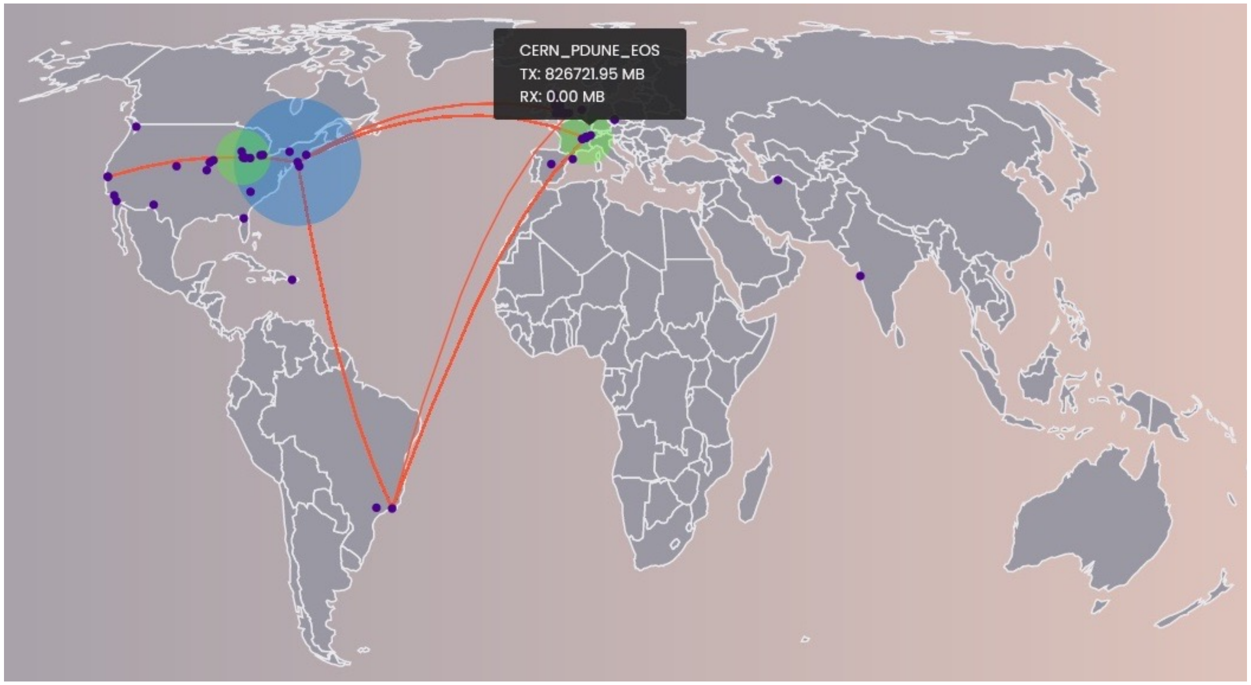


DUNE COMPUTING UPDATE

H. SCHELLMAN (OREGON STATE)
FOR THE COMPUTING CONSORTIUM

CDR progress

- Major design documents being finalized
  - Frameworks requirements – DONE -> HSF
  - Hardware data base – In production
  - SAM replacement
    - Metadata catalog – prototype
    - Rucio – in progress
    - Data dispatcher – design
    - Workflow dispatcher - design
  - DAQ-Offline interface requirements
  - Use cases
    - ProtoDUNE
    - Analysis

Two parts to Computing Infrastructure

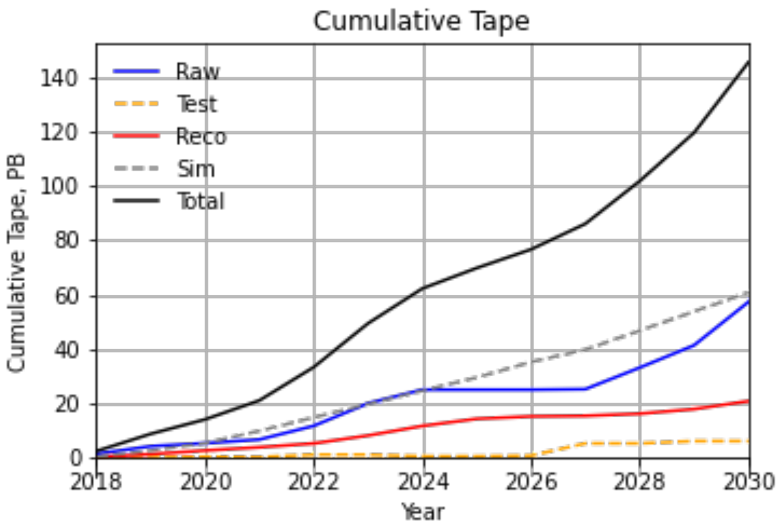
Development and Operations

Institution	Country
York University	Canada
CERN	CERN
IN2P3	France
Edinburgh	UK
Manchester	UK
RAL/STFC	UK
Queen Mary Univ. London	UK
Argonne*	USA
BNL*	USA
Colorado State*	USA
Fermilab*	USA
LBNL	USA
Minnesota*	USA
Oregon State University*	USA
Wichita State*	USA

* US computing consortium

Hardware contributions

Facility	Country
CBPF	BR
CA_Victoria	CA
CERN	CERN
FZU	CZ
PIC/CIEMAT	ES
CCIN2P3	FR
TIFR	IN
SURF/SARA	NL
JINR	RU
GridPP	UK
OSG	US
FNAL	US
BNL	US



CDR - Resource estimates to 2030

2 copies of raw data on tape (6 months on disk)

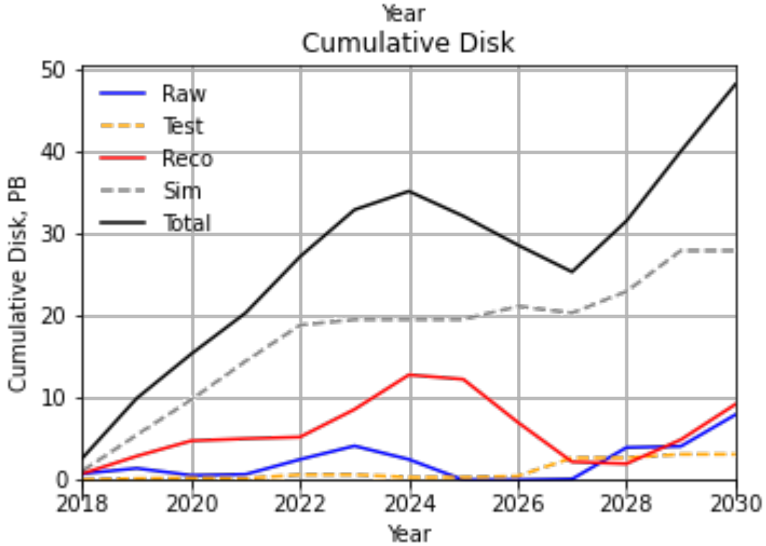
1 copy of "test" data stored for 6 months

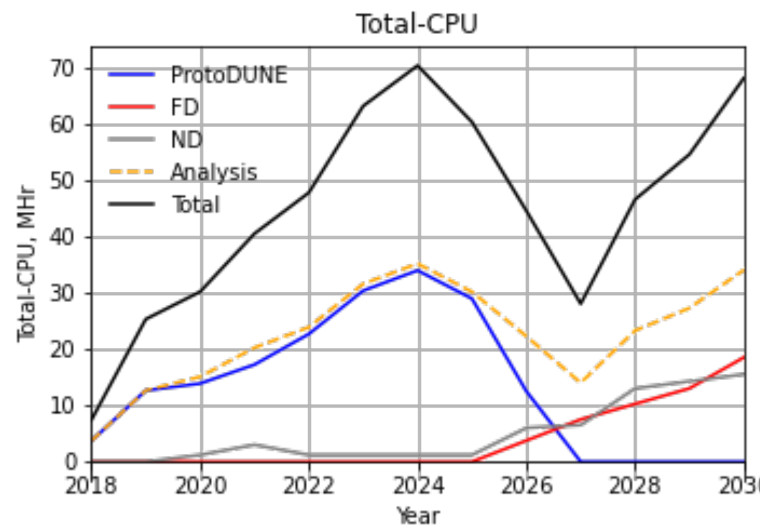
1 copy of reco/sim on tape

- Currently assume 1 reco pass over all data and 1 sim pass/year
- Assume reco/sim resident on disk for 2 years

Assume 2 disk copies of reco and sim

- impose shorter lifetimes on tests and intermediate sim steps.





Longer term projections

VD assumed to be similar to HD, raw data may be larger due to longer drift.

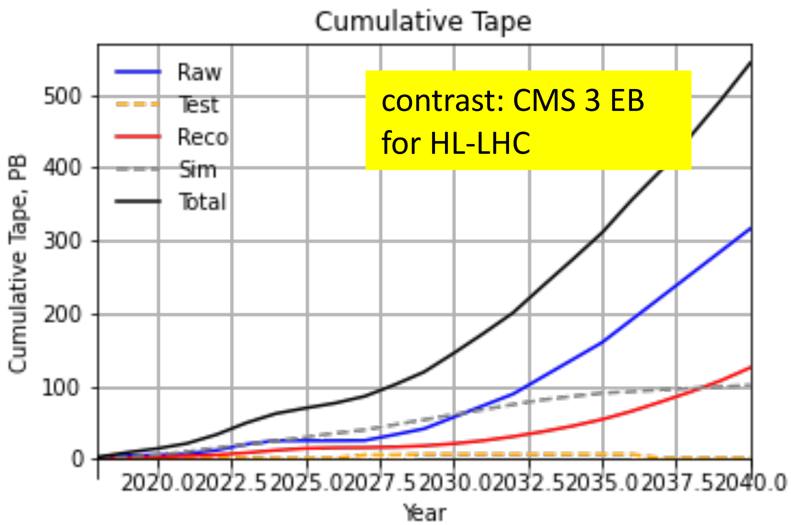
2 copies of raw data on tape (6 months on disk)

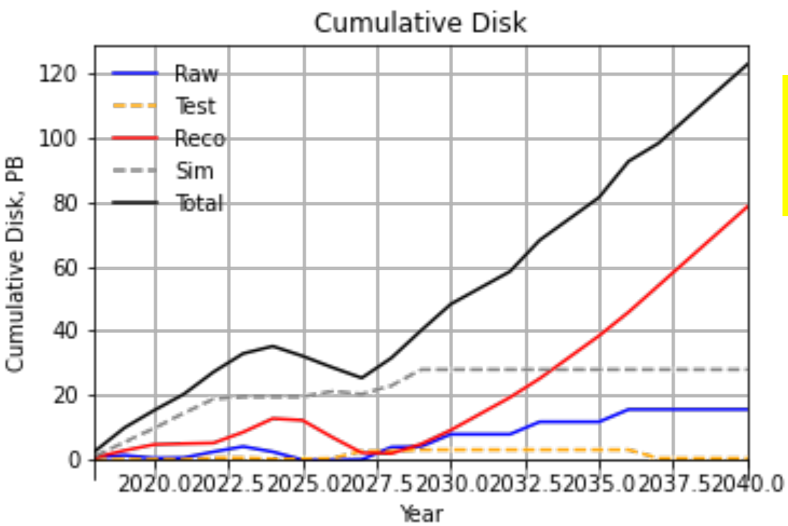
1 copy of "test" data stored for 6 months

1 copy of reco/sim on tape

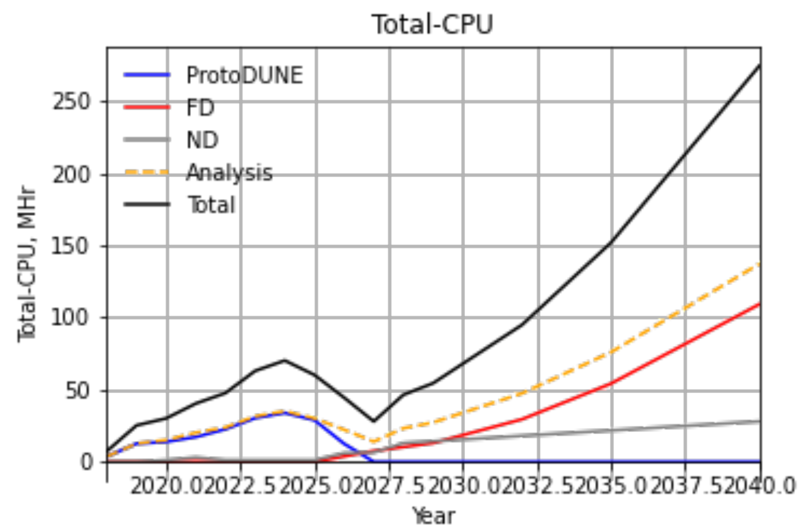
- Currently assume 1 reco **pass over all data** and 1 sim pass/year
- Assume reco/sim resident on disk for 2 years

Assume 2 disk copies of reco and sim



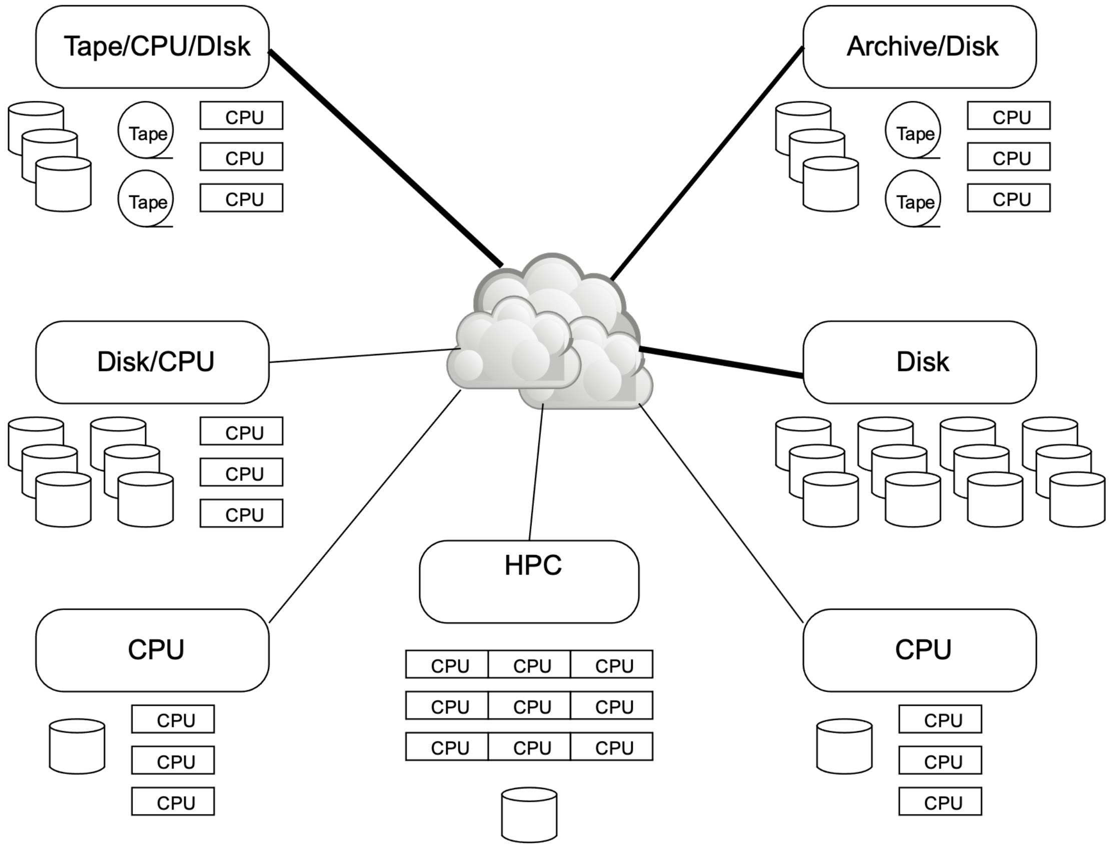


contrast: CMS 150 PB->2.2 EB in HL-LHC



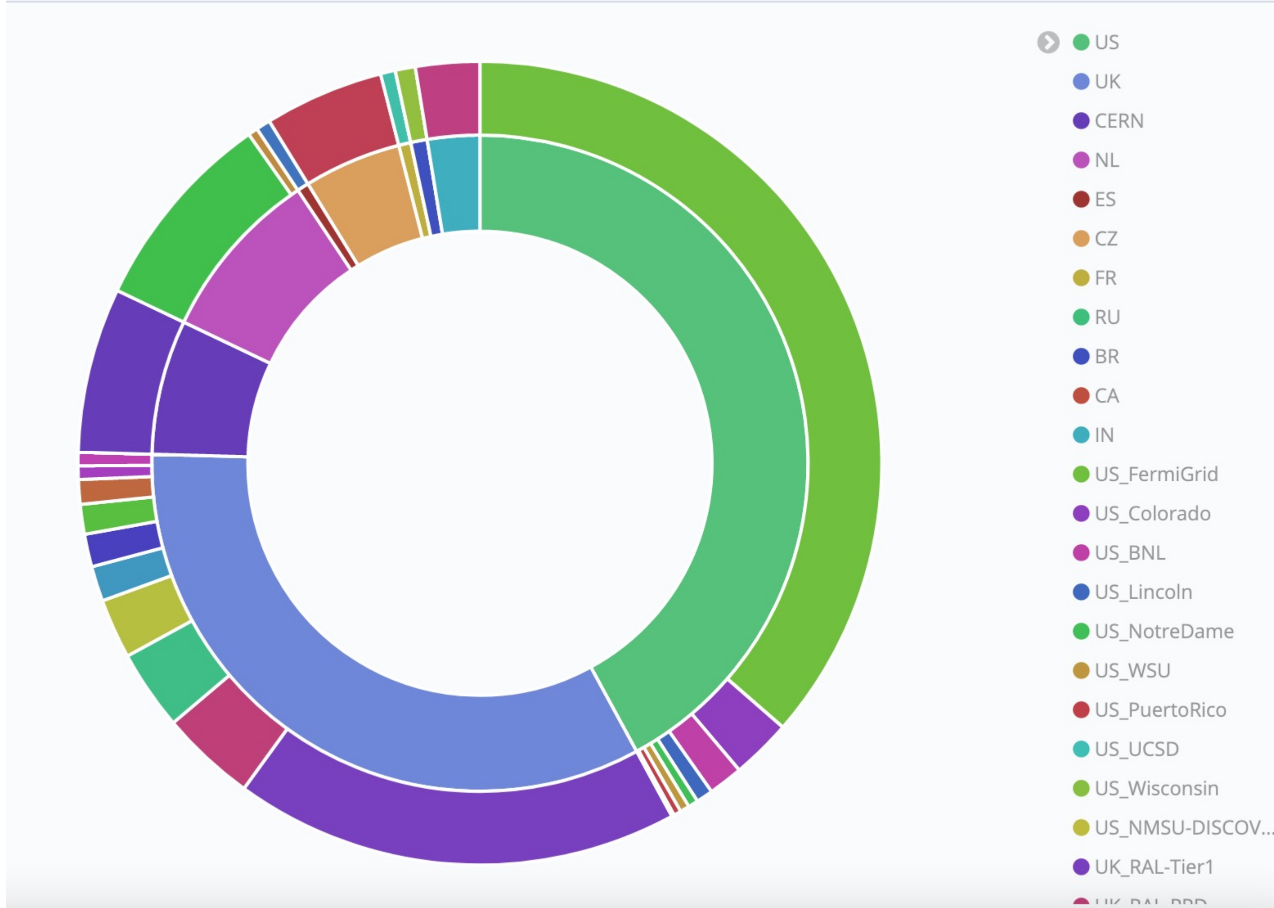
CDR - Distributed computing model
Does this work?

- Less “tiered” than current WLCG model
- Collaborating institutions (or groups of institutions) provide significant **services** (disk/CPU/archival)
- **Rucio** places multiple copies of datasets
- Workload/Data management system match data with appropriate delivery method
  - File already near local CPU
  - smart file location info
  - Direct copy to local cache
  - xrootd stream ← what we do now!
- Assumes good network connectivity
  - Currently working for 8,000 concurrent reconstruction jobs
  - Working with ESNET and European networking



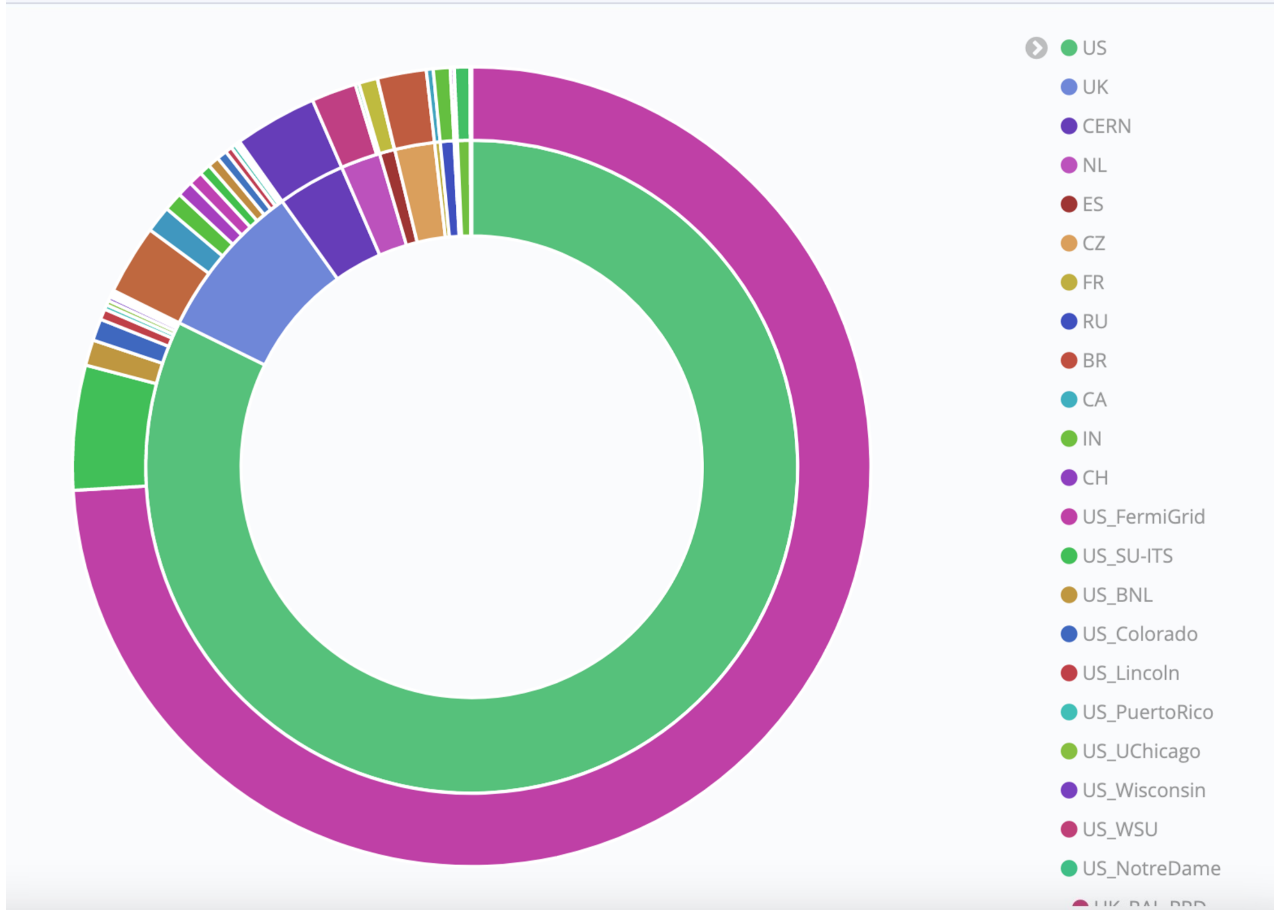
Where we are now: Production pass 4a

- This shows August 2021 contributions to production
- TIFR/India (new!) has lots of memory/core so is contributing substantially to simulation



Where we are now: Analysis is moving offsite

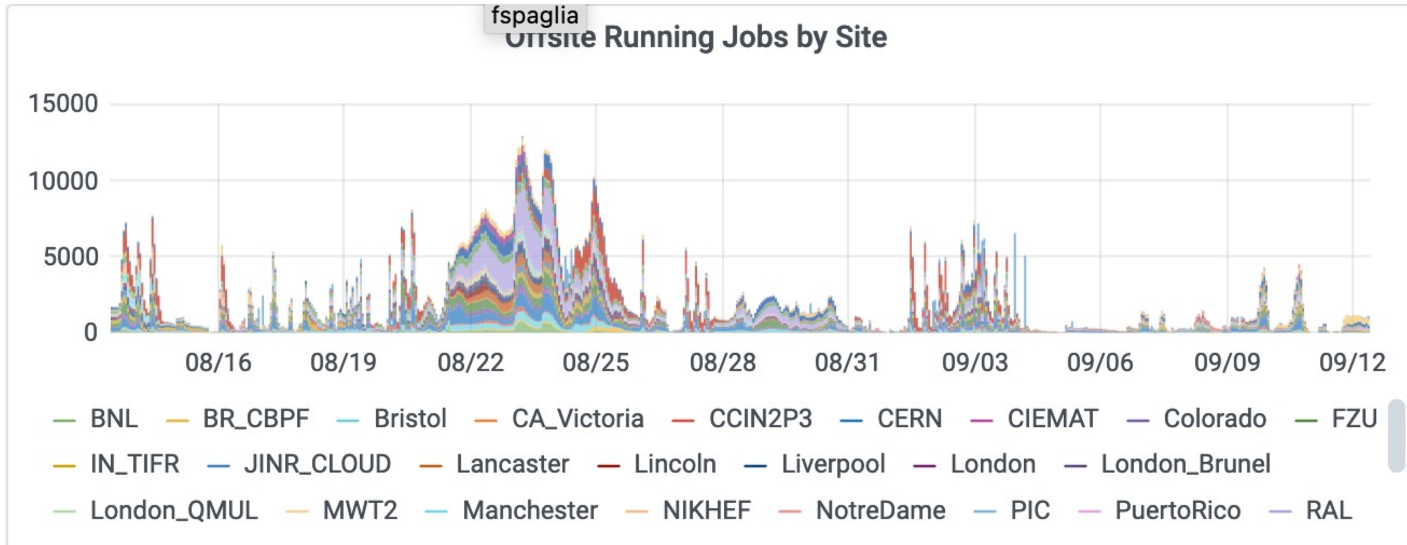
- This shows August 2021 contributions to analysis
- US sites are contributing as are many sites worldwide

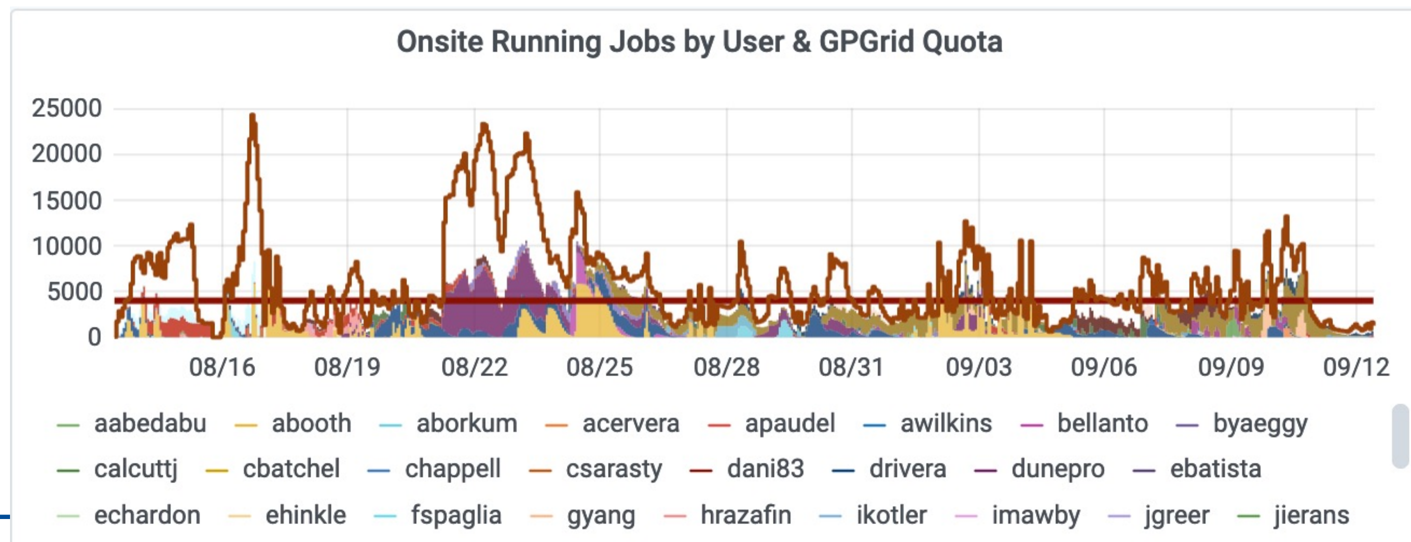


Last month

Mainly user analysis

Production team gaining new members



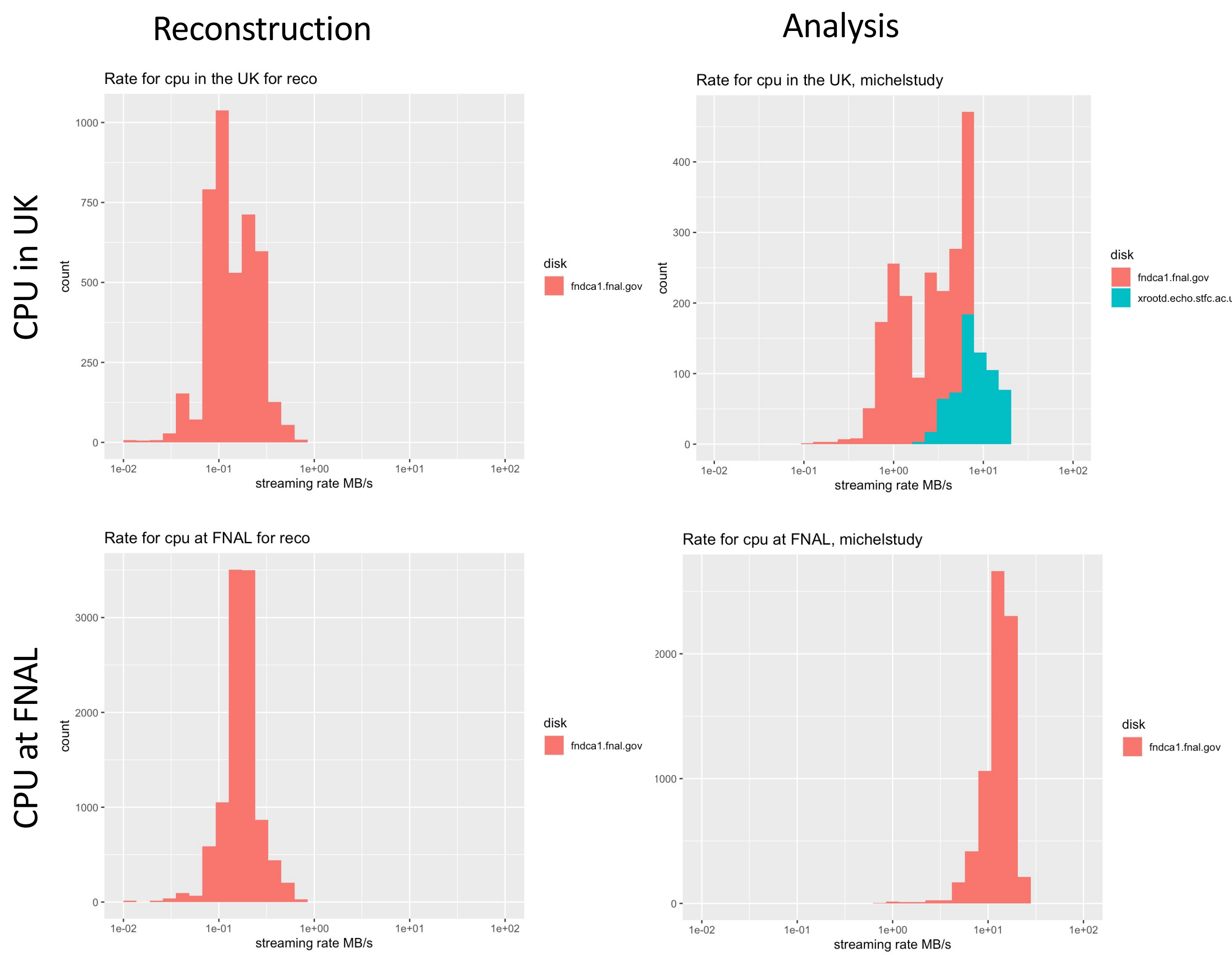


Responsibilities

- Tape storage
  - raw data – 2 copies – 1 at FNAL
  - sim/reco - 1 copy
- CPU
  - FNAL 25%, collaboration 75%
- Disk storage
  - National contributions 5-20% of the total from many countries
  - Pledges for 2021/2022 now being collected
- Network:
  - Working with ESNET on SURF->FNAL networking
  - Discussions with international partners (DUNEONE) on offshore compute network
  - Significant monitoring efforts underway

Data location studies

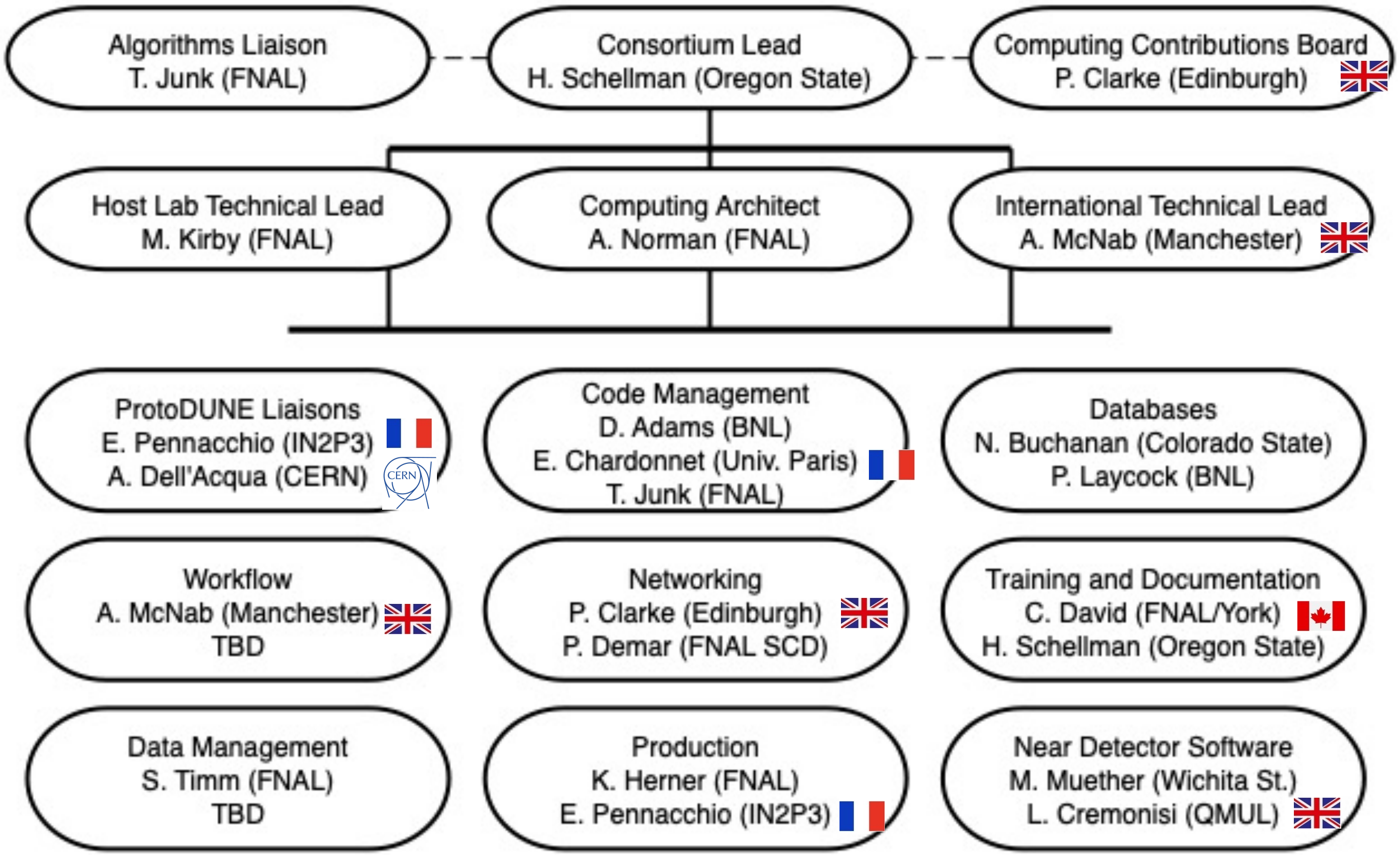
- OSU/FNAL project
- Mine sam event logs to study streaming rates vs RSE, CPU site, and application
- Red is FNAL-dCache
- Teal is RAL-Echo
- (x axis is log10(rate/MB) from 0.01 MB/s to 100 MB/s



Conclusion on hardware resources

- We have identified and are formalizing resource contributions from collaborators worldwide.
  - CPU resources (looks good for protoDUNE)
    - shared resources from WLCG/OSG give us **lots of flexibility here**
  - disk resources (needs both contributions and code development)
    - **For analysis, collocating disk and CPU within a region helps a lot**
    - **Our model emphasizes keeping heavily used samples on disk**
  - tape resources (FNAL and CERN for now – will be come a big issue after 2030)
    - tight controls on data volumes and retention policies for intermediate steps.
  - networking – working with ESNET as part of their planning exercise
  - Move to DUNEONE network, away from LHCONE for PD expt traffic.

Development Organization



Development Scope

- Use common tools (ESNET, rucio, WLCG …) where possible
- Detector is new
  - **New databases** need to be designed for conditions/calibrations
- DUNE events are very big and getting bigger 70 MB (PD) --> 3 GB (FD) → 115 TB/module (Supernova)
  - New framework
  - Memory management is … interesting …
  - HPC adaption
- Collaboration is large
  - Support (and train) large # of users
  - Need to monitor and coordinate large # of sites (32 already)
  - **support thousands of simultaneous connections to DB and data stores**
- Needs to be **ready** at small scale in **2022 for PD-II**, large scale between **2026-2029 for FD/ND**

Since March: Framework review

- We asked the HSF to review our framework requirements and got detailed responses
  - Our descriptions of use cases and time scales needs work – this helps the CDR
  - HPC use case is interesting
    - Most current frameworks run on HPC but do so in “single node” modes so they do not leverage the machine’s capabilities
    - The amount of GPU acceleration is also new and current frameworks have this bolted on to older CPU only scheduling models.
  - MPI work distribution and “event” splitting are completely new
    - Committee likes this but notes it is novel
  - Memory Needs. (6GB event from FD?)
    - This is HIGHLY dependent on actual workflow topologies
    - This could be huge or tiny depending on how we sequence our analysis and how we break events up over processing ranks
  - Don’t be constrained to use same framework for reconstruction and analysis

Frameworks review by HSF

Unique DUNE requirements:

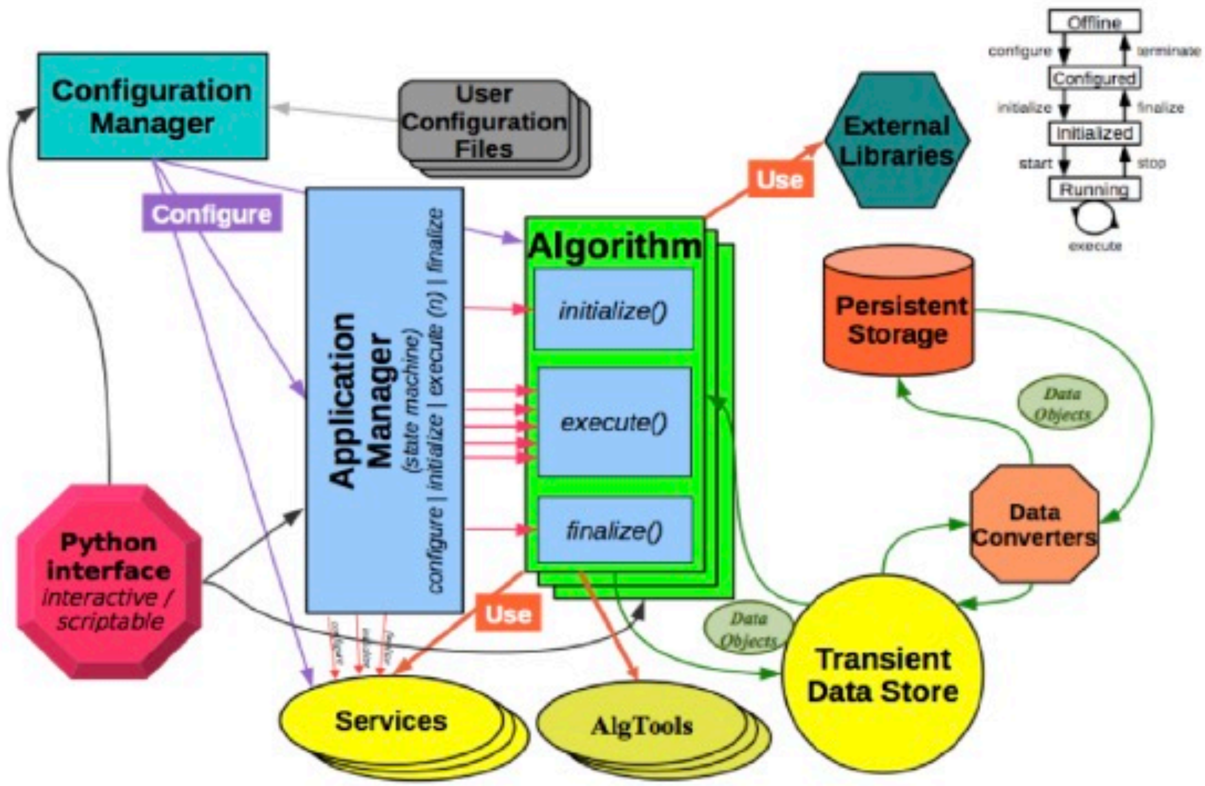
TPC/PD Data are simple on small scales → HPC

But 6 GB-100 TB readouts drive memory management requirements:

- Separate persistency/transient
- Precise tracking of provenance so parts can be reassembled
- multi-threading
- coherent processing across multiple architectures/sites

Can existing frameworks be modified to meet requirements? (major work)

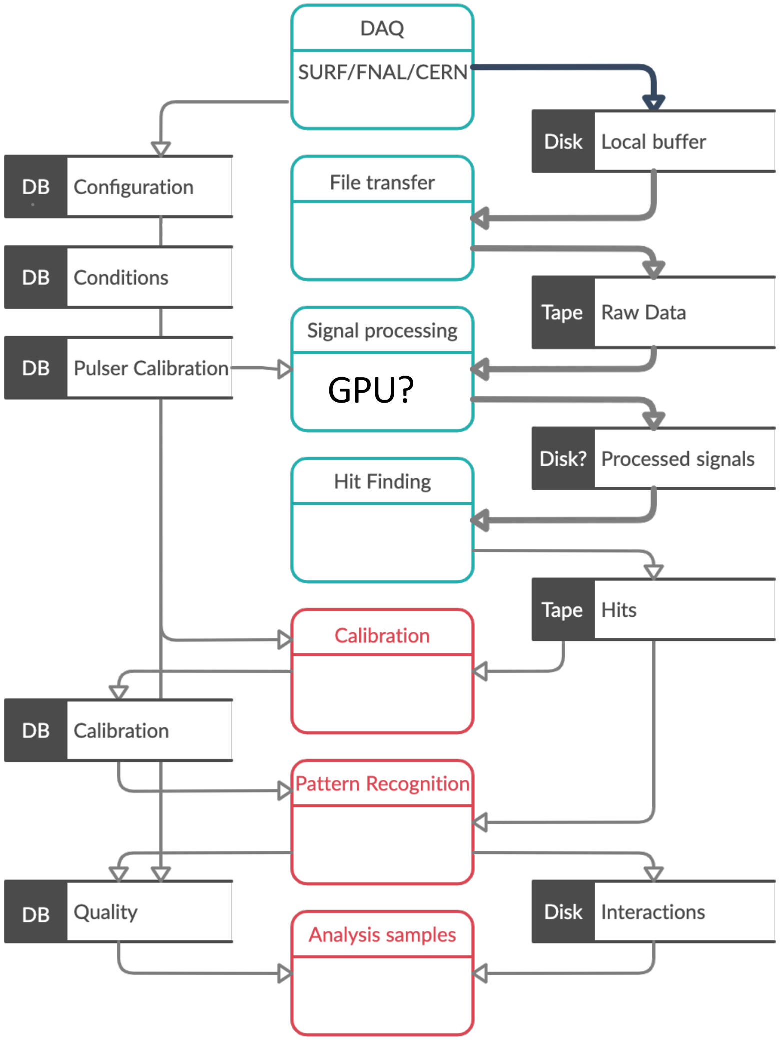
Are reconstruction and analysis frameworks the same? (probably no)



LBNF/DUNE

Data processing flow for FD/PD

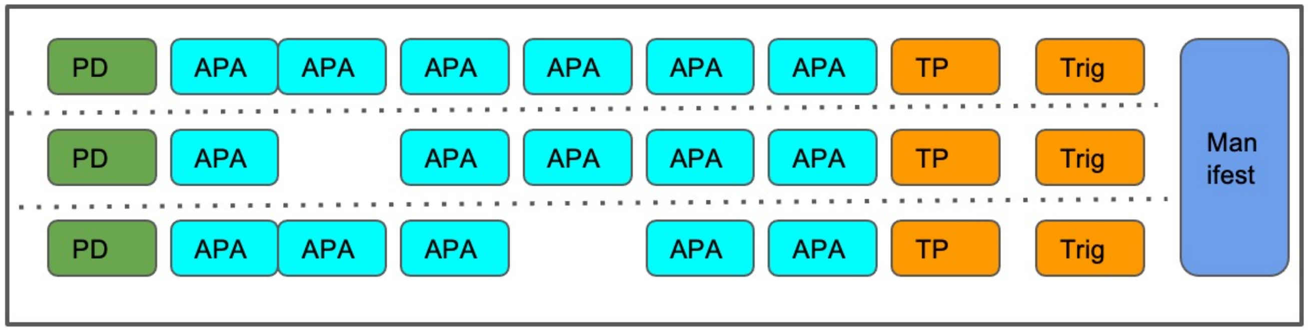
- Signal processing
  - Large scale operations on arrays
  - HDF5 instead of root structures?
  - Output can be zero-suppressed?
  - Separate framework
- Hit finding
  - If statements and fitting
  - Reduces output by x10-100
- Calibration
  - Some calibration samples are very large
- Pattern recognition
  - ML algorithms
- Analysis



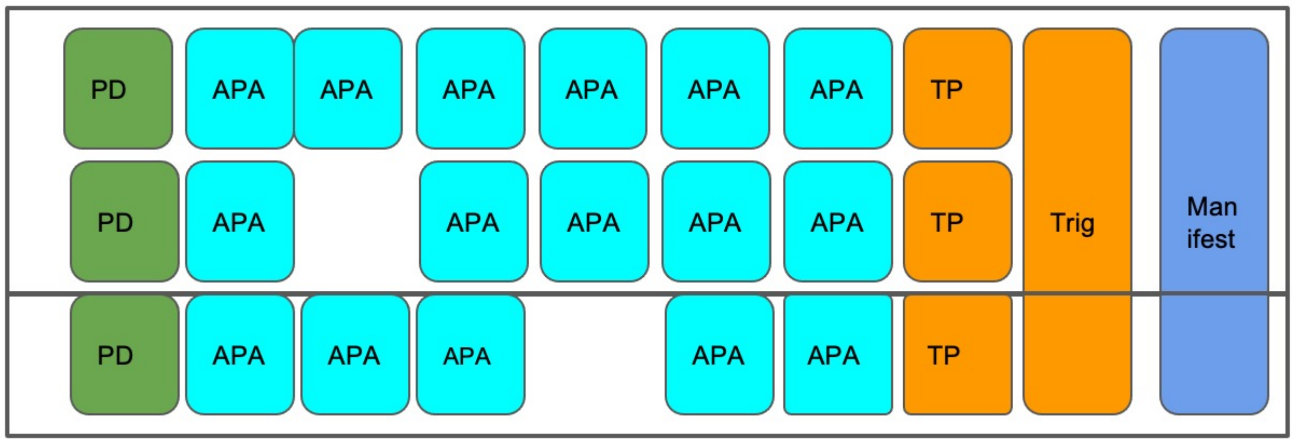
Frameworks and data structure

- PD and FD "events" may be GB scale.
  - Split to 30 MB APA chunks in memory
- Supernova are ~100TB scale
  - Need to overlap time segments offline or online
  - May be able to analyze interactions independently without "event" building.

Localized readout aggregate (cosmics/beam)



Extended (SNB) readout aggregate



LBNF/DUNE

HSF comments on novel requirements

- Multi-node Processing
  - This is new for HEP. Other event frameworks go through many "small" events and do not have need to break the event over multiple nodes
    - We have small number of HUGE events and computational scaling pushes towards multi-node processing especially at HPC centers
  - MPI for data/memory and reduction.
- Overlapping processing atoms
  - Subsetting the data at the framework level is "novel"
    - Example: splitting by APA or down to interaction candidate level
- Fluidity in data processing hierarchy
  - Anything that is not run/subrun/event is new (so time series like data and interactions within drifts are some what problematic from current frameworks)

Development: Frameworks review by HSF

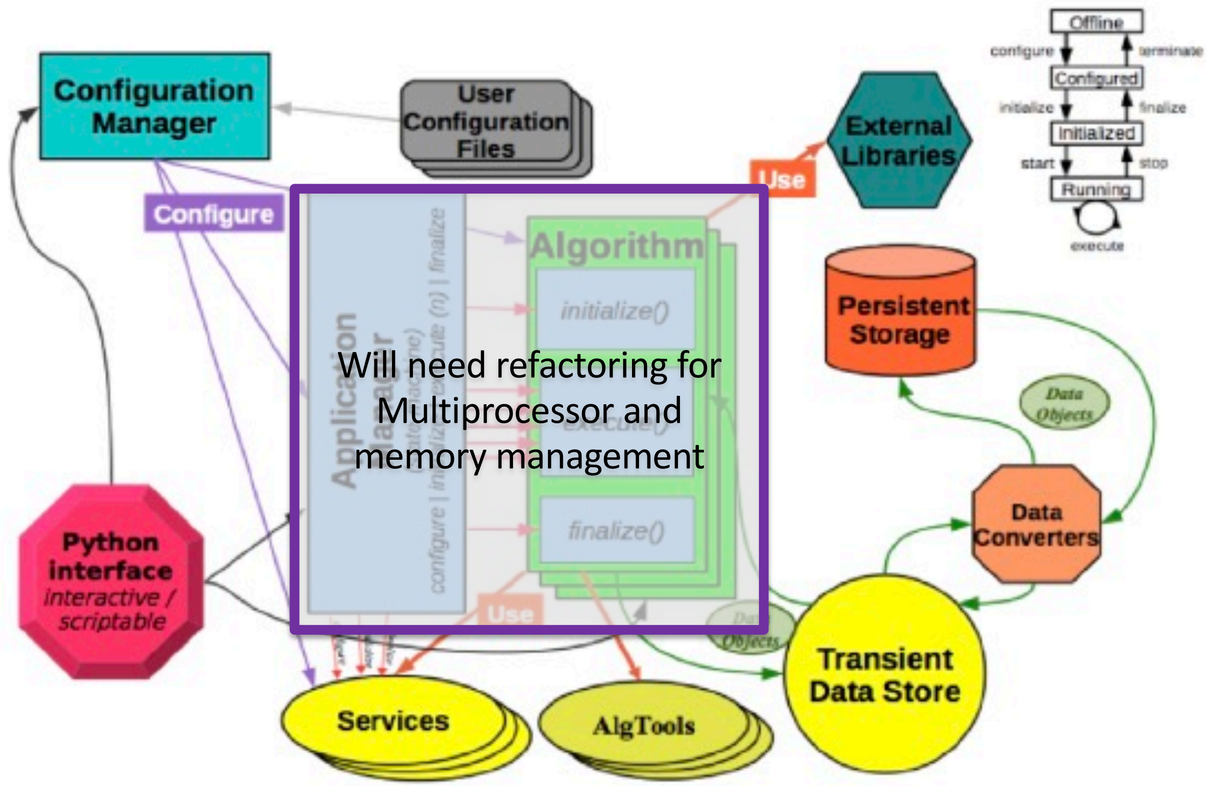
Unique DUNE requirements:

TPC/PD Data are simple on small scales → HPC

But 6 GB-100 TB readouts drive memory management requirements:

- Separate persistency/transient
- Precise tracking of provenance so parts can be reassembled
- multi-threading
- coherent processing across multiple architectures/sites

Can existing frameworks be modified to meet requirements?
Are reconstruction and analysis frameworks the same? (probably no)



Vertical Drift technology implications for computing

- Similar (or 32% larger for 3D) TPC channel count, smaller PD channel count
- Longer drift → more time samples
- Otherwise algorithmically similar to APA technology
- CRP readout already integrated into reconstruction chain.
- For now we are assuming data sizes will be similar to the HD modules.

Near Detector

- More similar to other HEP experiments – many detector systems, normal sized events
- About to start large simulation production
- Not using the LarSoft framework yet

Development example: Databases

Now: Need substantial updates for PDUNE II to incorporate conditions/calibrations cleanly

- ✓ Beam database
- ✓ Hardware database for FD
  - Already needed yesterday
  - Going into production now….

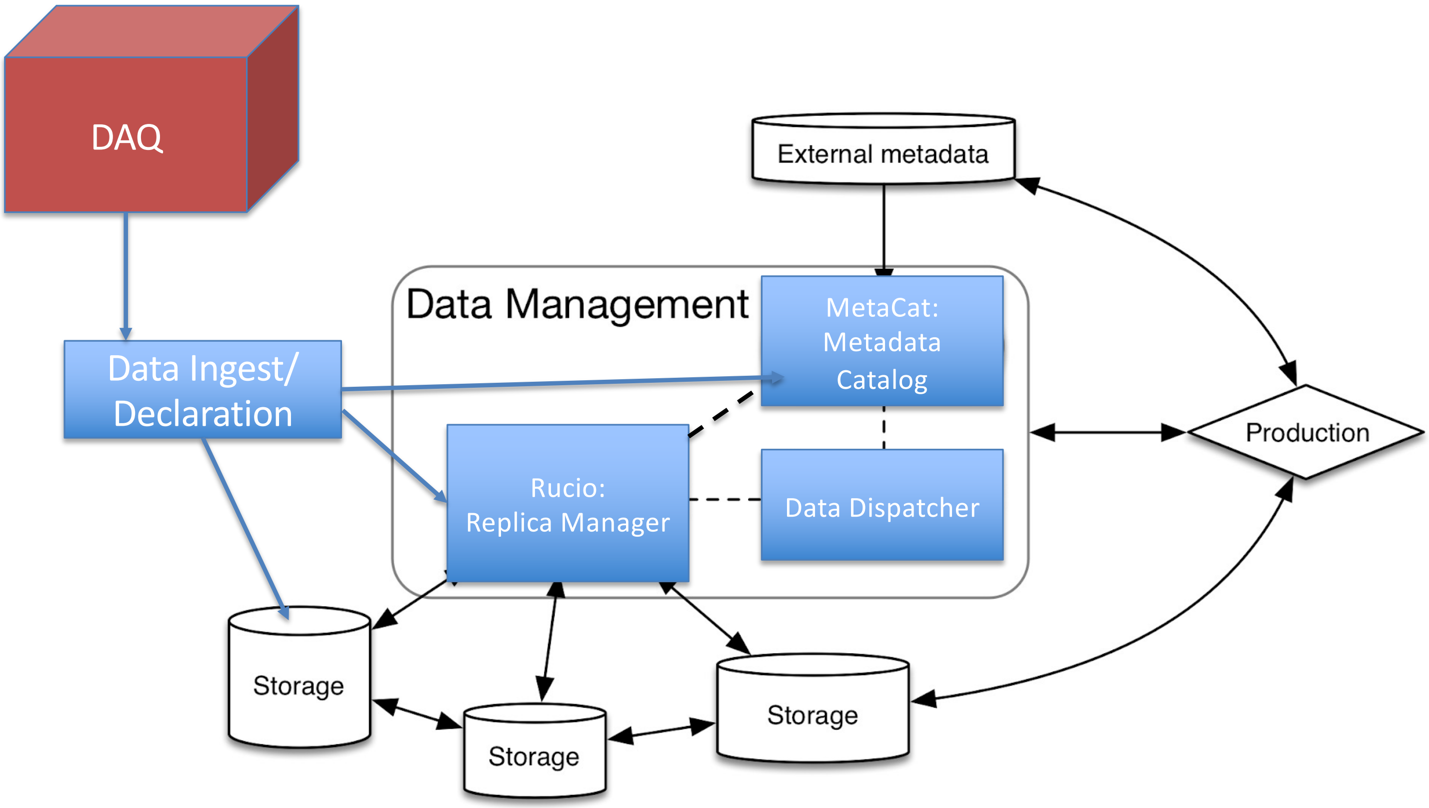
2021-2024: ProtoDUNE-II run and analysis

- ❑ Data Catalog
  - Developing more efficient MetaCat and Data tracking db's
- ❑ Compute systems monitoring
- ❑ Conditions/Slow controls/Calibrations

2024-2027: 2nd iteration to go to full scale for DUNE

- ❑ Calibration/Conditions at full scale for ND/FD
  - FD/ND will have many more channels than existing IF DB's were built to support
  - ~400,000 channels/FD module, many more for ND.
  - needs significant effort early on for design
  - will need significant horsepower to serve information to 10,000 cores worldwide.

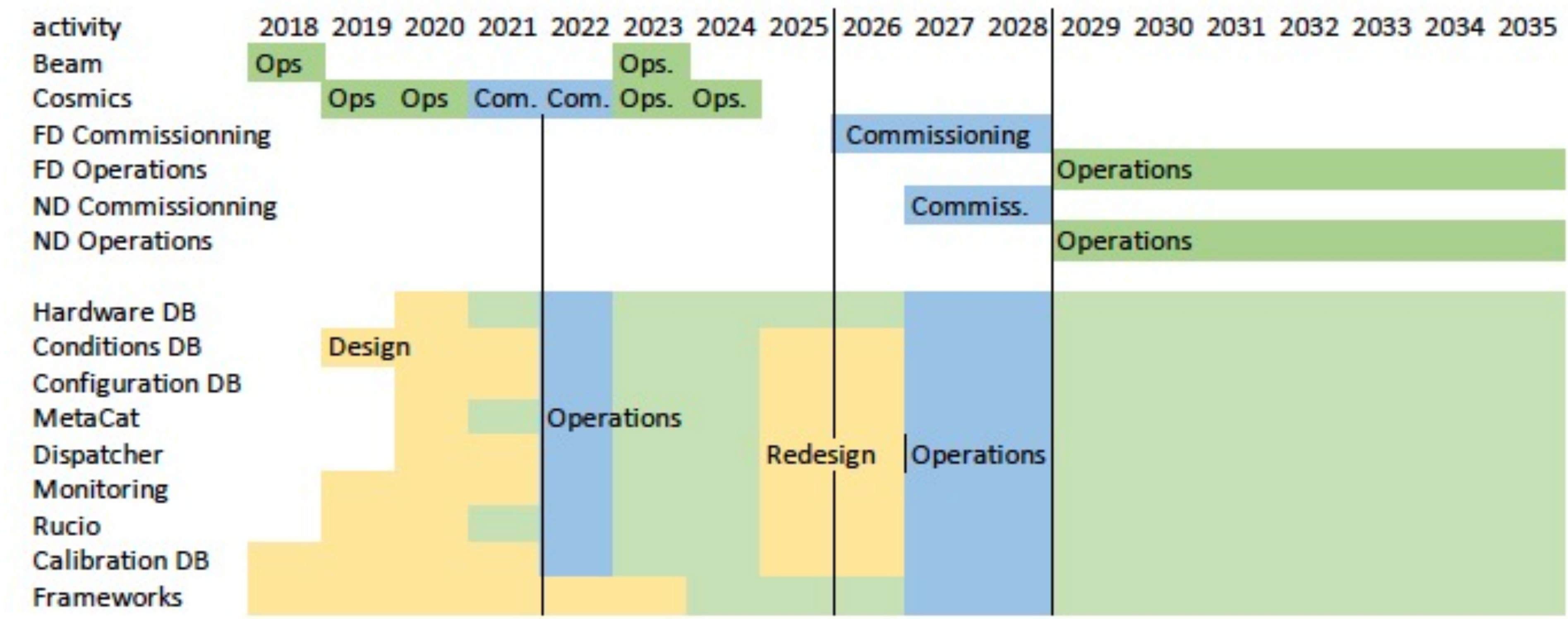
DUNE Data Management System



DUNE Data Management Projects

- **Development:**
  - **Rucio** Logical to Physical File mapping for Tape sites (non-deterministic) [James Perry, Edinburgh]
  - **MetaCat** moving towards deployment. [Igor Mandrichenko, FNAL]
  - **Data Dispatcher**—rework of SAM project functionality—[Brandon White, FNAL] Start summer 2021
  - **Data Ingest Daemon**—(Rework of FTS-Light functionality) - Start Fall 2021
  - **Data Transfer Daemon**--(Rework of FTS functionality to declare to Rucio/Metacat) – Start Fall 2021
- **Operations:**
  - Backloading all data into Rucio—[S. Timm, FNAL, Edinburg] – In progress
  - CTA testing @ CERN—[Wenlong Yuan, Edinburg] -- In progress
  - Rucio daily testing—[S. Timm, FNAL + Oregon State CS students] – In progress
  - Rucio transfer speed monitoring / Sam-Xrootd monitoring [Oregon State students]
  - Deployment of production OKD-based Rucio Server. [B. White, FNAL]

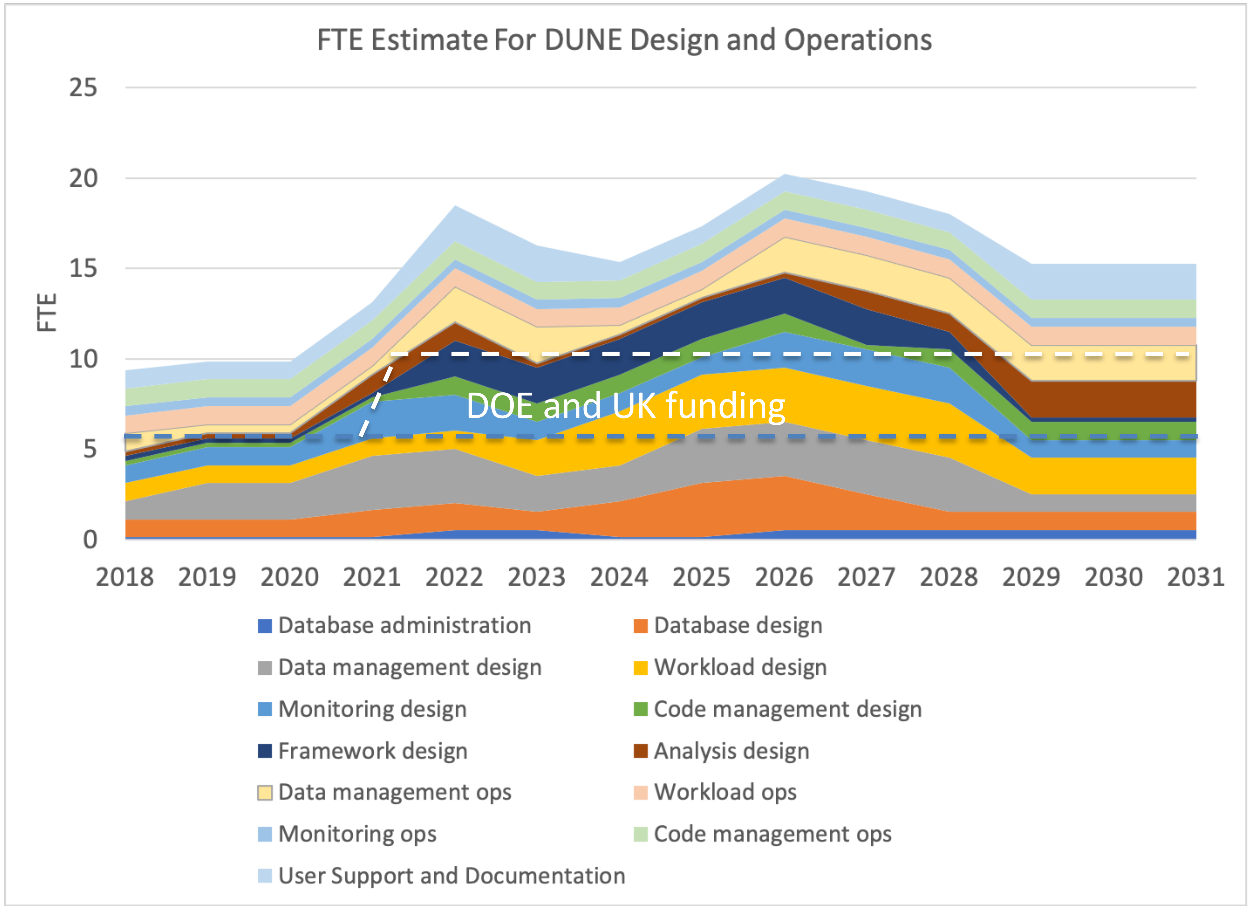
Development: Rough timeline



LBNF/DUNE

FTE estimate. Does not include shared facility (storage etc.) costs

- Some effort (mainly operations – pastels at top) can be trained collaboration physicists.
- Rest requires experts
- Currently have around 5 FTE experts (FNAL + collab), all in-kind contributions except UK DUNE funded personnel.
- Expert need is greatest for ProtoDUNE 2 and pre-operations in 2024-2028. 5-10 FTE > 50% US



Conclusions

Significant collaboration contributions to hardware and development effort have been identified

Storage contributions are high priority

More expert effort for development is needed for protoDUNE II in 2021-2022 and pre-operations starting in 2024-2027

Title	Platform for editing	Docdb for reference
DUNE Software Framework Requirements Taskforce Report	DocDB	DocDB
Near Detector Data Model	Overleaf	?
Data Tracking	GoogleDocs	?
Metadata Catalog requirements	GoogleDocs	?
ESNET report	docdb-20816	docdb-20816
Database description and definitions	Overleaf	?
Database hardware database requirements	Overleaf	?
Sites and Centres Model	GoogleDocs	Docdb-22984
Computing CDR	Overleaf (R only) (R/W)	?
DAQ/Computing Metadata	GoogleDocs	Docdb 22983
MetaCat Documentation	html	?
Raw Data Decoder Requirements/Spec	GoogleDocs	?