DUNE Rucio Experience

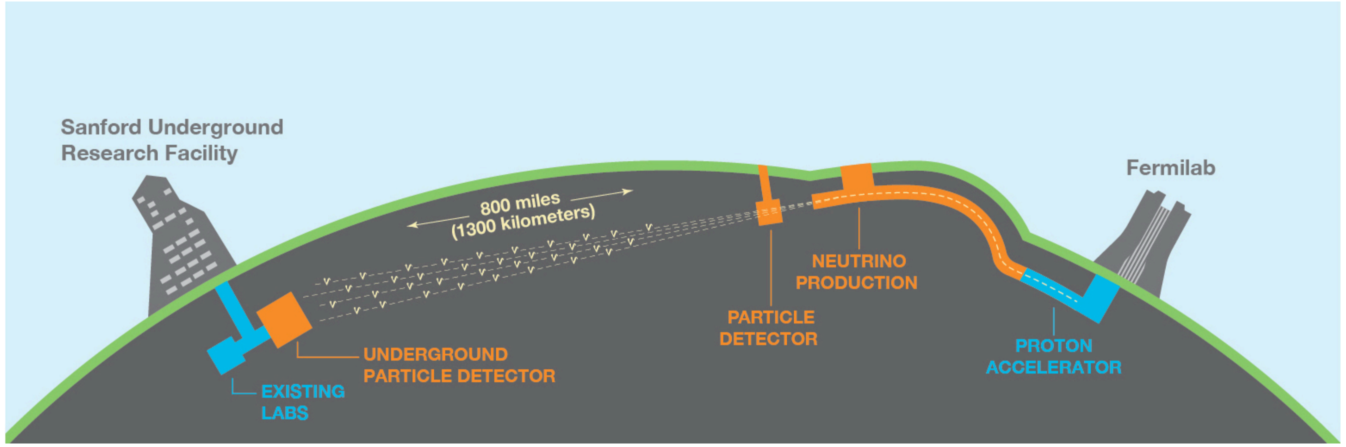
Michael Kirby, Fermilab/Scientific Computing Division

April 7, 2020

Jefferson National Accelerator Facility Software & Computing Roundtable

Outline DUNE Data Management Experience

- Introduction to the DUNE Experiment and physics goals
- Near & Far Detectors and ProtoDUNE
- DUNE Event Data and Computing Model
- Unique DUNE data challenges
  - SN and HPCs
  - single APA data access
- storage & compute projections
- past DUNE Data management
  - a little bit of history
  - F-FTS, SAM
  - dCache, ENSTORE



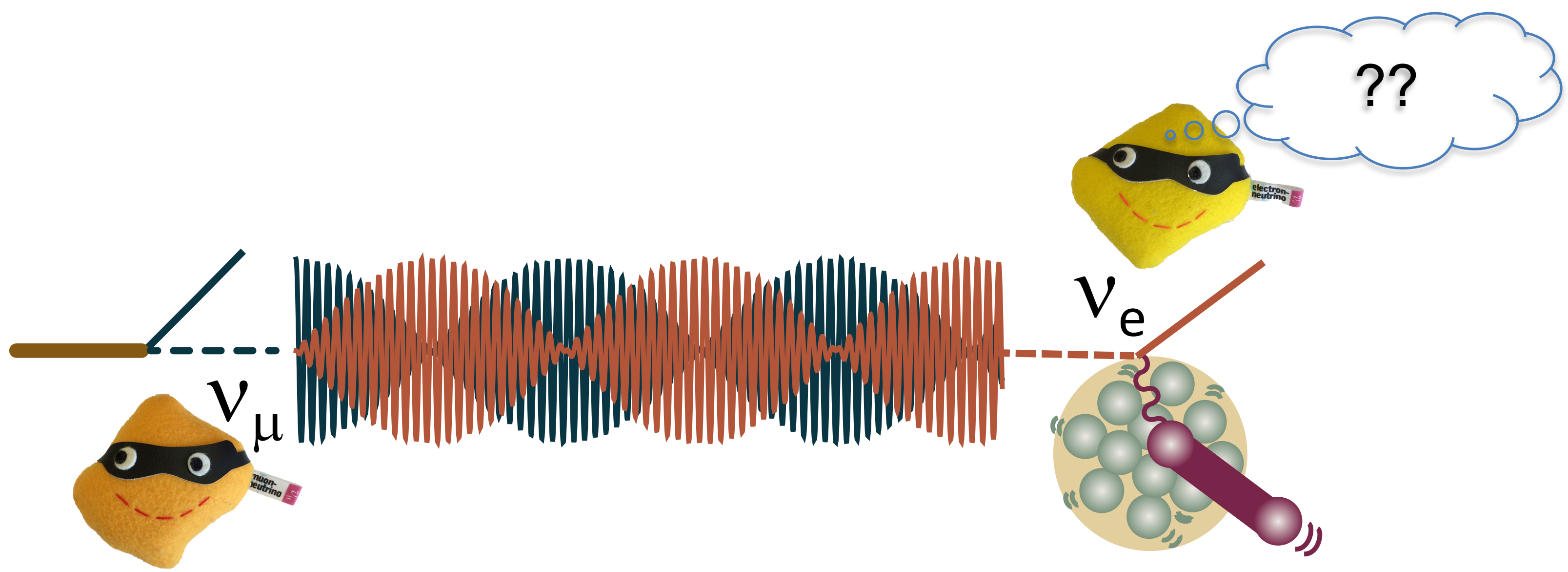
- Forward looking Data Management requirements
- Current status of DUNE Rucio
- integration with legacy software
- Rucio features requests
- replacement of legacy software

Thanks to Robert Illingworth, Heidi Schellman, and Steve Timm

DUNE Experiment Physics Goals

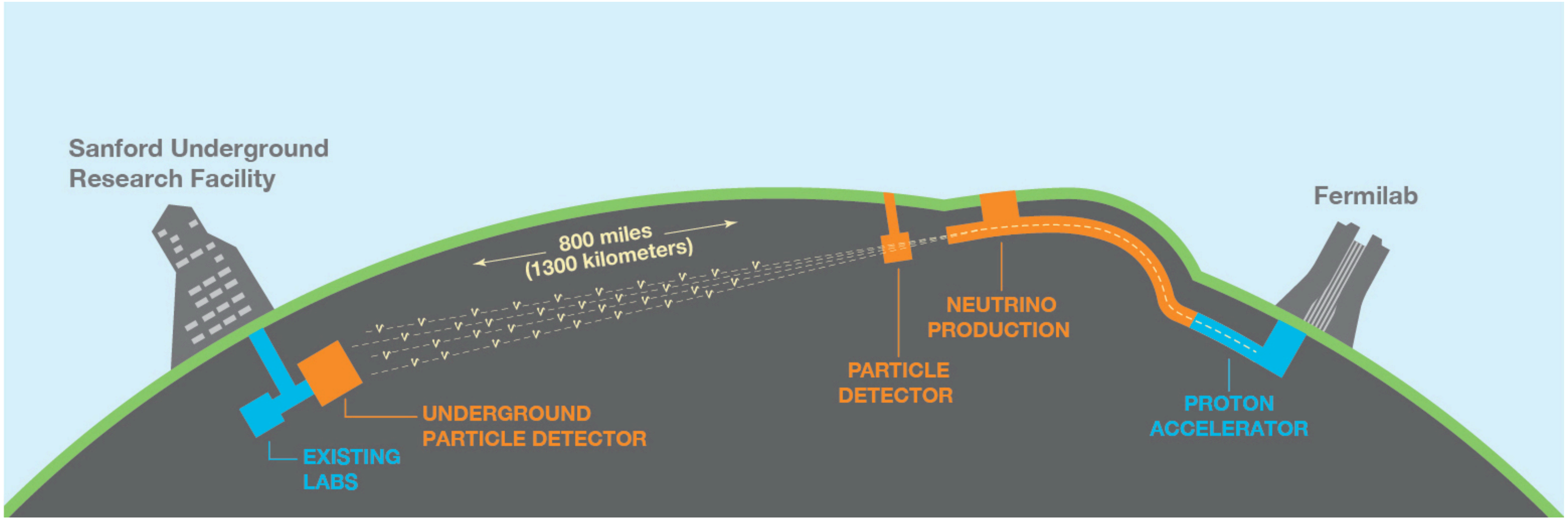
The quantum wavelength of a 2 GeV muon neutrino is ~ 10^{-16} m
But it is actually a superposition of the 3 mass types of neutrinos which have slightly different wavelengths – the beat wavelength between the types is about 2000 km.

Bottom line – propagation can change a muon type neutrino 
into an electron type neutrino 



Fermilab
Slide courtesy of Heidi Schellman

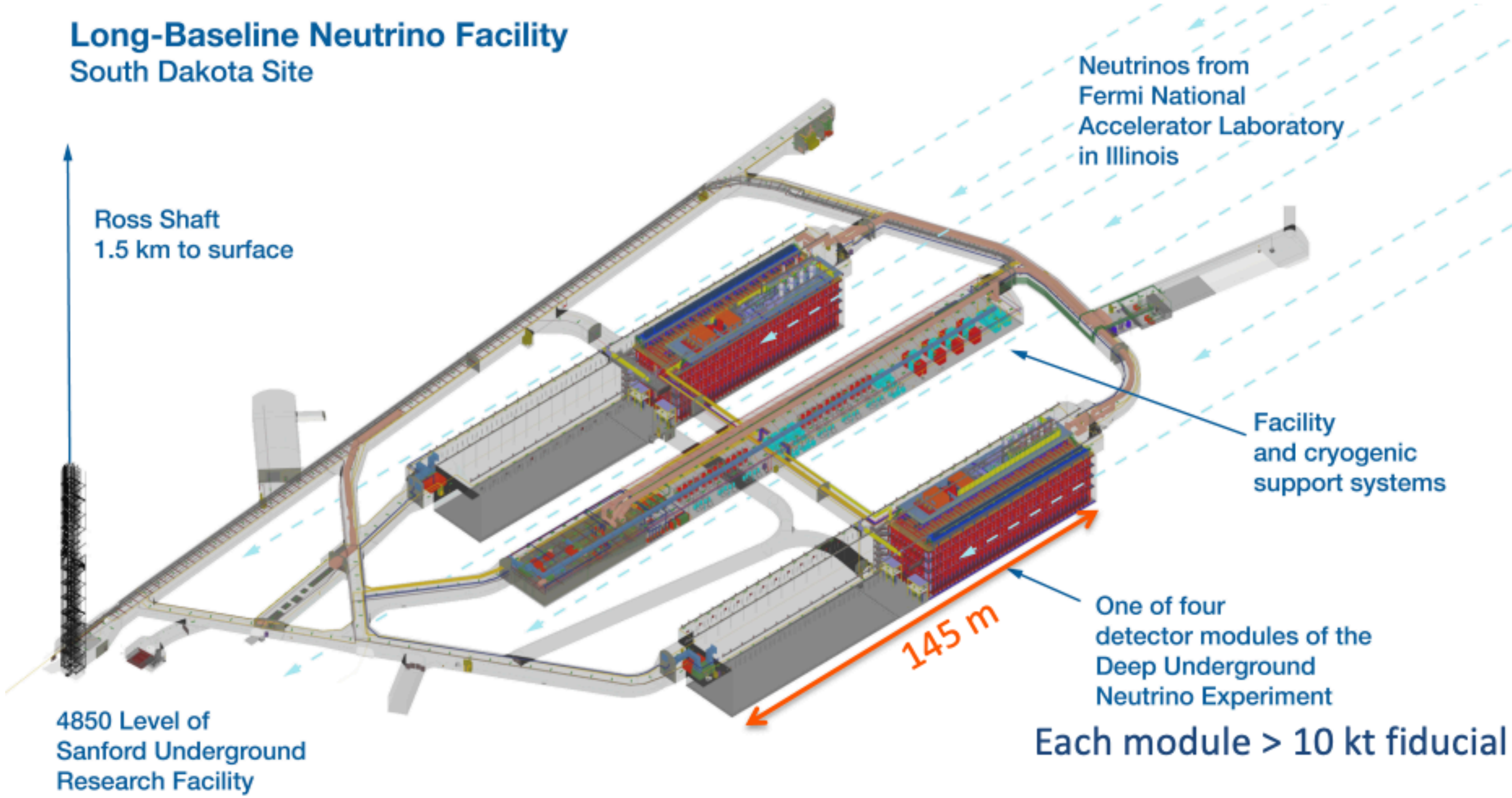
DUNE Experiment Arragement



- neutrino experiment looking for neutrino oscillation parameters (mass ordering, matter vs antimatter asymmetry, unitarity), proton decay, supernova neutrinos, and more.
- 40 kT LAr TPC detectors at 4850 ft underground in Lead, SD (Homestake Mine)
- Near Detector (still in design) at Fermilab near the neutrino production
- Two prototypes at CERN - (ProtoDUNE Single Phase - ProtoDUNE Dual Phase)

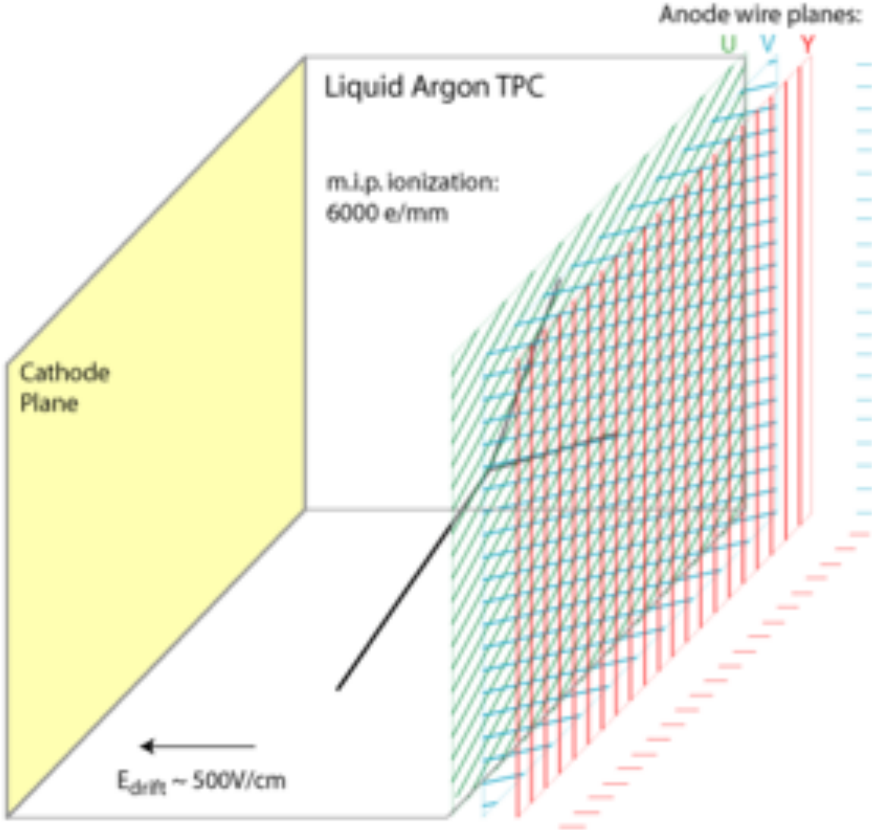
DUNE Far Detector Design

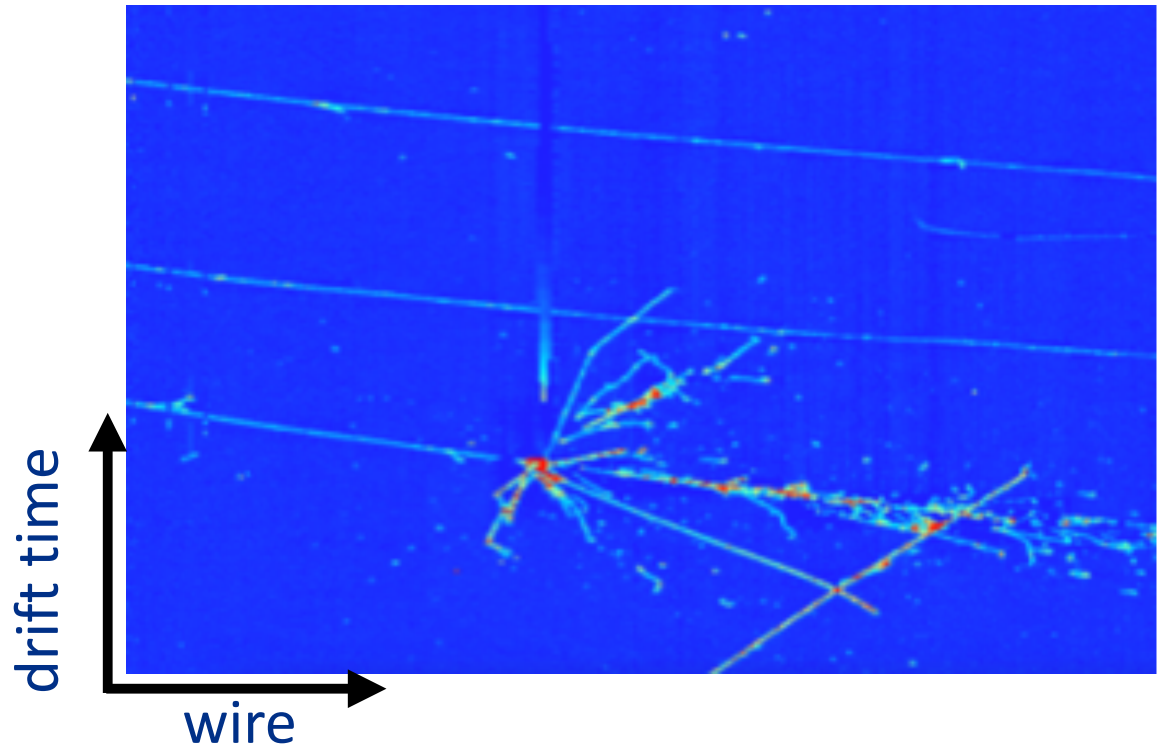
Slide: Ed Blucher



Each module > 10 kt fiducial

Liquid Argon Time Projection Chamber Detectors

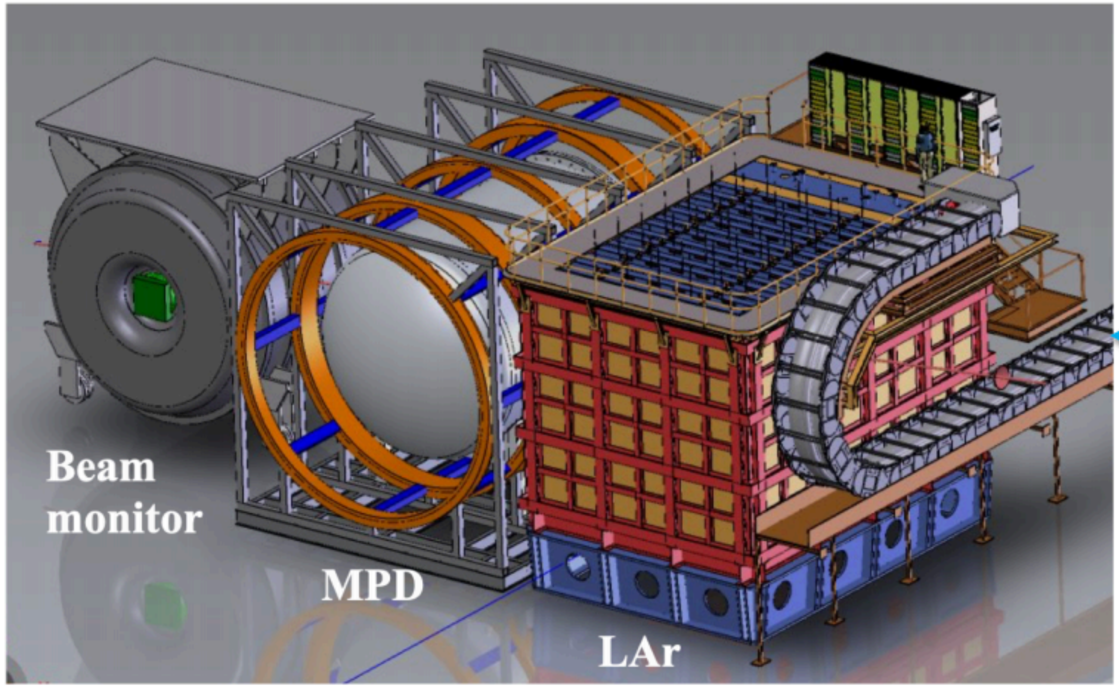


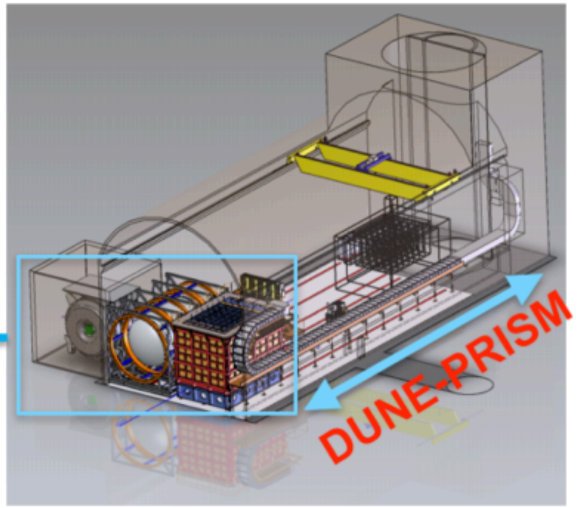


- DUNE Far Detector will be constructed of 4 LArTPC modules (68kT Argon w/ 40 kT active)
- High spatial and calorimetric resolution
- prototyping underway with both a single-phase and dual-phase protoDUNE at CERN

(Proposed) Near Detector Design

Slide: Ed Blucher



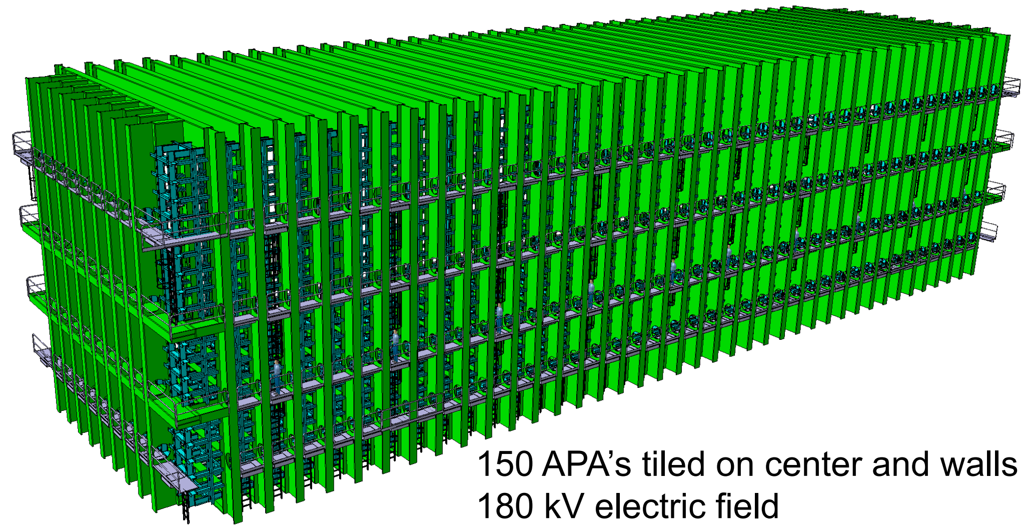


- LAr: Highly segmented LAr TPC (ArgonCube)
- MPD (Multi-purpose detector): High Pressure Gas Argon TPC, Calorimeter, and muon system magnetized by superconducting coils
- Beam monitor: High density plastic scintillator detector with tracking chambers and calorimetry in KLOE magnet
- DUNE-PRISM: Movement of LAr+MPD transverse to the beam, sampling different E_ν

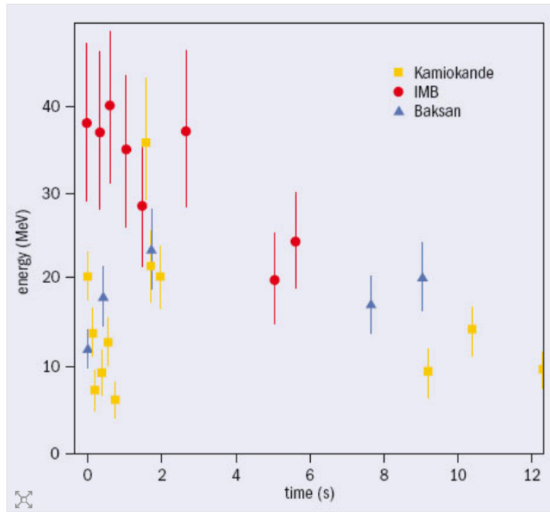
Unique DUNE Computing Challenges

- DAQ agreed to constraint of 30 PB raw data per year from FD
  – includes triggered data, supernova readout, calibrations, etc
  – FD has much greater bandwidth, but reduced with trigger, zero suppression, and compression of data
- Time-extended trigger records present unique situation
  – normal neutrino-beam trigger record is 5.4 ms
  – time-extended trigger record could be as long as 100 s
  – after zero-suppression estimated to be 184 TB
- 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 events is done to minimize data volume and facilitate additional processing
- Actively running two experiments
  – ProtoDUNE-SP - beam (2018) and cosmic-ray operations
  – ProtoDUNE-DP - cosmic-ray operations
  – more than 4 PB of raw data for SP and DP

Far Detector SP Module



150 APA’s tiled on center and walls
180 kV electric field
17 kT of Argon



SN1987A neutrino events observed by Kamiokande, IMB and Baksan showed that the neutrino burst lasted about 13 s.

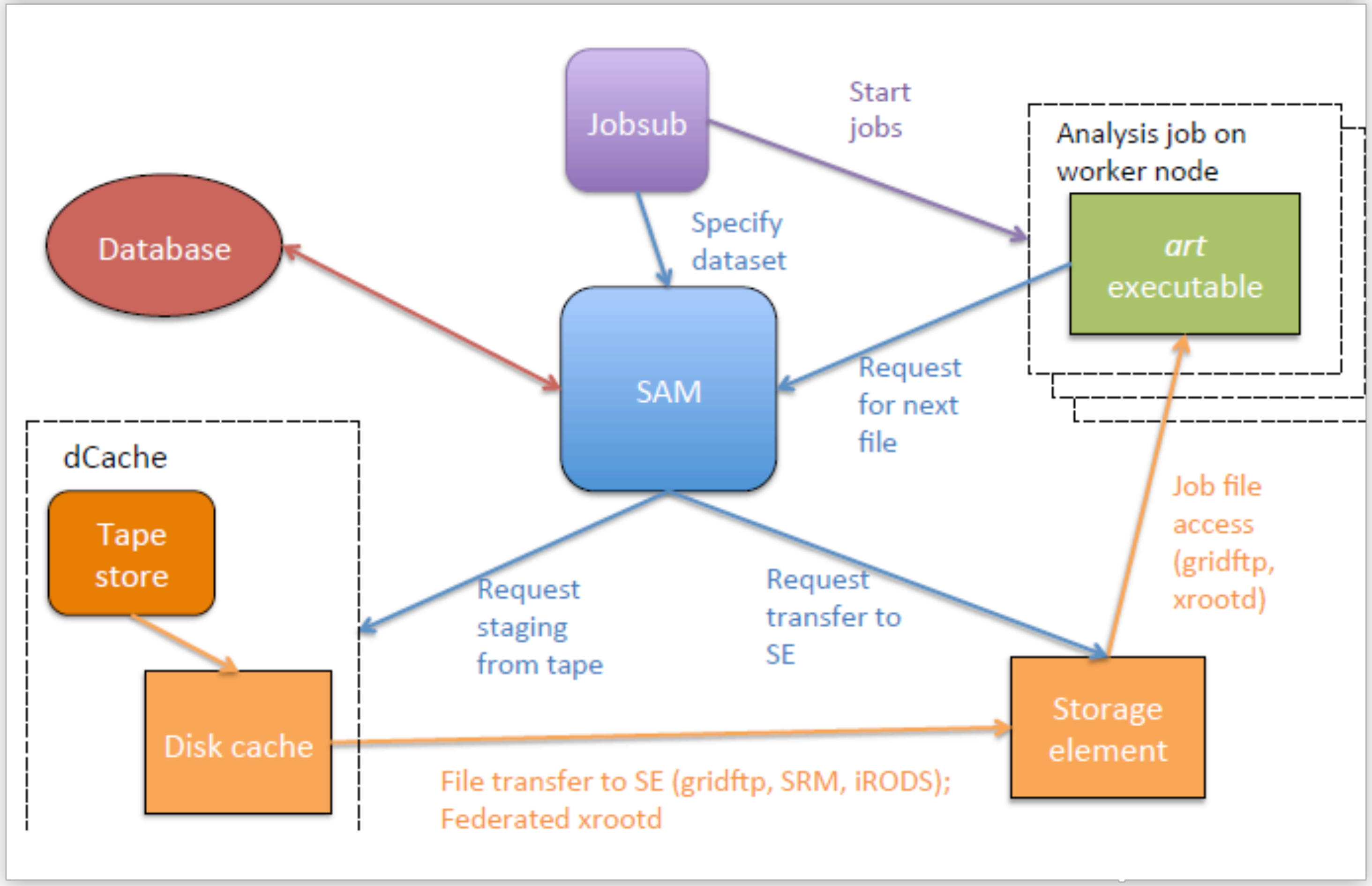
CERN Courier, Jan 2007

DUNE Data Management History

- DUNE circa 2017 had basically 1 storage element: FNAL dCache disk + ENSTORE tape
  - data was staged from tape onto disk on-demand and without restrictions
  - dCache cache was cleared based upon a LRU policy and approximately 2-3 PB r/w pool
  - no replication and all data management was done using FNAL FIFE software stack
  - Fermilab FTS, Serial Access via Metadata
- With initial operation of protoDUNE SP, started to utilize CERN EOS and CASTOR (thanks to CERN for providing those resources and support!)
- Formation of the DUNE Computing Consortium in Fall of 2018 expanded the resources available
  - recognized that Fermilab could not supply all of the resources for DUNE computing
  - additional sites and resources needed to be integrated into computing
  - development of DUNE Computing Model along with Event Data Model

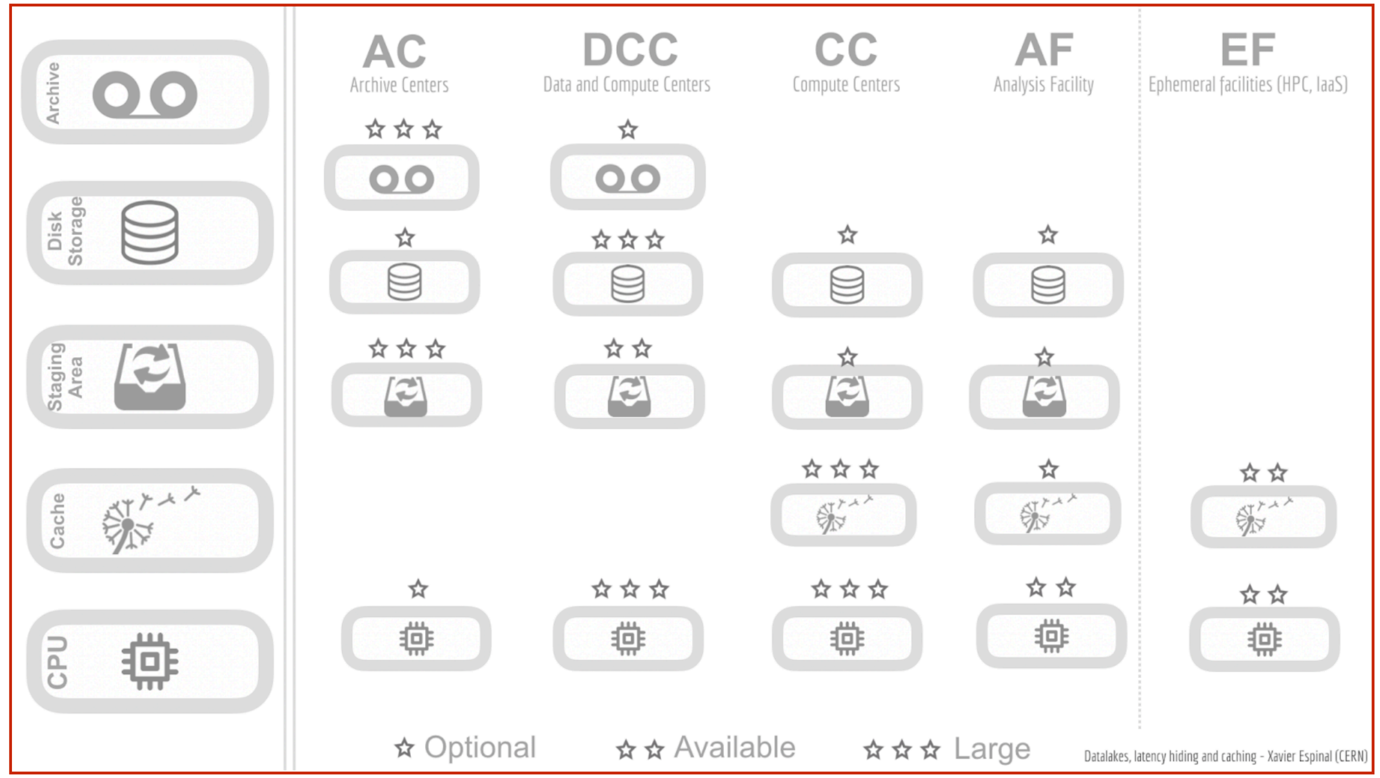
Serial Access via Metadata (SAM) and Fermi File Transfer Service (F-FTS)

- Developed more than 20 years ago for Tevatron experiments
- based upon a late-binding model of data processing on the grid
- user defined datasets using file metadata and selection
- DUNE still using for processing and workflow management
- DUNE has never used the data management/replica utilities
- F-FTS useful for moving data to SE, but doesn't have full management tools
- not seen as a forward looking solution (too much additional development and support needed going forward)



DUNE Computing Model for Institutional Sites

- Simplified terms for current DUNE sites
  - Tape Site - tape/staging
  - Disk Site - disk + CPU
  - Compute Site - CPU + cache
  - Analysis Site - cpu + cache
  - HPC - (HPC, IaaS)
- Goal is to have resource split between FNAL and other institutions - 25%/75%
- FNAL has some custodial responsibilities from the Dept of Energy that make this not possible for tape
- additional



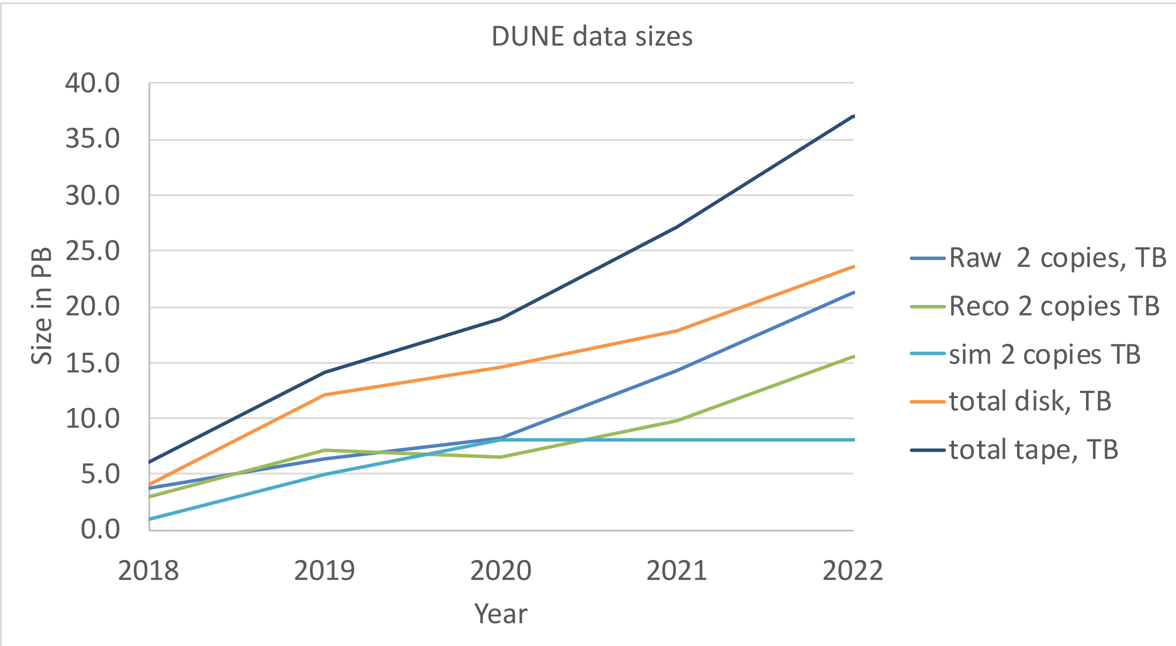
Data Access in DOMA, HSF/OSG/WLCG Joint Workshop J-LAB Newport News, VA 19-23 March 2019

Computing Model Policies

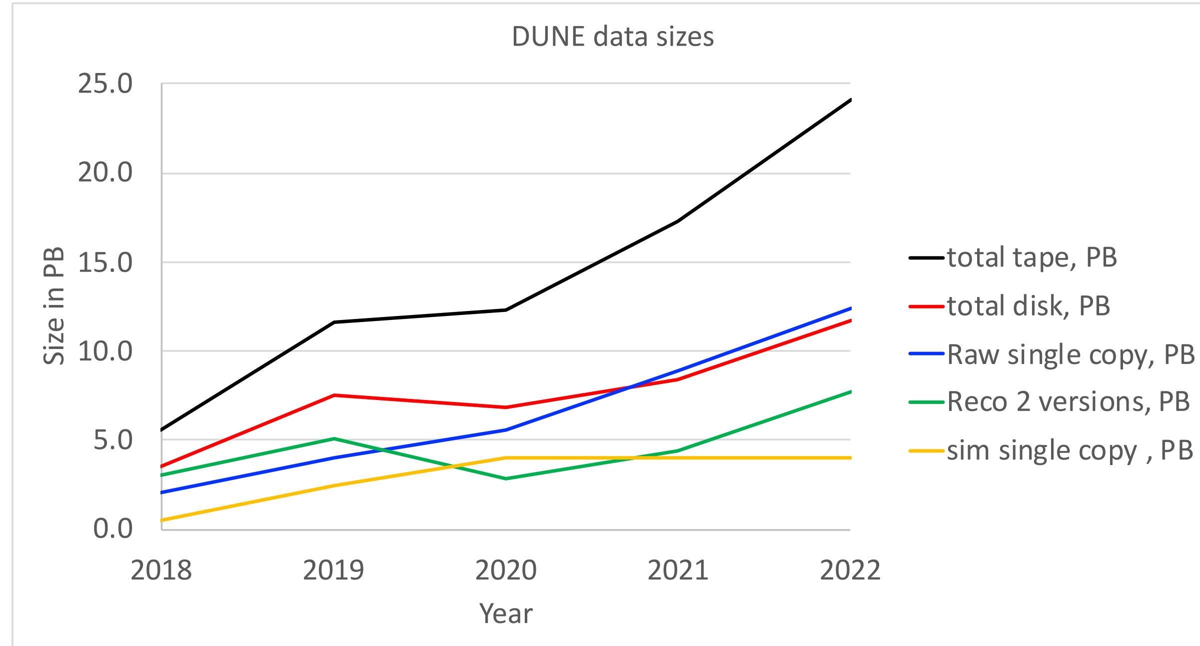
- Tape Storage
  - two copies of all raw data for security
  - FNAL provides storage for an archival copy of all raw data for DUNE (ND, FD, protoDUNE)
  - Rucio Storage Elements (RSEs) around world provide storage for 2nd copy
  - FNAL provides storage of derived datasets with lifetime of 2 years
  - FNAL provides storage for single copy of simulated data
  - RSEs around the world provide storage for second copy of simulated data
- Disk Storage
  - two or three copies of every active derived dataset on disk at any time
  - two derived datasets will be active at any one time
  - latest two active derived dataset staged to disk at FNAL
  - two or three copies of every active simulated dataset on disk at any time
  - two simulated datasets will be active at any one time (matching active derived dataset)
- From these policies can development estimates for resource needs

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

Data Management Needs for DUNE

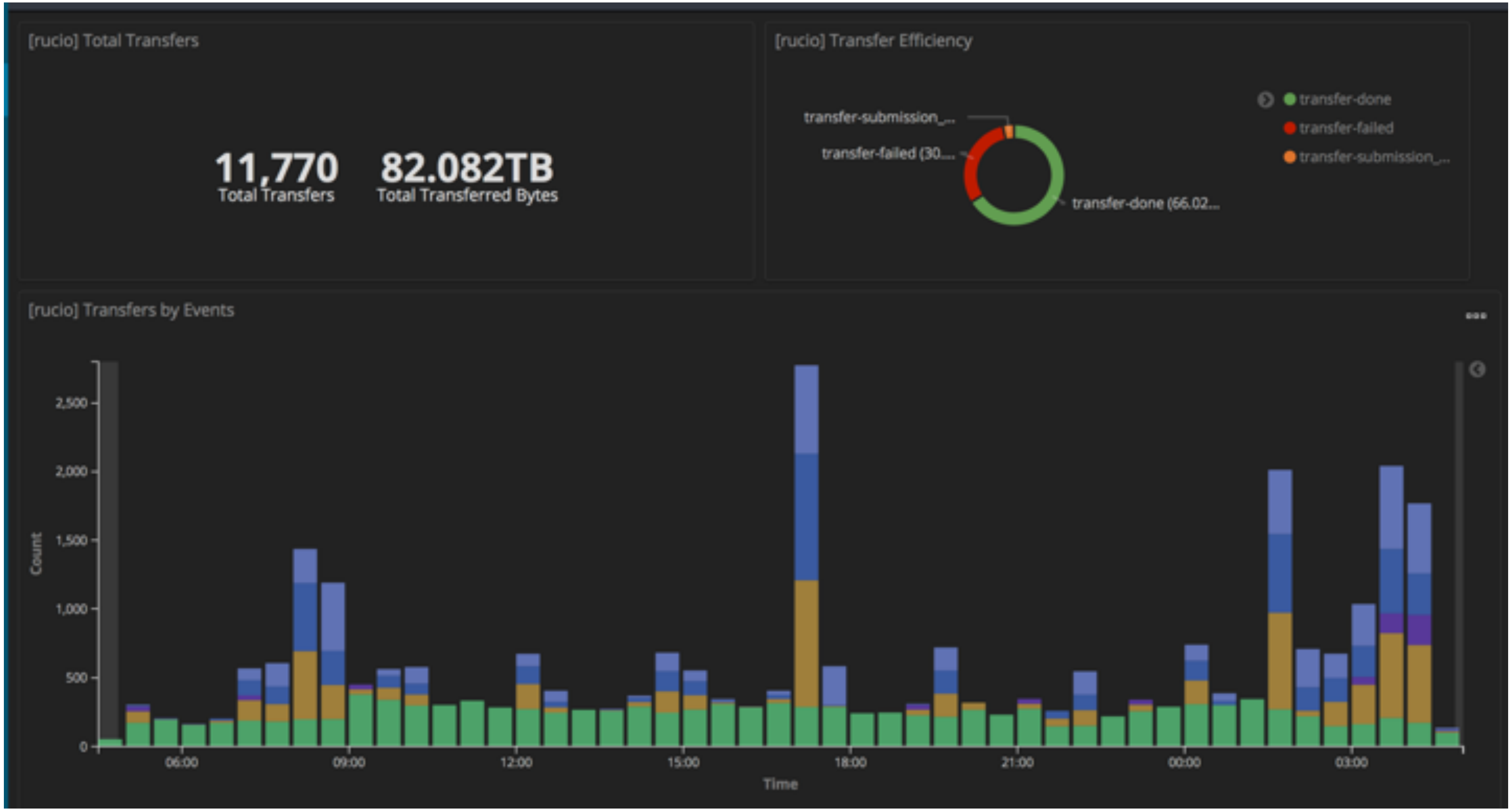
- manage multiple storage elements across the world/federations/countries
- integration of multiple transfer technologies and eventually token based authentication
- movement of data and datasets automagically
- control the replication of data and datasets
- allow creation of rules for automated processing
- monitoring of data movement and replication
- (eventually) management of workflow processes and data delivery
- (eventually) data discovery for production processing and analyzers analysis
- (eventually) integrate with future technology and storage solutions

SCIENTIFIC DATA MANAGEMENT
RUCIO

DUNE Rucio Instance at Fermilab

- DUNE has been running Rucio since Fall of 2018.
- Fermilab Scientific Data Storage Dept.
  - containerized Rucio server
  - using Postgres DB on back end
  - Centrally danaged all schema and software
- DUNE Data Management Team
  - Rucio clients move data from point A to point B. (asking for help if things get stuck)
  - Creation and declaration of new Rucio replicas
  - Onboarding new remote Rucio storage elements
- Interaction with remote sites for transfer
- Ingest of ProtoDUNE raw data still done with legacy system (File Transfer Service)
- Legacy system gets data from CERN EOS to CERN Castor and FNAL dCache/Enstore Tape
- script run to declare it to Rucio
  - (one dataset per run, large containers of related datasets)
- Rucio is used to send it everywhere else
- Rucio is also used to manage limited disk space on CERN EOS
- SAM (Sequential Access with Metadata) is used to tell us what it is, and where it is
- 15 commissioned Rucio Storage Elements (RSE)
  - 14 PB under Rucio management
  - 1.2 Million DID’s
  - 2.7 Million replicas

Kibana based Monitoring

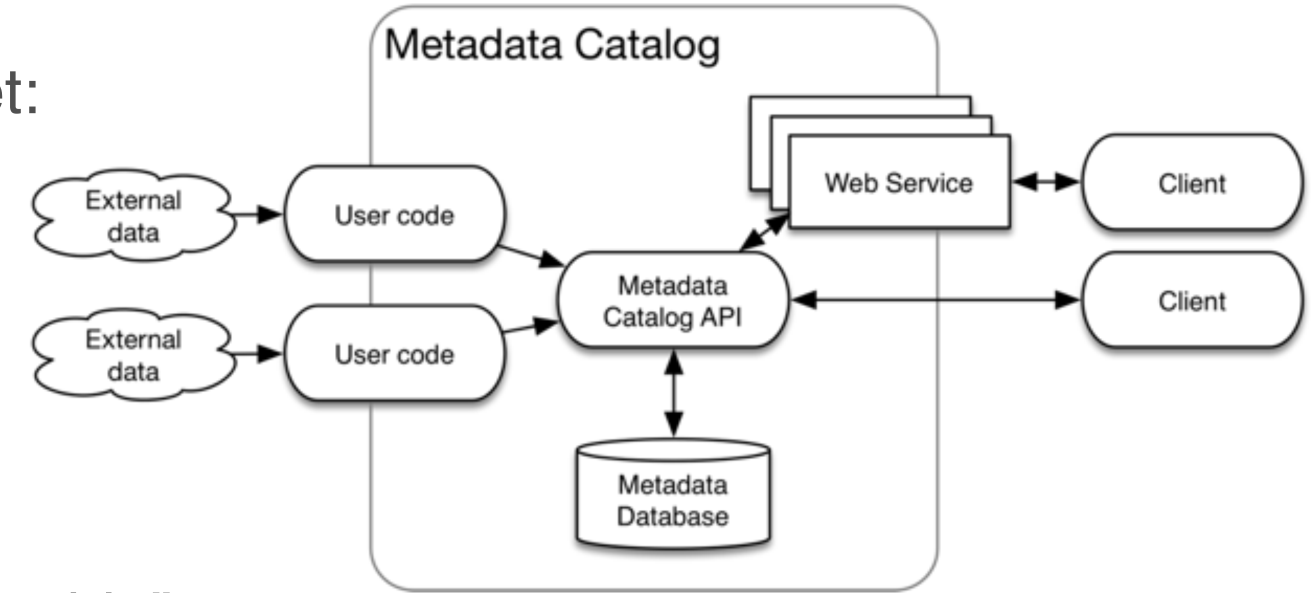


Additional Features to Rucio

- Quality of Service
  - know when a file is online or on tape.
  - Detection of condition when Fermilab dCache has the file in online storage and prefer that as a source.
- Deterministic vs. non-deterministic
  - "Deterministic" is for disk sites—files stored in a hashed path
  - "Non-deterministic" used on tape sites
  - human-readable path constructed from metadata fields
    - Pending a new feature to serve the path to Rucio
- Lightweight client for REST API
  - Stock client has lots of dependencies
- 3rd party copy doesn't work between all possible RSE's
  - 3rd party xrdcp doesn't work at some sites
  - Others don't support gridftp anymore
  - https (webdav) is coming but not supported everywhere yet.
  - SRM is going away, but on tape sites still only reliable way to stage

Metadata Catalog for DUNE Data Discovery

From Igor Mandrichenko talk at DUNE Dec 2019 DB Workshop



- Users are used to :
  - Defining a dynamic data set
  - Running a project across the whole data set:
    - Each job says “give me the next file”
    - Each job notifies the project manager when it’s processed successfully.
    - Recovery jobs can be generated.
- Rucio has fixed data sets.
  - Some important ones we will declare “immutable”
  - Others we will declare “monotonic” to which you can add but not delete
- Metadata server API will allow queries of the metadata plus user callouts to the conditions database.

Fermilab

Longer Term Requests

- Rucio working with JWT Tokens (SciTokens,Indigo IAM, etc)
- Data Delivery Microservice
  - Similar to the ATLAS service—can get objects as small as a single data unit
- Alternative File Formats
  - HDF5 of interest
  - Interested in object stores in general, particularly for raw data processing.

Conclusions

- DUNE faces a few unique data access challenges
- Creation of the DUNE Computing Consortium and expansion of the facilities/resources available necessitated new Data Management tools
- Rucio seemed the natural choice with feature set, support model, and wide community support
  - there are some features and updates that DUNE is working to help develop (e.g. QoS)
  - DUNE hoping to continue to expand our contributions to the Rucio development
  - Current instance of DUNE Rucio operating well and dramatically helped with management of 15 RSEs across the world
- there are additional tools that will need to be developed for a complete set of data access services - working on those with some prototypes already designed
- Thanks!