Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Design of the protoDUNE raw data management infrastructure

To cite this article: S Fuess *et al* 2017 *J. Phys.: Conf. Ser.* **898** 062036

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 062036 doi:10.1088/1742-6596/898/6/062036

Design of the protoDUNE raw data management infrastructure

S Fuess[1], R Illingworth[1], M Mengel[1], A Norman[1], M Potekhin[2] and B Viren[2]

[1] Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
[2] Brookhaven National Laboratory, Upton, NY 11973, USA

E-mail: potekhin@bnl.gov

Abstract. The Deep Underground Neutrino Experiment (DUNE) will employ a set of Liquid Argon Time Projection Chambers (LArTPC) with a total mass of 40 kt as the main components of its Far Detector. In order to validate this technology and characterize the detector performance at full scale, an ambitious experimental program (called "protoDUNE") has been initiated which includes a test of the large-scale prototypes for the single-phase and dual-phase LArTPC technologies, which will run in a beam at CERN. The total raw data volume that is slated to be collected during the scheduled 3-month beam run is estimated to be in excess of 2.5 PB for each detector. This data volume will require that the protoDUNE experiment carefully design the DAQ, data handling and data quality monitoring systems to be capable of dealing with challenges inherent with peta-scale data management while simultaneously fulfilling the requirements of disseminating the data to a worldwide collaboration and DUNE associated computing sites. We present our approach to solving these problems by leveraging the design, expertise and components created for the LHC and Intensity Frontier experiments into a unified architecture that is capable of meeting the needs of protoDUNE.

1. Introduction

The protoDUNE program will help validate various DUNE technology aspects before proceeding with the construction of the large-scale principal DUNE detectors at the Sanford Underground Research Facility [1, 2]. The program is designed to make a series of measurements on the interaction of charged particles in the Liquid Argon medium. These measures will be performed in a test beam provided by a new dedicated target station and beamline transport system at the CERN SPS accelerator complex. This "neutrino platform" facility has the potential to be an important resource for the characterization of the DUNE/protoDUNE Liquid Argon Time Projection Chamber (LArTPC) detectors and for providing realistic electromagnetic and hadronic shower response functions for different particles interacting in these detectors. The experimental program includes two separate large LArTPC prototypes, one based on a "single-phase" (liquid) technology and one based on a "dual-phase" (liquid/gaseous) TPC readout technology, with CERN experiment designations **NP04** and **NP02** respectively. Both detectors are scheduled to be deployed at CERN in 2017, with beam data taking in 2018.

The two detector apparatuses, when deployed, will share parts of their computing infrastructure. Their data acquisition (DAQ) systems and corresponding data buffers will operate in an independent manner. The focus of this paper is on the data management for

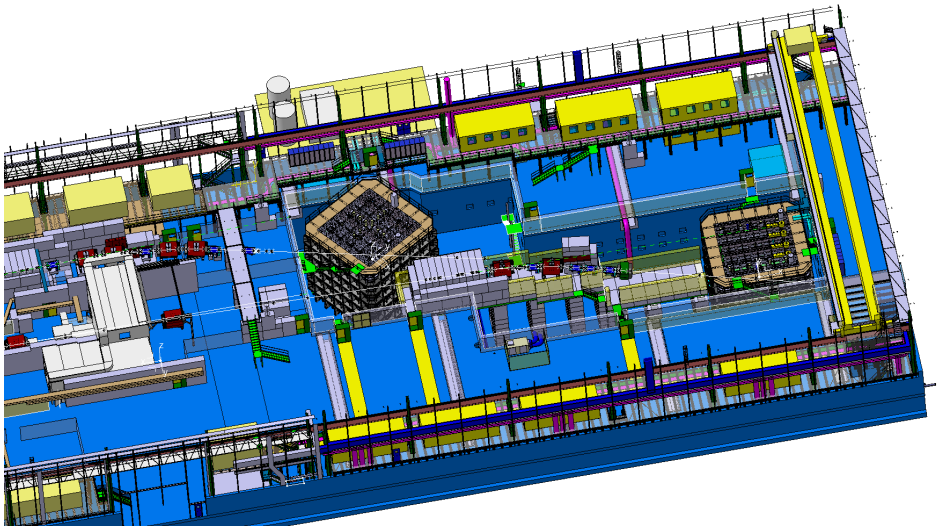

Figure 1. Diagram of the layout of the CERN north area with staging of the protoDUNE dual phase detector (center) and single phase detector (right).

the single-phase LArTPC (NP04). This detector functions without amplification in the Liquid Argon medium and is in essence a large single volume ionization/drift chamber. The drift volume is equipped with a large number of readout electrodes (wires), each with its own electronics chain. In this "cold electronics design" employed by **NP04**, the front-end electronics is situated within the cryostat and operates at cryogenic temperatures in order to minimize noise.

Both NP02 and NP04 detectors will be placed in an extension of the CERN North Area Experimental Hall. The general layout of the experimental area is shown in Fig. 1. Within these areas there will be enclosures with sufficient power and cooling capacity to house the elements of the protoDUNE DAQ readout and computing infrastructure which needs to be in close proximity to the physical detectors. These enclosures are shown schematically as yellow blocks in the upper-right portion of Fig. 1. From these counting houses, each prototype detector will be have a dedicated 20 Gb/s network connection over optical fiber to the CERN central storage facilities located in the West Area campus of CERN.

2. LArTPC as a Data Source

DUNE/protoDUNE LArTPC design has a fine spatial resolution due to the high channel count and spatial granularity, wire pitch and geometry, of the readout planes. The readout design also takes advantage the relatively slow drift velocity of ions in the Argon and the 2 MHz digitization clock to achieve similarly fine spatial and temporal resolution along the drift axis of the detector. The readout window is set to 5 ms which is driven by the electron drift time across the active volume (which is slightly less). These characteristics are typical of similar modern TPC detectors.

In the absence of zero suppression these factors do however result in a large data volume that needs to be read out and processed for a single readout. A DUNE/protoDUNE readout event may be compared to sequential stack of a thousand of digital images of the signals collected from the electrodes. The instantaneous triggered data rate and the data rates averaged over the beam's spill and duty factors are expected to be primarily limited by the network bandwidth available within the DAQ infrastructure and the I/O bandwidth available for recording the data. Lossless data reduction/compression (e.g. Huffman encoding) techniques are applied to the data

Table 1. Detector, beam, DAQ and data rate parameters for the single phase protoDUNE experiment (NP04).

Detector Parameter	Target
TPC channel count	15,360
Digitization frequency	2 MHz
Readout window	5 ms
SPS spill time	4.8 s
SPS cycle	22.5 s
Trigger rate (target minimum/maximum)	25–100 Hz
Single readout size (per trigger)	230.4 MB
Target compression factor	×4
Instantaneous data rate (in-spill)	1440 MB/s
Average data rate	576 MB/s
Total data recorded (beam + cosmic)	2.5 PB
Buffer to store 3 days worth of data	300 TB

within the DAQ readout chain and are projected to result in a factor of 4x reduction to the raw data. Based on these projections, the NP04 experiment is expected to accumulate 2.5 PB of physics beam data during 3 months of running. The total data volume accumulated by the experiment is expected to be larger when data from detector commissioning prior to the beam run, and cosmic ray data from after the beam running are included. These rates are detailed in Table. 1.

3. Raw Data Flow in protoDUNE

3.1. DAQ and the Online Buffer

Details of the DAQ design and Online Monitoring in protoDUNE are outside of the scope of this document but it is helpful to note a few details relevant for the topic at hand.

The DAQ has a few interconnected layers of processes running on dedicated computers. The "outer layer" of the system where the data is put together in a format suitable for writing into files consists of *Event Builders*. The Event Builders are hosted on a few machines equipped with quad 10 Gbps NICs and connected to a high-bandwidth switch. Their function is to assemble complete readout frames from data fragments obtained from the inner layer of DAQ, the *Board Readers*.

The Event Builders create the necessary file-level header information and write the resulting data to the *Online Buffer* which is attached to the protoDUNE DAQ. The design of this buffer is inspired by a system sucessfully employed in ATLAS experiment and is based on COTS storage hardware from DELL which features a high degree of redundnacy. It is structured as three identical high-capacity storage systems, i.e. scales vertically. Each system contains

- Two front-end nodes (DELL R620 or similar)
- Storage Controller (DELL MD3420 or similar)
- Four expansion shelves (DELL MD1220 or similar)
- A total of ∼120 HDDs

In addition to redundancy, the high HDD count is the main factor which allows to commit protoDUNE data to the disk at the high rate according to the experiment plan (Sec. 2) and ensure robust operation.

A backup (auxiliary) option is to implement the buffer as a a XRootD [3] cluster that may be deployed on existing general purpose hardware provided by the *CERN Neutrino Platform* orgranization. A prototype of such system has been tested.

3.2. The Data Flow pattern in protoDUNE

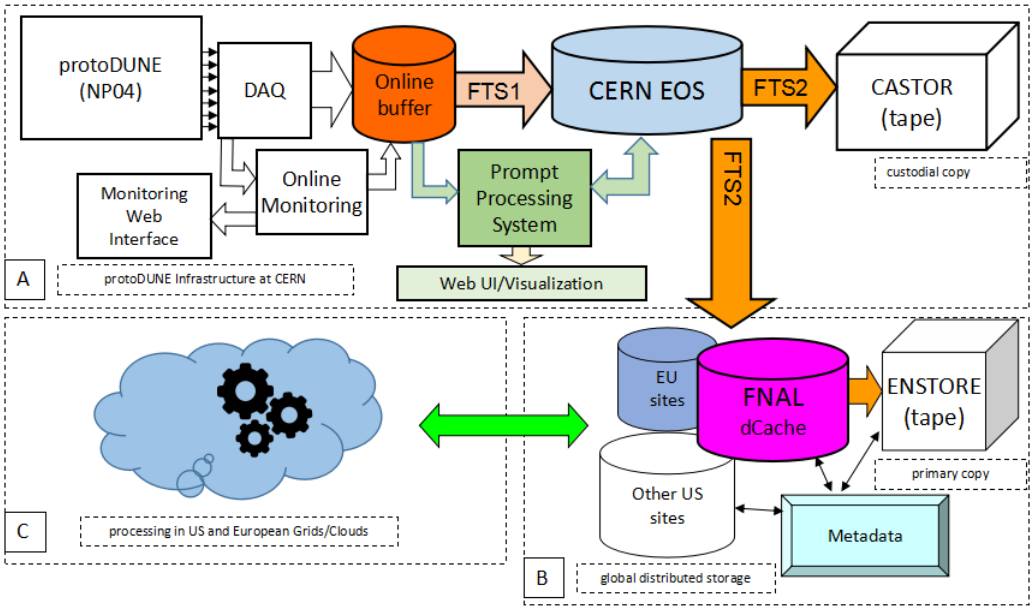

Figure 2. Conceptual diagram of the flow of raw data in protoDUNE

In the following, we shall consider handling and management of the raw data after it is written by the Event Builders to the Online Buffer. A conceptual diagram of the raw data flow in protoDUNE is presented in Fig.2 which shows the general pattern of data flow and also reflects the central role of the EOS system at CERN [4] in the raw data management scheme. Reliance on EOS is motivated by the experience and architecture of the LHC experiments as well as the protoDUNE data characteristics presented above in Sec. 2.

The elements in this diagram which contain "FTS" in their label correspond to components and instances of the *Fermi File Transfer Service* – FTS for short [5, 6] – which transports data between predefined endpoints. Fermi FTS was designed originally for the NOvA experiment [7] DAQ system and has a proven track record of moving over 7.7 PB of data for NOvA and 4.7 PB of data for MicroBooNE [8]. The Fermi FTS contains functionality which allows for completely automatic operation including error handling, transmission retries, monitoring etc.

There is more than one instance of FTS in the protoDUNE data transmission chain. One instance (labeled "FTS1" in the diagram) is responsible for moving the data from the Online Buffer into EOS. The second instance – "FTS2" – copies the data after it has arrived to EOS to tape storage at CERN (the CASTOR system [4]). It is also tasked with transmitting the data to Fermilab and other peer institutions. At FNAL the data are placed in a dCache [9] storage cluster. A secondary tape copy is then created at Fermilab using the data resident in dCache as the source. Other participating data centers receiving their copies of raw data may function differently e.g. use other methods of local storage federation.

3.3. The "Dropbox" Logic

In both instances an FTS agent is triggered by arrival of files to a designated storage location which is expected to be accessible in a POSIX-like mode. It effectively serves as a "dropbox",

i.e. files are selected for transfer asyncronously based on certain criteria (such as the filename pattern etc) and subjected to following operations:

- Identified
- Registered
- Scheduled for transmission
- Verified
- Cleaned up after transfer

3.4. Metadata

An important feature of the FTS is its interface to and integration with the *SAM* Metadata system deployed at FNAL which serves the needs of about a dozen experiments in both Energy Frontier and Intensity Frontier domains. In addition to essential file catalog functionality, SAM has extensive storage management capabilities covering both disk (e.g. dCache) and tape (e.g. FNAL Enstore) types of storage. The most common way to associate metadata with a file in SAM is to package it as an auxiliary file in JSON format, following a certain naming convention. This file is then automatically detected by FTS and records in SAM are created in accordance with its content.

Operations related to metadata present certain challenges in the high-volume and high-rate environment of the protoDUNE data flow. For example, for the metadata to be useful, it may be necessary to extract some of its elements by reading a large fraction or even all of the data file written to the buffer. In the very high I/O bandwidth scenario of protoDUNE this puts substantial additional load on storage systems. Another example is checksum operations as the checksum is often considered a crucial part of the file metadata. If checksums are required from the very beginning of the data transfer chain, complete files must be read from the Online Buffer. In addition, handling checksums requires a non-negligible amount of CPU.

The current plan is to utilize the checksum calculation capabilities provided by XRootD during both legs of the transfer: it is first calculated by the "FTS1" instance (see 3.2) and added as a part of the metadata. It is then used in further transfers by "FTS2" to CASTOR and dCache.

3.5. Transfer Protocols

A number of protocols are supported by both EOS and FTS inclusing XRootD, gridFTP and third-party transfers. Prime candidate for protoDUNE is XRootD which has proven scalability and reliability.

In addition to serving as the principal staging area from which the data is copied to tape storage at CERN and from which it is transmitted to FNAL, EOS will also be used to provide transparent access to data to the protoDUNE prompt processing system in order support the Data Quality Moniotring (DQM) in protoDUNE as explained in Sec. 4. Its XRootD interface makes it a particularly attractive option.

4. Data Quality Monitoring

Data Quality Monitoring (DQM) plays an important role in protoDUNE. Its goal is to generate time-critical information on a short time scale needed to ascertain the condition and performance of both the detector and the DAQ, in order for operators to quickly detect problems, take action and prevent loss of useable data and/or beam time.

DQM consists of two parts, the low latency Online Monitoring (OM) system which is engineered as a component of DAQ, and the "prompt processing system". This is reflected in the diagram in Fig.2. These systems are complementary to each other but operate in

different enviroments and on a different time scale. The benchmark turnaround time for "prompt processing" jobs is set roughly at 10 min, with actual numbers depending on finalized content of the payloads. Jobs will be organized in stages and subject to prioritization to guarantee timely production of DQM visual and data products. A prototype system has been developed at the time of writing. The final stages of prompt processing will include simplified event reconstruction based on LArSoft framework developed at FNAL.

Compared to prompt processing, Online Monitoring aims to provide response on the scale of seconds and in general under a minute. The prompt processing system processes only a small fraction of the data but can potentially perform more sophisticated calculations since it is easier to scale out its CPU capabilities. Both OM and prompt processing system will be located in the vicinity of the NP04 detector. It may be necessary to scale out parts of prompt processing to the CERN central batch facility so the system is designed to support this mode of operation as well.

5. Summary

The single-phase protoDUNE experiment will generate data at a very high rate with a few PB of raw data to be recorded in mass storage during the time of its operation. To meet the challenges associated with handling these data the DUNE Collaboration opted to leverage designs, software and components from a few HEP and Intensity Frontier experiments. This includes

- The design of the Online Buffer
- The principal data transfer system (Fermi FTS) and the Metadata system (Fermilab SAM)
- XRootD storage federation technology
- LArSoft framework for Liquid Argon detector simulation and reconstruction

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science. The Fermi National Accelerator Laboratory is operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy. Brookhaven National Laboratory is operated by Brookhaven Science Associates under contract No. DE-SC0012704 with the United States Department of Energy.

References

[1] R Acciarri et al. "Long-Baseline Neutrino Facility (LBNF) and Deep Underground Neutrino Experiment (DUNE) Conceptual Design Report Volume 1: The LBNF and DUNE Projects". e-Print: arXiv:**1601.05471**
[2] R Acciarri et al. "Long-Baseline Neutrino Facility (LBNF) and Deep Underground Neutrino Experiment (DUNE) Conceptual Design Report, Volume 4 The DUNE Detectors at LBNF". e-Print: arXiv:**1601.02984**
[3] L Bauerdick et al. "Using Xrootd to Federate Regional Storage." *J. Phys.: Conf. Series.* Vol.**396**. IOP Publishing, 2012.
[4] L Mascetti et al. "Disk storage at CERN." *J. Phys.: Conf. Series.* Vol.**664**. IOP Publishing, 2015.
[5] R A Illingworth "A data handling system for modern and future Fermilab experiments." *J. Phys.: Conf. Series.* Vol.**513**. IOP Publishing, 2014.
[6] Norman A 2014 *The Fermilab File Transfer System.* e-Print: FNAL CD-DocDB-5412
[7] R Plunkett et al. "Status of the NOvA Experiment." *J. Phys.: Conf. Series.* Vol.**120**. IOP Publishing, 2008.
[8] B Jones et al. "The Status of the MicroBooNE Experiment." *J. Phys.: Conf. Series.* Vol.**408**. IOP Publishing, 2013.
[9] A Millar et al. "dCache, agile adoption of storage technology." *J. Phys.: Conf. Series.* Vol.**396**. IOP Publishing, 2012.