Design of the Data Management Infrastructure for the protoDUNE experiment

S. Fuess[1]
O. Gutsche[1]
R. Illingworth[1]
M. Mengel[1]
A. Norman[1]
M. Potekhin[2,3]
B. Viren[2]

1 Fermilab
2 Brookhaven National Lab
3 presenter

Overview

- The **D**eep **U**nderground **N**eutrino **E**xperiment: DUNE
  - experiment overview
  - the Liquid Argon TPC (LArTPC)

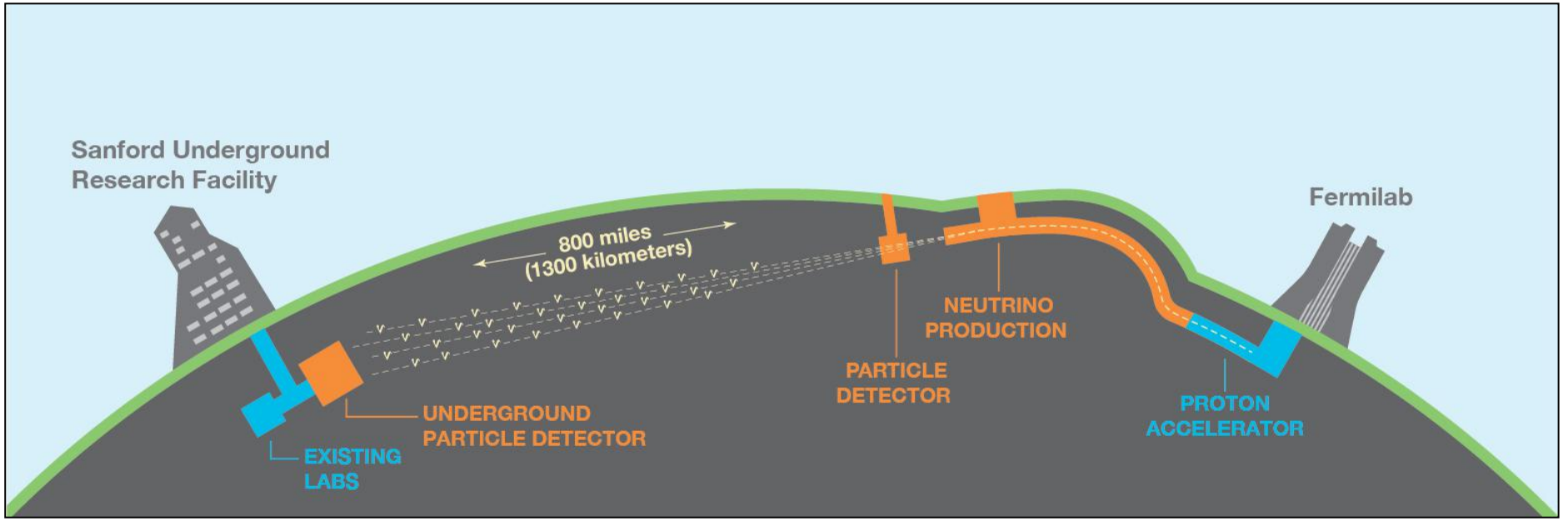
- protoDUNE
  - experimental program at CERN involving two large LArTPC prototypes
    - This presentation focuses on the Single-Phase prototype NP04
  - data volumes, characteristics, and challenges

- Data management for protoDUNE
  - scale and requirements
  - general design
  - components reuse
  - status and plans

DUNE: the Primary Science Program

- Precision measurement of neutrino oscillation parameters
- Search for proton decay in several modes, for example $p \rightarrow K\nu$
- Detection and measurement of the neutrino flux from core-collapse supernovae in our galaxy (should any occur during the lifetime of the experiment)

DUNE components

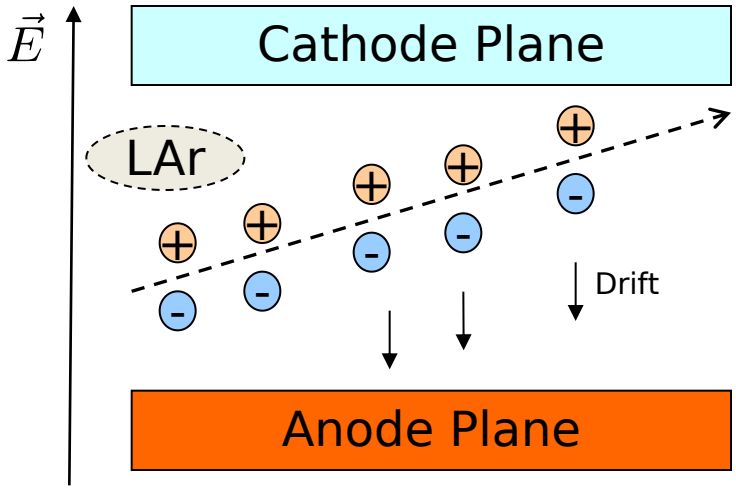


DUNE has been conceived around three central components:

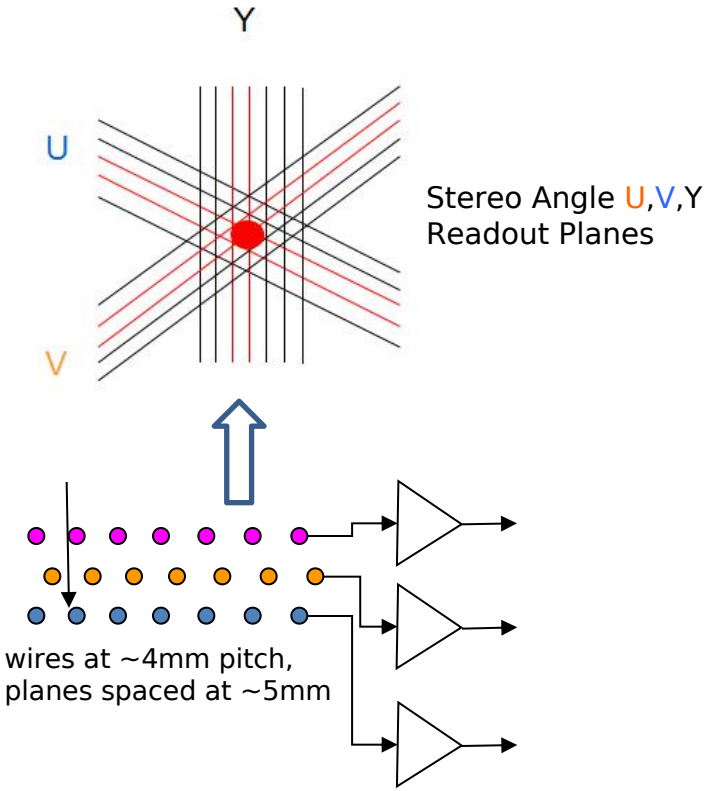
- an intense 1.2MW wide-band neutrino beam originating at FNAL
- a capable fine-grained near neutrino detector close to the neutrino source
- a massive 40kT Liquid Argon time-projection chamber deployed as a far neutrino detector 1,300 km from FNAL and 1.5km underground

Single-Phase LArTPC (DUNE and protoDUNE)

- One of the *two design options* for the modules of the DUNE Far Detector TPC
- Liquid Argon serves as both the target and the sensitive medium. LArTPC is essentially an ionization chamber with multiple sets of electrodes (wires)
- Planar arrays of sensor wires are grouped in the **anode assembly**, including two *induction planes* with wires at a **stereo angle** and the *collection* plane.
- Two coordinates (in the plane) are determined via stereo projections on three planes, and the third (along the drift) via the time measurement

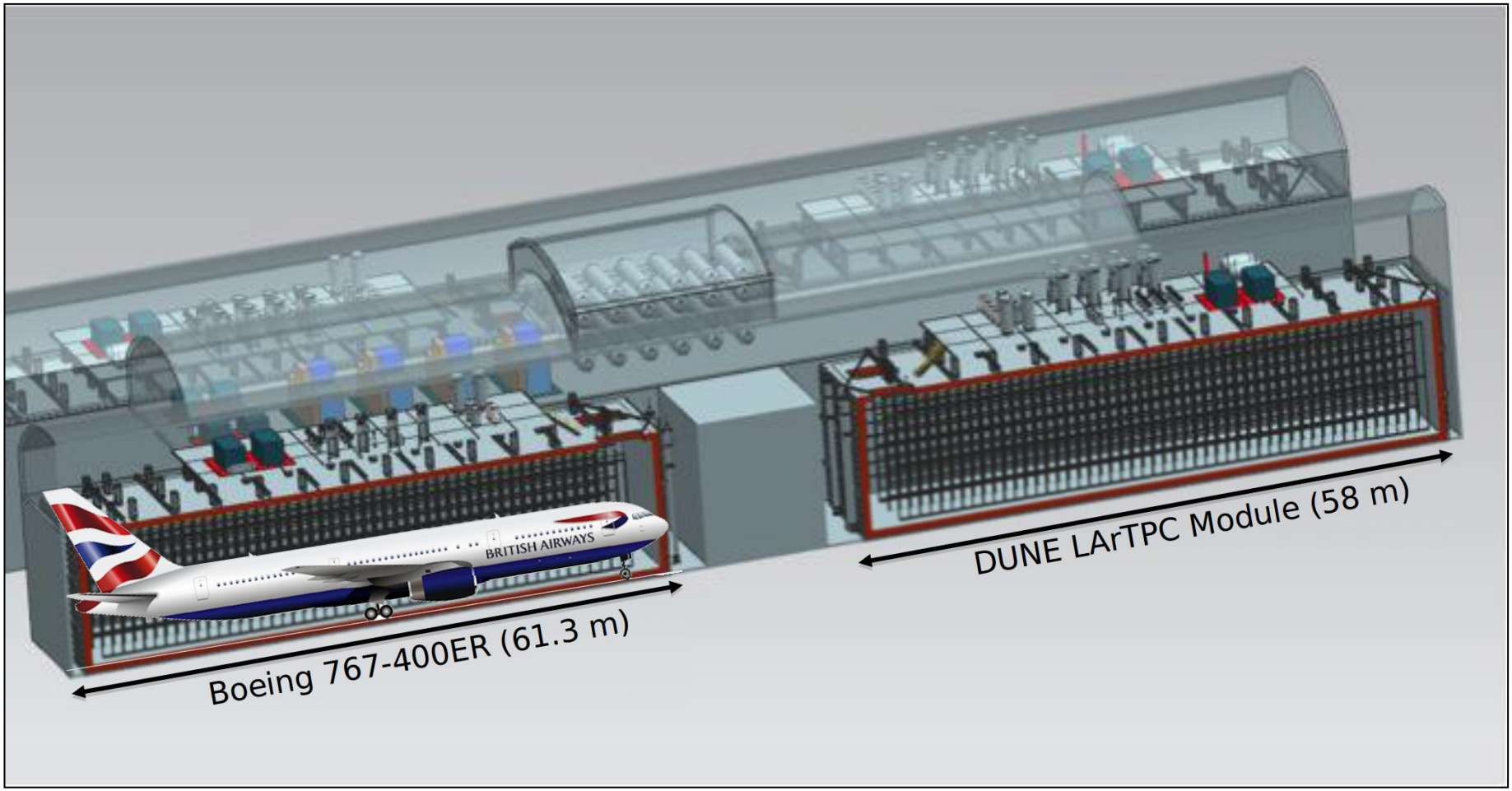


Two *induction planes* with wires at a stereo angle, and the collection plane: each wire connected to a single amplifier and readout circuit



DUNE LArTPC scale

- Four 10kt TPC modules (each 58m long)
- 1,536,000 TPC channels
- Integrated photon detector

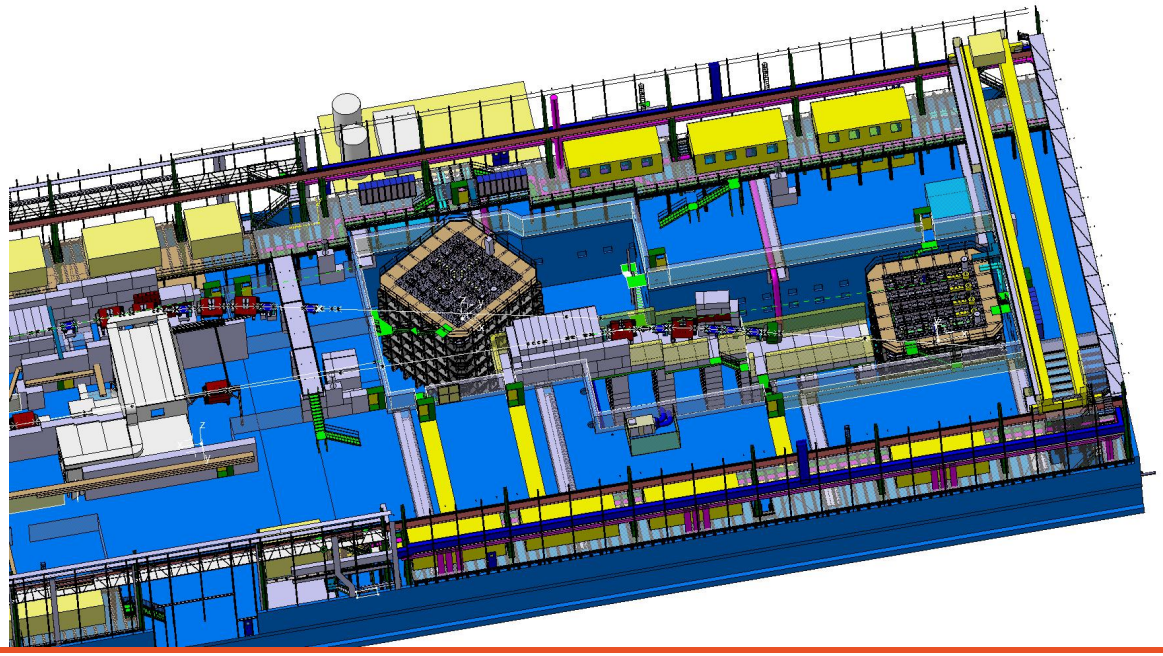


The protoDUNE program at CERN

- Engineering prototypes of Single- and Dual-Phase Liquid Argon TPCs (CERN designation NP04 and NP02 correspondingly)

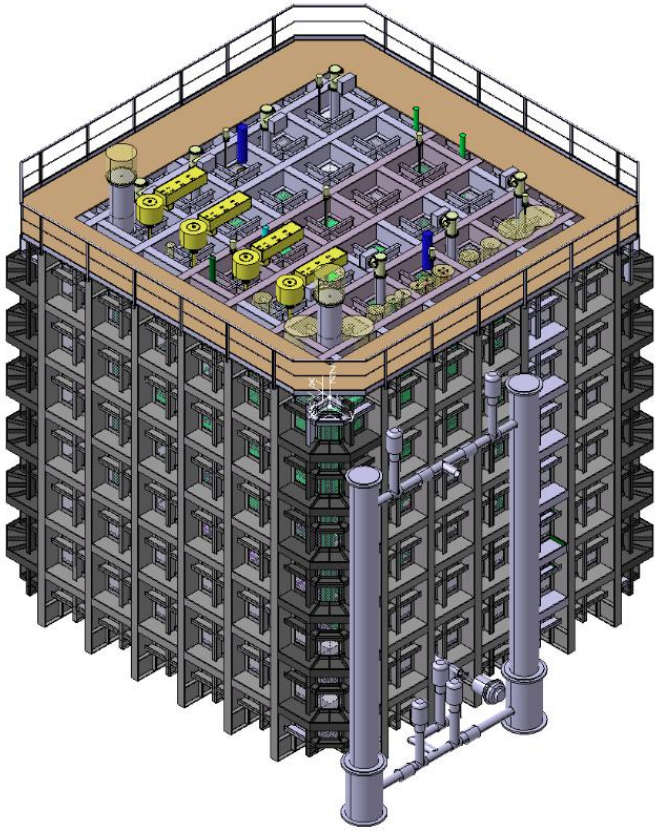
- Test-beam facility under construction in the extension of the North Area Hall, with a purpose-built tertiary beam from the SPS (H4) to provide various particle types

- In addition to validating the design of the detectors, protoDUNE will provide a unique opportunity for detector characterization and evaluation of reconstruction techniques in controlled test-beam conditions and with varying event types. Beam and cosmic ray triggers.

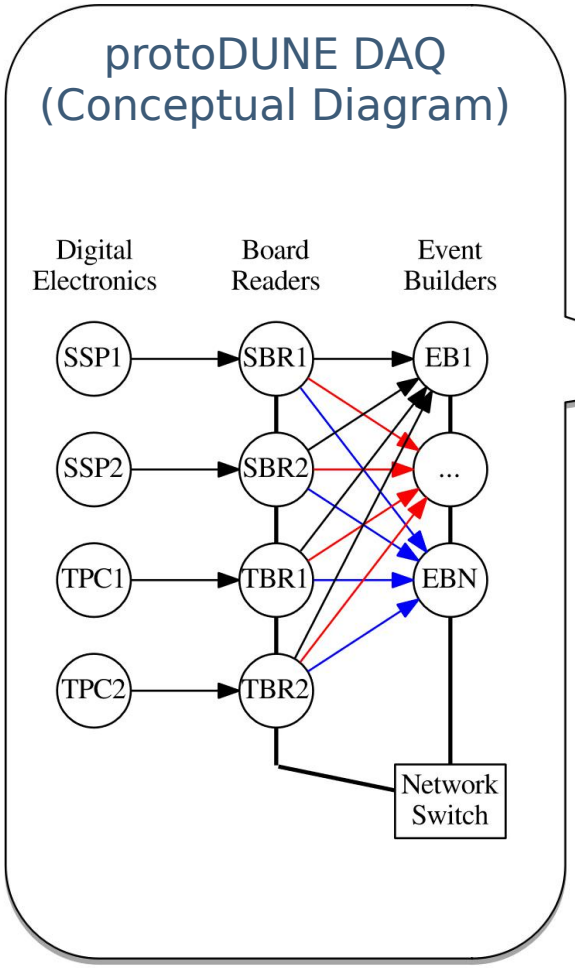


Single-Phase protoDUNE (experiment NP04)

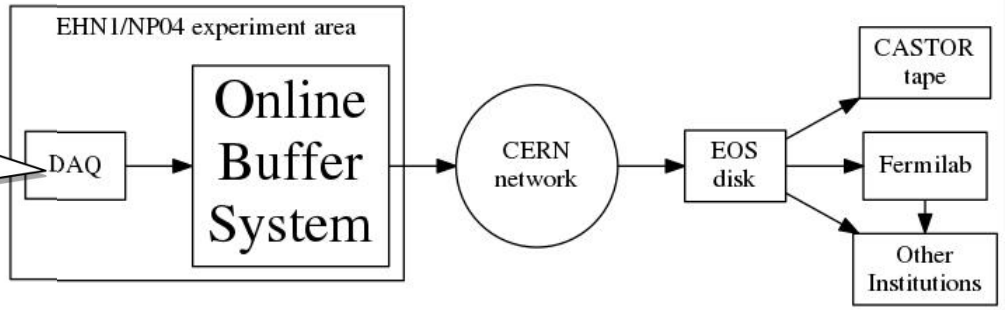
- Includes full-scale elements of the DUNE LArTPC: $2.3\times6.2m^2$ each
- TPC volume: $7.3\times7.4\times6.2m^3$
- External cryostat dimensions: $\sim11\times11\times11m^3$
- TPC channel count: 15,360
- Digitization frequency: 2MHz
- Nominal readout window: 5ms
- Nominal beam trigger rate: 25Hz
- Single readout size: 230MB
- Lossless compression factor: 4
- Post-compression peak data rate: 1.4GB/s
- Contingencies
  - higher trigger rate (50-100Hz)
  - lesser compression
  - larger number of cosmic triggers
- To size up the data handling system, the benchmark of 3GB/s rate is assumed
- Nominal 20Gbps network bandwidth from the experiment to CERN central storage
- Up to 100TB daily data volume
- ~1.5PB of data to be collected during the run



Elements of the protoDUNE data flow



The online buffer is necessary to satisfy the CERN requirement for experiment-local storage and to absorb high instantaneous data rate coming out of DAQ



The online buffer characteristics:

- Nominal capacity: 300TB (3 days data)
- Data ingestion rate (peak): 3GB/s

High bandwidth implications:

- Requires raid storage/high-speed media
- SSD prohibitively expensive at this volume
- multiple HDDs required; cf. at nominal 50MB/s write access 60 spindles will be needed

The protoDUNE online buffer: design options

- Two design options under consideration
  - Primary design is storage appliance

Storage Appliance

- High-bandwidth storage appliance combined with a storage array
  - e.g. DELL PowerVault MD3 series
  - Small number of servers for computational tasks (e.g. DELL R630s)
- Fault tolerance built into storage array hardware
- Vertical scalability (scale-up):
  - Scales discretely in capacity and bandwidth
  - Two systems required
  - Large incremental cost
  - High unit cost
- New hardware to be purchased from DELL with support contracts

Distributed XRootD Cluster

- XRootD cluster with ~**60** nodes
- Fault tolerance achieved through redundancy of nodes
- Horizontal scalability (scale-out):
  - Add nodes elastically as needed
  - Small incremental cost
  - Low unit cost
- Older commodity servers
- Re-purposed/refurbished with HDD upgrades

Data transport − component reuse

- Requirements have been formulated for data transmission in protoDUNE which include fault tolerance, recovery, creation of metadata, accounting, monitoring etc.
- There requirements are met by a combination of systems developed and currently in use at FNAL:
  - F-FTS: the "Fermilab File Transfer System"
  - SAM: the file metadata and catalog system
- Both system successfully utilized in Intensity Frontier and HEP experiments
- F-FTS supports automation, recovery, monitoring, is compatible with a variety of protocols and can orchestrate third-party transfers
- protoDUNE online data handling infrastructure mimics the systems used in major LHC experiments, i.e.
  - The CERN EOS distributed disk storage which receives data from the online buffer
  - CASTOR - the principal central tape storage facility at CERN
- F-FTS will be used to manage data transfer between the buffer, EOS, CASTOR, FNAL and other sites (see next slide)
- Principal storage at FNAL will consist of dCache and "Enstore" - the FNAL tape mass storage
- The dCache instance at FNAL is accessible to outside clients via XRootD

Fermilab-FTS

- Designed originally for NOvA Experiment's DAQ Systems
- Proven track record of moving over 7.7 PB of data for NOvA and 4.7 PB of data for MicroBooNE.
- F-FTS fulfills all of the basic requirements of a typical HEP and/or IF experiment
- It includes automation, error recovery, monitoring, interaction with the metadata system
- Supports a variety of 3rd party transfer protocols
- Most common use case is "dropbox" logic:
  - Designated data ingest area ("dropbox")
  - Newly arrived files in "dropbox" are
    - Identified
    - Registered
    - Scheduled for transmission
    - Verified
    - Cleaned up after transfer

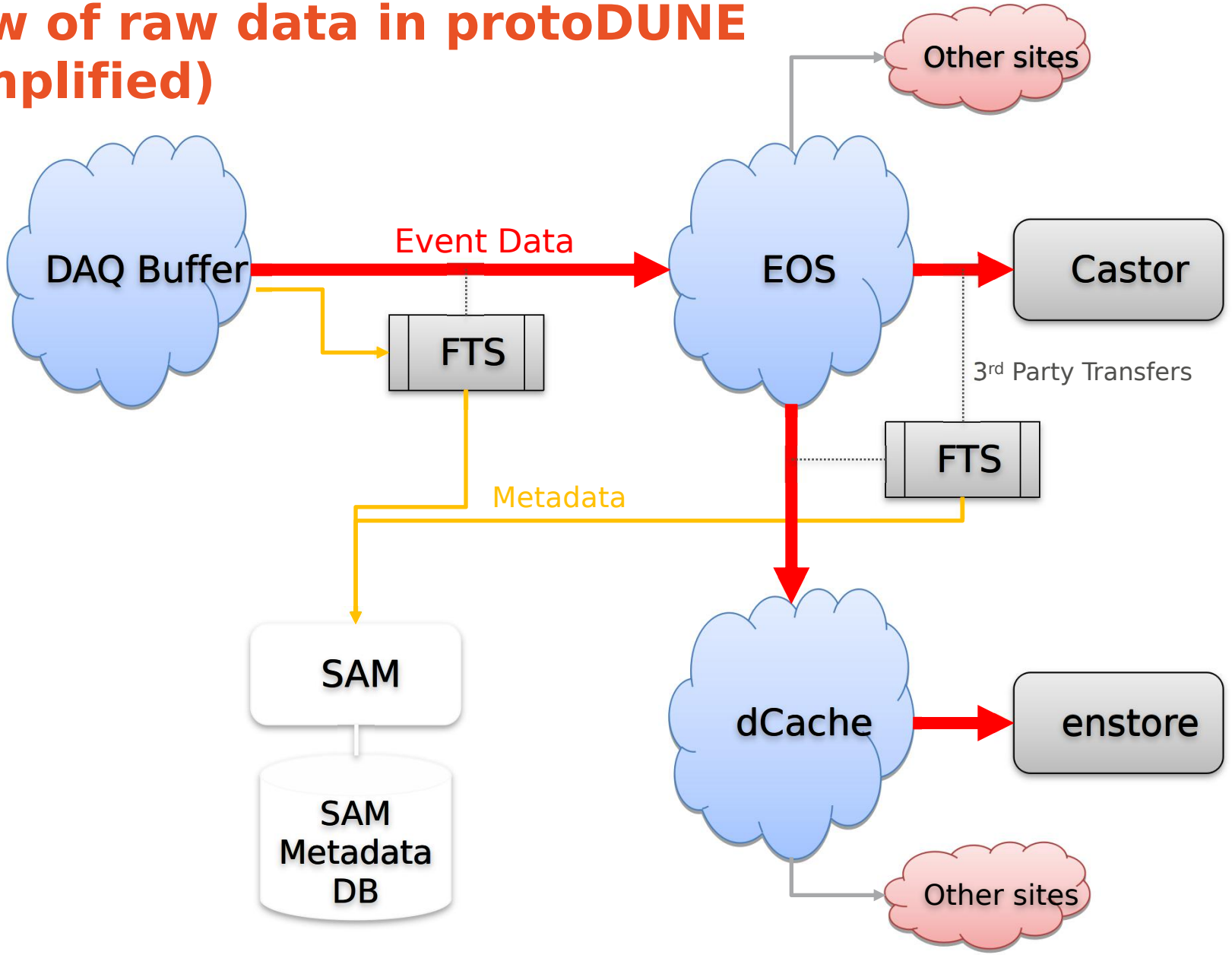
Challenges - metadata

- For the metadata to be useful, it may be necessary to extract some of its elements by reading a large fraction or all of the data file written to the buffer
  - in the very high I/O bandwidth scenario of protoDUNE this put additional stress on storage
  - checksum is often considered a crucial part of the file metadata
  - if checksums are required from the very beginning of the data transfer chain, reading of the complete files is necessary
  - checksums require a non-negligible amount of CPU

Challenges - prompt processing

- In adddition to the basic monitoring there are plans for "prompt processing" - express streams for data QA
  - it is defined as processing of a portion of the raw data which takes place on the time scale of minutes rather than hours after data is taken
  - the scope of the prompt processing is now work in progress
  - depending on the eventual decision and available resrouces, it can be limited to calculating the channel noise spectra (FFT), or in the high-end scenario include complete reconstruction of a subset of events
  - will require basic workflow management with high-velocity data
- CERN EOS appears to be the ideal platform for staging the data
- Pilot-based workload distribution will solve the intrinsic latency of a batch queue system

Flow of raw data in protoDUNE (simplified)



Status and plans

- Characteristics of the protoDUNE data have been estimated and documented, and will drive the design and scale of the systems to handle raw data
- protoDUNE will leverage the CERN storage infrastructure in a way similar to the LHC experiments (cf. the data rate and volume comparable to ATLAS in Run 1)
- Data transport and metadata systems (F-FTS, SAM) developed and maintained by FNAL will be used for data transfer and distribution.
- The online buffer is a critically important component which must operate at a very high data rate (up to 3GB/s) and provide substantial storage volume
- Various design options for the buffer have been investigated and DELL PowerVault platform has been chosed as the primary option and will be tested at CERN in 2017
- Prompt processing is in the initial phase of development
- Deployment of the data handling system at CERN will start in May 2017