Fermilab and OSG: Perspectives from neutrino, muon, and astronomy experiments

Ken Herner, Alex Himmel, Robert Illingworth, Mike Kirby

September 3, 2020

The non-CMS Landscape at Fermilab

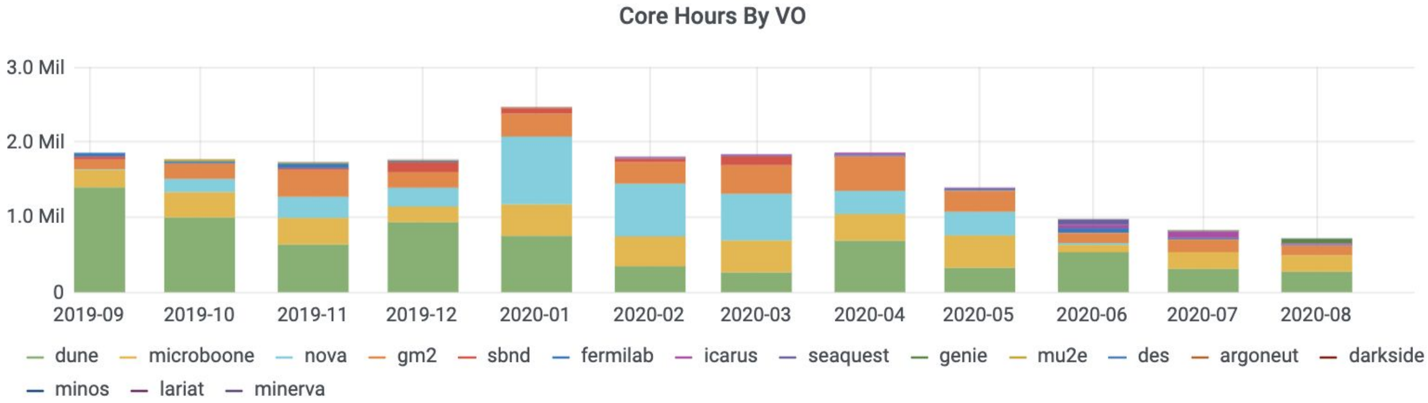
- Fermilab is the host laboratory for DUNE
- Hosts a number of other neutrino, precision muon, and astrophysics experiments
- The FIFE Project aims to provide a common, modular toolset for these experiments (and others who wish to use them, e.g. EGP), including
  - Job submission and monitoring
  - Workflow management software
  - Data management and transfer tools
  - Continuous Integration service
- Regular users of StashCache (see backup)

Some experiments currently using FIFE tools

FABRIC FOR FRONTIER EXPERIMENTS

Remote resources via the GWMS factory

- Like CMS, most FNAL experiments use GlideinWMS for job access
  - Shared pool for non-CMS experiments (DUNE may create its own global pool soon)
- Major success in 2019-2020 has been the automatic Singularity rollout
  - Invoked via GlideinWMS, users only need set the SingularityImage classad if they don't want default image (the default is made to look like FNAL WNs)
  - Removes the major impediment to user jobs outside of FNAL: worry about missing/incompatible system libraries



Payload hours outside of Fermilab, past 1 year

The Short-Baseline Neutrino Program

Important precursor to DUNE

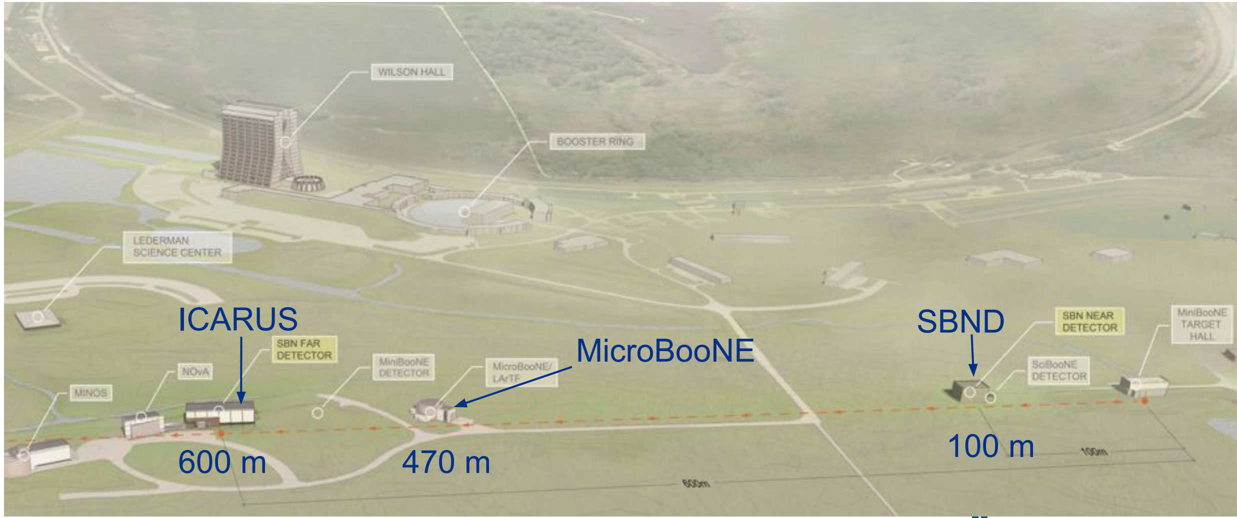
Consists of **MicroBooNE** (running), **ICARUS** (commissioning), **SBND** (late 2020); all LArTPC

Similar near-far detector concepts as NOvA/DUNE

ICARUS very interested in adopting Rucio ASAP

Common challenges:

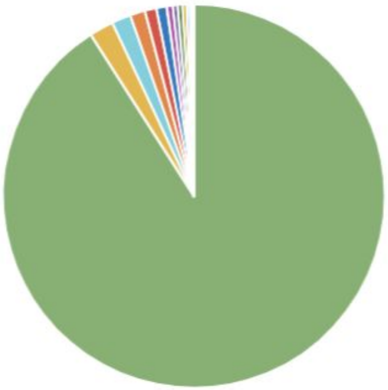
- Memory footprint
- Learning how to effectively use multiple cores =>
- Effectively using HPC resources
- Storage/data model (all are multi-PB experiments)



The Fermilab Muon Program

g-2: Probe hints of anomalous magnetic moment of the muon

Currently taking data



	total
FermiGrid	17.5 Mil
UColorado_HEP	395 K
UKI-NORTHGRID-LIV-HEP	311 K
UKI-NORTHGRID-LANCS-HEP	244 K
UKI-NORTHGRID-MAN-HEP	190 K
Nebraska-Lincoln	167 K
ICC-SLATE-HTC	95.7 K
NWICG_NDCMS	82.1 K
GLOW	77.1 K
WSU - GRID_ce2	67.8 K
MWT2 ATLAS UC	47.1 K
INFN-PISA	30.5 K
Nebraska-CMS	20.4 K
New Mexico State Discovery	9.73 K
Clemson-Palmetto	4.97 K
AGLT2	1.290 K
Caltech CMS Tier2	973
T3_US_NERSC	923
SU ITS	227
Nebraska-Omaha	12

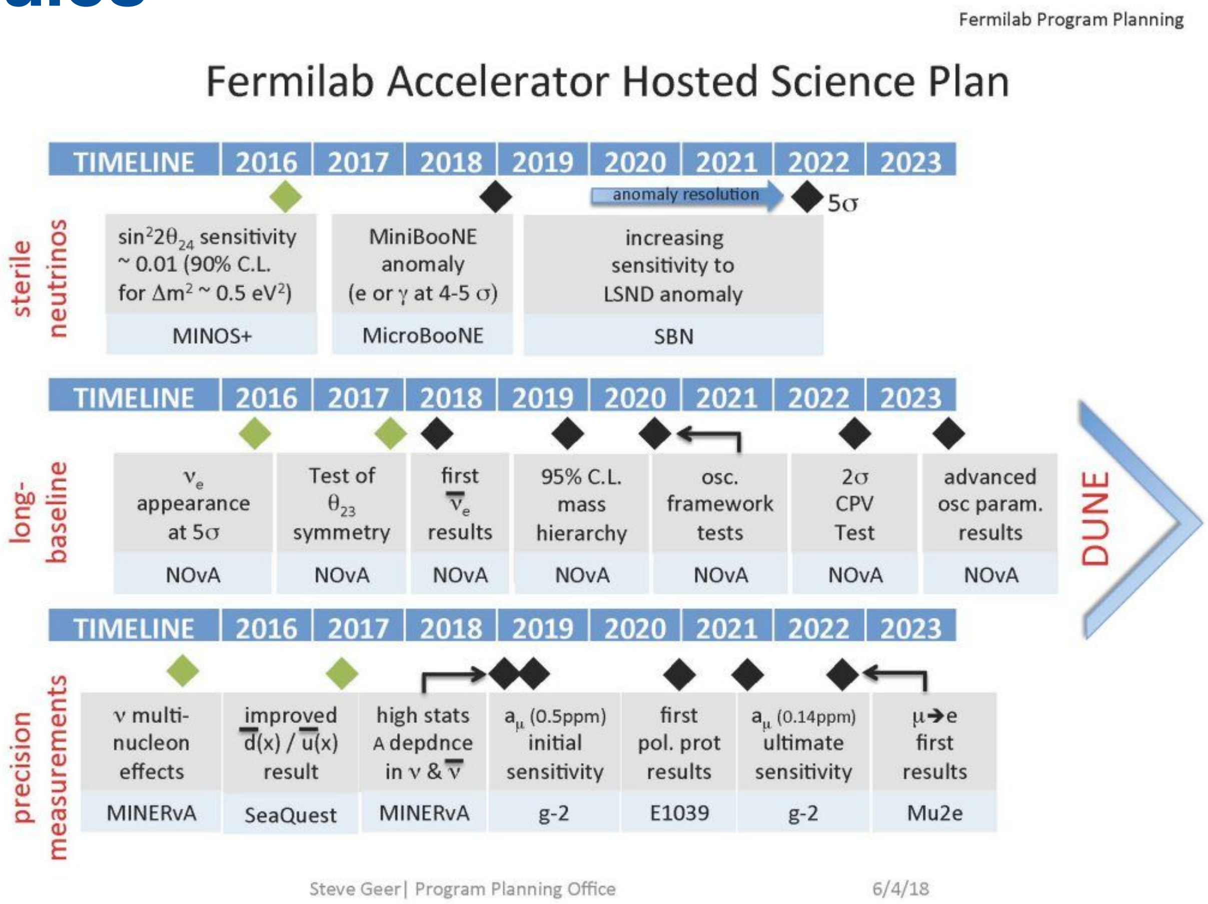
Mu2e: Look for muon -> electron decay (lepton flavor violation)

Design and construction continuing

One of the earlier experiments to get onto OSG (50 M+ hours in 2015-2016)

Experiment schedules

- Significant overlap between running experiments in 2019-21
- Additional challenges include DUNE Near Detector design/construction

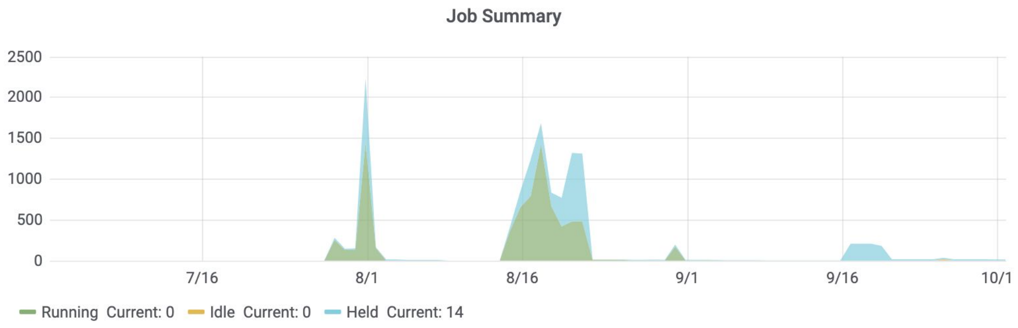


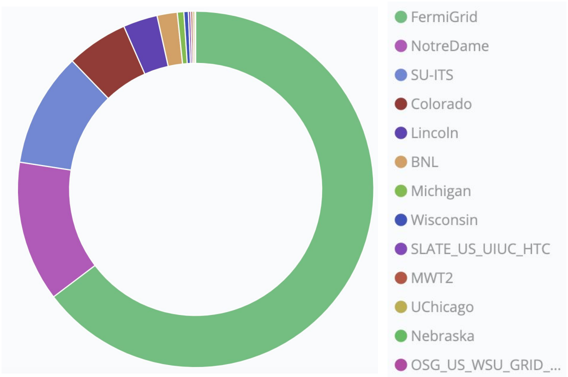
Astronomy and Astrophysics

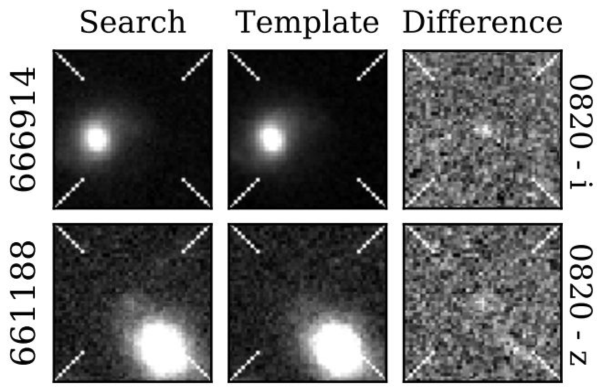
Dark Energy Survey- Gravitational Waves get important OSG resources

Search for optical counterparts to GW events: low demand most of the time, but rapid provisioning is the key (i.e. access to many sites)

Restructuring pipeline over next 1-2 years (was SL6/Python 2, so Singularity was important)



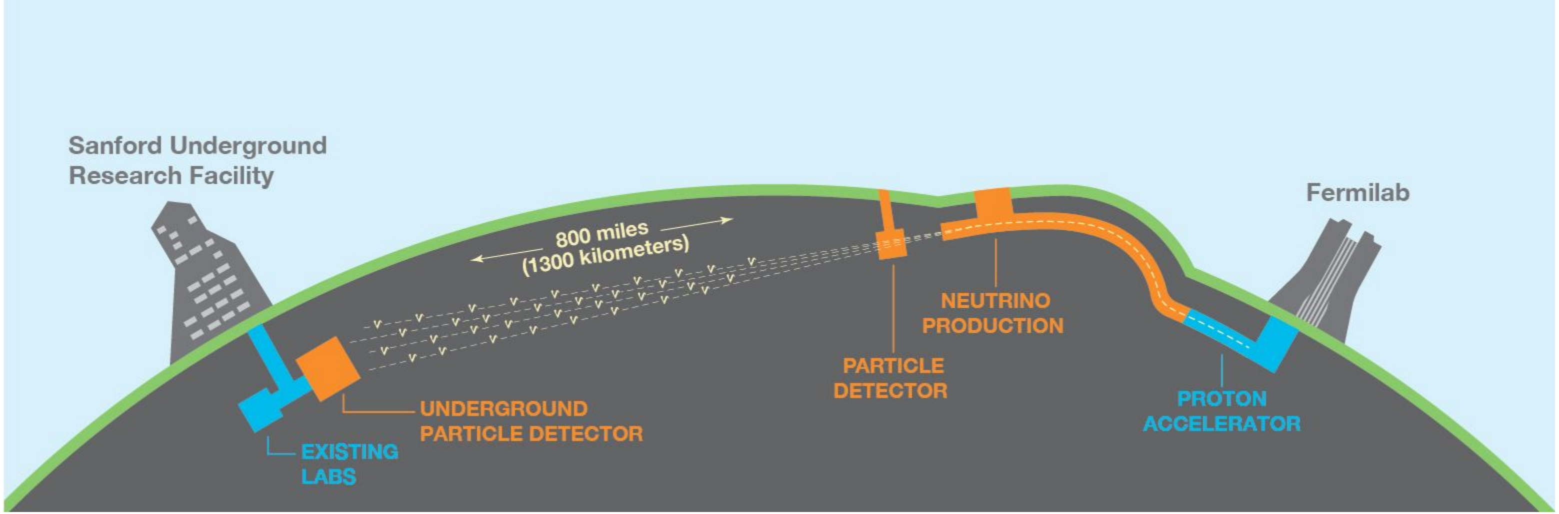




Astronomy and Astrophysics

Coming Later in 2020:
Supernova studies for DES Y3 Cosmology Analysis
(Can run embarrassingly parallel with some modifications)
OSG will provide a very important relief valve for pressure on the DES NERSC allocation

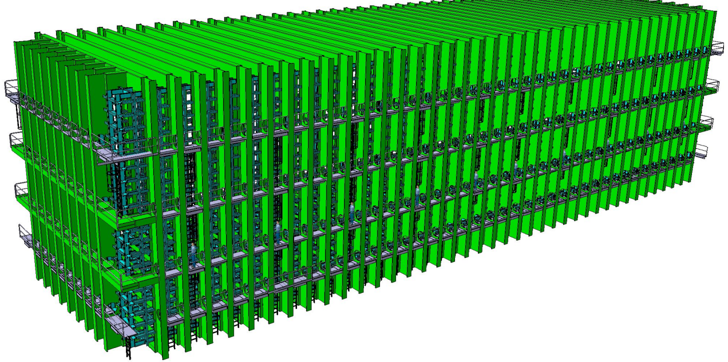
DUNE Experiment Overview



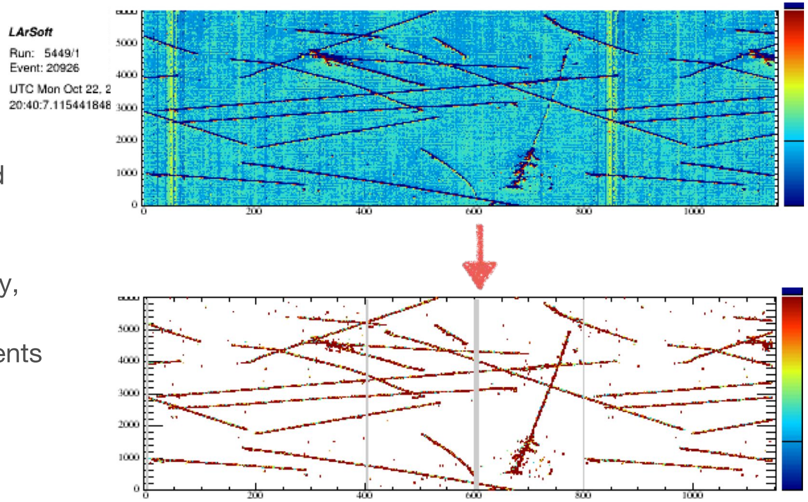
- neutrino experiment measuring neutrino oscillation parameters (mass ordering, matter vs antimatter asymmetry, unitarity), proton decay, supernova neutrinos, and more.
- Far Detector consists of 4 LAr TPC modules at 4850 ft underground in Lead, SD (SURF)
- Near Detector (proposed design) at Fermilab near the neutrino production
- baseline of 1300 km and neutrino beam optimized for oscillation measurement sensitivity
- Two prototypes at CERN - (ProtoDUNE Single Phase - ProtoDUNE Dual Phase)

DUNE Far Detector Design

Far Detector SP Module



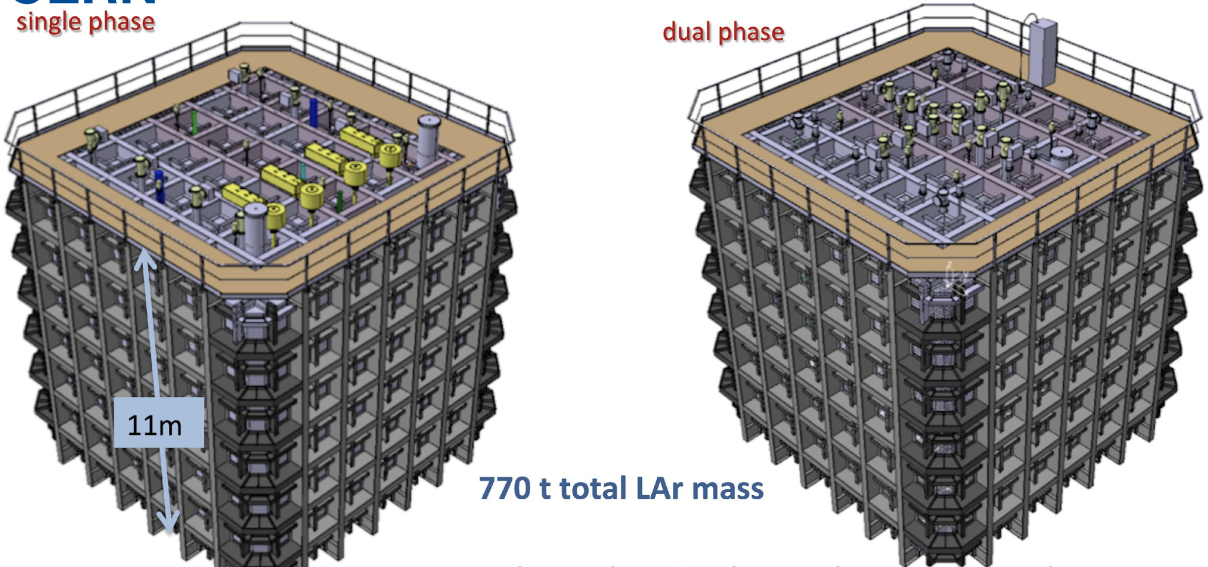
- First Far Detector build will be a single phase LAr TPC
  - 17 kT of liquid Argon
  - 150 Anode Plane Assemblies tiled on center and walls
  - 180 kV electric field across each drift volume -> 5.4 ms drift time
- Beam timing triggered readout for oscillations physics analysis
  - normal neutrino-beam trigger record is 5.4 ms
  - 12-bit ADC sampled every 0.5 μs
  - 2560 channels per APA
  - 150 APAs 6 GB uncompressed or 2-3 GB compressed
  - 5000 trigger records per day -> 5-10 PB/year/module
- time-extended readout window of far detector module varies greatly
  - continuous readout (SuperNova), calibrations, etc
  - DAQ designed with greater bandwidth, but reduced with trigger, zero suppression, and compressed data format
- goal of 30 PB/year from Far Detector dominated by calibration data
- reconstruction of signals and hits spatially independent within an Anode-Plane Assembly, but 2D deconvolution and FFT require time stitching
- processing of a single trigger record can generate multiple “events” - consider these events to be causally separable regions of interest
- creation of analysis events to minimize data volume and facilitate additional processing



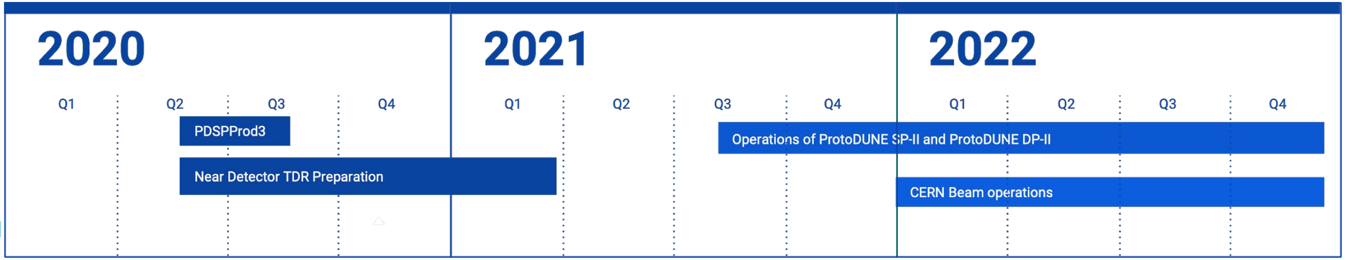
ProtoDUNE Detectors at CERN

- Construct small single-phase and dual-phase LAr TPC (~770 tonnes LAr)
- Single-phase constructed with 6 APAs (instead of 150)
- Dual-phase constructed with 4 Charge Readout Planes

Figure by Thomas Kutter

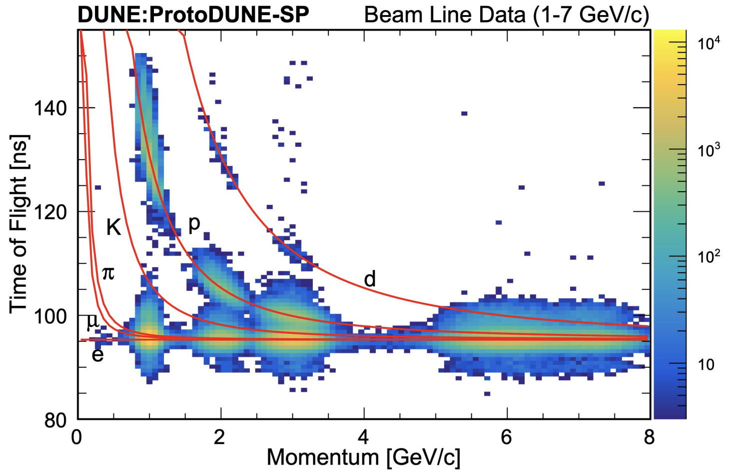


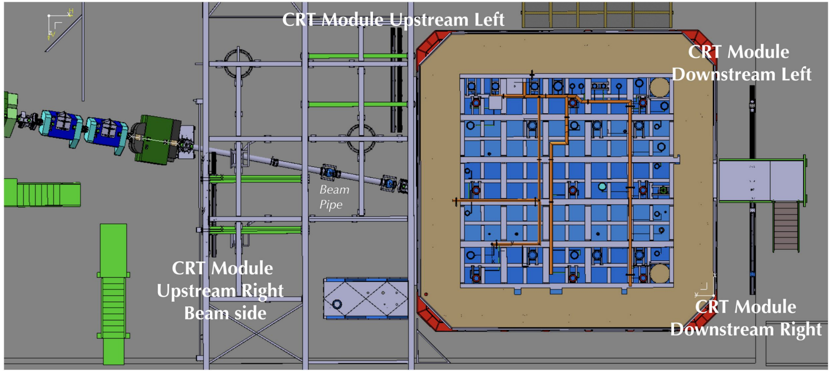
Internal: 7.9 m (Transv) x 8.5 m (Parallel) x 8.1 m (Height)
External: 10.8m (Transv) x 11.4 m (Parallel) x 11.0 m (Height).



Implementation of Computing Model with Data from ProtoDUNE - SP

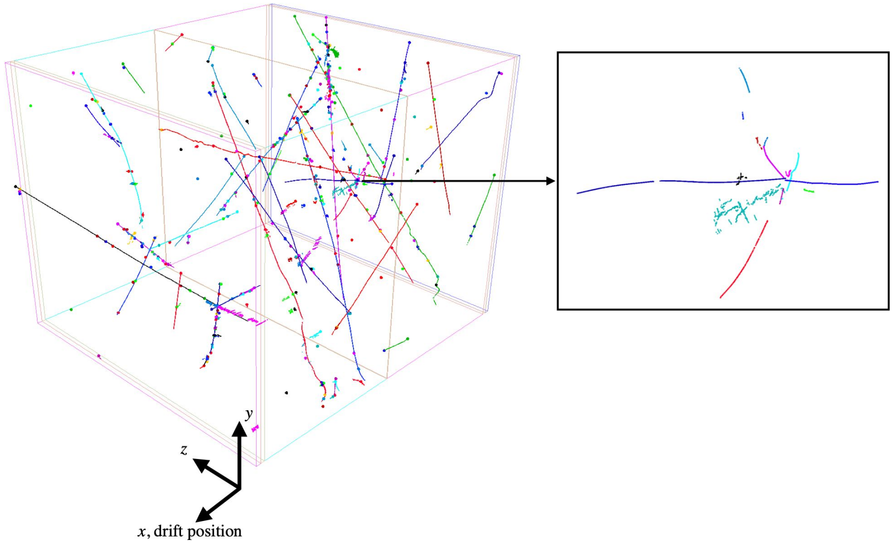
- Single Phase beam Oct - Nov 2018
- Time-of-Flight and Cherenkov tagged events
  - 300k π - Momentum 1, 2, 3, 6-7 GeV/c
  - additional e, p, K events
  - 8M total beam events - 600 TB raw data
- Additional > 50M cosmic events
  - 2 PB raw data (more data since recorded)
  - varying the purity, HV, Xenon doping
- utilize this large sample of data to test the Computing Model for DUNE
- Slides from Robert Illingworth next - “DUNE Data Management Experience with Rucio”





ProtoDUNE Reconstruction Processing

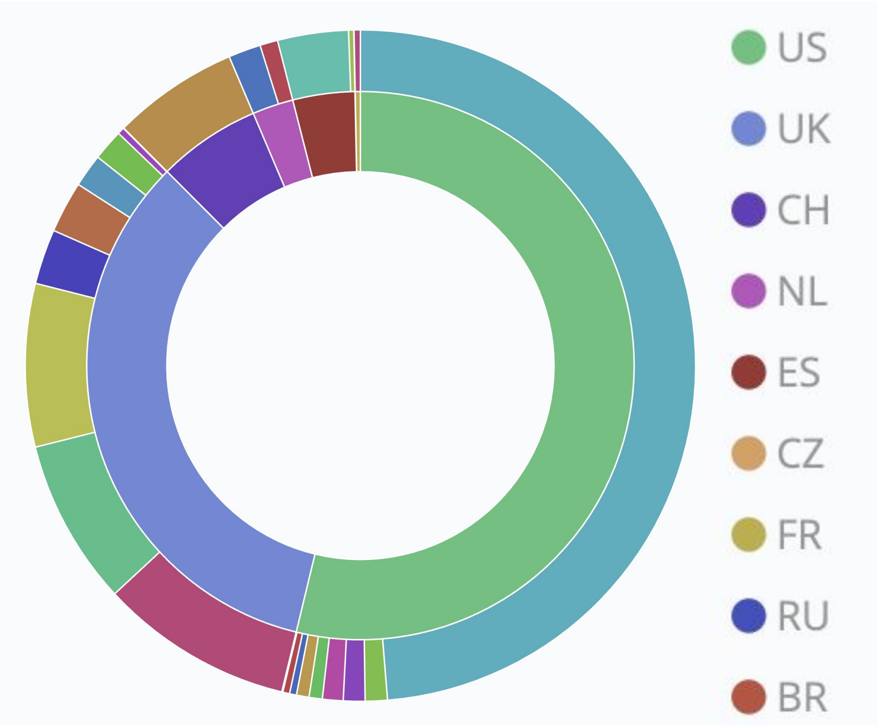
- Based on art framework and LArSoft physics software with ProtoDUNE specific modules
  - WireCell, Pandora libraries utilized extensively
- Processing of a 100 event (8 GB) file takes ~500 sec/event (80 sec/APA)
  - Signal processing is < 2 GB memory
  - Pattern recognition is < 4 GB memory
  - ProtoDUNE SP event is 75 MB/evt
- reduction to 20 MB/evt - 2 GB files
- 25 Hz trigger rate volume is equivalent to DUNE Far Detector beam-trigger stream
- "First results on ProtoDUNE-SP liquid argon time projection chamber performance from a beam test at the CERN Neutrino Platform" https://arxiv.org/abs/2007.06722



ProtoDUNE Production Processing

- DUNE doing excellent job incorporating new Compute Elements (CE) and Storage Elements (SE) using OSG and WLCG infrastructure
  - Continue to add resources from sites around the world - 36 sites
  - Addition of Storage Elements continues - 13
- Soon undertake ProtoDUNE Single Phase Production version 3 (PD-SPProd3)
- Data processing on distributed computing
  - (FNAL ~50% similar to previous usage)
- Utilizing NERSC SuperComputer Cori allocation through HEPCloud for simulation generation (10000 simultaneous jobs running ~40% of total DUNE CPU hours)
- Anticipate using 80 - 100 M CPU hours/year in during ProtoDUNE II operations

CPU Hour fraction Feb 1 - May 20, 2020

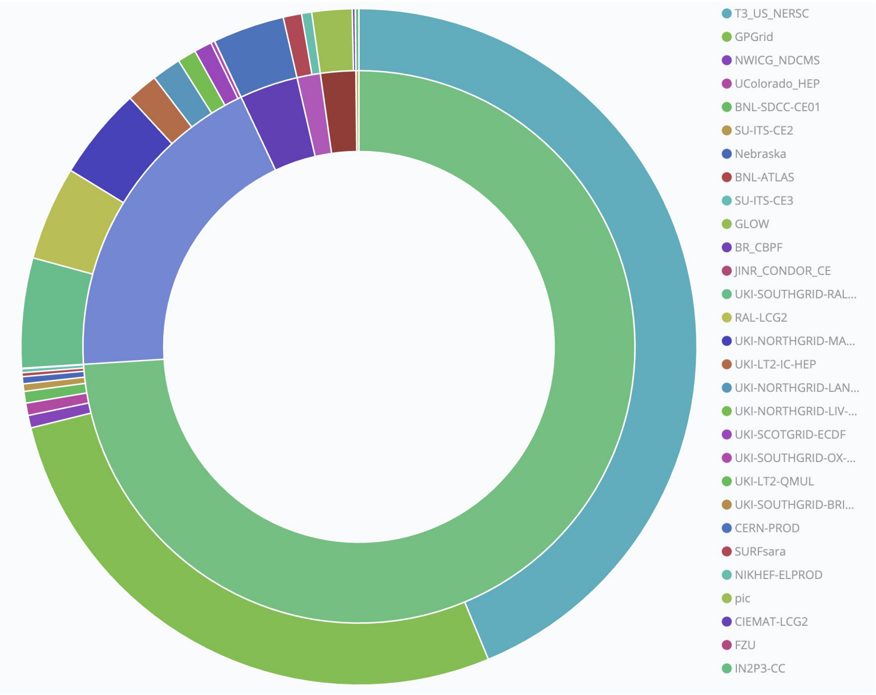


Inner Circle is country

ProtoDUNE Production Processing

- DUNE doing excellent job incorporating new Compute Elements (CE) and Storage Elements (SE) using OSG and WLCG infrastructure
  - Continue to add resources from sites around the world - 36 sites
  - Addition of Storage Elements continues - 13
- Soon undertake ProtoDUNE Single Phase Production version 3 (PD-SPProd3)
- Data processing on distributed computing
  - (FNAL ~50% similar to previous usage)
- Utilizing NERSC SuperComputer Cori allocation through HEPCloud for simulation generation (10000 simultaneous jobs running ~40% of total DUNE CPU hours)
- Anticipate using 80 - 100 M CPU hours/year in during ProtoDUNE II operations

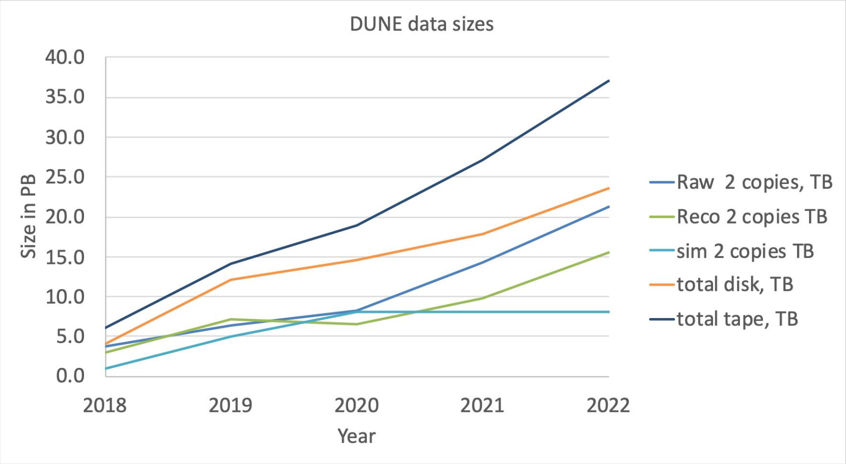
CPU Hour fraction Feb 1 - May 20, 2020



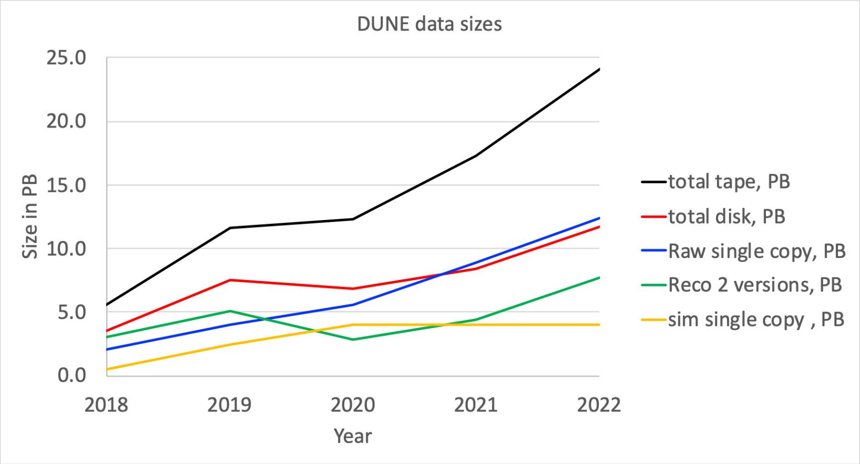
Adding in HPC @ NERSC Cori

Tape and disk storage 2018-2022

Total DUNE Storage



FNAL DUNE Storage



- Computing Model for DUNE Storage
  - 2 archival copies of raw, derived, and simulated data - 1 copy at FNAL, second copy distributed institutions
  - production processing of SP and DP data and matching simulation twice per year
  - 2 or 3 copies of active derived and simulated datasets on disk - dataset stays active for 1 year
  - late binding of between sites and jobs means that streaming of data to jobs is dominant data access pattern

DUNE Data Management Currently

- Two ProtoDUNE detectors at CERN, each 5% of full detector size
- ~10 PB of data accumulated thus far
- Roughly half raw data, half reconstruction output products, small amount of MC. Event size ~60MB
- 36 compute sites around the world
- 13 disk only sites, 4 disk+tape sites.
- Data streamed via xrootd from the closest location.
- Very similar to other experiments that use the grid.

DUNE Data Terms

SUBDETECTOR	SD_1	SD_2	SD_3	SD_4	SD_5	SD_6
Trig Record	1	1	1	1	1	1
Trig Record	2	2	2	2	2	2
Trig Record	3	3	3	3	3	3
	…	…	…	…	…	…

SUBDETECTOR	SD_1	SD_2	…	SD_{149}	SD_{150}
Trig Record	1	1		1	1
Trig Record	2	2		2	2
Trig Record	3	3		3	3
	…	…	…	…	…

ProtoDUNE 6 subdetectors
60 MB compressed

Trigger record: Output of the DAQ for a single trigger

Data Unit: 1 Subdetector for 1 trigger record

Data Object: Collection of data units

Full DUNE: 150 subdetectors
6GB/readout (5 ms)

Trigger records will be split by subdetector across many different data objects.
The MANIFEST tells us what is where.

DUNE data challenges

- DUNE far detector output 30PB per year, plus more from near detector
- Subdetectors from the same trigger record may end up in multiple files
- Time slices from the same trigger record may end up in multiple files.
- In case of supernova burst readout we will have to split both by subdetector and by time slices, and get a very large volume of data out fast.

- Exploring HDF5 format to store the data instead of root
- Exploring non-file-based object stores in general.
  - A object based approach may make more sense than the arbitrary assignment to files

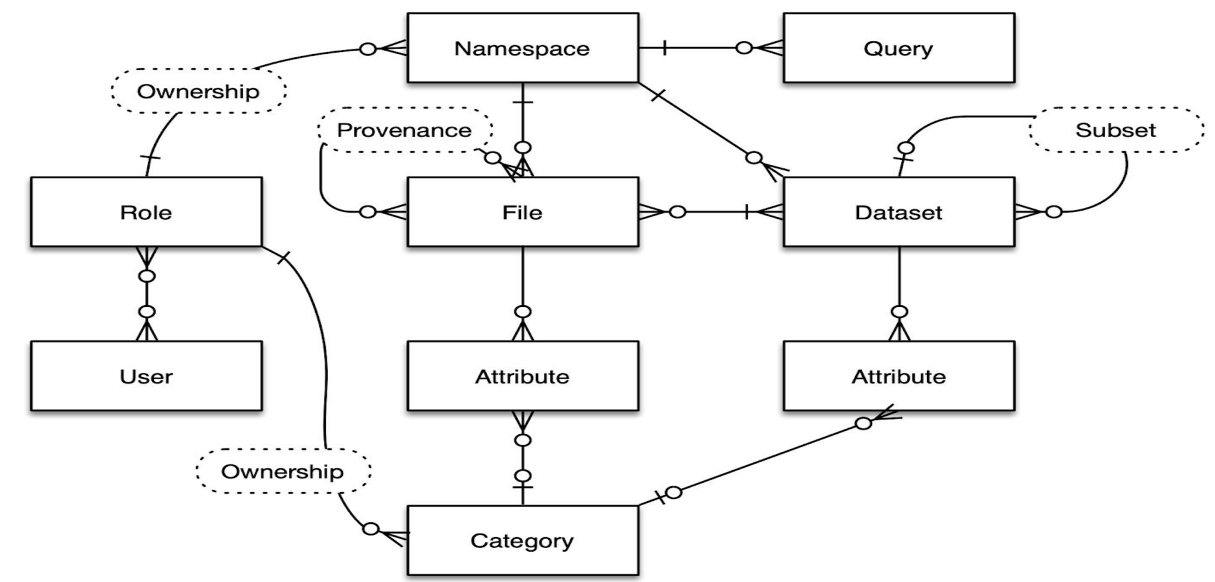
Migrating to Rucio for data management

- In the process of migrating to use Rucio for data management

- Need to replace 3 main functions of monolithic legacy system
  - Replica manager (Where is the file)-> Rucio
  - File Provenance (Metadata)
  - Data Delivery / project tracking

- Three projects needed to get there:
  - New Data Ingest service replacing legacy data transfer system.
  - New Metadata service replacing legacy metadata database
  - New Data Delivery Client service for workflow management

Metadata Server: MetaCat



Data model

- Metadata sets very similar to Rucio model
- Flexible queries and ability to combine with other databases (ie runs, beam conditions)
- Requirements are complete
- Reference implementation in testing

Monitoring and data locations

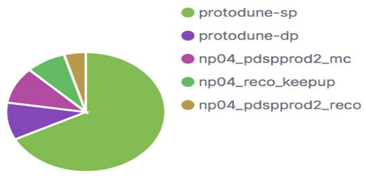
[rucio] Mock _Total dids

1,379,448 DIDs | 5.8PB Total bytes

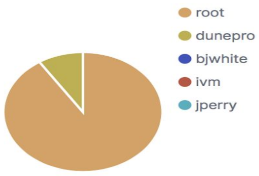
[rucio] total replicas

3,097,440 Total replicas | 13PB Total bytes

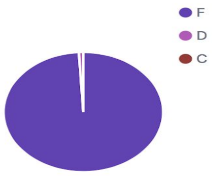
[rucio] DIDs per scope



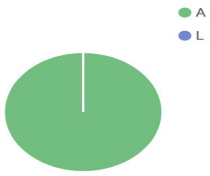
[rucio] DIDs per account



[rucio] DIDs per did type



[rucio] DIDs per availability

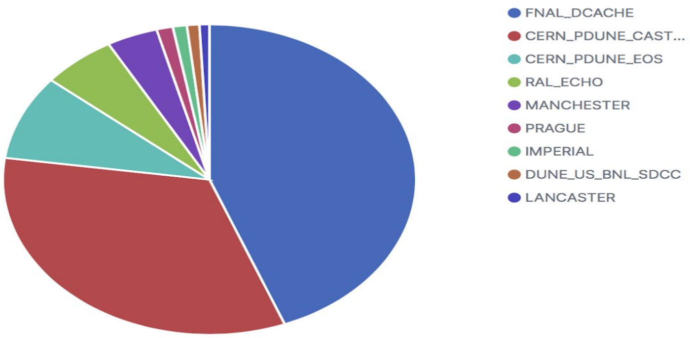


[rucio] Replicas per site

FNAL_DCACHE	1,364,606	5.8PB
CERN_PDUNE_CASTOR	1,027,176	4.4PB
CERN_PDUNE_EOS	273,658	870.8TB
RAL_ECHO	179,633	844.9TB
MANCHESTER	123,689	463.9TB
PRAGUE	39,558	272.7TB
IMPERIAL	33,894	243TB
DUNE_US_BNL_SDCC	30,148	66.7TB
LANCASTER	24,045	122.3TB

Export: Raw Formatted

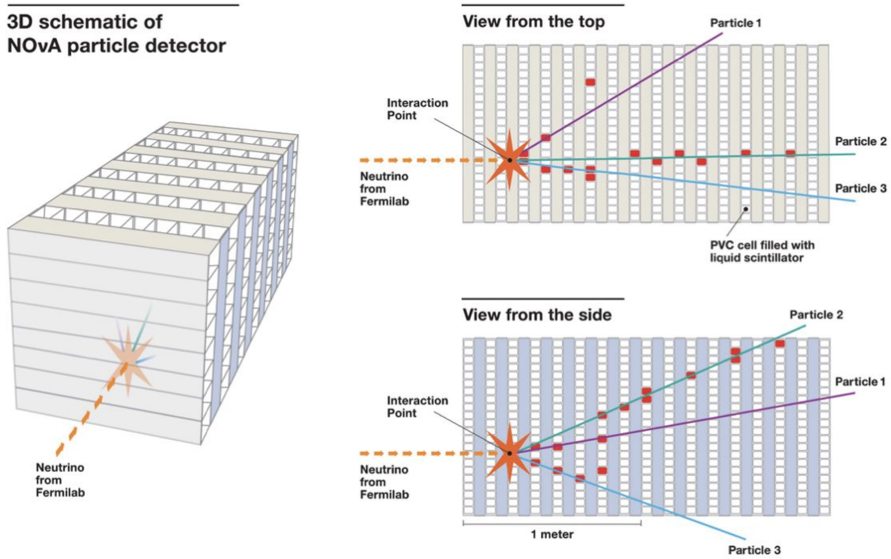
[rucio] Replicas pie



Fermilab

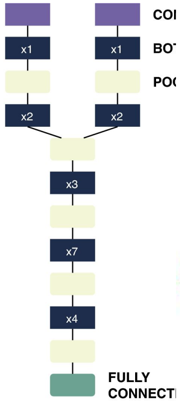
The NOvA Experiment

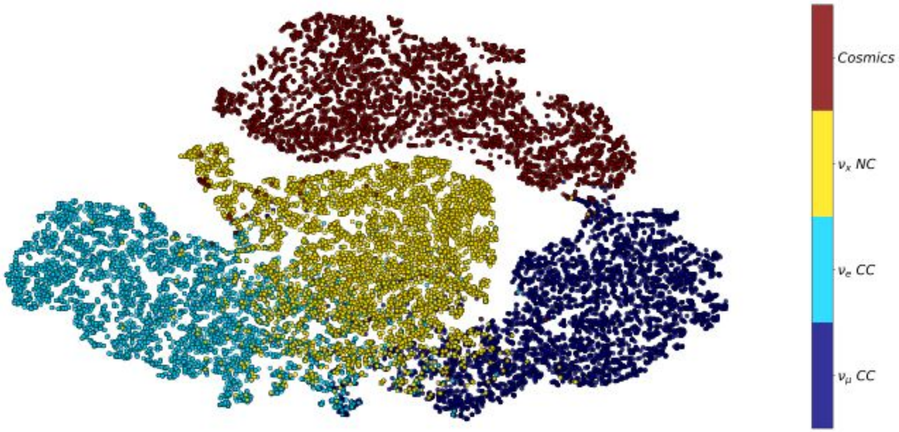
- NOvA is a long-baseline neutrino experiment, measuring oscillations, cross section, and exotic physics.
- NOvA detectors are homogenous, segmented, tracking calorimeters with 2 orthogonal views and few cm-scale cells.
- This detector is well-suited to many ML algorithms, and ML now plays a critical role in many NOvA analyses.

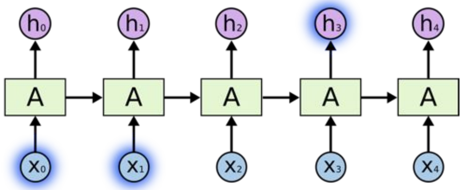


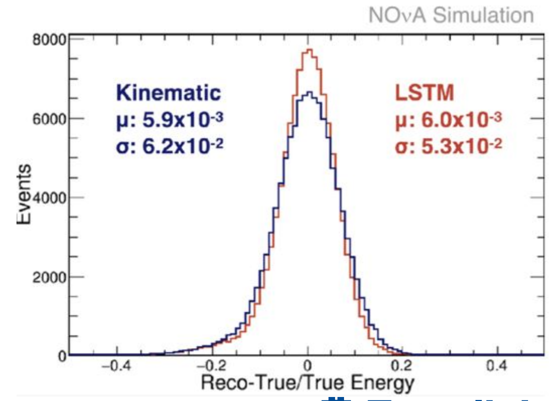
Deep Learning Applications in NOvA

- Classification with convolutional neural networks:
  - Now quite popular, NOvA was the first to apply this technique to a published physics measurement.
  - One network is used to identify the flavor of the neutrino which interacted.
  - Other networks are designed to identify individual particles in the final state.
- “Offline L1” cosmic rejection with a CNN
  - Use just whether cells are hit or not, independent of reconstruction and calibration which change over time.
- Energy estimation with multiple techniques:
  - Convolutional neural networks using event “images”
  - Recurrent neural networks (LSTMs) using a variable number of reconstructed objects.









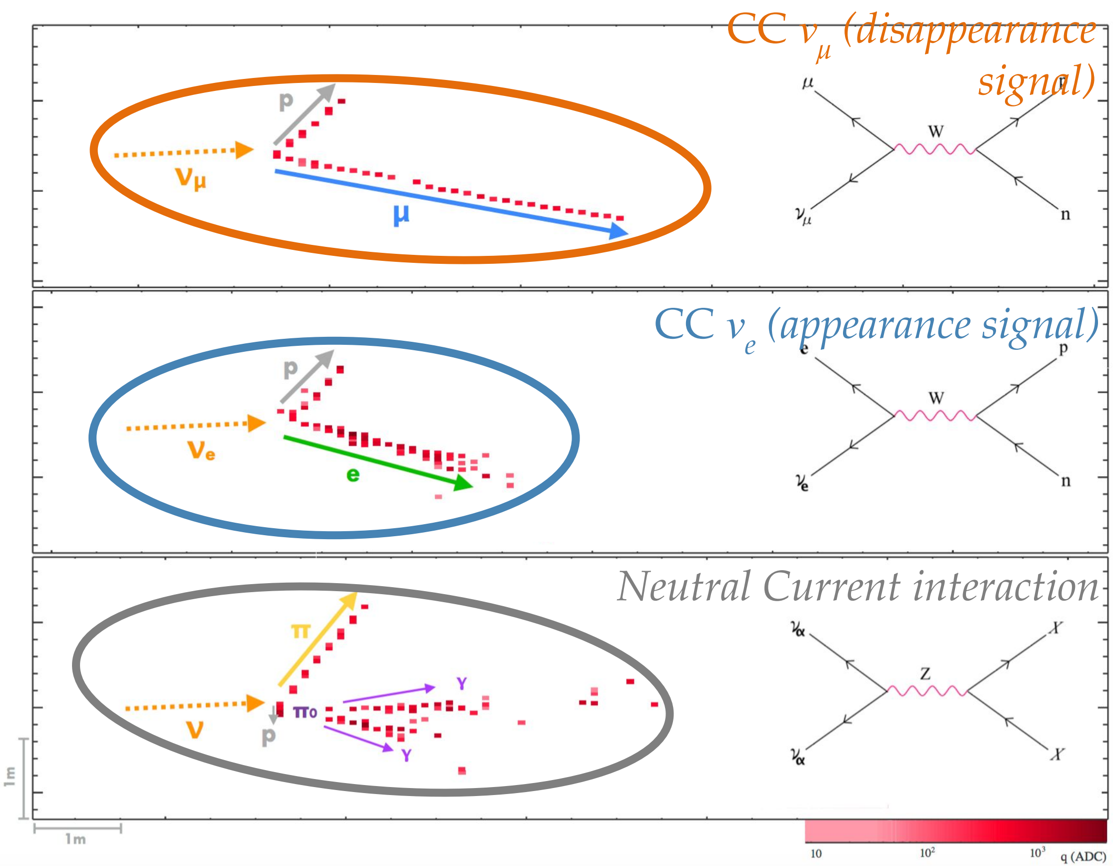
A Convolutional Neural Network Neutrino Event Classifier: https://arxiv.org/abs/1604.01444

Context-Enriched Identification of Particles with a Convolutional Network for Neutrino Events: https://arxiv.org/abs/1906.00713

Making Deep Learning Operational

Training Networks

- For the most part, we have not found that resources for *training* are too constrained.
- We have used:
  - Fermilab's Wilson Cluster
  - ANL Machines
  - University GPU clusters (Indiana, Minnesota)
- Those resources all had additional hoops relative to the OSG…
  - Special account requests outside standard Fermilab computing.
  - Different architectures and disk systems which required some additional training.
- ...but this is not a problem because training is a discrete task handled by a small number of individuals.



Fermilab

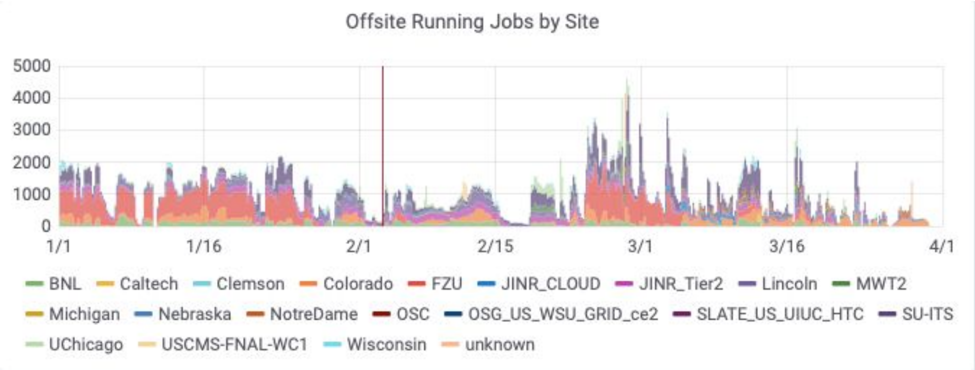
Making Deep Learning Operational

Running deep learning as a part of standard reconstruction:

- Needed to run as part of the same process that we used for our standard reconstruction.
  - Too large a scale to re-invent production infrastructure for new sites.
  - Existing production team needed to be able to run.

- The big challenge: evaluation time.
  - Our first CNNs were extremely computationally intensive (>1/2 of our total reconstruction time).
  - However, it was inefficient to try to use limited GPU resources since there are many more CPU.
  - Slow speeds limited our ability to validate and to take risks on interesting new ideas.

- Our most recent development prioritized CPU evaluation speed.
  - Change frameworks (Caffe to TensorFlow) gave us dramatic speed-ups.
  - Moved to MobileNet v2, which gave equivalent performance but much faster evaluation time.
  - Overall, sped up reconstruction by an order of magnitude.

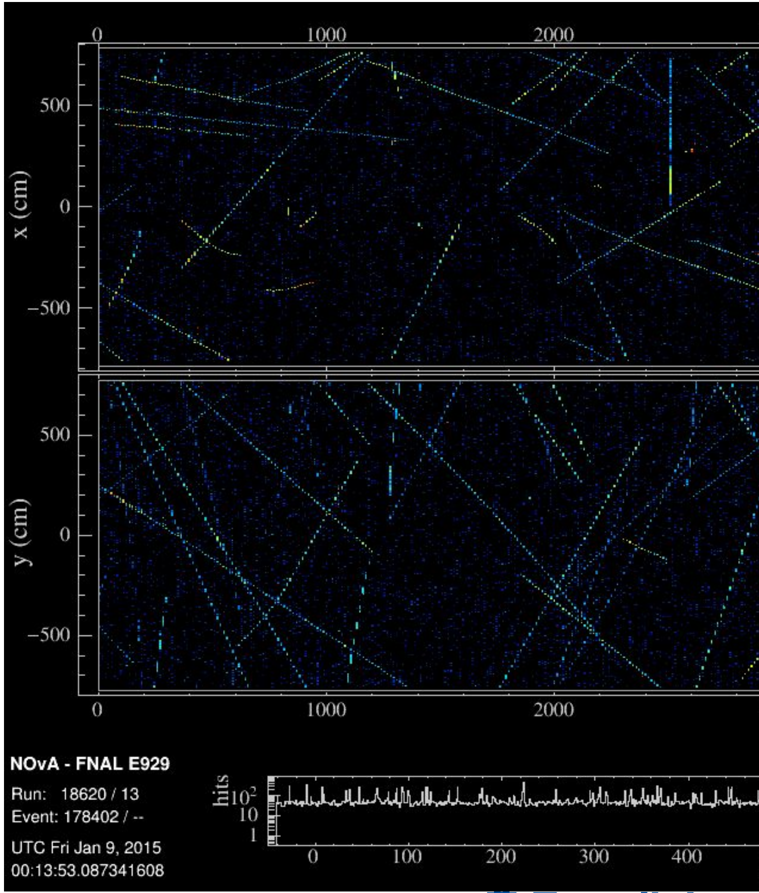


Fermilab

Making Deep Learning Operational

Running with OSG GPUs:

- This only made sense to try with our cosmic rejection network:
  - This is a workflow that is almost entirely ML, rather than having a large "standard reco" part.
  - The faster networks like MobileNet did not work well for this application which works with larger images.
- When we ran, we could only get access to 2 sites:
  - Syracuse had high availability of older, slower nodes
  - Nebraska had low availability of newer, faster nodes
  - We have since gotten UCSD working, but not exercised it extensively.
- However, availability was such that we could only process about 10% of our dataset in the 2 months available for the processing.
- When applying this filtering to our neutrino data stream, we found 0 GPU availability and needed to run filtering on CPUs.
  - The dataset was ~similar in size to the 10% of cosmics and completed in a similar amount of time due to the much higher CPU availability.
  - But, the CPUs are a "competing" resource since all our production jobs use CPUs but cosmic filtering was unique in needing GPUs.

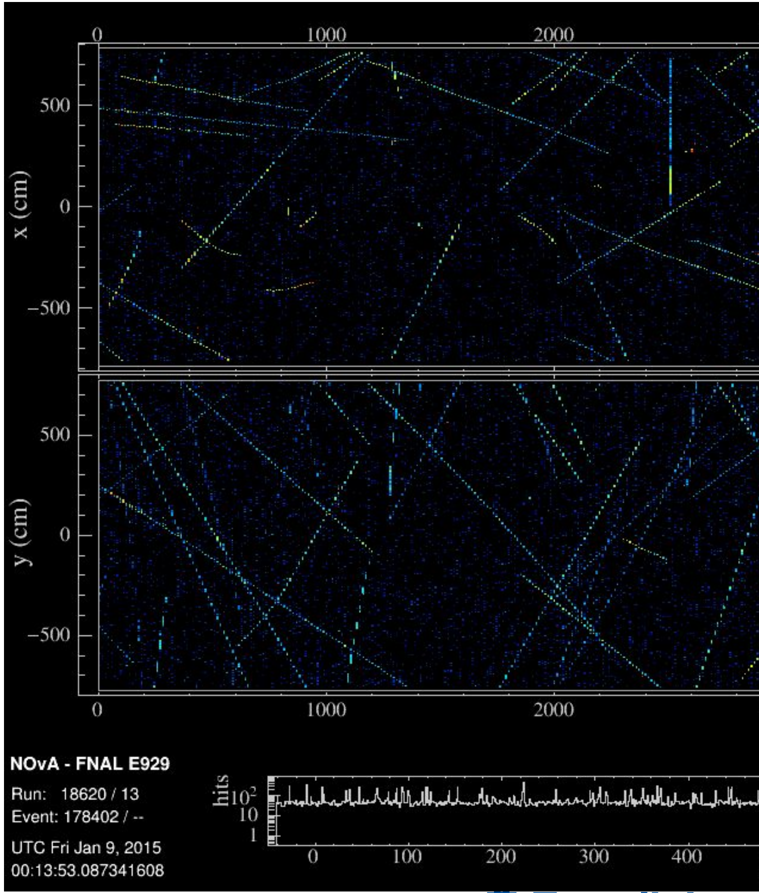


Fermilab

Making Deep Learning Operational

So, what now?

- It was pretty clear to us after this experience that we could not rely on the OSG to handle the large backprocessing.
- So, we are now exploring using HPC facilities at ANL to handle the large backprocessing.
  - HPCs with GPU accelerators were not available last year but are are now starting to come online.
  - An HPC with large burst resources is better suited to this task.
- We intend to continue using the OSG, but only for “keep-up” processing of the data as it comes in.
  - Opportunistic availability is much better suited to processing a small amount of data as it comes in.
  - In keep-up, we can also tolerate a few weeks of backlog due to low availability and then recover later.



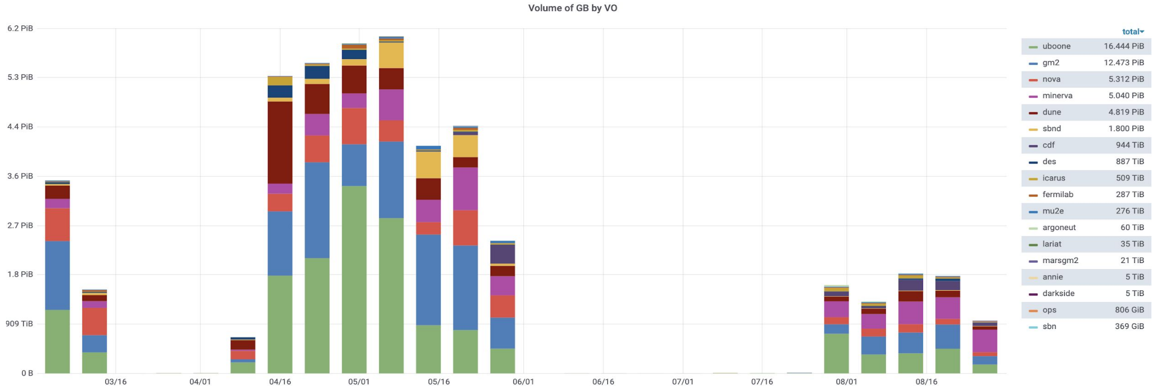
Summary

- FNAL neutrino, muon, and astro experiments greatly benefit from OSG infrastructure and expertise
- Affected by broader trends in the field (tokens, GPUs, HPC); will have to evolve accordingly

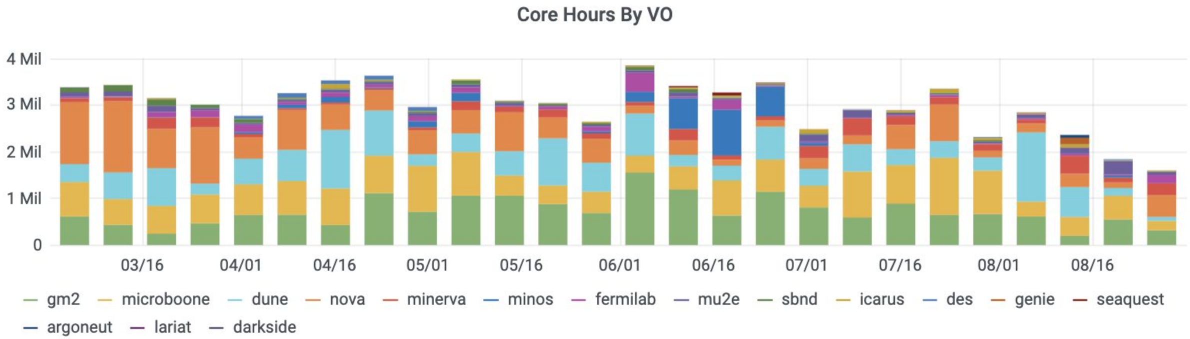
BACKUP

Supported Experiment Data and Job Volumes

- 35 experiments; more than 600 unique users
- 30+ sites in 11 countries
- About 160K jobs per day; 170M hours per year
- **Combined numbers approaching scale of LHC (6-7x smaller wrt ATLAS+CMS)**



Total weekly data transferred by experiment, last 6 months

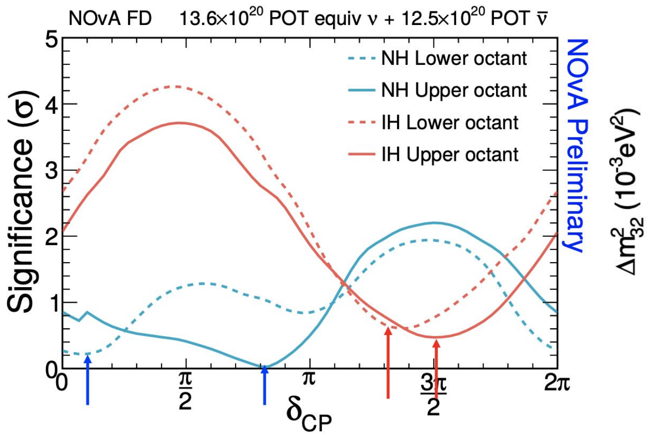


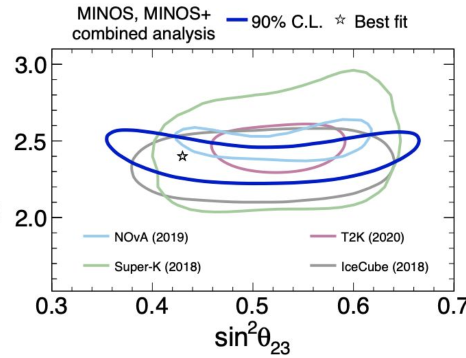
Total weekly wall time by experiment, last 6 months

Fermilab

Selected Science Results since last AHM

Stolen from roadmap talk; space to advertise results; feel free to add

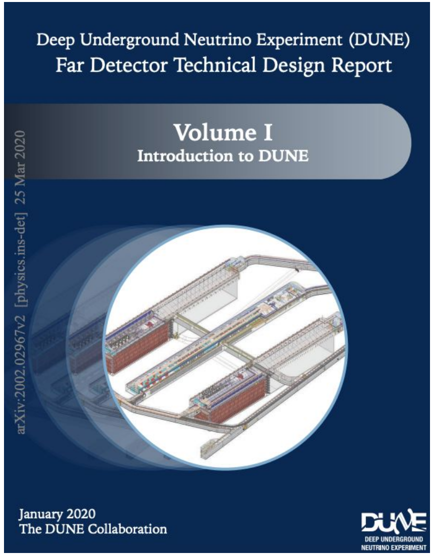




PHYSICAL REVIEW D **101,** 112006 (2020)

Search for multimessenger signals in NOvA coincident with LIGO/Virgo detections

M. A. Acero,[2] P. Adamson,[12] L. Aliaga,[12] T. Alion,[39] V. Allakhverdian,[26] N. Anfimov,[26] A. Antoshkin,[26] L. Asquith,[39] A. Aurisano,[6] A. Back,[24] C. Backhouse,[43] M. Baird,[20,39,44] N. Balashov,[26] P. Baldi,[25] B. A. Bambah,[17] S. Bashar,[42] K. Bays,[4,19] S. Bending,[43] R. Bernstein,[12] V. Bhatnagar,[32] B. Bhuyan,[14] J. Bian,[25,30] J. Blair,[16] A. C. Booth,[39] P. Bour,[9]



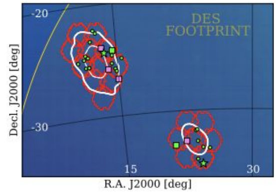


Figure 1. Summary of exposures taken and candidates identified by the DESGW pipeline. DECam pointings are shown

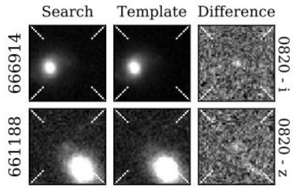


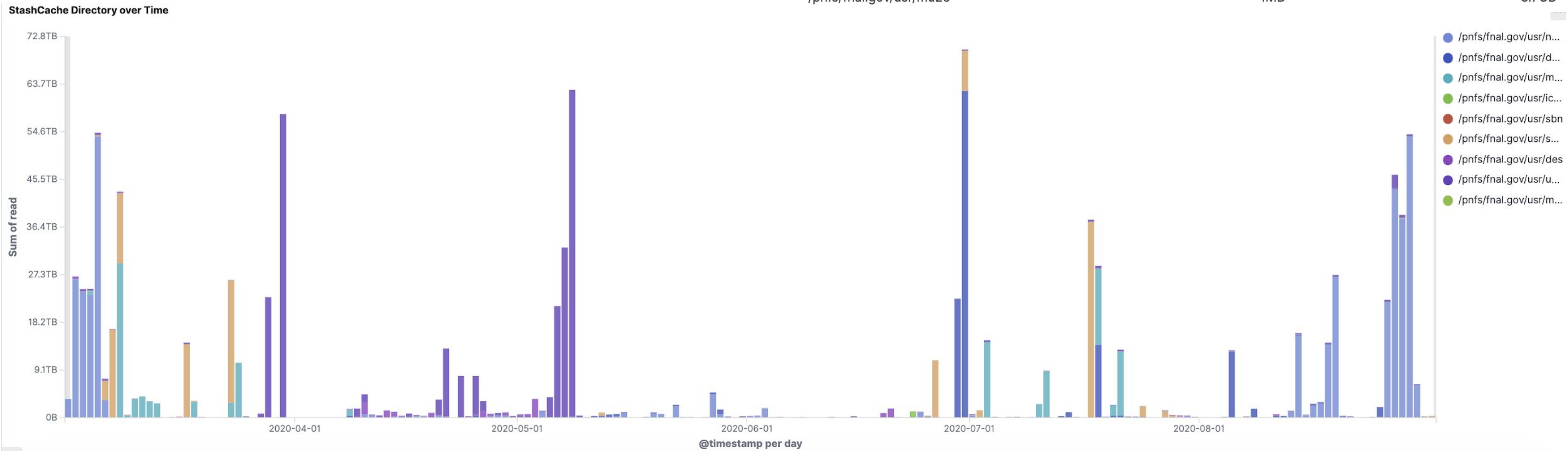
Figure 2. Images of objects passing all selection criteria before machine-learning classification. For each object, the set of images for the night with the least noisy difference image is displayed. All images are centered on the detected transient. The DESGW ID of the object is listed on the

StashCache Usage

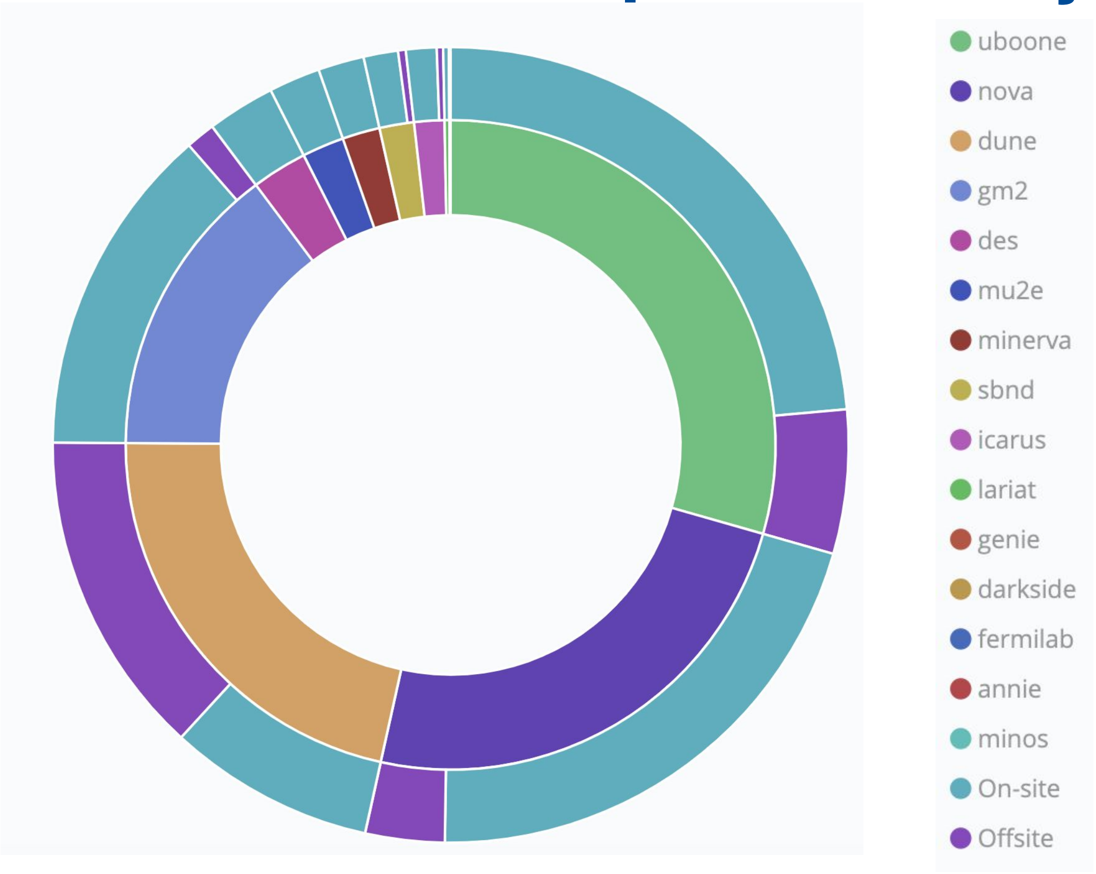
FNAL expts access StashCache over CVMFS

DUNE has second-highest reread factor (over 5000) of any group in past 6 months

Directory	Working Set	Data Read
/pnfs/fnal.gov/usr/nova	187.2GB	382.3TB
/pnfs/fnal.gov/usr/uboone	409.6GB	251.9TB
/pnfs/fnal.gov/usr/sbnd	71.1GB	135TB
/pnfs/fnal.gov/usr/dune	24.2GB	130.6TB
/pnfs/fnal.gov/usr/minerva	350.4GB	116.4TB
/pnfs/fnal.gov/usr/des	730.4MB	17.4TB
/pnfs/fnal.gov/usr/icarus	1.4GB	1.5TB
/pnfs/fnal.gov/usr/sbn	127MB	437.9GB
/pnfs/fnal.gov/usr/mu2e	1MB	3.7GB



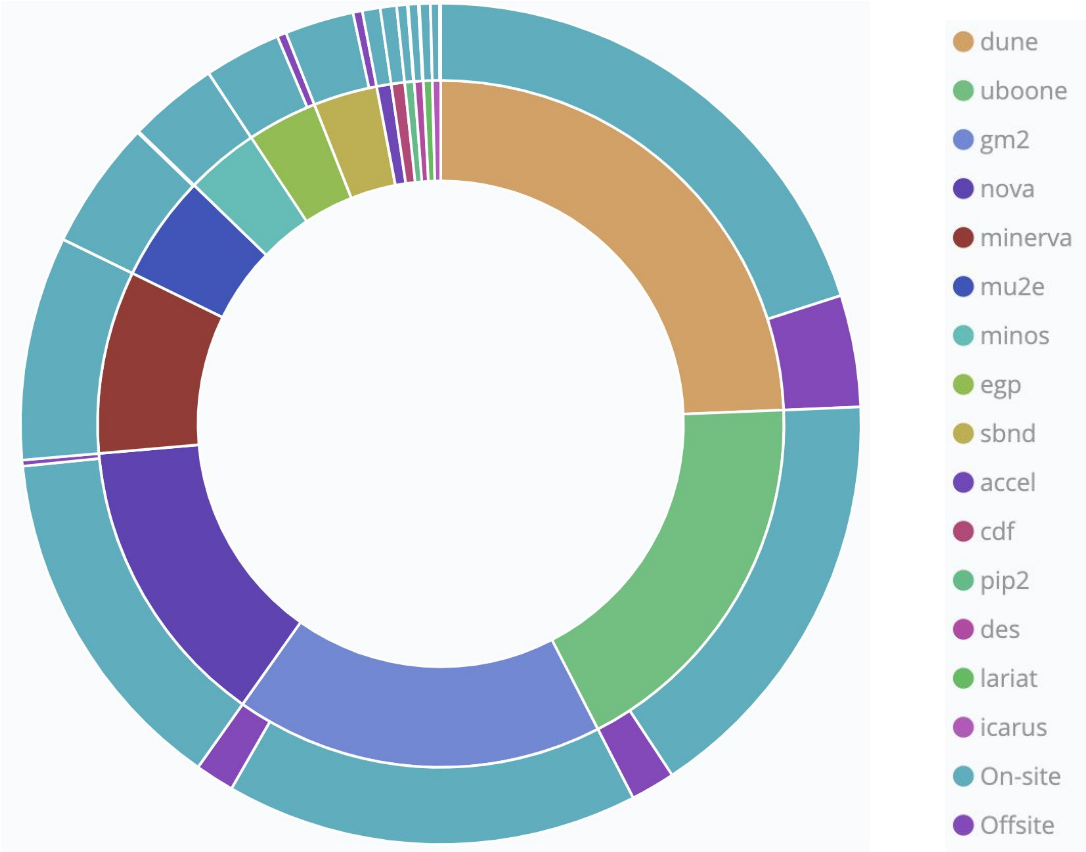
Site distribution of production* jobs, past year



[Link to generate plot]()

* Defined as owned by the <experiment>pro accounts.

Site distribution of analysis* jobs, past year



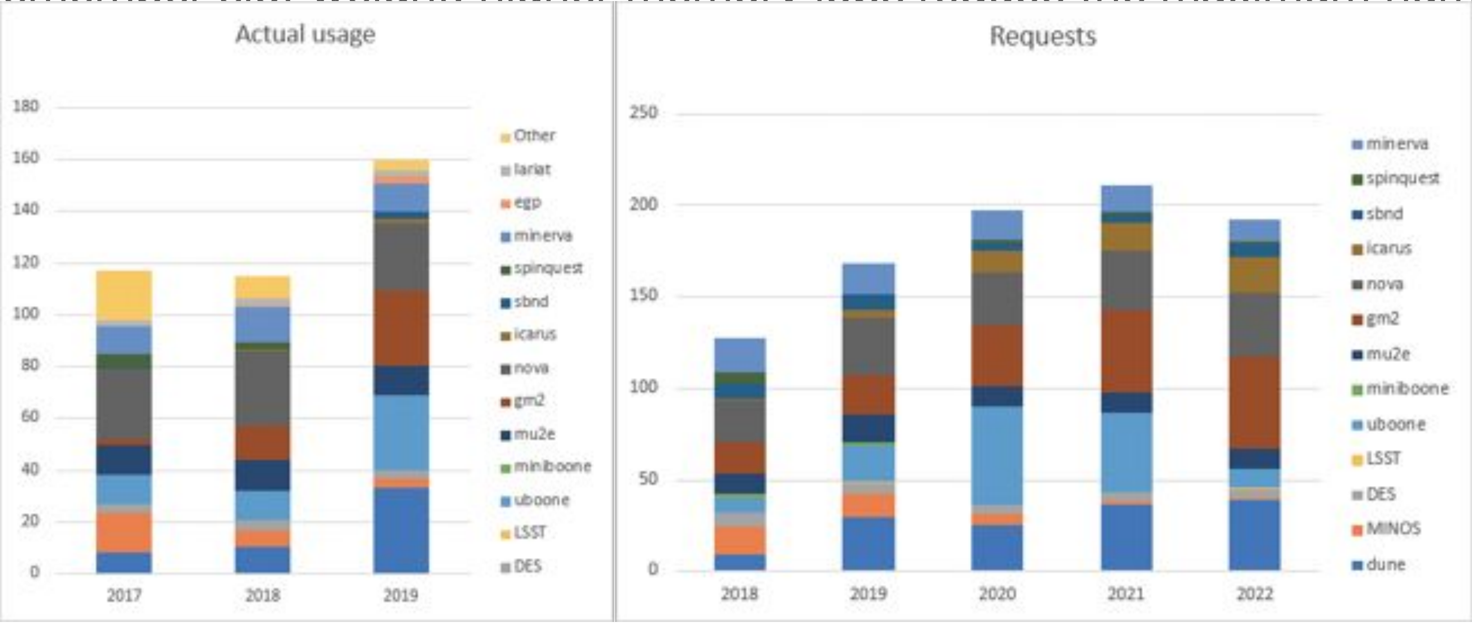
[Link to generate plot]

* Defined as not owned by the <experiment>pro accounts.

Experiment Requests and resources

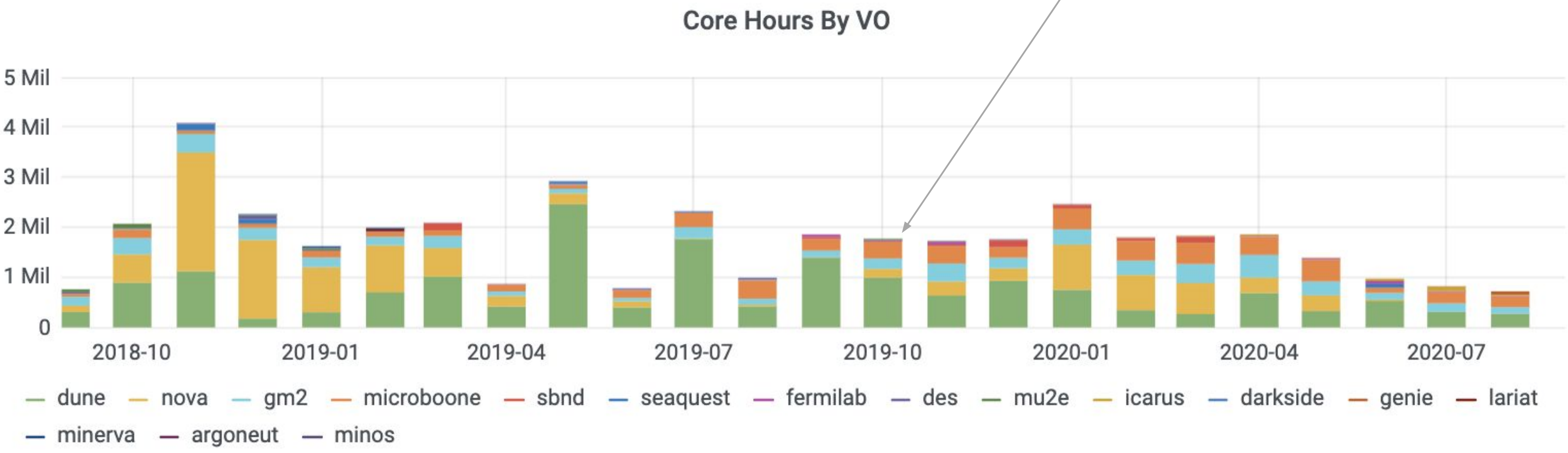
FermiGrid has expanded but significant chunks are out of warranty. Can cover about 240M slot-hours, but that’s over the year and expt peaks tend to overlap

Also remember slot weight: higher memory jobs reduce the maximum number of concu



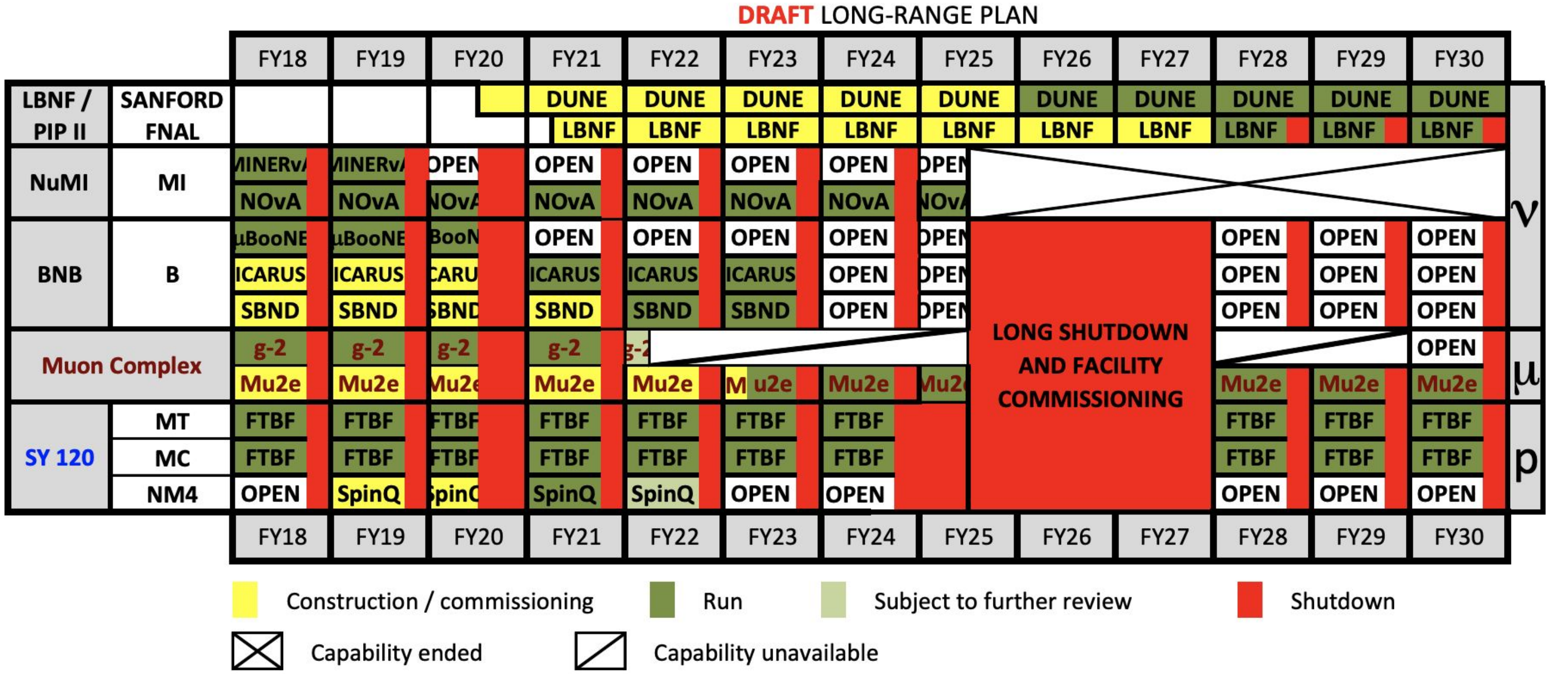
Singularity Invocation Effects

Singularity via GlideinWMS rollout



Draft Long-range FNAL plans

Office of the CRO June 2020



NOTES

1. This draft long-range plan is updated annually, typically following the summer PAC meeting.
2. The FY20 Summer shutdown began early in response to the COVID-19 pandemic. Summer shutdowns will typically last about 4 months during the construction of LBNF/DUNE and PIP-II. The timing and length of the 2-year Long Shutdown associated with the major construction activities at the lab will become clearer as the projects are baselined. Optimized commissioning and physics startup plans will be developed.