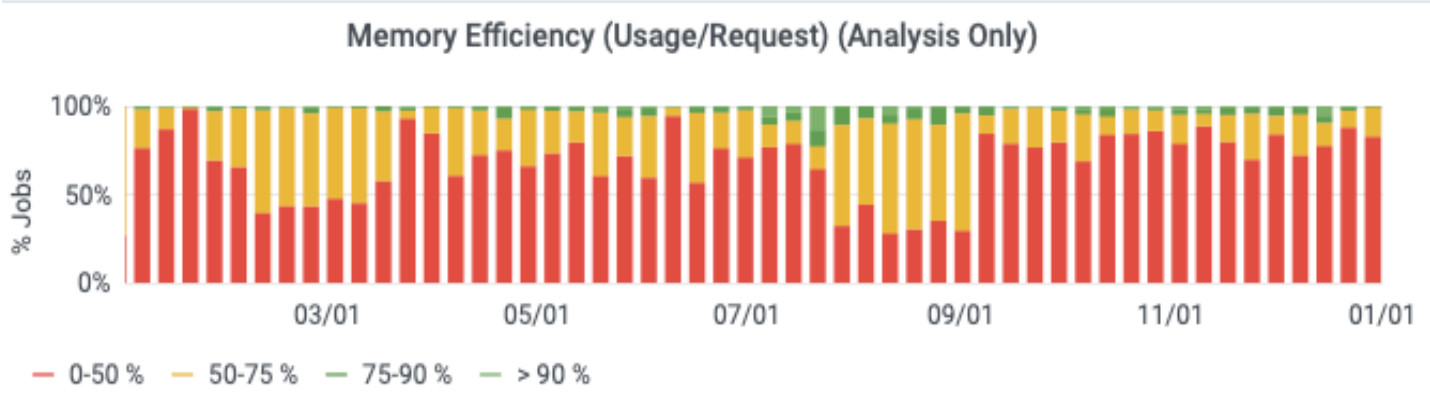


Memory: Production Only, Calendar 2021



Memory, Analysis Only, Calendar 2021





Most of the
Green is MARS

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

	2020	2021	2022	2023	2024
Requested	25 (FNAL) (36 Total)	29 FNAL	2*16 +17 FNAL =49 MSlotHrs +49*X offsite	2*26 +17 FNAL =69 MSlotHrs +77*X offsite	2*28 +17 FNAL = 73 MSlotHrs +84*X offsite
Actual Used	29 (GPGrid) 42(WLCG+GPGrid) 3.85 (NERSC)	36 FNAL 20 OSG+WLCG 2 NERSC	14.1 FNAL YTD	N/A	
Efficiency	%	%80	N/A	A	

Offsite often has more memory/slot so hard to compare

Now includes MARS in request. Assume 2 slots/process except for MARS

Aim is 25% FNAL, 75% offsite. MARS and 2 GB limit make this hard

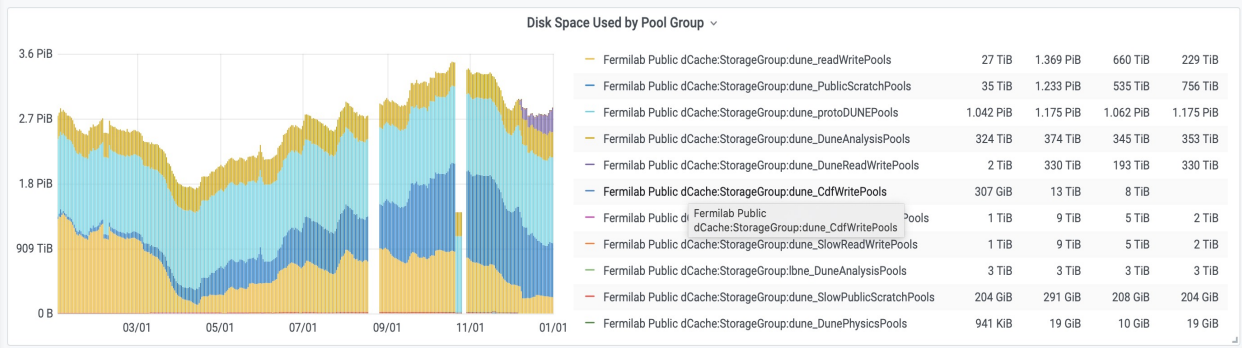
CPU Adaptations Going Forward

How can experiment use OSG/HPC/Cloud/HEPCloud going forward? **Also WLCG!**

- **DUNE needs ~4 GB/process for reco and many analysis jobs.**
- > 50% and sometimes 75% of computing is done offsite.
- We have cached ~4 PB of reconstructed/simulated data in Europe to make better use of European compute resources. This increased efficiency in Europe substantially and relieved pressure on FNAL tape drives.
- DUNE has used NERSC mostly for MC simulation through joint FIFE allocation via HEPCloud. Will continue.
- Plenty of GPU use cases, have some GPU allocation on Perlmutter this year
- Have used GPUs in the cloud for inference as a service testing in collaboration with MIT

Disk: dCache Usage and Predictions (in TB)

Persistent + Other = Pledge



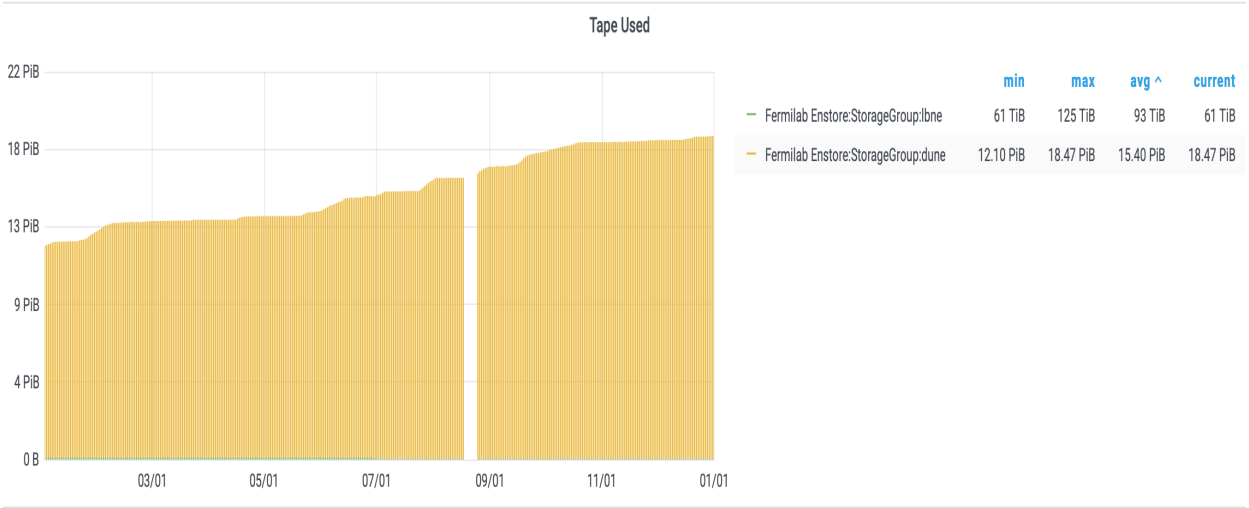
Numbers are from the model

Current:
Total r/w (tape backed): 5100 TB
Total scratch: 700 TB
Total persistent: 900TB

Will not track cache usage, but need to know of unusual requests

	Analysis **(Persistent)**	**Dedicated (Write)**	**Total**
Current	900 TB (actual)	5.1 PB (actual)	6.0PB
2022	900 TB	6.7 PB	7.6 PB
2023	2000 TB	9.8 PB	11.8PB
2024	2000 TB	9.9 PB	11.9PB

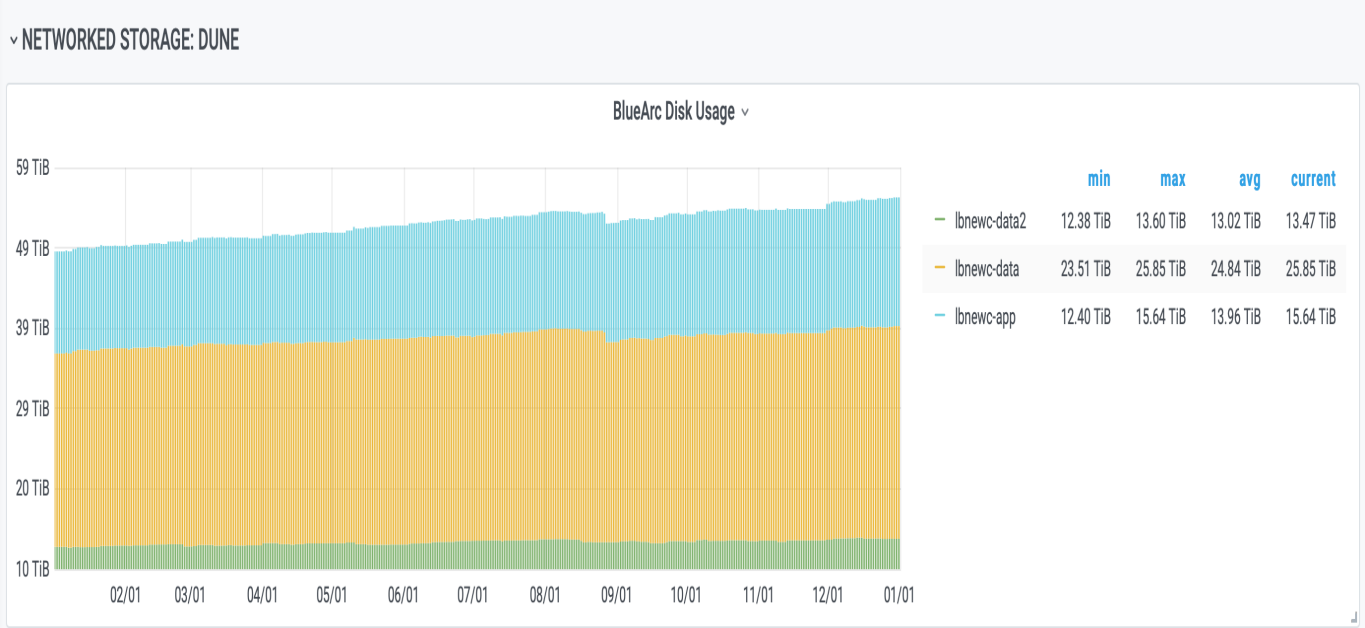
Tape - Usage and Predictions (in PB)



	Total Added By End of Year
At end 2021	18.5 PB (actual)
2022	22.6 PB
2023	36.9 PB
2024	48 PB

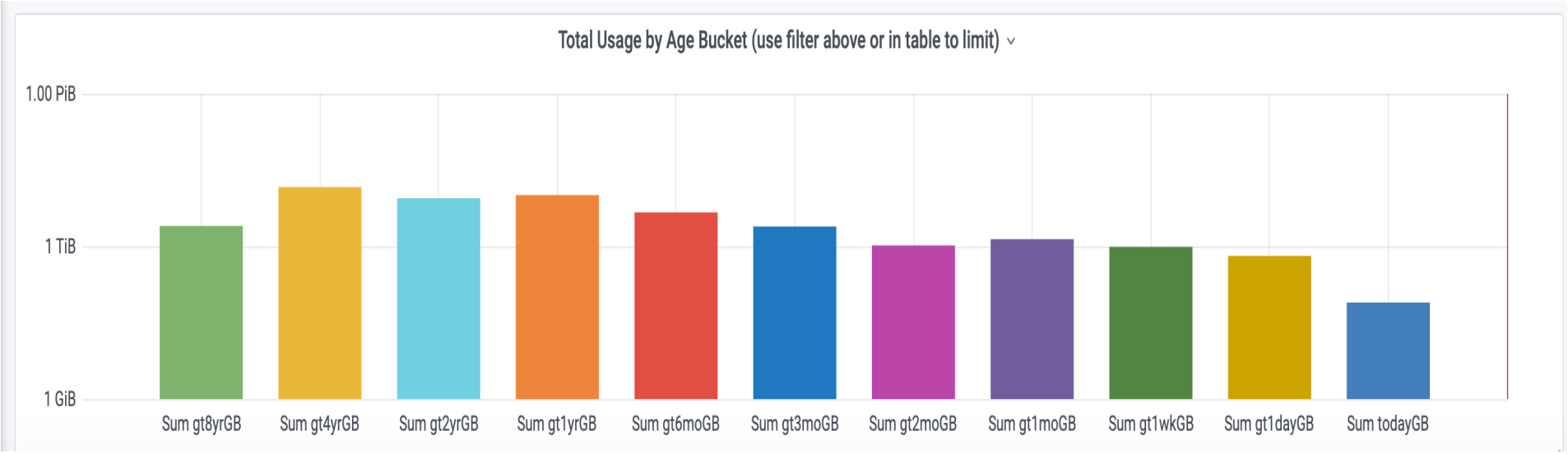
Numbers are from the model. ProtoDUNE will write a lot of data in 2023

Disk: NAS Usage and Predictions (in TB Units)



	App	Data
2022	16.3TB	60.6TB
2023	20.0TB	62.0TB
2024	20.0TB	62.0TB

Age of files in NAS



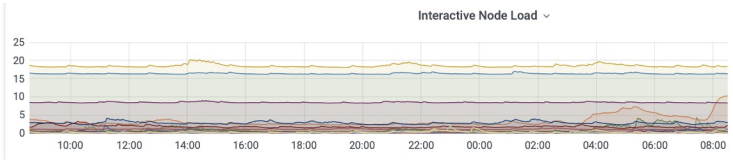
Data Lifetimes

- All new data added to DUNE data management will have a retention class and a retention lifetime from the beginning. Existing data retention strategy is based on data_tier and data_stream(type)
- The model includes tape retention times for raw (100 yr), reconstructed (15 yr) and test data (0.5 yr)
- Negotiating with current ProtoDUNE physicists to put short lifetimes on commissioning data (noise studies, pedestal studies, cold box data, etc.)
- Hope to move major user samples to cataloged rucio controlled areas
- There is a significant amount (~2PB) of old reconstructed data output that is scheduled to be purged off tape some point this year.

What Do You Want to Achieve in Computing Over Next Three Years

Goals	Where does the experiment need to contribute	Where does SCD need to contribute
Transition to new data and workflow management systems in production	Requirements and effort	Rucio quality of service, data transfer, monitoring, extensions to POMS and possibly Jobsub
Framework capable of handling very large data objects	Requirements and effort	High level expertise
Complete analysis and simulation of ProtoDUNE	Algorithms and code management, database design	Database support, networking, storage and personnel in DUNE leadership
Continue to support TDR efforts for near detector and vertical drift	Joint effort	Joint effort
Transition to SPACK packaging	Knowledge of software stack	SCD personnel expertise

Analysis facilities for > 1000 users



Currently using 17 GPVM's for

- LArSoft algorithm development
- end stage root analysis of user generated samples
- command-line grid submission
- last 6 months, 50 active users submitting grid jobs

Need GPVM's for

- High memory testing of analysis – need more memory!!!
- Debugging
- Event display
- Db access
- Xrootd access to data

Current Analysis Facility

- Still in development - early test of Elastic Analysis Facility are promising
- Need enough fast disk for user samples

Future Analysis Facility

- Want to build on the successful LPC Analysis Facility model within the FNAL Neutrino Physics Center
- Important to have specific effort to support users and provide tutorials
- large, fast disk with user quotas (not 100 PB over xrootd)
- potentially combination of both HTC queue and dynamic, jupyter-based cluster (COFFEA style, Elastic AF)
- GPU/accelerator availability

Anything else?

Strong international effort to pool resources.

Production and now analysis are largely running offsite

Need lots of disk and >= 4GB/core!