DUNE Computing Challenges

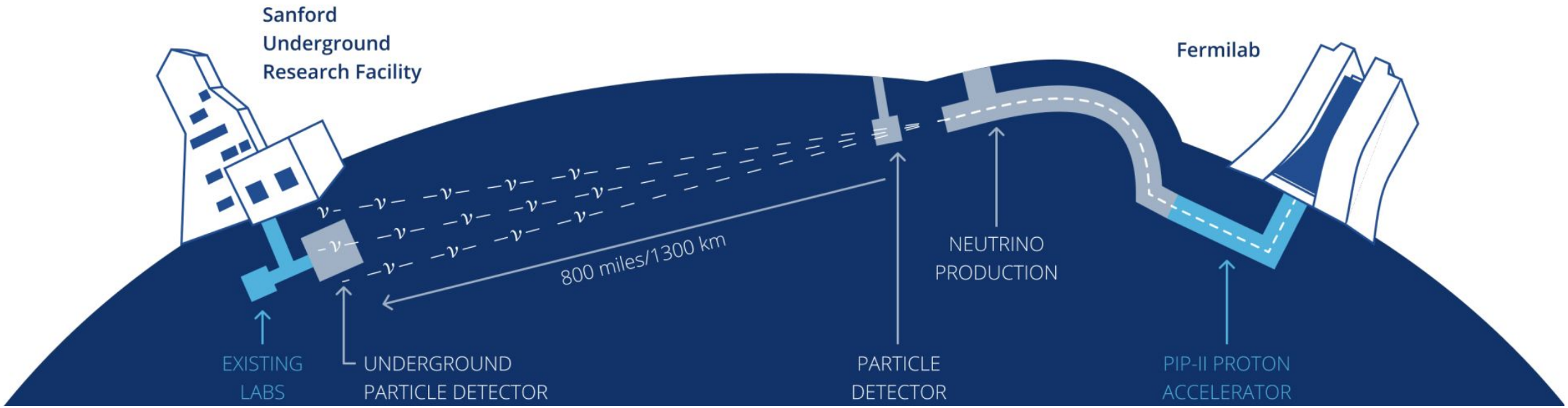
Meghna Bhattacharya, Fermilab
for the DUNE Collaboration

6th Global Research Platform Workshop

September 15, Chicago

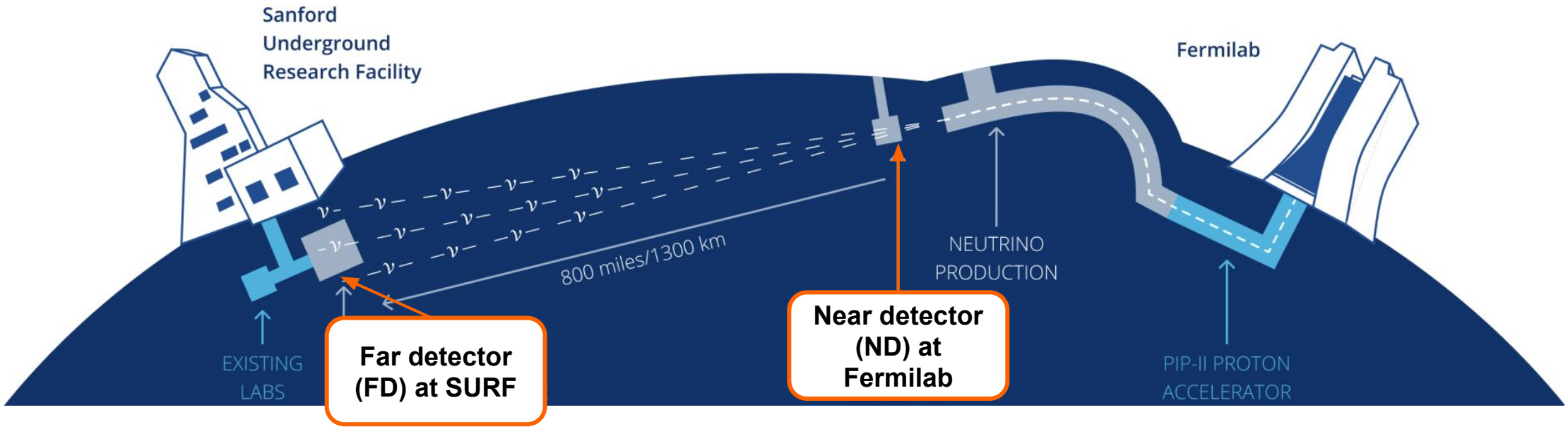
DUNE: Deep Underground Neutrino Experiment

- **Next-generation international neutrino experiment hosted in the US**



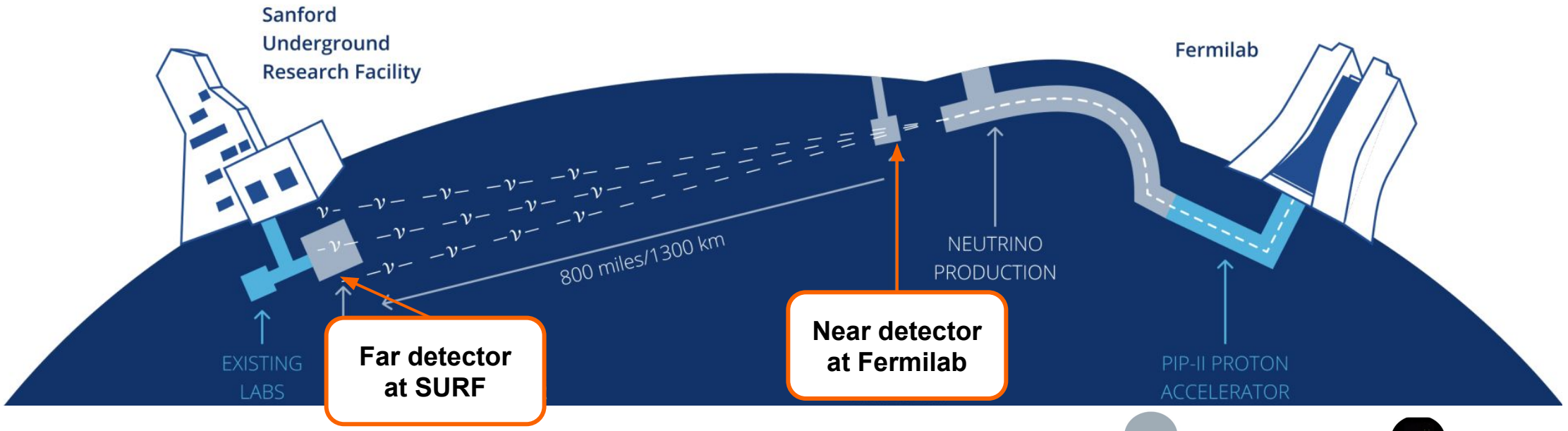
DUNE: Deep Underground Neutrino Experiment

- **Next-generation international neutrino experiment hosted in the US**

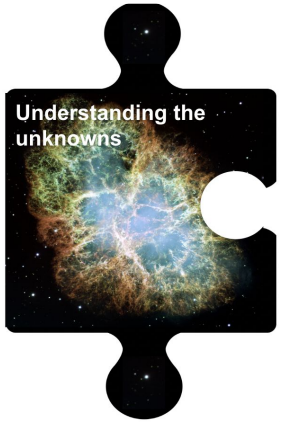


DUNE: Deep Underground Neutrino Experiment

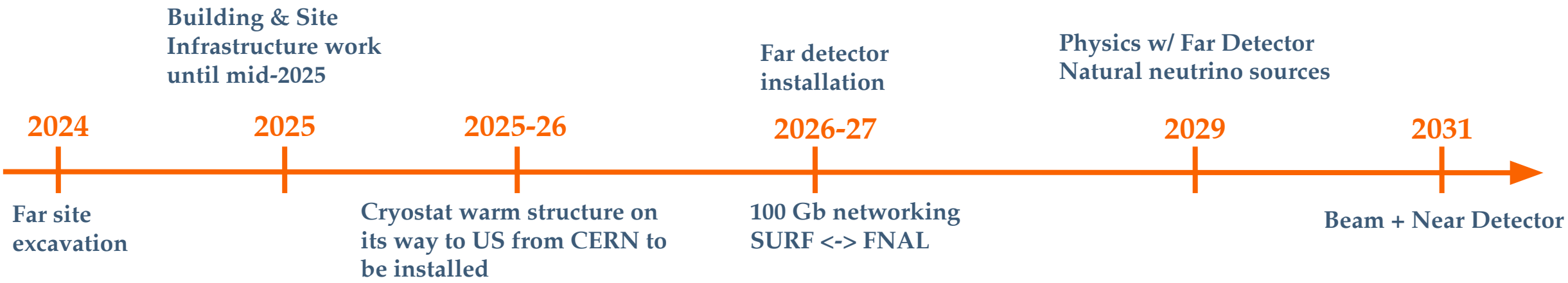
- **Next-generation international neutrino experiment hosted in the US**



- **A multi-purpose experiment designed to probe nature at different levels**



DUNE Timeline



DUNE Timeline

2024
Far site excavation

2025
Building & Site Infrastructure work until mid-2025

2025-26
Cryostat warm structure on its way to US from CERN to be installed

2026-27
Far detector installation

100 Gb networking SURF <-> FNAL

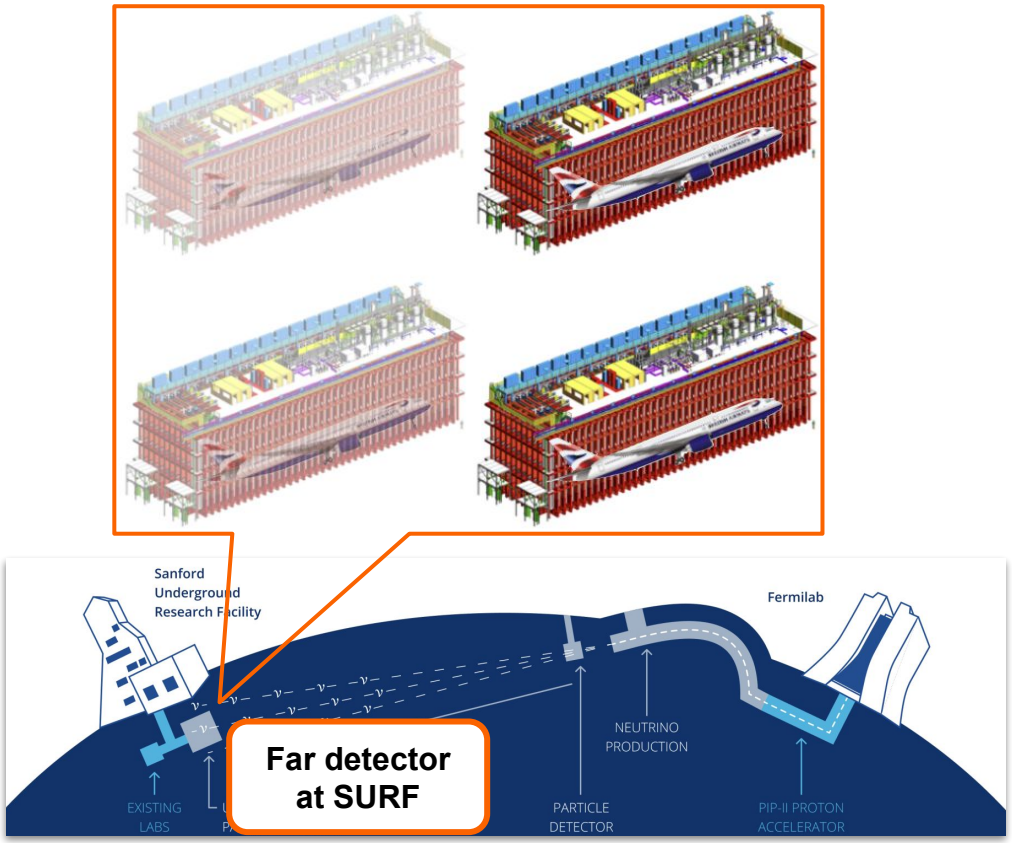
2029
Physics w/ Far Detector
Natural neutrino sources

2031
Beam + Near Detector

2024: Far site Excavation Complete

DUNE Computing Challenges: Far Detector

~1.5 km underground in South Dakota

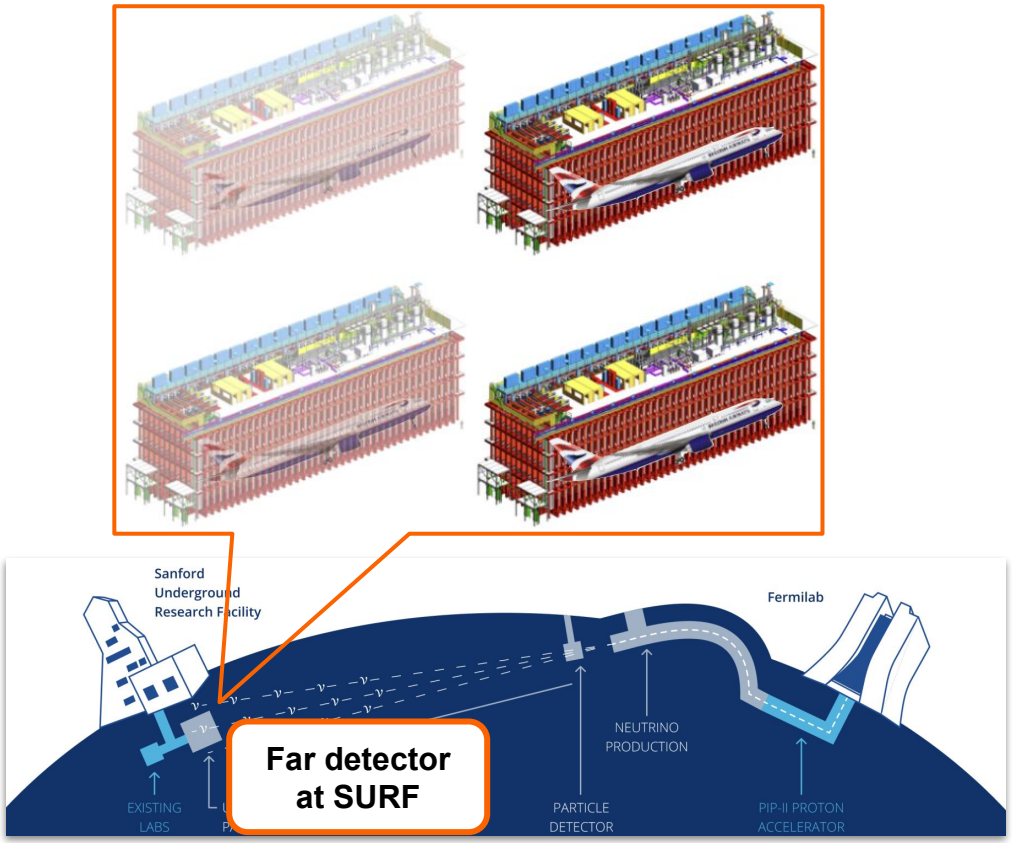


Far Detector designed to ensure sensitivity for extensive physics program

- Beam events
- Cosmic rays
- Calibrations
- Solar neutrino
- Supernova

DUNE Computing Challenges: Far Detector

~1.5 km underground in South Dakota

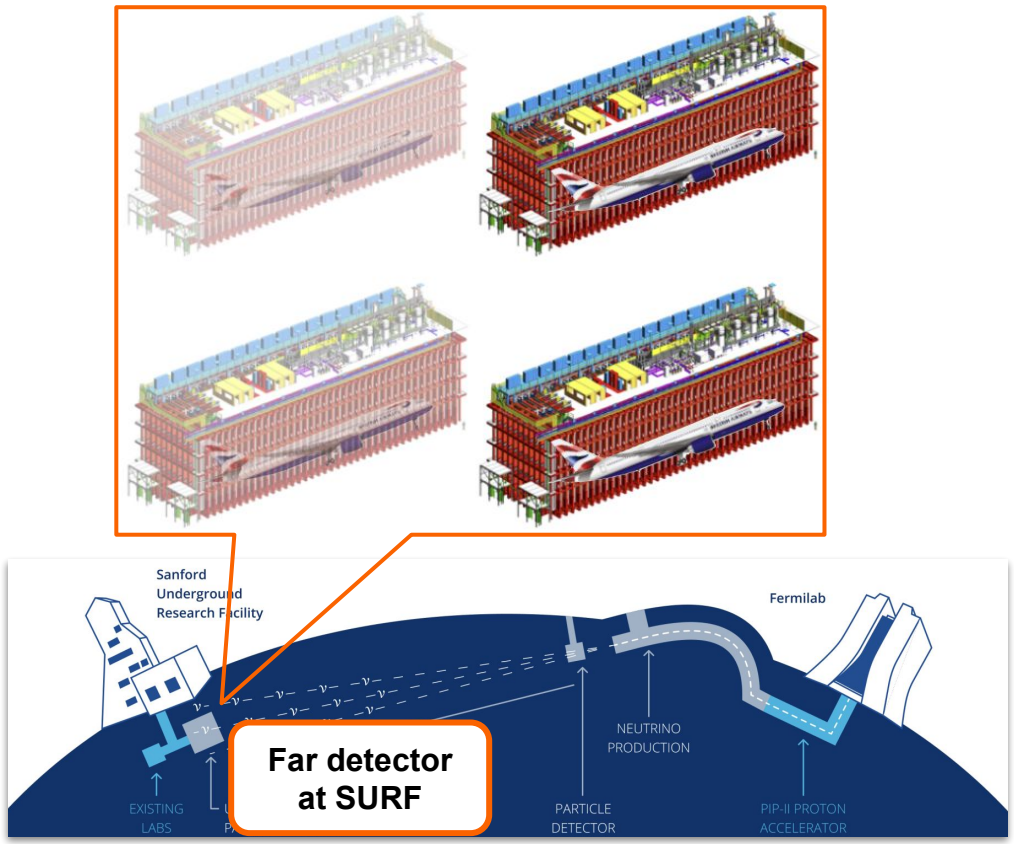


Far Detector designed to ensure sensitivity for extensive physics program

	Rate/module	size/instance	size/module/year
Beam events	41/day	3.8 GB	30 TB
Cosmic rays	4500/day	3.8 GB	6.2 PB
Calibrations	2/year	750 TB	1.5 PB
Solar neutrino	10,000/year	3.8 GB	3.8 GB
Supernova	1/month	140 TB	1.7 PB
Total			9.4 PB

DUNE Computing Challenges: Far Detector

~1.5 km underground in South Dakota



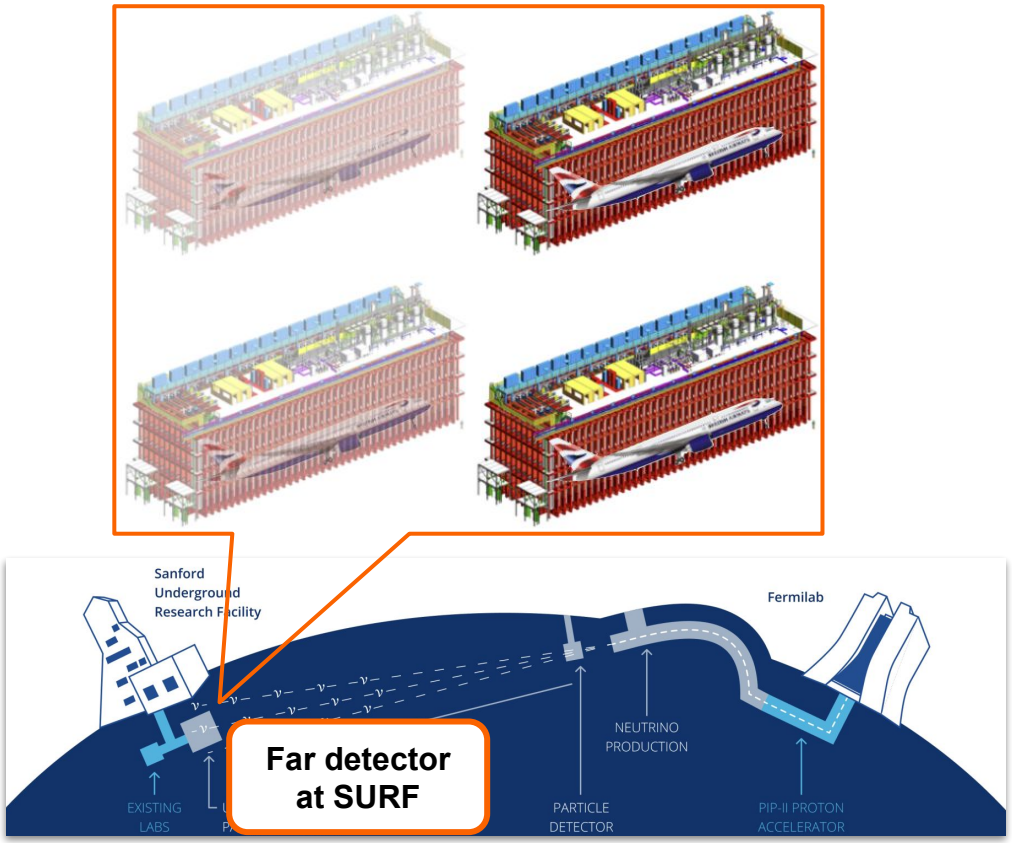
Far Detector designed to ensure sensitivity for extensive physics program

	Rate/module	size/instance	size/module/year
Beam events	41/day	3.8 GB	30 TB
Cosmic rays	4500/day	3.8 GB	6.2 PB
Calibrations	2/year	750 TB	1.5 PB
Solar neutrino	10,000/year	3.8 GB	3.8 GB
Supernova	1/month	140 TB	1.7 PB
Total			9.4 PB

- 1 FD module produces ~4 GB/drift window
  - Interactions - ns
  - TPC sampling rate - μs
  - drift time - ms
- beam coincidence readouts large compared with information density, relatively rare

DUNE Computing Challenges: Far Detector

~1.5 km underground in South Dakota

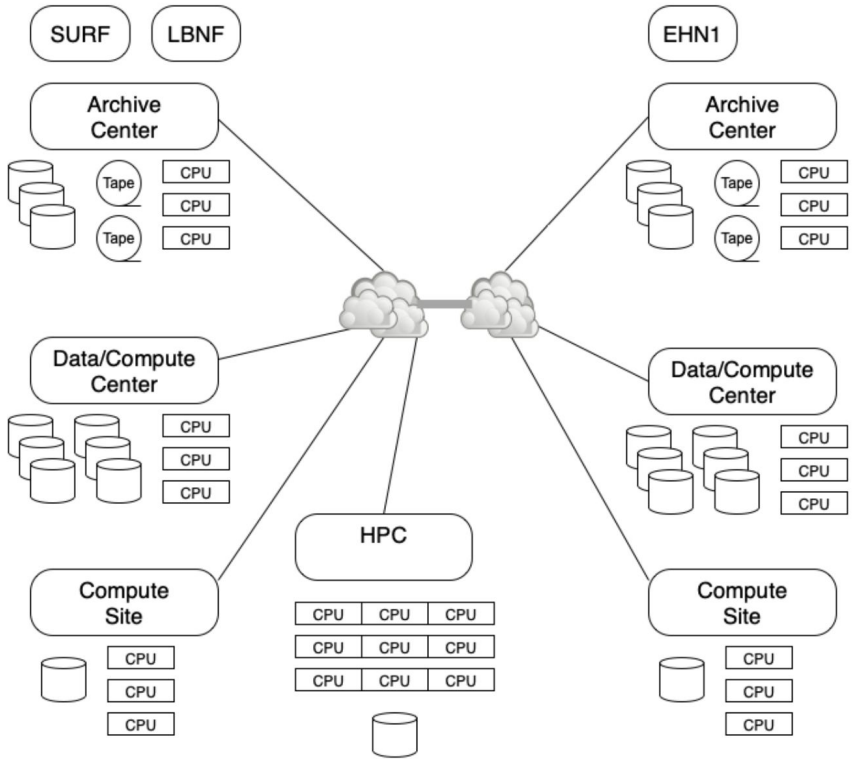


Far Detector designed to ensure sensitivity for extensive physics program

	Rate/module	size/instance	size/module/year
Beam events	41/day	3.8 GB	30 TB
Cosmic rays	4500/day	3.8 GB	6.2 PB
Calibrations	2/year	750 TB	1.5 PB
Solar neutrino	10,000/year	3.8 GB	3.8 GB
Supernova	1/month	140 TB	1.7 PB
Total			9.4 PB

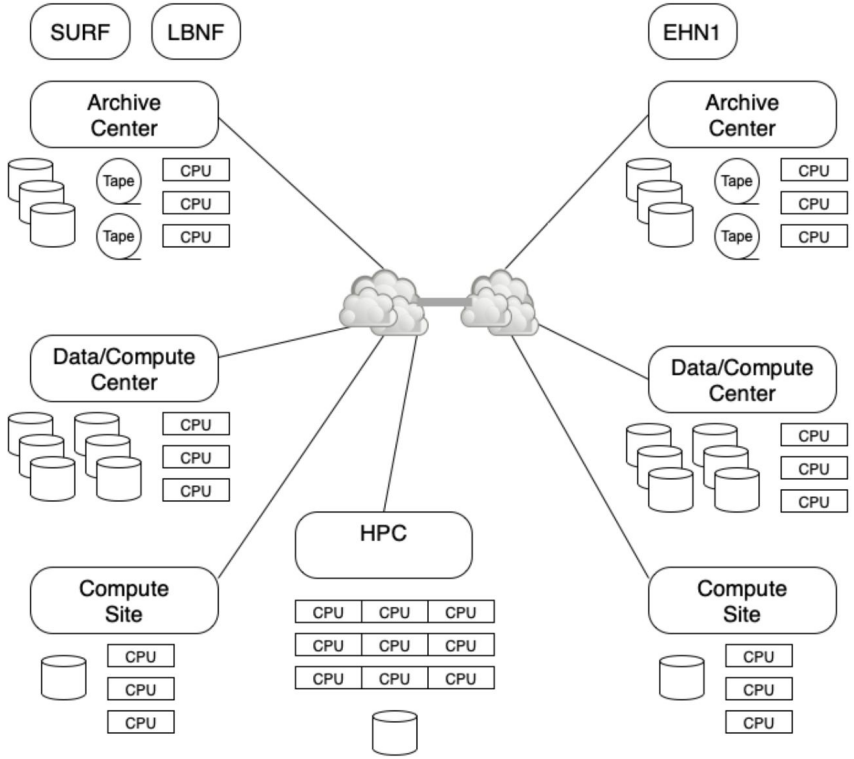
DUNE requirement - less than 30 PB/year total to archival storage from all active FDs

Distributed Computing Tools



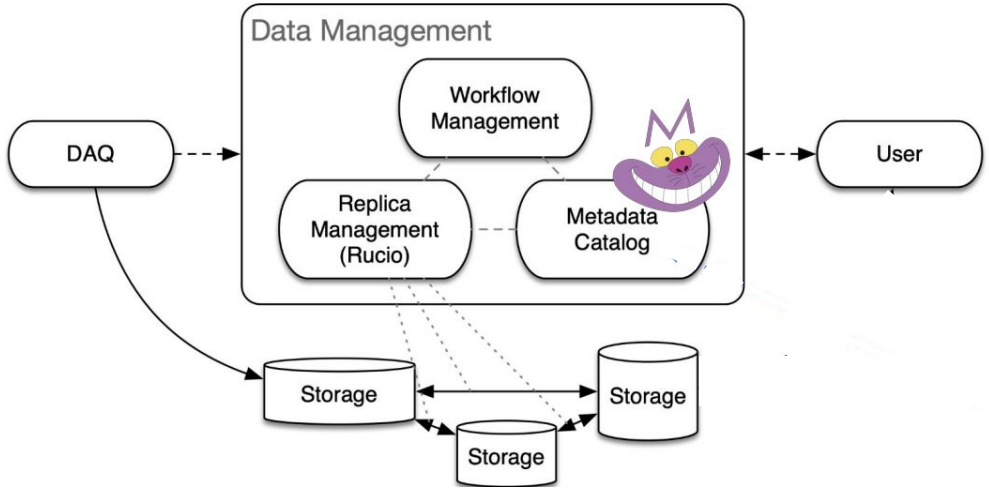
- High Throughput Computing, similar to LHC experiments
  - More service-based than a tiered model
  - Sites provide one of more services (standard compute, HPC, storage/compute, archiving, user analysis, etc.)
- Use variety of sites; mix of dedicated resources and opportunistic access through OSG
- Stream input data over network with XRootD

Distributed Computing Tools



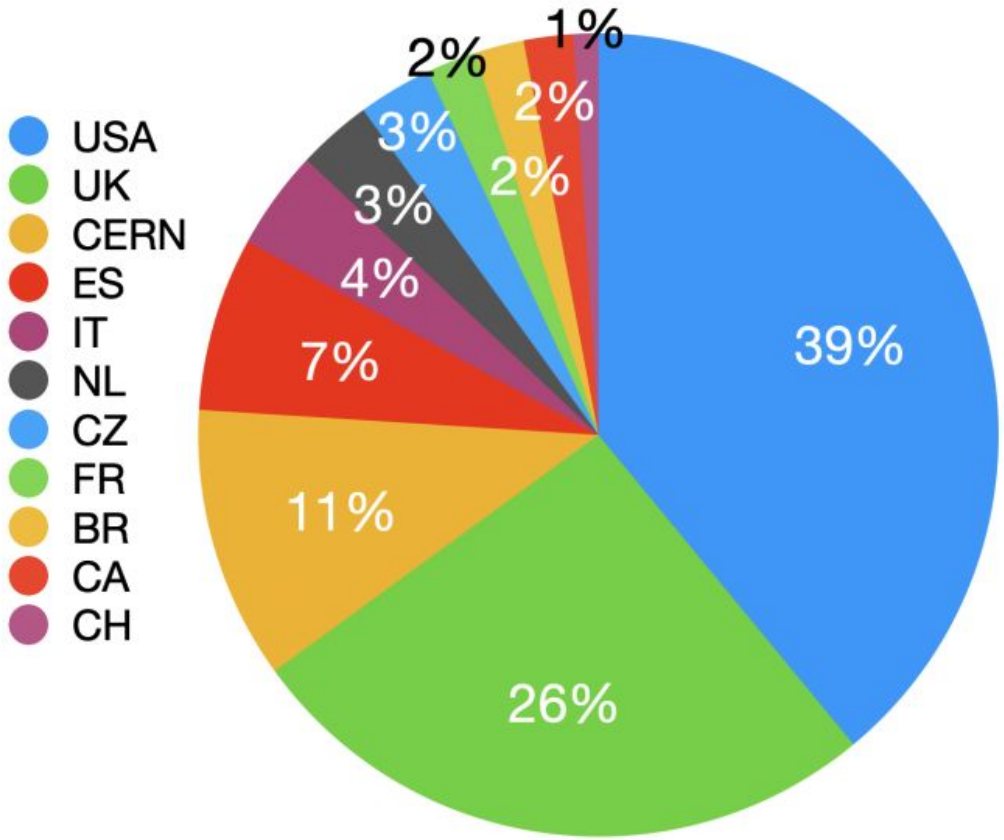
- Data management utilizes
  - Metacat (Metadata Catalog)
  - Rucio for dataset transfer and replication
  - justIN workflow management system

- High Throughput Computing, similar to LHC experiments
  - More service-based than a tiered model
  - Sites provide one of more services (standard compute, HPC, storage/compute, archiving, user analysis, etc.)
- Use variety of sites; mix of dedicated resources and opportunistic access through OSG
- Stream input data over network with XRootD



DUNE Computing: a global effort

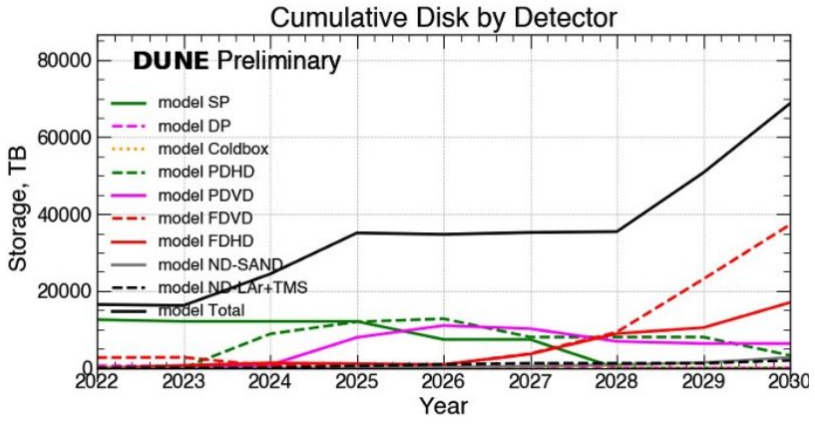
- DUNE production campaign includes reconstruction and simulation for DUNE Far Detectors and ProtoDUNE
- Largest single national contribution from the US, but total European contribution > 57% of CPU processing
- Considerable challenges with large memory consumption and data volume movement
  - FD beam neutrino samples
  - FD solar and supernova simulations
  - ND simulation with overlay
  - Successfully utilizing compute sites and RSEs from around the world



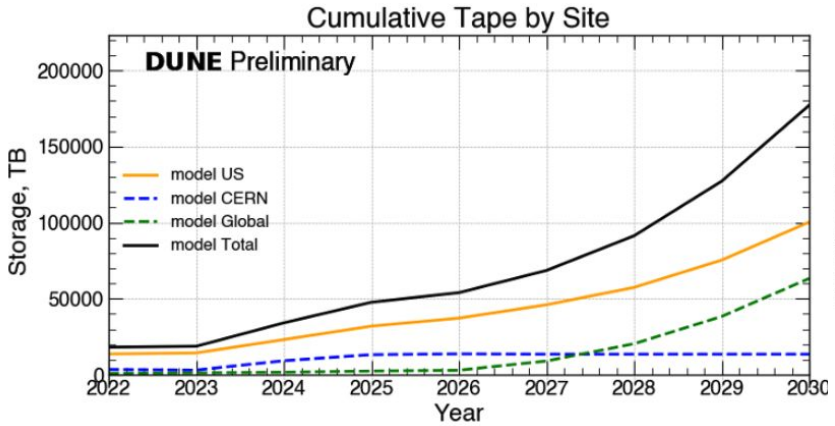
Fraction of CPU resources utilized in each nation for DUNE production processing campaigns during 2024

DUNE Computing: Storage & Requirements

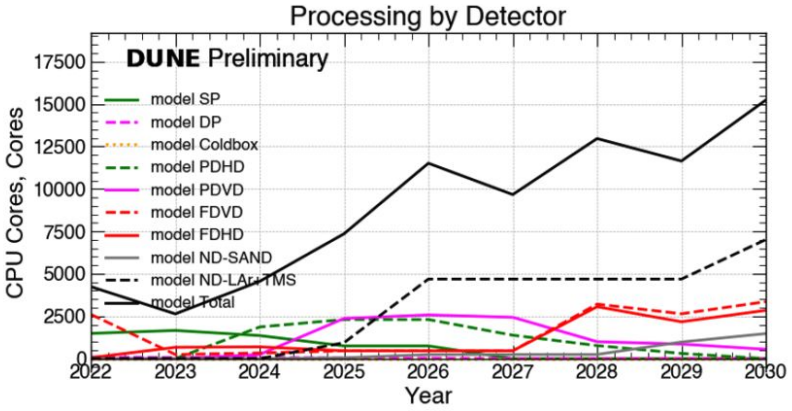
- Combination of disk cache and tape for archiving
- ~10% of LHC experiments in 2030s
- 2 physically separate copies of raw detector data
- Knowledge gained from prototype operations has helped refine the computing model projections



Disk storage needs based upon data volume projections from prototype detectors



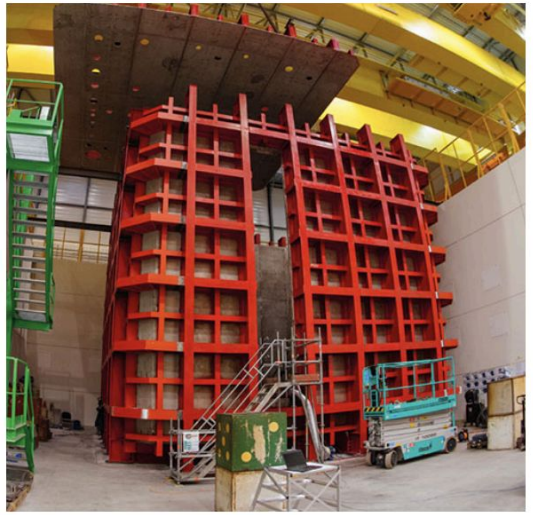
Tape storage projections

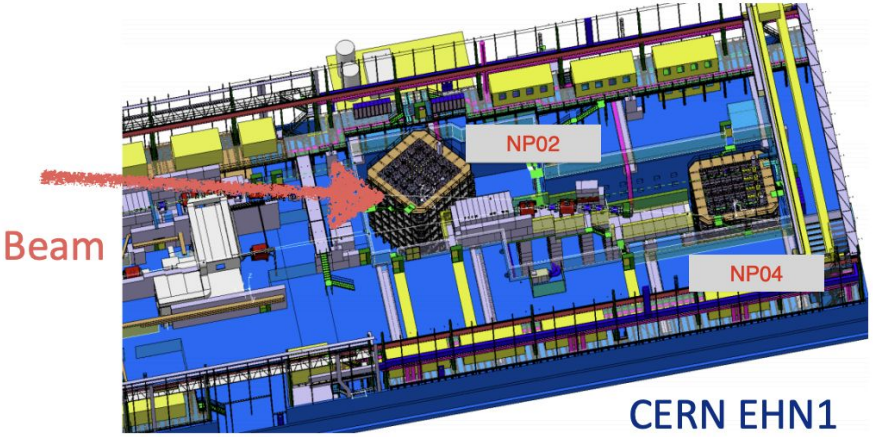


CPU needs projections

ProtoDUNE

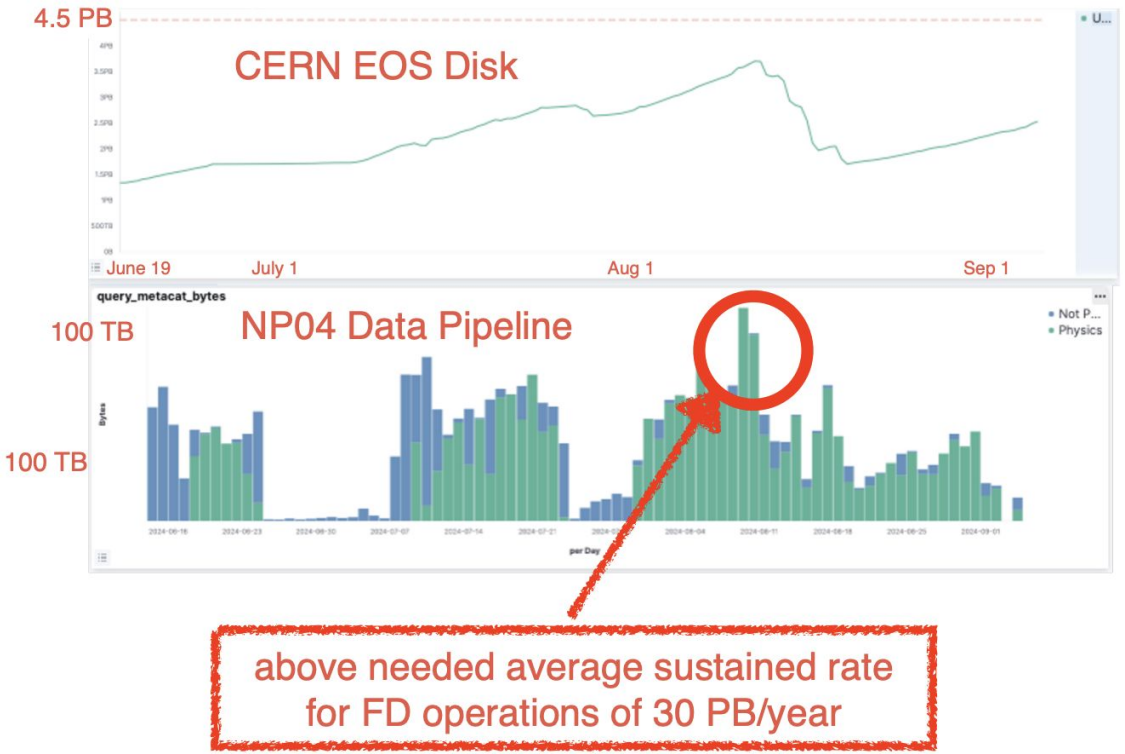
- CERN Neutrino Platform hosting ProtoDUNE II - Horizontal Drift (NP04) and Vertical Drift (NP02)
- Phase I: transfer from CERN to Fermilab repeatedly to mimic DAQ dataflow
  - Total data was 500 TB
- Phase II: Reconstruct dataset, aiming for concurrency levels expected during beam running (a few thousand cores to keep up with DAQ)
  - Use as much new file delivery and metadata catalog infrastructure as possible
- ProtoDUNE-HD (NP04) operated in 2024
- ProtoDUNE -VD (NP02) currently operating
  - finishing 8 weeks of beam program today



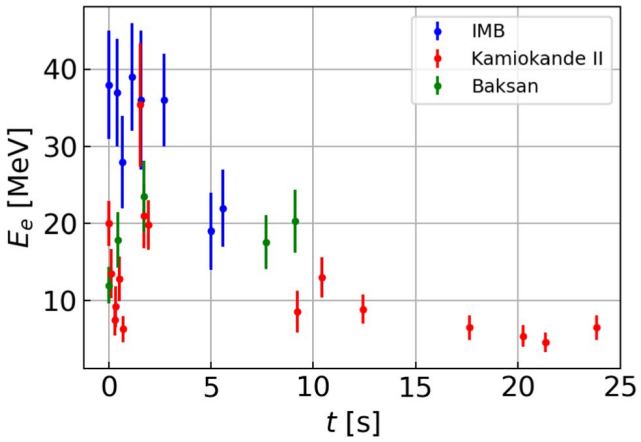


Data Management and Data Pipeline: NP04

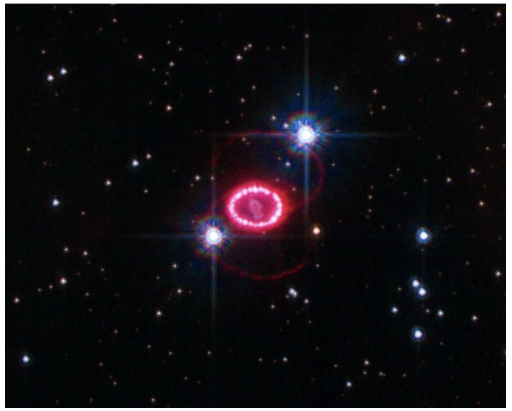
- Pipeline utilizes sequences of tools (FTS3, Metacat, Rucio, and custom ingest/declaration daemons)
- Early estimates anticipated 2 PB of beam data to be written
- Aug 10-11 consecutive days 100+ TB
- Wrote 4.5PB of data
    - distributed 2PB of raw data to other disk sites in one week
    - files moved to BNL, PIC, NIKHEF, FNAL, PRAGUE, RAL-PP, SURFSARA
- Similar current performance for NP02 (PDVD)



First and only Supernova Neutrino Burst so far



25 neutrinos detected in
13 seconds by 3 detectors

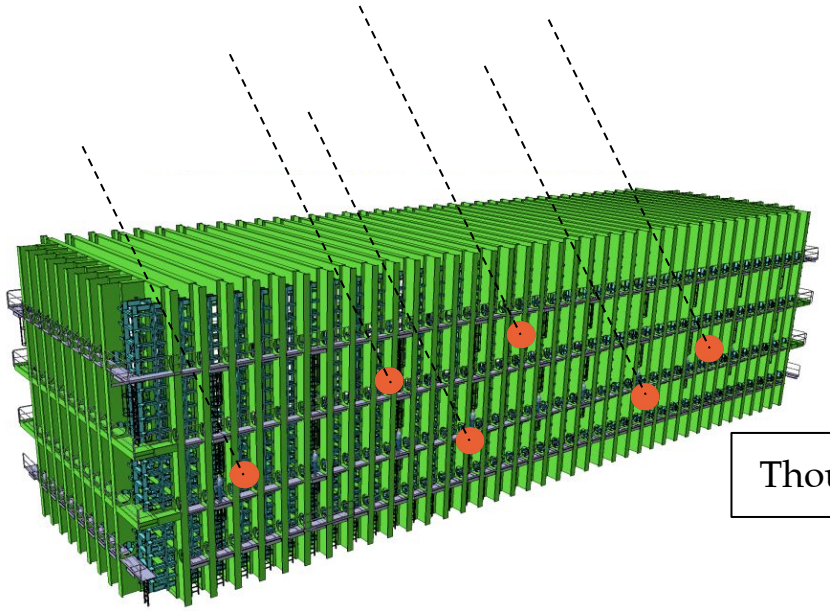


SN1987A remnant

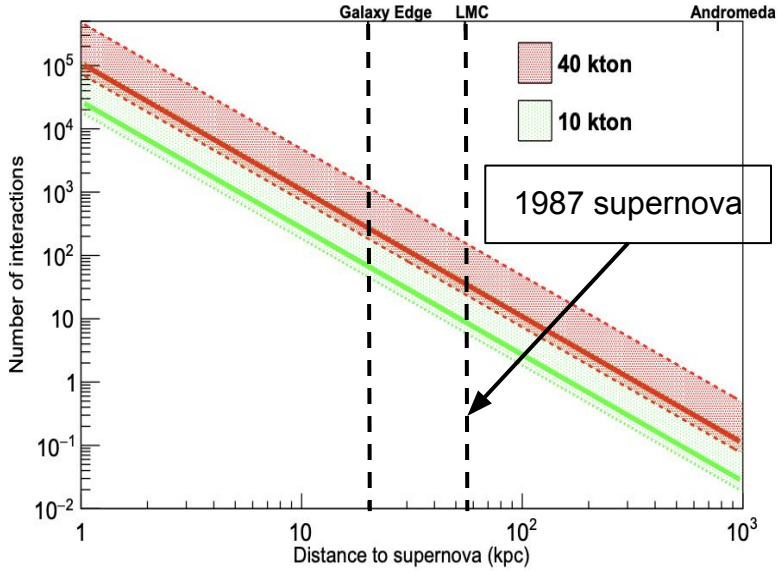
~3 hours delay between arrival of neutrinos and optical light

- Confirmed our basic understanding of core-collapse
  - Rich in physics, ~1 citation per day for last 20 years
- World wide effort to prepare for next one
- Galactic rate of supernovae is ~few / century
- Potentially the most time critical data from DUNE Far Detector

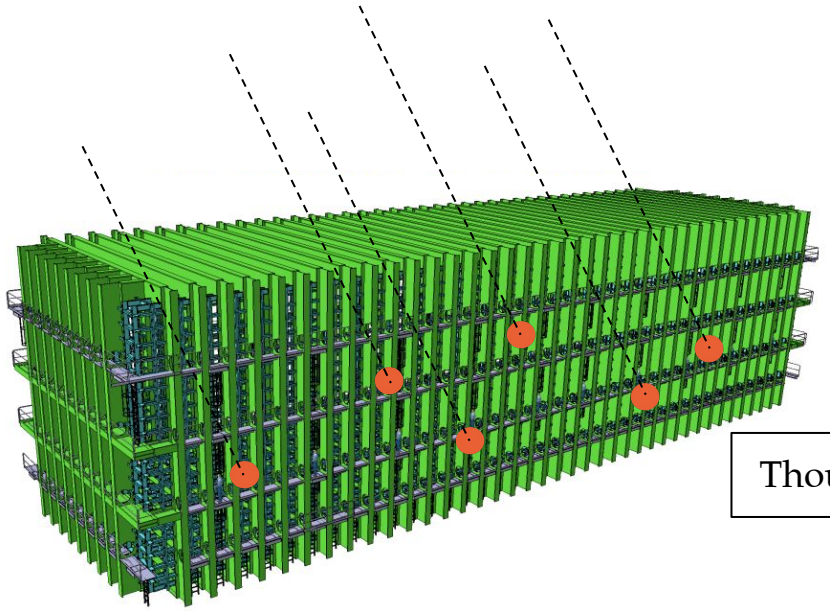
DUNE's Supernova Challenge



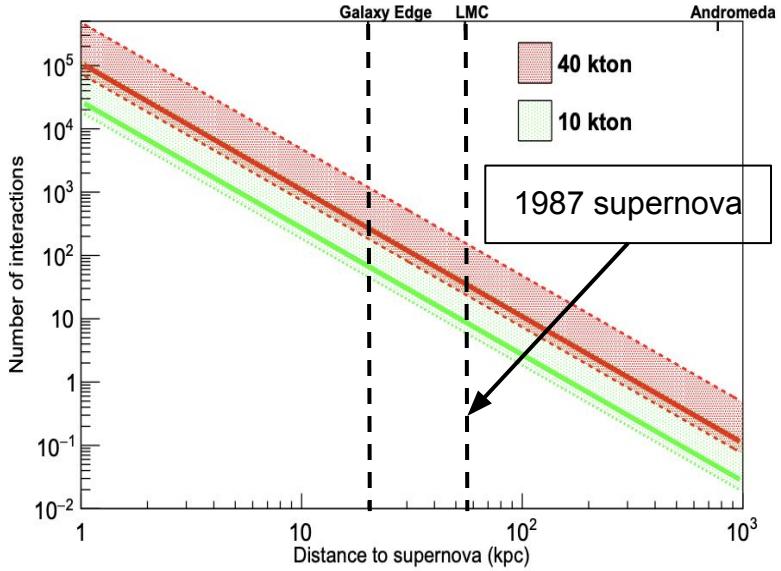
Thousands of neutrino events within seconds



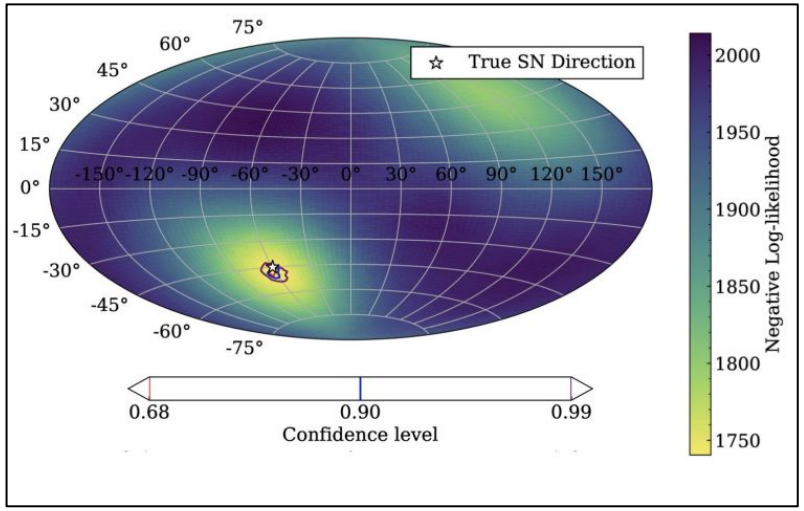
DUNE's Supernova Challenge



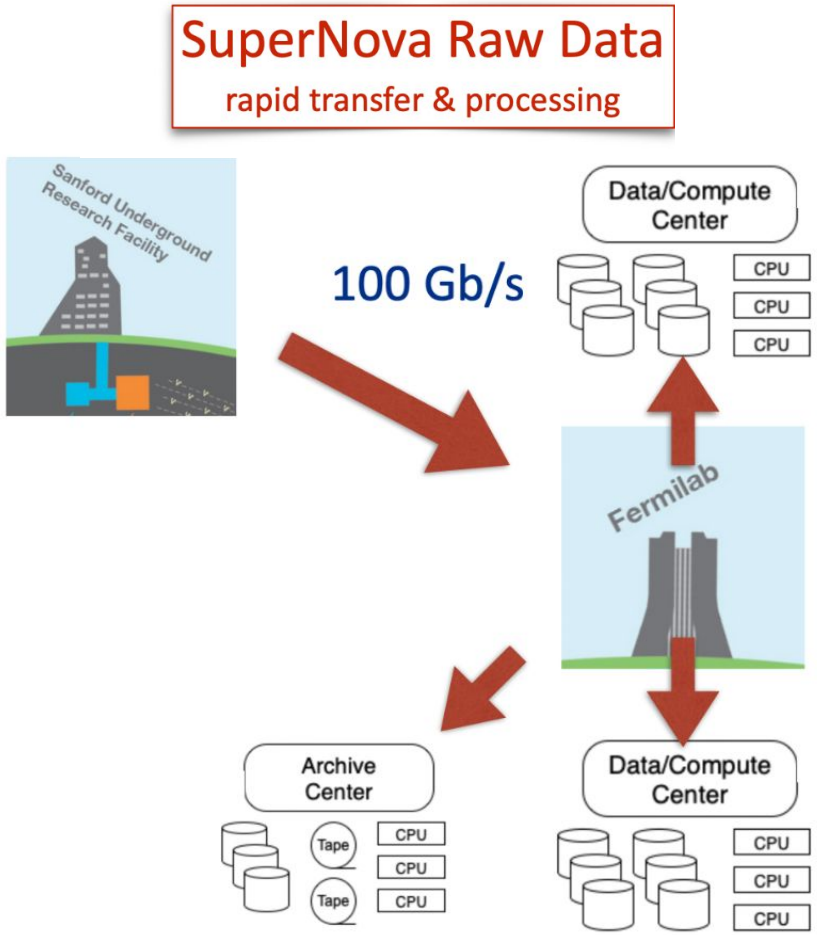
Thousands of neutrino events within seconds



- Detector read out in a continuous mode for ~100 seconds
- Expected data: 140 TB (X 4)/100 seconds
- 1 SN trigger/month (including false alarm)
- The vital role of neutrino detectors → Alert astronomers
  - Requires fast data transfer and data processing

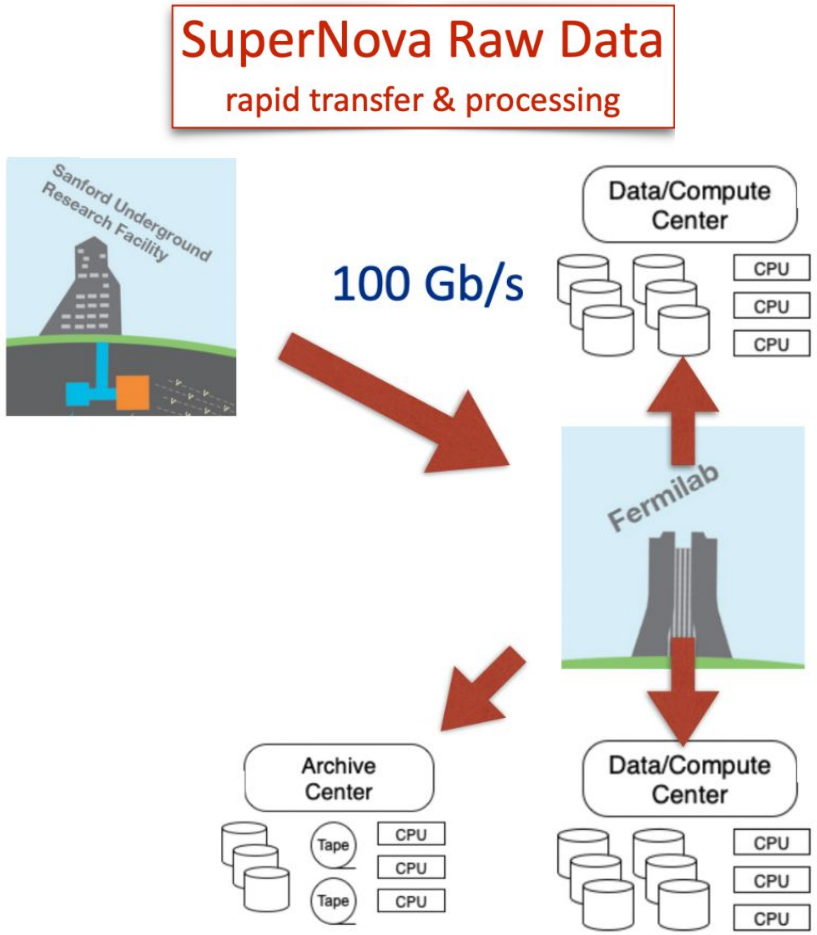


DUNE's Supernova Challenge



not to scale, not a technical design
it's just a cartoon

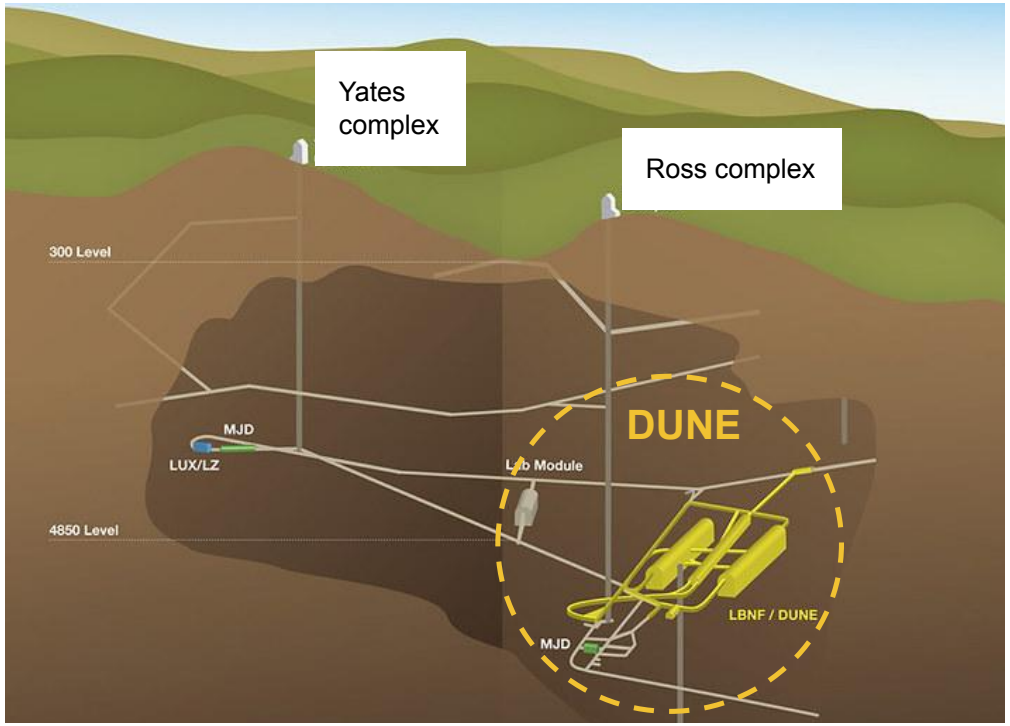
DUNE's Supernova Challenge

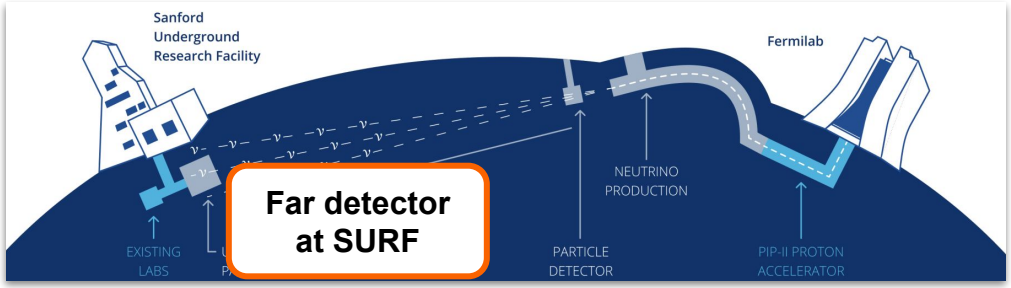


not to scale, not a technical design
it's just a cartoon

- Limited space and infrastructure (i.e. cooling) at far site → no bulk processing on local farm
- 10,000 – 40,000 present-day CPUs needed for reconstruction → 4-8 hours
  - HPC centers
    - Concern → data transfer in and out
    - Entire workflow - stitching data, output of reco - failure modes - efficiency vs. accuracy trade off
- Must be able to handle large input stream as well as output at similar rate

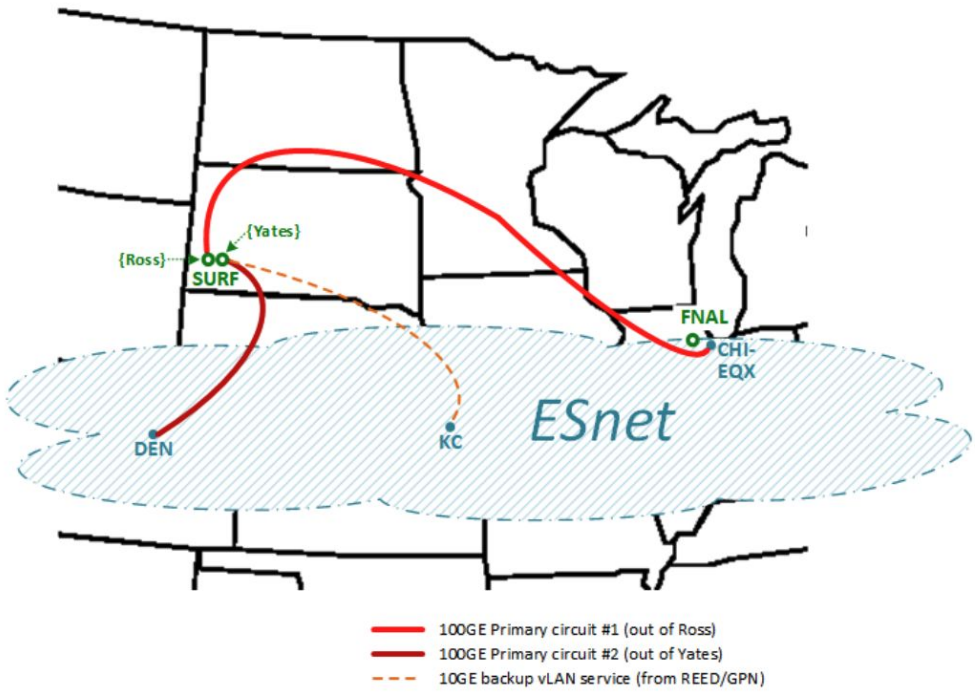
DUNE Networking: LAN

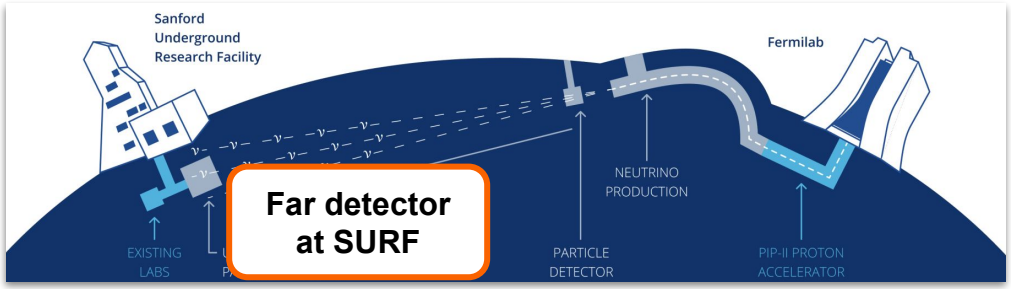




- Highly resilient network infrastructure and services
  - Core & perimeter network devices in Ross Dry MCR
    - Redundant devices/services in Yates
  - Redundant vertical fiber paths up both Ross & Yates shafts
  - Out-of-band management network to ensure remote recovery during outages

DUNE Networking: WAN





- Two geographically-diverse 100GE circuits to ESnet in Chicago & Denver
- On-demand Secure Circuits and Advance Reservation System (OSCARS) circuits complete the path(s)
- Managed end-to-end by ESnet
- Current plan is to utilize both circuits with Equal-Cost Multi-Path (ECMP) load balancing
- SURF to FNAL WAN is extremely important for a fast turnaround of transient signals such as Supernova triggered data along with beam data

DUNE Offline Networks

- Networking between compute sites varies; leveraging existing setups at LHC compute sites where possible (e.g. LHCONE), other large-scale infrastructure (e.g. ESnet, GÉANT, NRENs, etc.)

- Monitoring tools include perfSONAR

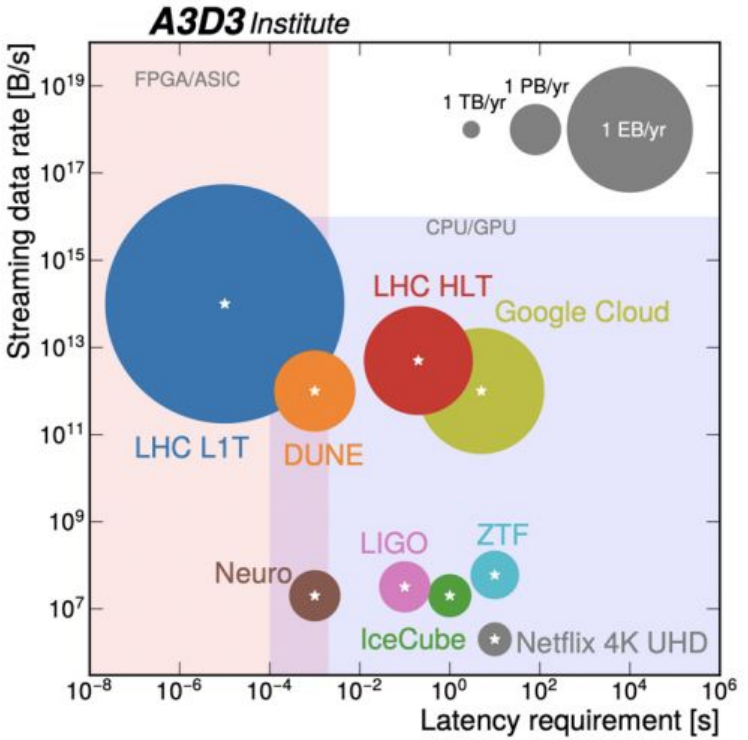
A Paradigm Shift

- HEP faces unique high-throughput computing challenges from massive data rates
- Advanced computing techniques
  - Enable deeper insights and improve performance
  - Improve operational efficiency
  - Ultimately accelerate time-to-physics and discovery

Evolve HEP computing infrastructure
Storage technologies, analysis facilities, heterogeneous computing (e.g. GPUs)

Leverage multidisciplinary computational & domain science expertise
Federal HPC facilities and commercial cloud, specialized services, modern software stacks

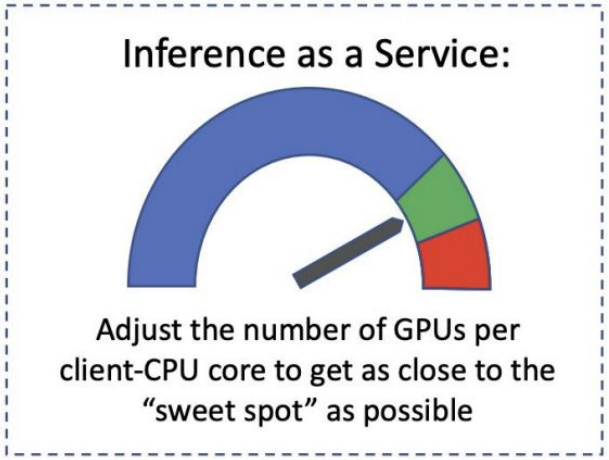
Embrace AI/ML for HEP and also HEP for AI/ML
Develop AI capabilities for HEP science, support HEP contributions to broader AI advances

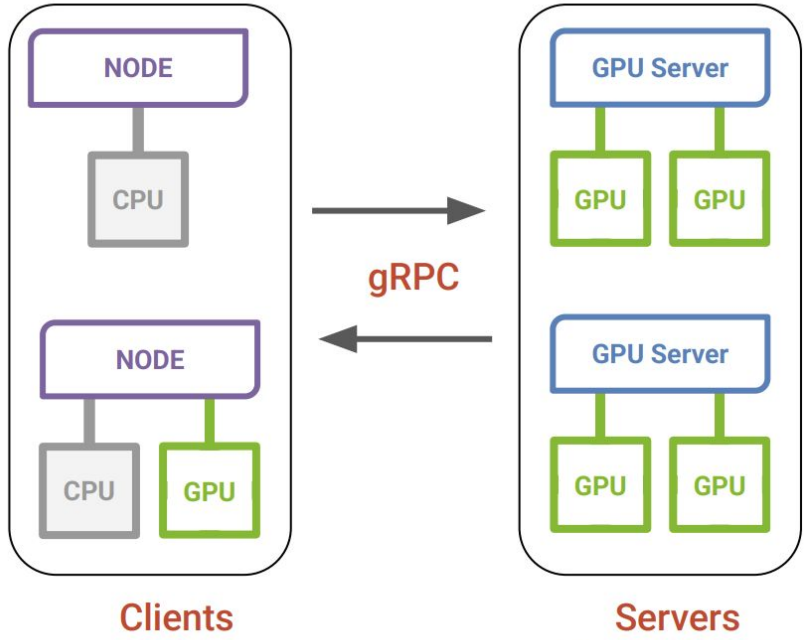


DUNE Software R&D efforts to explore HPC integration

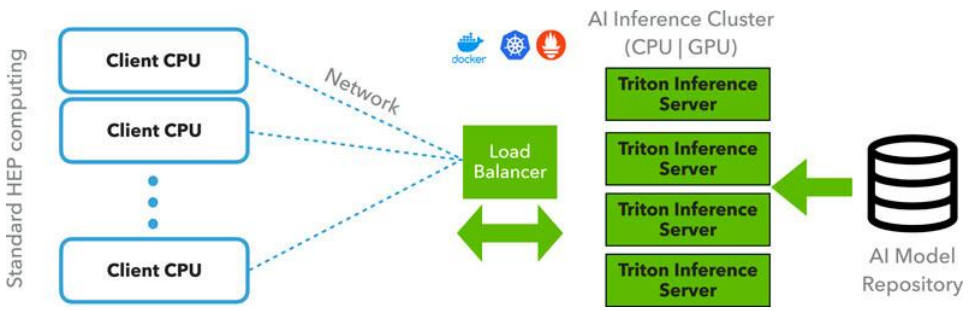
GPUaaS in DUNE

- Inference as a Service provides a flexible, alternative deployment scheme where machines with coprocessors host an inference server and remote clients send inference requests via network connections





GPUaaS in DUNE

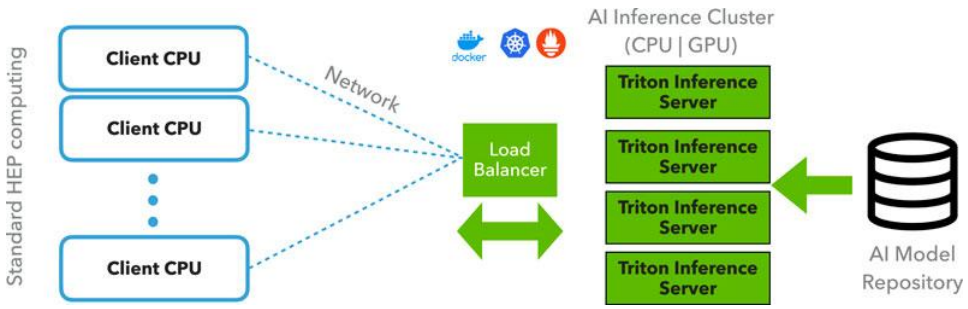


M.Wang et al., Front. Big Data 3 (2020)

	ML module	non-ML modules	Total
	Wall time (s)		
CPU only	220	110	330
CPU + GPUaaS	13	110	123

- Track reconstruction ML modules are CPU bottlenecks in DUNE workflows
- Use Triton inference server and gRPC in job for communication, allows many-to-1 model of CPU jobs to single GPU
- Many CPU jobs can share one GPU; rest of workflow remains CPU-only → portable to any site with network access
- **>10× faster** with GPUaaS

GPUaaS in DUNE

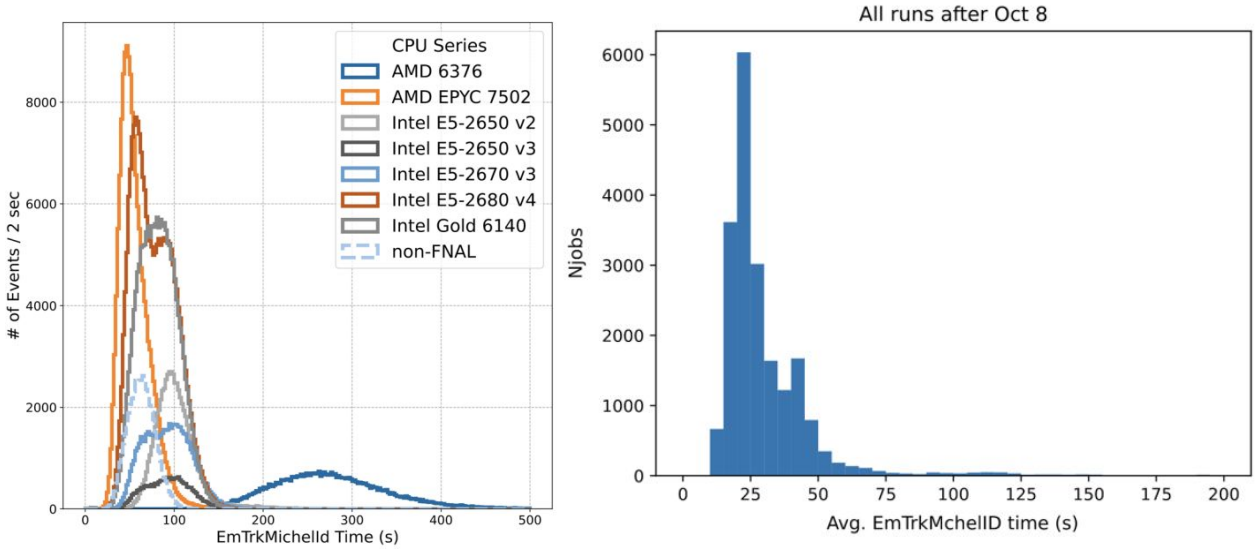


M.Wang et al., Front. Big Data 3 (2020)

	ML module	non-ML modules	Total
	Wall time (s)		
CPU only	220	110	330
CPU + GPUaaS	13	110	123

- ProtoDUNE beam data in 2021 with cloud-based GPU server
- Overall speed increase, but amount of data movement 10X wrt CPU-only version
- Must take care not to saturate network capacity

- Track reconstruction ML modules are CPU bottlenecks in DUNE workflows
- Use Triton inference server and gRPC in job for communication, allows many-to-1 model of CPU jobs to single GPU
- Many CPU jobs can share one GPU; rest of workflow remains CPU-only → portable to any site with network access
- **>10× faster** with GPUaaS



T. Cai et al., Comput. Soft. Big Sci. 7, 11 (2023)

Summary

- DUNE will deliver groundbreaking results in fundamental physics
- A supernova in our galaxy will lead to a wealth of information, but processing detector data in a timely fashion will be one of DUNE's biggest challenges
- DUNE's geographically distant location requires robust networking to deliver all the groundbreaking science!
- Active R&D efforts in the areas of AI/ML and HPC integration into computing models

Thank you!

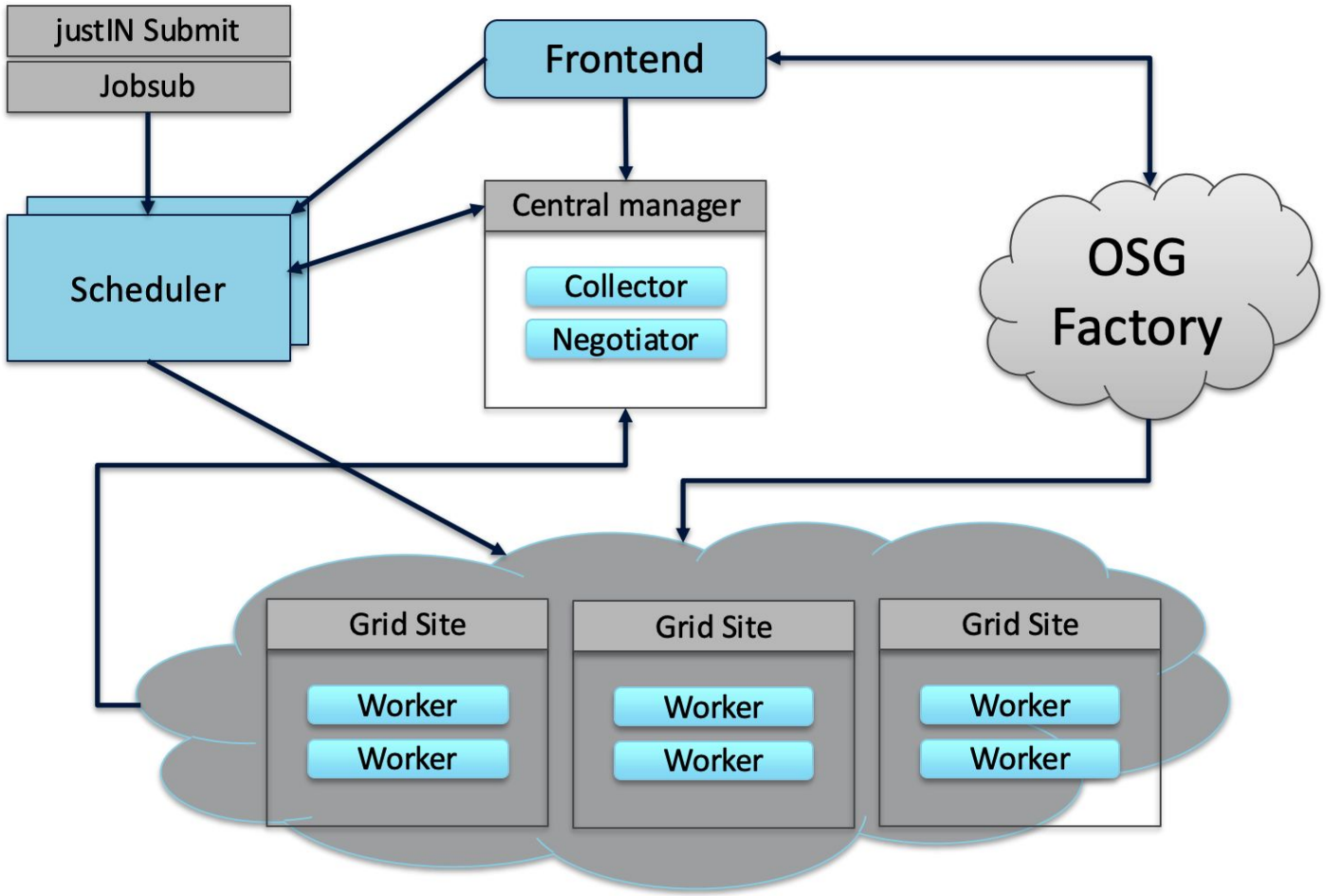
>1400 collaborators

207 institutions at Africa, Asia, Europe, North and South America as of July 2024

Backup

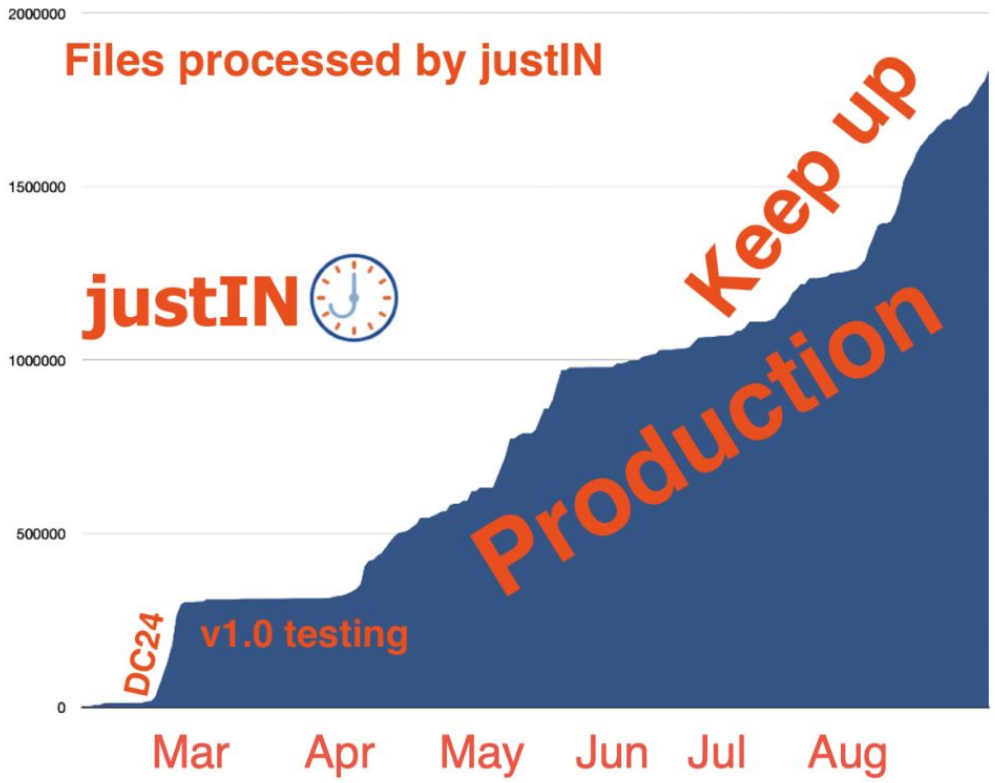
Scaling the DUNE Global Pool

Current Infrastructure map



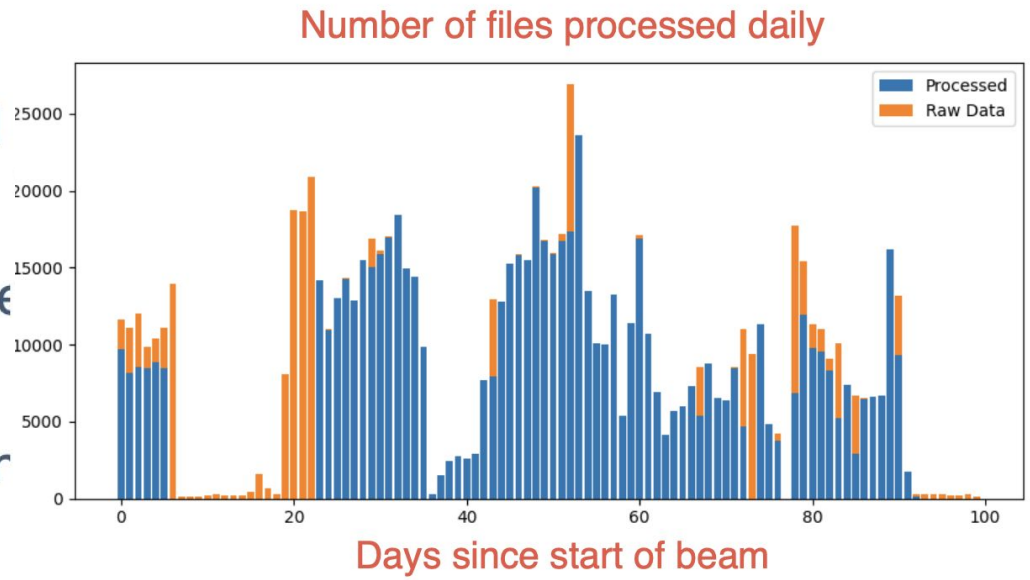
justIN Workflow Management System

- justIN ties together MetaCat, Rucio, and GlideInWMS
  - runs jobscript on dataset specified with an MQL metadata query
  - directs jobs to the optimal sites and handles all the Rucio storage operations
  - justIN successfully tested during WLCG Data Challenge 2024
  - v1.0.1 is now the basis of official DUNE Productions
- Andrew McNab's poster 402 Wed poster session



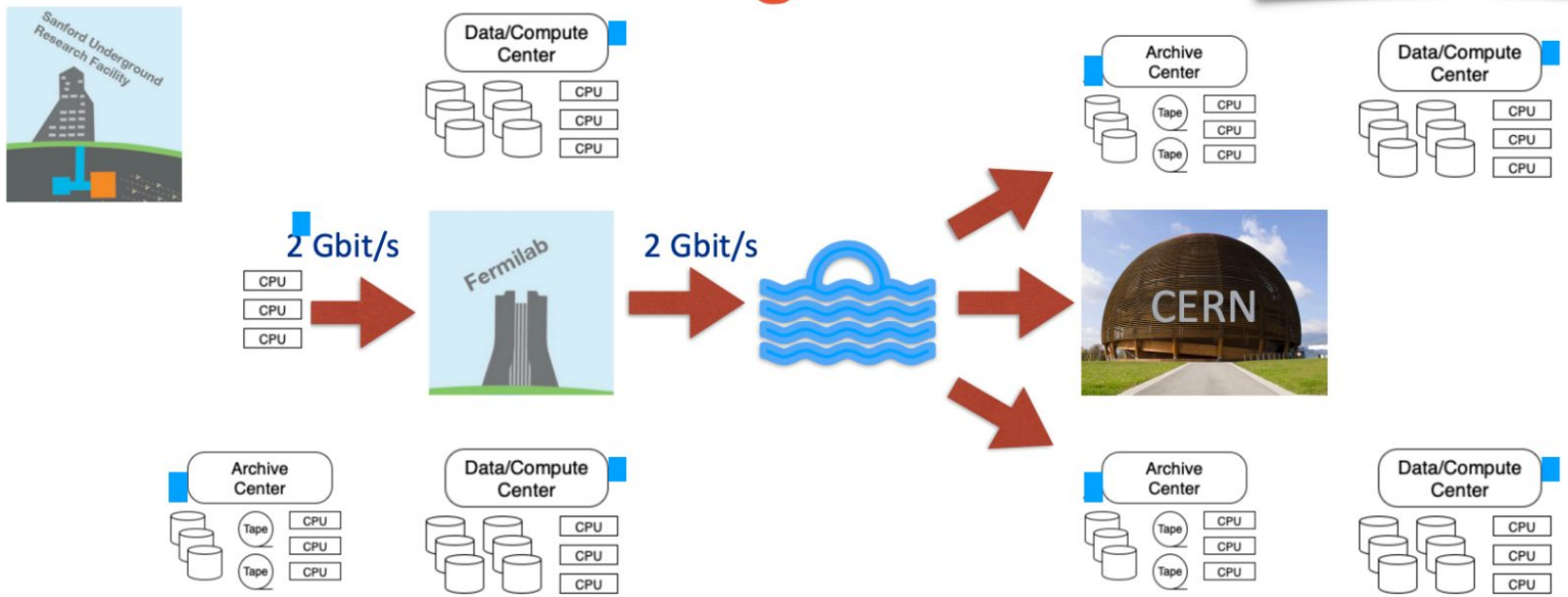
HD ProtoDUNE keep up processing

- beam ran on-and-off 06/19 – 09/16
- automated submission (twice a day)
- 650+ TB of reco files have been produced (reco2 stage)
- need to resubmit for times where site issue occurred
- experience will inform design of production systems to more easily automate (i.e. recovery jobs) & logging
- using new justIN workflow system developed and improving feature set



DUNE WLCG Data Challenge 2024

"FD" Raw Data
archival storage

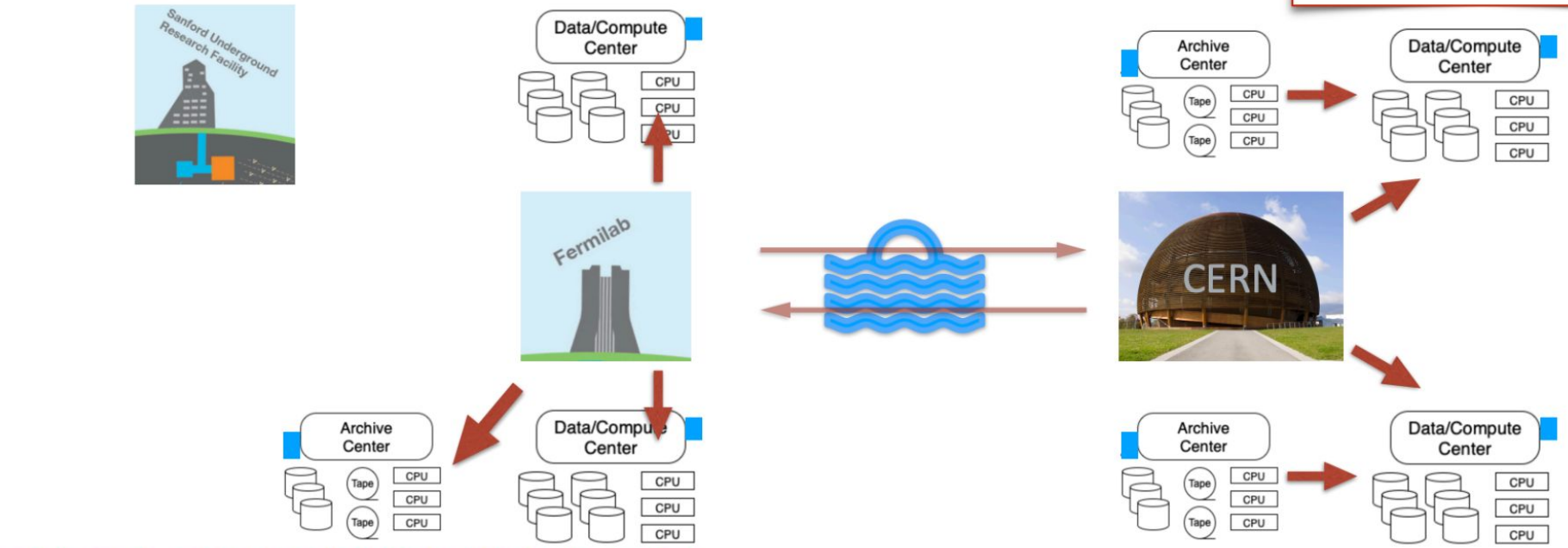


not to scale, not a technical design, it's just a cartoon

- Simulate the archival of 25% of the raw data rate from the Far Detector
  - translates to 2 Gbit/s from SURF to FNAL
  - replicate that "FD" raw data to archival storage facilities around the world
  - replicate the "FD" raw data to disk storage elements around the world for prompt access from compute elements
- Both job submission and RSE to RSE w/ token authentication/authorization

DUNE WLCG Data Challenge 2024

"FD" Raw Data
raw processing

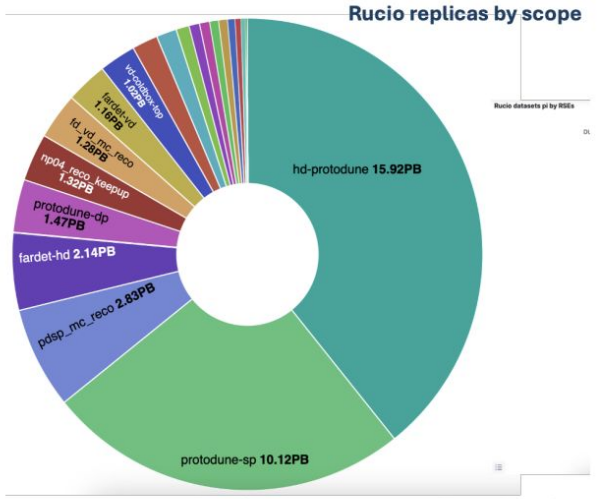


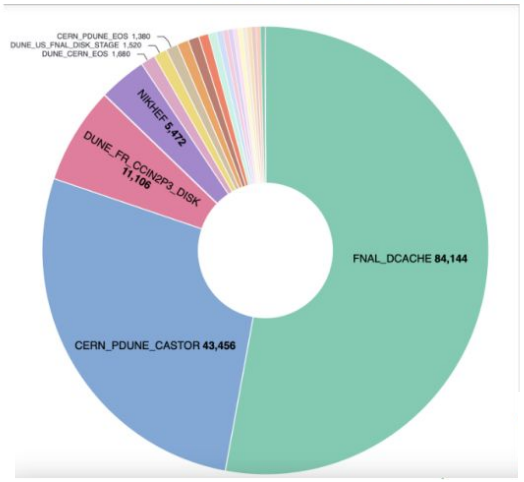
not to scale, not a technical design it's just a cartoon

- Maintain continuous processing workload at distributed sites commensurate with 25% "FD" raw data rate
  - utilize compute elements across the WLCG and OSG
  - match the locality of jobs with locality of data at nearby RSEs
- Both job submission and RSE to RSE w/ token authentication/authorization

Data Management and Data Pipeline from NP04 (PDHD)

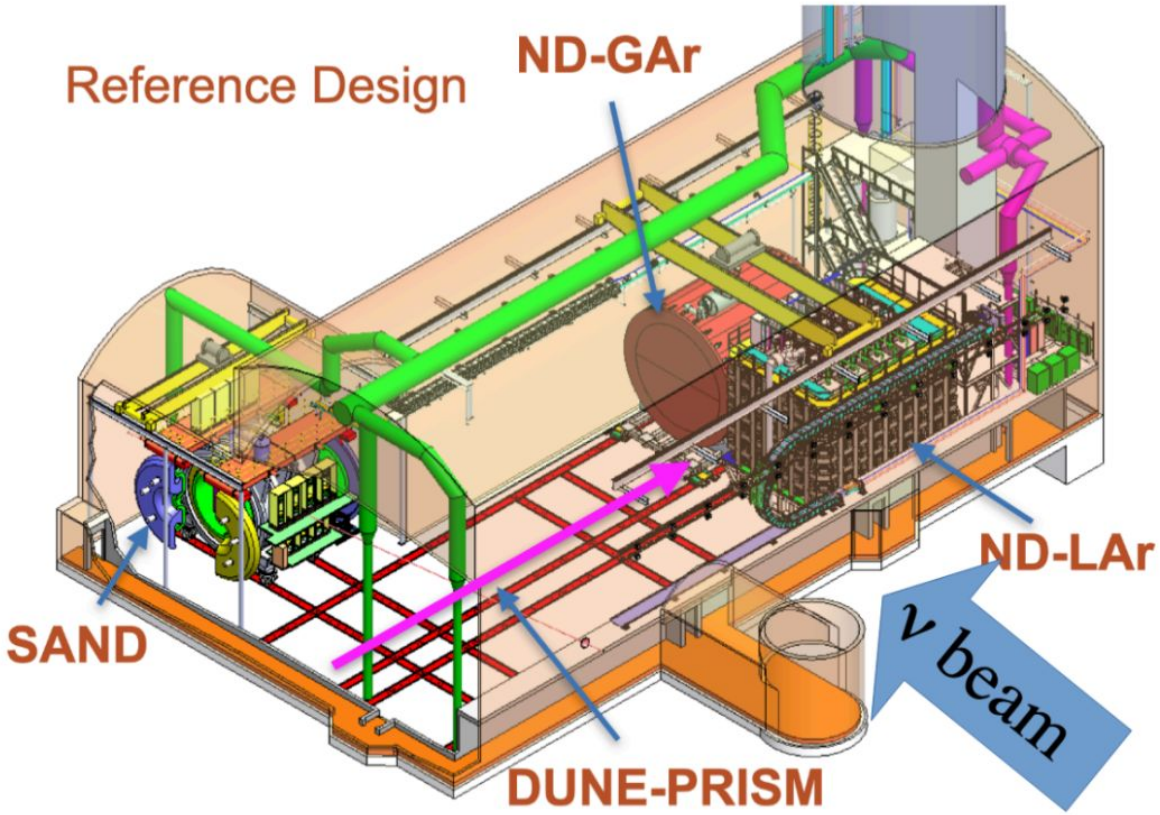
- Pipeline utilizes sequences of tools (FTS3, Metacat, Rucio, and custom ingest/declaration daemons)
- Early estimates anticipated 2-3PB of beam data to be written
- Aug 10-11 consecutive days 100+ TB
- wrote 4.5PB of data
- Moved about 2PB of protodune-hd raw data to other disk sites in one week to accommodate additional data taking
- files moved to BNL, PIC, NIKHEF, FNAL, PRAGUE, RAL-PP, SURFSARA





Near Detector Design and Data Flow

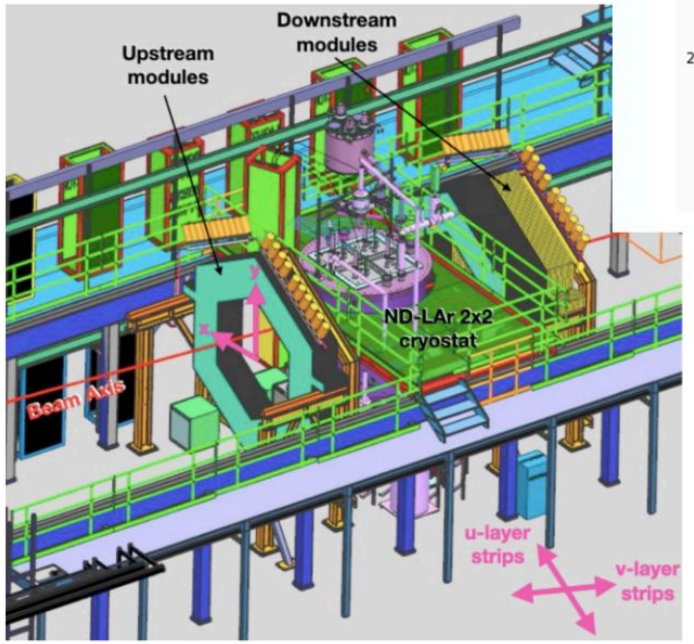
- designed to constrain neutrino beam flux, precisely measure cross sections, minimize detector response uncertainties
- three subdetectors: ND-LAr, Muon Spectrometer, and SAND
- intense neutrino beam dictate different detector design >15 overlapping ν-interactions
- physics and detector design precipitate move towards GPU focused software for reco and sim
- despite very different data occupancy, signatures, and structure - need to be treating in the same way for smaller systematics

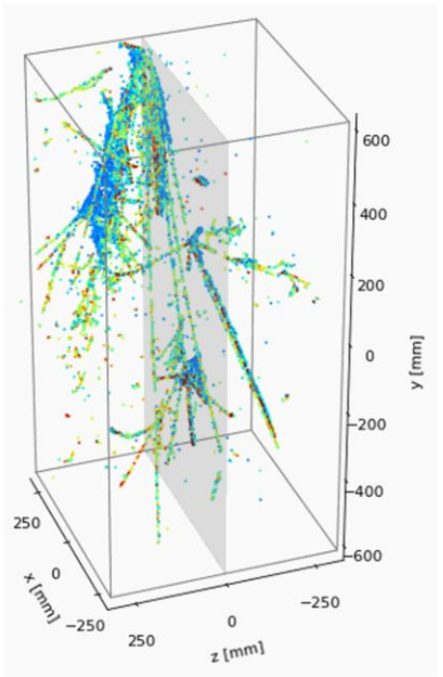


DUNE Near Detector Facility

Near Detector 2x2 Demonstrator

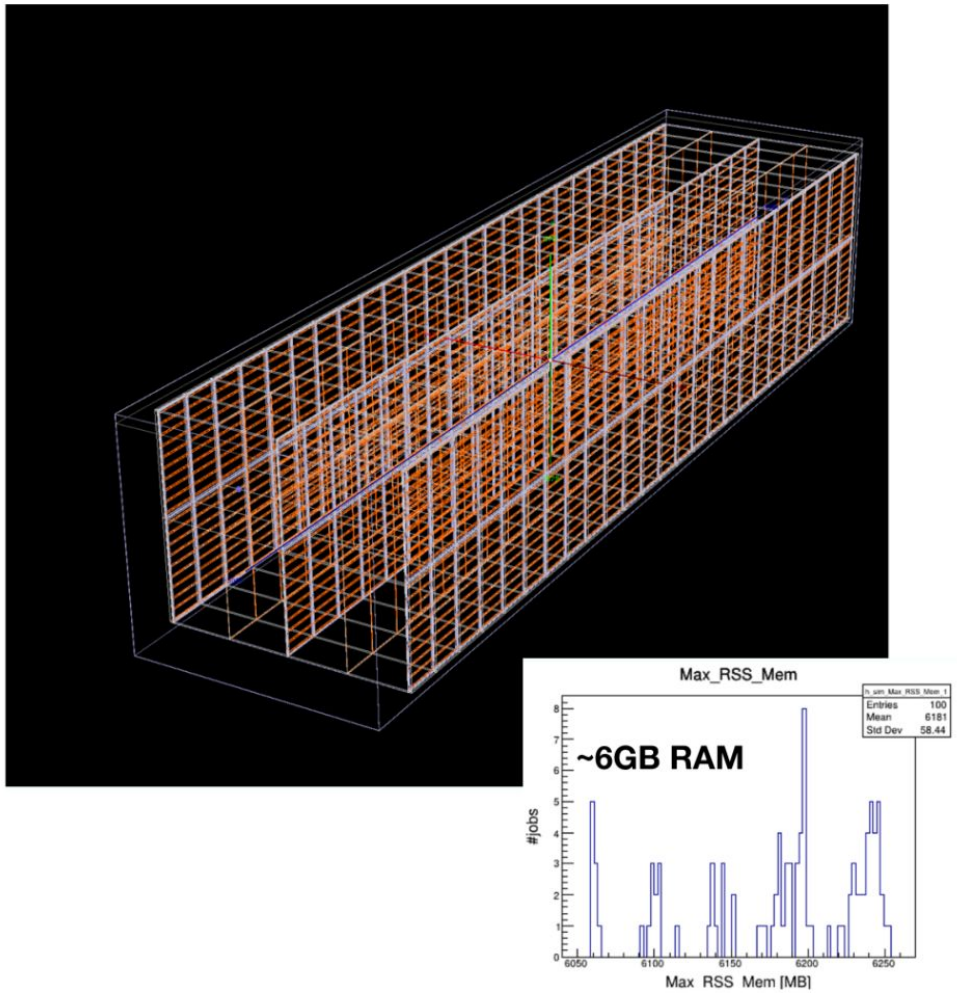
- prototype using NuMI neutrino beam
- perform neutrino physics measurements at DUNE neutrino energy and on an Argon target
- data pipeline developed for ProtoDUNE was replicated at Fermilab for 2x2-MINERvA LAr prototype
- approx. 1 week of good data taken with (anti)neutrino beam just before July shutdown
  - 1.2 TB of Minerva chambers data
  - 13.2 TB of LAr Light readout
  - 0.6 TB of LAr charge readout.
- working with the 2x2 team to more fully integrate Data Pipeline/Rucio/Metacat into operations at NERSC





FD Detector Simulation

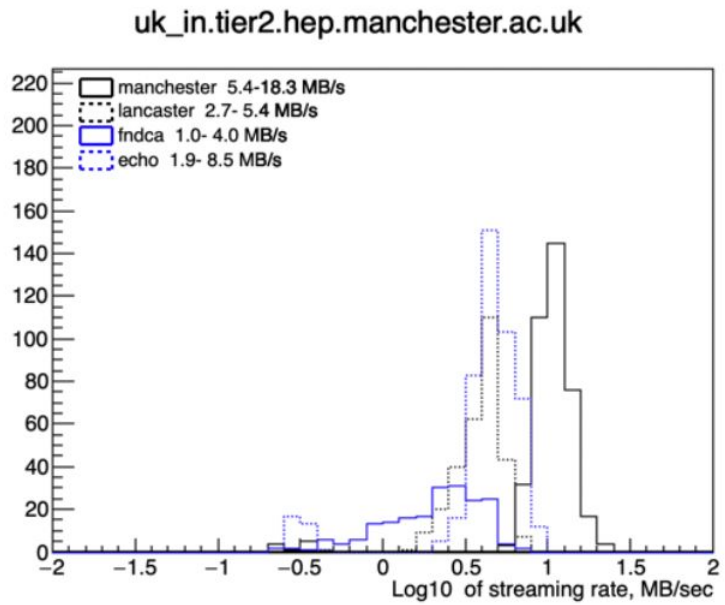
- large, open, and homogenous FD is a strength of DUNE detector but presents many challenges for current framework
  - simulation of interactions across many orders of magnitude
  - very large memory footprint from treating readout as a single chunk
  - lends itself to utilization of GPUs which isn't native to art
- currently possible by separating at the APA level, but "bypasses" framework level hooks
- optical photon simulation on CPUs is computational impractical

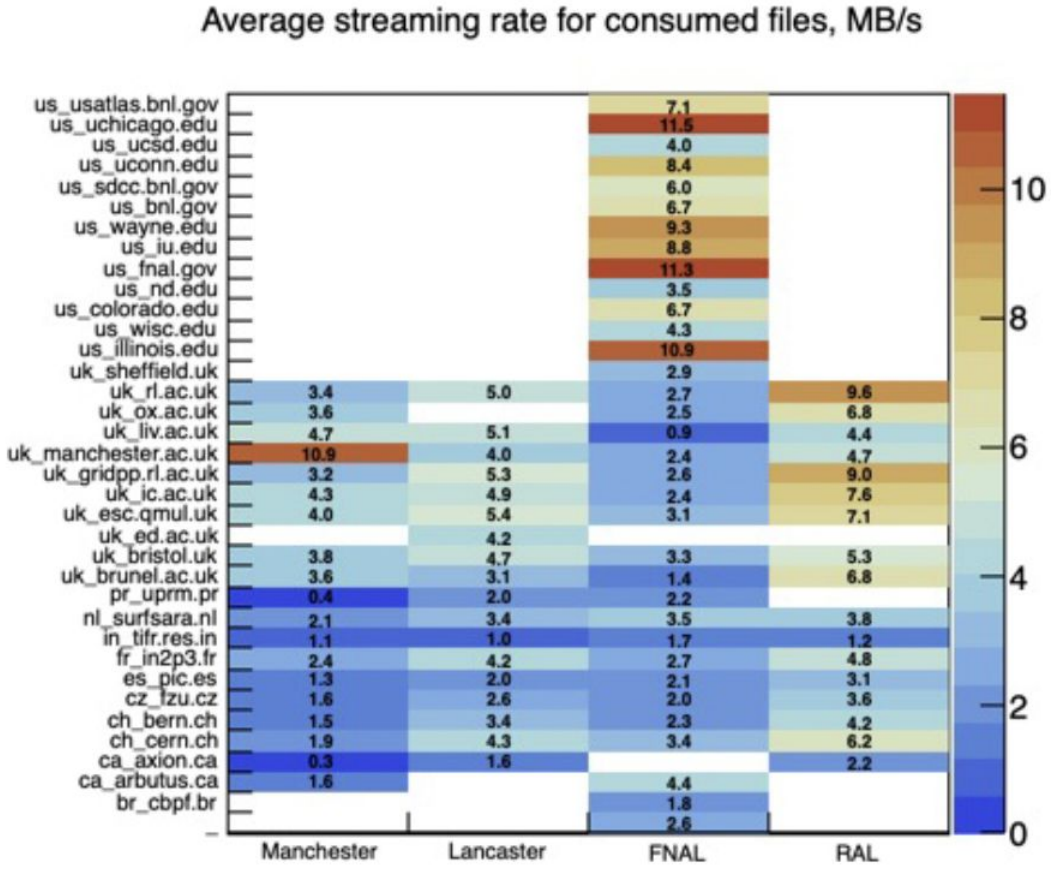


memory footprint ill suited for WLCG slots

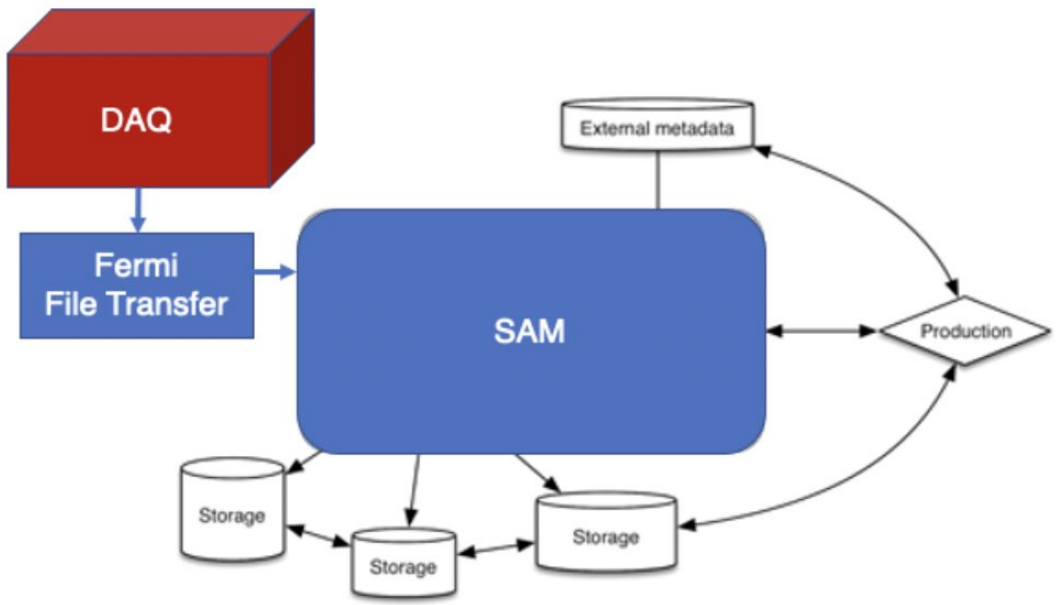
Network latency effects

- Clear improvement in streaming rates (and event throughput) when reading from "nearby" sites. Similar workflow in these tests
- Important for file delivery mechanisms to take into account and choose the "closest" file

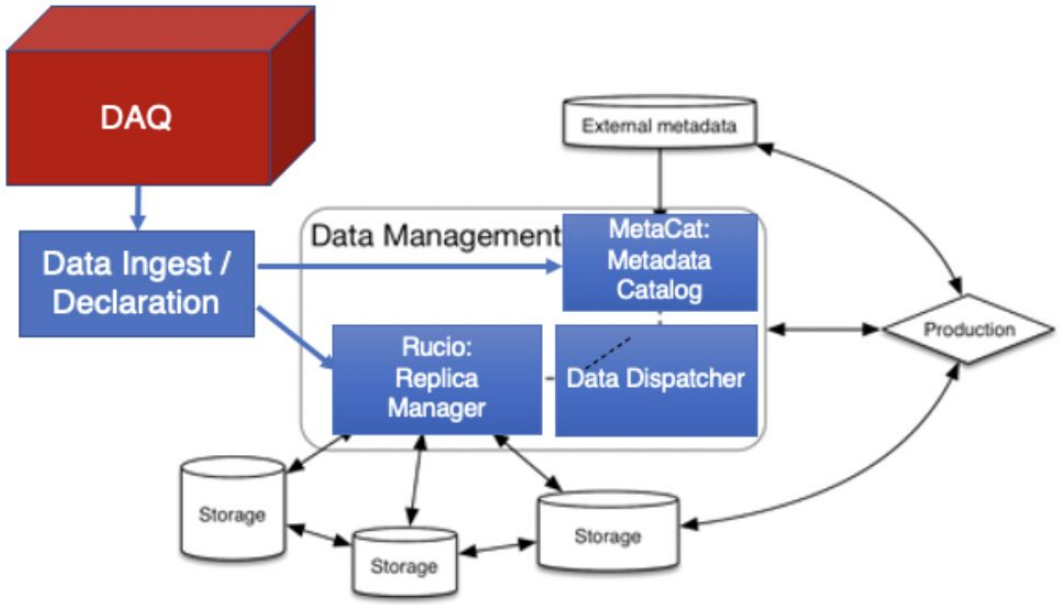




File Delivery and Dataset Replication



Legacy system (SAM)
Handles file delivery, metadata catalog and replica catalog; does not choose "best" replica to deliver to job



Updated system
Separate services for metadata, file delivery (DD), and replica management. DD has more features for replica choice; MetaCat more metadata and dataset definition features