A Proposal for the Single-Phase protoDUNE Experiment Data Challenges in 2017-2018

R.Pordes, M.Potekhin, R.Sulej

August 1, 2017

Abstract

It is important to ensure 100% readiness of the end-to-end protoDUNE software and computing complex during the detector commissioning period and throughout data taking in 2018. We propose to conduct two Data Challenges in order to identify and address potential issues before they can impact the experiment.

Contents

1 About this document

The goal of this document is to provide concise information about the proposed protoDUNE Data Challenges to the individual working groups in order to coordinate effort and come to a consensus as to the scope, plans and schedule of the proposed Data Challenges. *It does not contain detailed descriptions and/or designs of the protoDUNE computing infrastructure elements* and the reader is reffered to existing documentation where needed, with references provided in the text.

2 The Scope of the Data Challenges

Information regarding the protoDUNE (NP04) computing is summarized in the protoDUNE-SP Technical Design Report [**?**]. The following components (including infrastructure, middleware and software) can be identified as relevant in the context of Data Challenges (which will be also referred to as *"DC"* in this document):

- The DAQ **Online Buffer** which stores raw data assembled by the Event Builders, as files on disk. The internals of the DAQ system itself are not within the scope.
- **Beam Instrumentation Interface**. Most of the Beam Instrumentation data is not included into the data stream handled by the protoDUNE DAQ (whose primary task is to capture data from the TPC and the Photon Detector), and is instead transmitted to the Beam Instrumentation Database at CERN. It is not available immediately for each triggered event, and is instead available in batches after each spill cycle of the SPS. Interfacing this system is a task that needs to be addressed for protoDUNE to be able to include these data in its offline processing in an optimal manner.
- The **Slow Control and Online Monitoring System** which provide control, monitoring and display of the data acquisition readout and detector parameters.
- The **Raw Data Management System** [**?**] which performs data transfers between a few endpoints at CERN and FNAL starting with the Online Buffer and which is also tasked with proper accounting and handling of Metadata by interacting with the SAM system at FNAL (see next item). The Data Handling System will inteface the disk-based mass storage at CERN and FNAL (i.e. EOS [**?**] and dCache) as well as tape systems (CASTOR and Enstore respectively). We anticipate using the *Fermi File Transfer System* (also abbreviated as F-FTS [**?**]) for most of this functionality.
- The **Metadata** – SAM system at FNAL which comprises the functionality of the file catalog, data storage and retrieval based on Metadata and can also be used for orchestration of production workflows.
- **Data Quality Monitoring** (DQM) which is tasked with running algorithms with turnaround time short enough to provide actionable results (under and hour) but which won't fit into the computational footprint of the Online Monitoring (a part of DAQ). To support DQM, the ProtoDUNE Prompt Processing System (**p3s**) [**?**, **?**] has been created and will run any type of DQM workload as formulated and programmed by the working groups.
- **Calibration and Production Software**. While it is not expected that these components will be finalized (and some even exist) at the time of the first Data Challenge it is important

to have a firm grasp of the required interfaces, data flow patterns and other crucial aspects of the protoDUNE Production Systems. Having prototypes in place is therefore crucial on the time scale of the Data Challenges. An important component of the overall production chain is the *Data Reduction* step along the lines described in [**?**]. It combines a few signal-processing procedures (cf. digital filtering and noise reduction), and it will be necessary in any version or architecture of the production software. Understanding the interfaces and practical implications of this component would be a useful part of the Data Challenge.

- **Analysis Suite Prototypes**. Same comment applies here as to the previous item, i.e. while there is no expectation for the final design and implementation of the analysis chain to exist early in the experiment (especially that it is by nature the most dynamic and fluid part of all software) it is important to put in place, document and optimize its interfaces with various components of infrastructure e.g. calibration databases, Metadata, software and data provenance controls etc.

- **Production Operations Management Service (POMS)**. This FNAL-based service [**?**] manages and automates jobs submission on distributed resources on the Grid. It will be the primary platform for protoDUNE on which to run production and analysis.

Relationship between the various infrastructure and other components listed above is schematically illustrated in Fig.**??**. This diagram relflects the fact that transmission of protoDUNE data is a multistage process and in particular there will be separate instances of F-FTS transferring the data from the Online Buffer to mass storage at CERN, and then from CERN to FNAL and elsewhere.

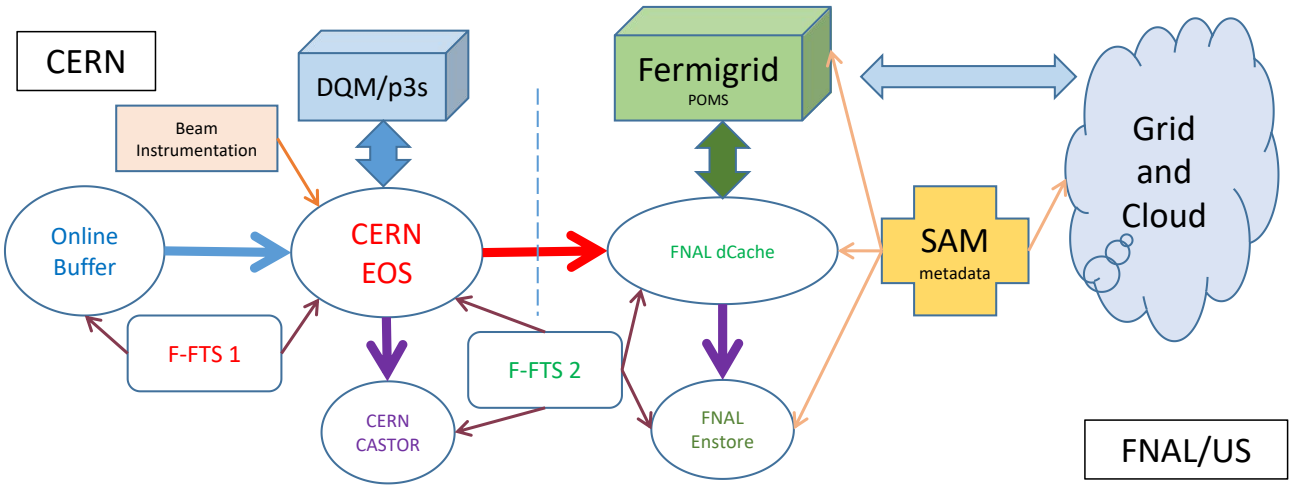

Figure 1: Schematics of data flow and processing in protoDUNE data challenges.

Data and production streams which belong to protoDUNE Monte Carlo simulation domain are not included in the scope of Data Challenges presented here.

3 Plans for Data Challenges

3.1 Work-up to the Data Challenges

Before any end-to-end data challenge can be meaningul, all individual components and interfaces must undergo their own functional testing. The following non-inclusive list illustrates the scope of

the pre-DC activity:

- **Slow Control and Online Monitoring**
- **Data Transfer and SAM interaction**. Both F-FTS1 and F-FTS2 components as schematically depicted in Fig. **??** must undergo end-to-end functional testing including registration in SAM and sinking data to tape storage both at CERN and FNAL. A test agent will be created which simulates creation of the raw data files by DAQ, which are written to the Online Buffer and picked up by the F-FTS for transmission.
- **Beam Instrumentation Interface**. At a minimum, prototype schemas for protoDUNE beam detectors must be created in the CERN Beam Instrumentation Database, and populated with dummy data. An extraction process must be put in place which "captures" these data and commits it either to a database managed by protoDUNE and/or as files on disk (EOS). A prototype of the merging process must be put in place. This will also require data access libraries to be utilized in the production process.
- **DQM/p3s**. The prompt processing system must be tested and validated to run with the primary data staging area being in EOS. The issues of deployment, updating and maintenance of LArSoft and other requisite software must be resolved.
- **Data Reduction Software**. A prototype of this software module must be tested with simulated data and packaged in a way that is portable across FNAL and facilities at CERN.
- **Calibration and Production Software**. Prototypes of this software need to be put in place. They will include components specific to LArTPC, the Photon Detector and the Beam Instrumentation data, which will also require merging and processing.
- **Production Management and Support**. This includes running prototypes of this software utilizing the FNAL POMS.
- **Analysis**. The main point in this exercise to to ready the interface of the analysis software to SAM and other components of the production system(s). There are no strict requirements to functionality and performance of the analysis software at this stage.

The work-up period must be completed no later than September 2017. It is expected that Data Reduction, Production, DQM and othe components will have substantial overlaps and/or commonalities, which must be exploited from the point of view of software sharing and reuse. One of the goals of the work-up period is to identify such items.

3.2 Timeline, Reporting and Documentation

Because of the significant CERN involvement in protoDUNE computing, and to facilitate communication with the CERN IT management it appears optimal to document the working plans and progress of Data Challenges on the CERN TWiki pages provided by CENF. More detailed technical documentation regarding components and systems will be managed on DUNE DocDB as before.

The work-up as described above should be completed by September 2017, and reports for each component provided to the protoDUNE leadership and the protoDUNE-CERN computing liaison representative. DC1 will take place in October 2017. DC2 will be conducted in February 2018. Each Data Challenge will result in a concise report which will be made available to the Collaboration and follow-up shall be organized as needed.

3.3 DC1 Scenario

- Simulated raw data is deposited into the Online Bufer by a specially created test agent.
- The data is detected by F-FTS1 and transmitted to EOS.
- F-FTS2 initiates data transfer to dCache at FNAL. Contact is made with the Metadata system (SAM) where the files are registered.
- Beam Instrumentation data is captured by protoDUNE systems and is stored either on a DB server or as files on EOS. These data do not need to be real at this point, as the purpose of this step is to test interfaces and protocols.
- Automated process initiates DQM streams in p3s. In DC1, full-blown processing on CERN Tier-0 [**?**] will not be required and at a minimum the operation of p3s will be validated on the Neutrino Platform cluster [**?**].
- Production team at FNAL submits production jobs using the newly arrived data as input. This includes prototype interfaces to software components of the LArTPC, Photon Detector and Beam Instrumentation systems.

3.4 DC2 Scenario

- DAQ sends raw data (such as pulser or simulated trigger data) to the Online Bufer.
- The data is detected by F-FTS1 and transmitted to EOS.
- F-FTS2 initiates data transfer to dCache at FNAL. Contact is made with the Metadata system (SAM) where the files are registered.
- Realistic Beam Instrumentation data is captured by protoDUNE systems and is stored either on a DB server or as files on EOS.
- Automated process initiates DQM streams in p3s using CERN Tier-0. At a minimum, ADC and noise spectra (FFT) are produced and delivered to the user by a Web service.
- Production team at FNAL submits production jobs using the newly arrived data as input. Interfaces between software components, data etc as well as actual functionaltiy are close to final.
- Analysis chain is activated.

4 Personnel and Responsibilities

This list is currently being revised.

- Online Buffer and the test agent – DAQ team
- Beam Instrumentation Interface – Beam Instrumentation Group, DB coordinator: J.Paley
- F-FTS – FNAL team (lead: A.Norman)

- p3s – prompt processing team (lead: M.Potekhin)
- DQM and production payloads – DRA team (co-leads: D.Stefan, R.Sulej)
- Calibrations – M.Mooney
- Integration consultant – B.Viren
- Metadata and SAM – FNAL SCD (T.Junk, S.Fuess)
- Production support at CERN and FNAL – TBD
- Analysis – TBD
- Coordination – R.Pordes

References

[1] DUNE DocDB 1794: *ProtoDUNE-SP Technical Design Report*
http://docs.dunescience.org:8080/cgi-bin/ShowDocument?docid=1794

[2] DUNE DocDB 1212: *Design of the Data Management System for the protoDUNE Experiment*
http://docs.dunescience.org:8080/cgi-bin/ShowDocument?docid=1212

[3] The CERN Exabyte Scale Storage
http://information-technology.web.cern.ch/services/eos-service

[4] The Fermilab File Transfer System
http://cd-docdb.fnal.gov/cgi-bin/RetrieveFile?docid=5412&filename=
datamanagement-changeprocedures.pdf&version=1

[5] DUNE DocDB 1811: *Prompt Processing System Requirements for the Single-Phase protoDUNE*
http://docs.dunescience.org:8080/cgi-bin/ShowDocument?docid=1811

[6] A Design of the Prompt Processing System for the Single-Phase protoDUNE experiment (NP04
)
http://docs.dunescience.org:8080/cgi-bin/ShowDocument?docid=1861

[7] Proposed Initial Data Reduction for protoDUNE/SP
http://cd-docdb.fnal.gov/cgi-bin/RetrieveFile?docid=2089&filename=
datamanagement-changeprocedures.pdf&version=1

[8] Production Operations Management Service (POMS)
https://cdcvs.fnal.gov/redmine/projects/prod_mgmt_db

[9] The CERN batch computing service
http://information-technology.web.cern.ch/services/batch

[10] Neutrino Computing Cluster at CERN
https://twiki.cern.ch/twiki/bin/view/CENF/NeutrinoClusterCERN