DUNE Rucio Plans

Robert Illingworth

GDB

11 September 2019

DUNE

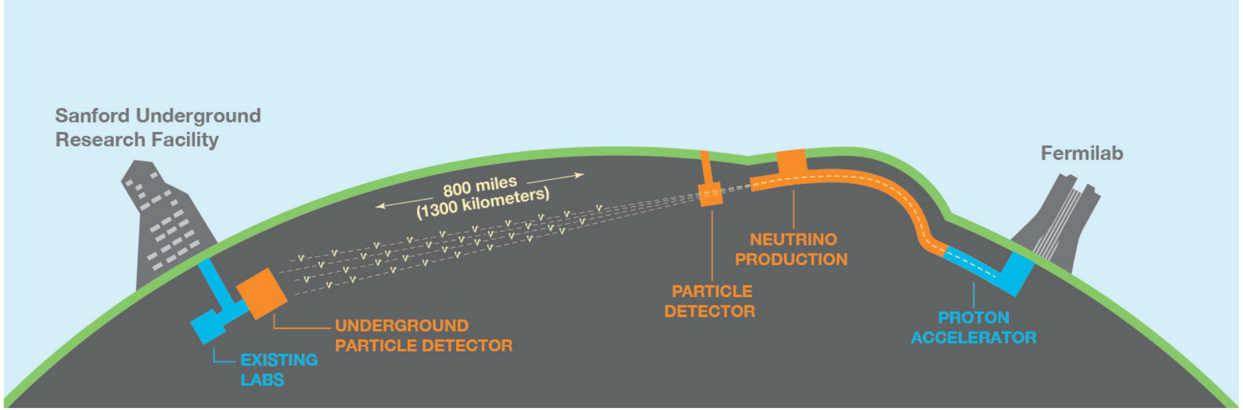
DUNE – Deep Underground Neutrino Experiment

http://www.dunescience.org/

Send a beam of neutrinos from Fermilab to South Dakota starting in ~2026

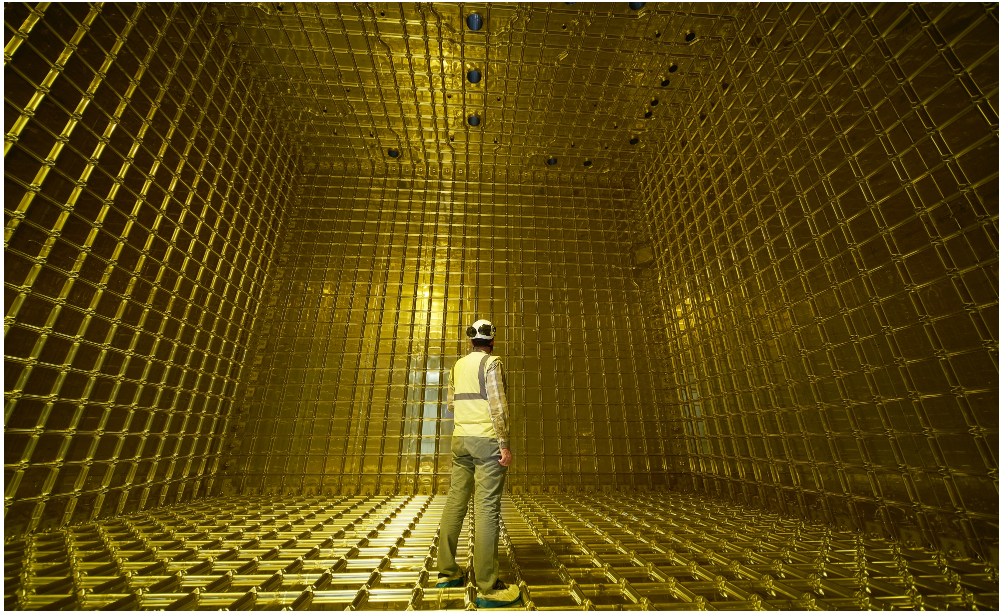
40kt Liquid Argon TPC Far Detector

Smaller Near Detector at FNAL (tracker/calorimeter)


Sanford Underground Research Facility
800 miles (1300 kilometers)
Fermilab
NEUTRINO PRODUCTION
PARTICLE DETECTOR
PROTON ACCELERATOR
UNDERGROUND PARTICLE DETECTOR
EXISTING LABS

ProtoDUNE

- Prototype detectors for DUNE located at CERN
  - Two of them utilizing different technologies
    - Single Phase (SP)
    - Dual Phase (DP)
- Ongoing cosmic data taking
- SP took test beam data in September 2018
- 6 PB of data + reconstructed output
- Proposed test beam run for both detectors in ~2021-22

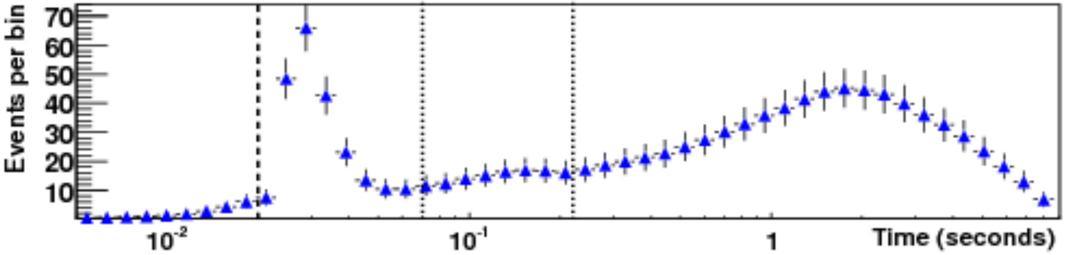


Data challenge

- Far detector data comes in **very** large chunks – 25xProtoDUNE
- Beam and cosmic interactions are 1-6 GB each per 10kt Module
  - Rates are ~ 5000/day/module dominated by cosmics
  - Need to read out 3-10 ms of data to get a full drift
- One 5.4 ms readout means
    - 1 tick = 12 bits
    - 1 channel = 10,800 ticks -> 16 KB
    - 1 APA = 2,560 channels = 41 MB uncompressed
    - 1 module = 150 APA's = 6.2 GB uncompressed
- All data types add up to about:
  - ~**12 PB/year/module (uncompressed)** x 4 modules
  - ~1.6 GB/sec for 4 modules, DC…
  - Compression could potentially reduce this by factor of 3-4 for SP
- ProtoDUNE-SP already ran at this rate, but for only 6 weeks.

Supernovas

- DUNE should be sensitive to nearby (Milky Way and friends) supernovae. Real ones are every 30-200 years but radioactive decays can make false alarms



- Supernova readout = 100 sec, one trigger/month
- 100 sec readout implies
  - 1 channel = 300 MB uncompressed
  - 1 APA = 768 GB uncompressed
  - 1 module = 115 TB uncompressed
  - 4 modules = **460 TB** … takes 10 hrs to read at 100 Gbs
- Some calibration runs will be similar in scope….

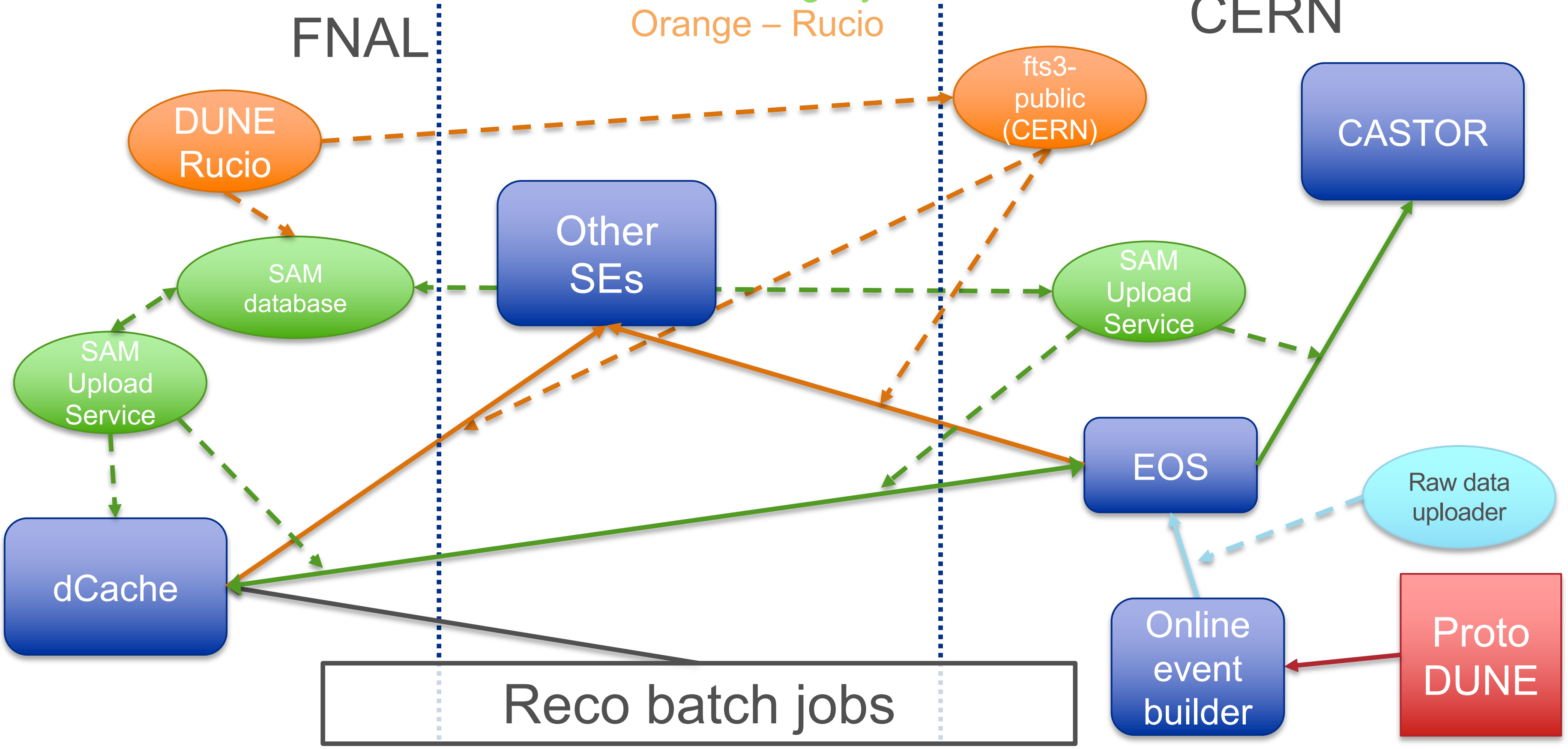
DUNE data management current status

- DUNE data management is currently SAM based (Tevatron Run II & FNAL IF data management system; default choice when we started)
  - Rich metadata catalogue
  - Replica catalogue
    - But relatively little in the way of transfer tools

- Currently running Rucio overlaid on the legacy system
  - Initial data upload and CERN->FNAL transfers still done by SAM
  - Rucio is used to manage CERN EOS disk usage (deletion)
  - Rucio does other site to site transfers; synced to SAM catalogue
  - But many files are now in two separate catalogues
    - Bound to get out of sync over time…

ProtoDUNE dataflow

Green – legacy
Orange – Rucio



DUNE data management current status II

- Production/analysis access to data is still via legacy SAM methods
  - Using existing tools from FNAL IF experiments
  - Normal users don’t see any change

- Any new SEs integrated will be solely Rucio managed

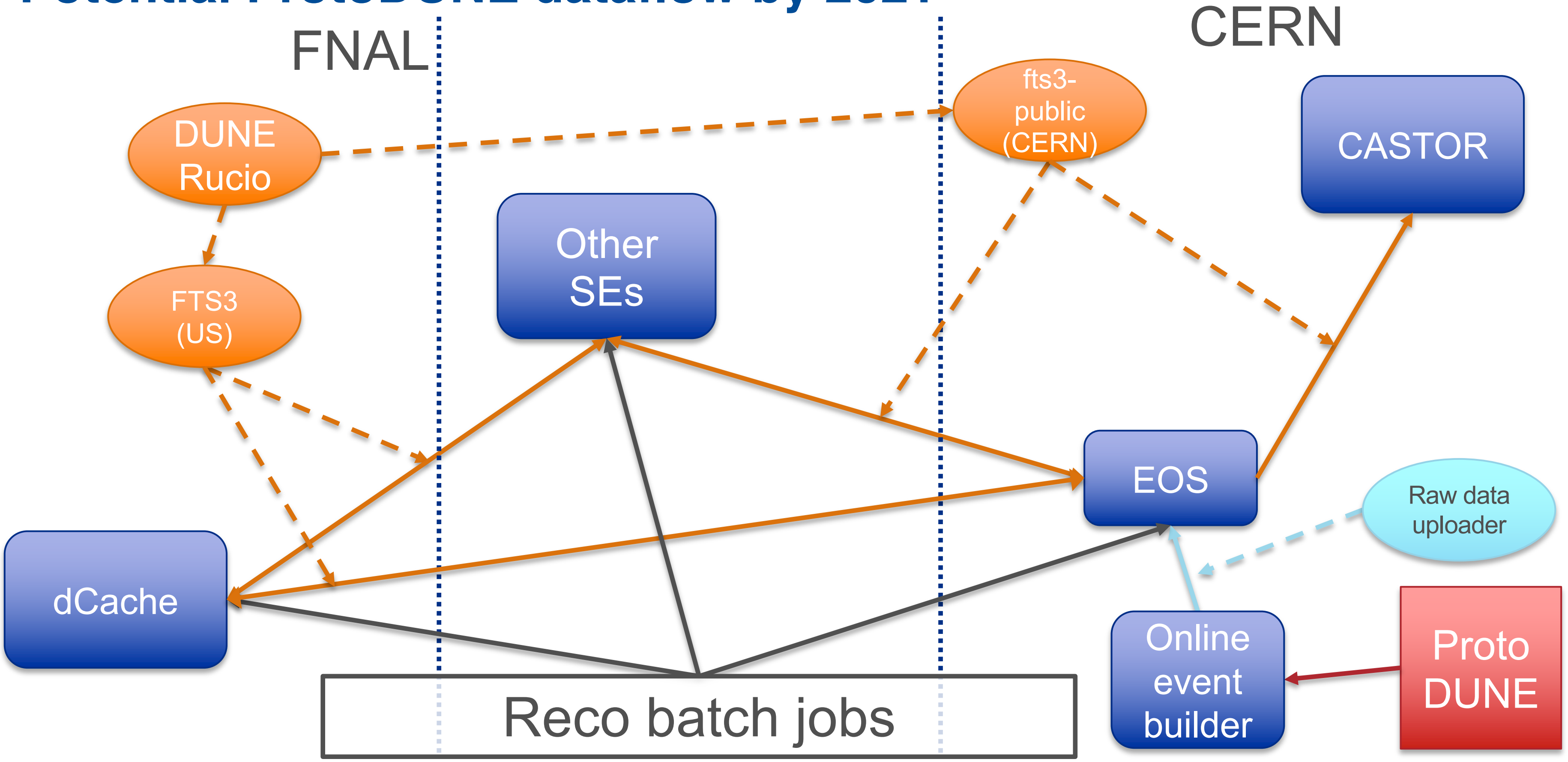
- Rucio is progressively becoming a production system for DUNE
  - But not something analysis users interact with yet.

Forthcoming plans for DUNE

- Move towards an entirely Rucio based system
  - Deprecate the SAM replica catalogue.
  - By 2021 ProtoDUNE run use Rucio for CERN -> FNAL copies
    - Need to set up transfers into tape system for this
    - Will also need to develop experiment expertise for operational support
      - Rucio documentation says what's there, but generally not why you might want to use it
      - Experiment needs to decide on dataset replication and deletion policies

- Longer term plans
  - Replace the SAM metadata catalogue with something new
  - SAM has complex metadata with powerful query facilities; current Rucio metadata capabilities are much simpler
  - "Data discovery" service tying data management metadata to experiment databases

Potential ProtoDUNE dataflow by 2021



Experiences with Rucio

- Rucio is a good fit for current DUNE requirements
  - Similar HEP use cases
  - Distributed data management is important for DUNE

- Improved ability to customize permissions/pfn mapping/etc is necessary
  - This is being worked on

- Better SE QoS handling (tiered storage) would be very beneficial
  - Current implementation matches ATLAS/CMS storage model very well; not too flexible beyond that
  - The FNAL dCache SE is being used as a single tape-backed cache with files recalled on demand
  - The SE is declared to Rucio as a tape RSE, this means that transfers trigger stage requests, but causes some issues as there is no knowledge that some of the data may already be on disk

Data movement (somewhat Rucio related)

- When we started we tried to look to the future and get away from SRM/gridftp
- Unfortunately we were a bit ahead of the curve, and webdav or xrootd TPC failed to work a pretty much anywhere

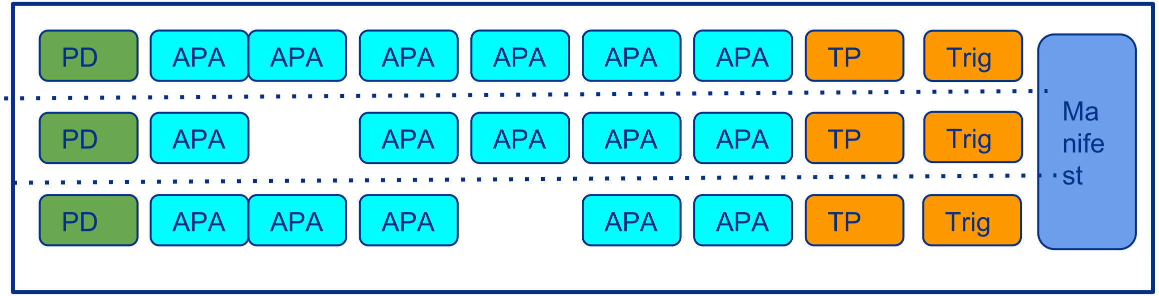
- Some of the issues
  - FNAL public dCache is a Rucio tape SE; only SRM works as a protocol because of the need to stage files
  - We tried to use RAL Echo via WebDAV; the S3 interface underlying Dynafed cannot handle files >8GB in one operation; most ProtoDUNE data files are 8GB in size

- The situation has improved and we need to revisit
  - We (DUNE) have been tracking DOMA-TPC activities, but not actively participating

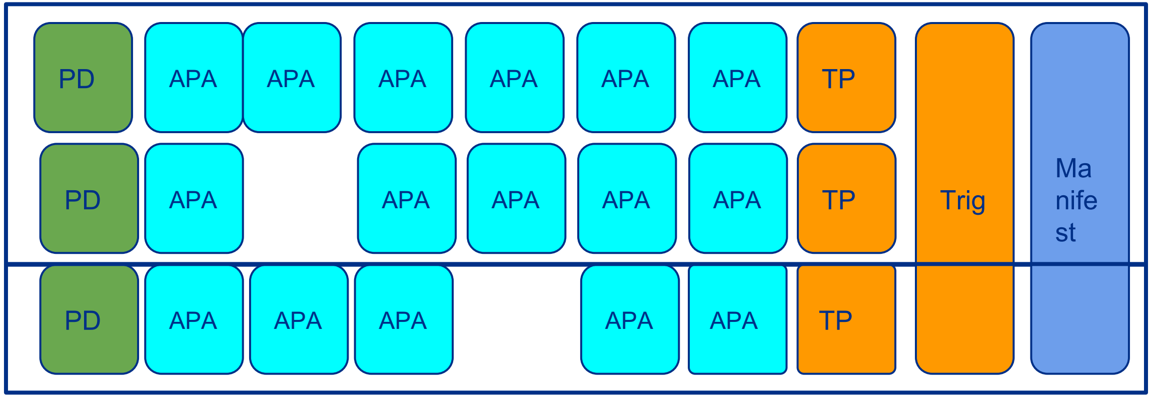
Future ideas

- Raw data consists of many identical readout modules each MB scale
  - They could be formed into files in different ways – for example by trigger (time localized); by module
  - Supernova readout is far too big to fit in a single file; has to be split

Time Localized readout (cosmics/beam/calibration)

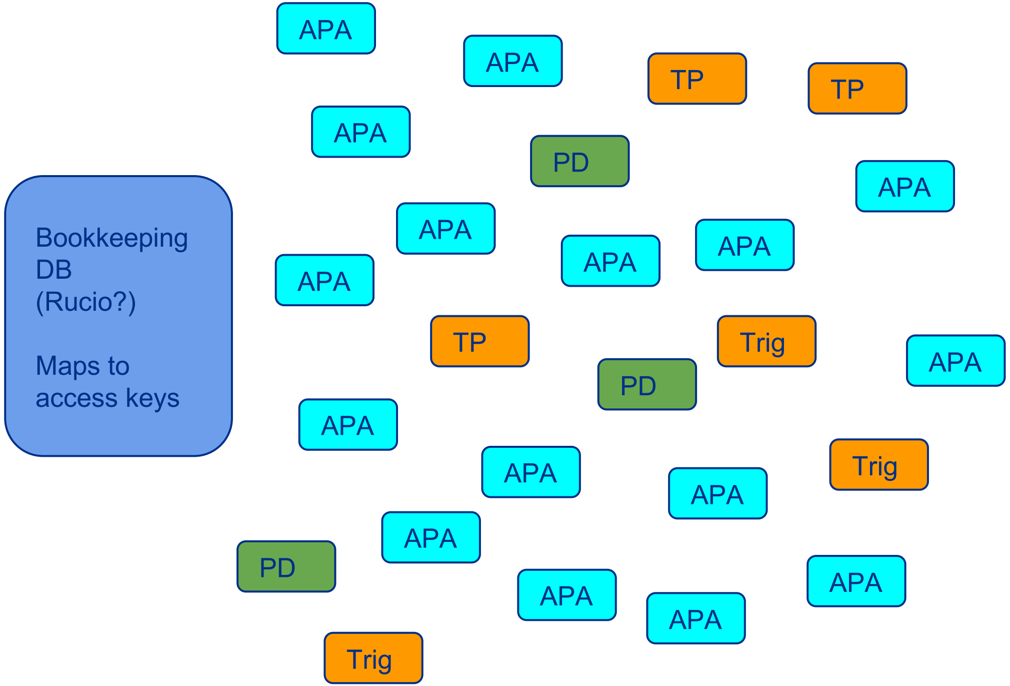

PD APA APA APA APA APA APA TP Trig
PD APA APA APA APA APA TP Trig
PD APA APA APA APA APA TP Trig
Manifest

Time Extended (SNB) readout aggregate


PD APA APA APA APA APA APA TP
PD APA APA APA APA APA TP
PD APA APA APA APA APA TP
Trig
Manifest

Object stores

- This structure would seem to fit an object store model
  - How could we fit this into the Rucio catalogue?
- No longer think about files
  - Split data into convenient chunks
  - Write and read data in any convenient order
- Involves multiple as yet unanswered questions
  - Cataloguing is more complex
  - Non-local access protocols?
  - Archive to tape, or to other SEs (probably can't avoid files after all)



Summary

- DUNE has started to use Rucio
- Progressively integrating into the existing DM system
  - Trying to avoid disruptive changes wherever possible
- Rucio features have been a good fit for DUNE requirements
  - More customization features would be good (and are being worked on)
- Long term plan is to exclusively rely on Rucio for DM and redo/replace other components to fully integrate with it
- Interested in potential use of new storage systems, such as object stores