Notes for the **protoDUNE** computing infrastructure WBS

M.Potekhin

December 13, 2015

1 Introduction

1.1 Scope of this document

This document addresses work items related to components and systems to be deployed for handling of the data in protoDUNE after its capture in DAQ and before it is made available for managed production and subsequent analysis at sites like FNAL. Accordingly, transfer of raw data from the site of experiment to mass storage at CERN, transfer monitoring, replication to the data centers in the US, including interaction with metadata and other systems as needed, are within the scope of this document.

Design of DAQ and online systems such as run control, slow controls, conditions DB and offline processing are not considered in this document.

1.2 Scale of protoDUNE data

According to current plans, the following parameters and metrics will apply to protoDUNE data:

- Zero-Suppression will be used for practically all recorded data. There may be a short periods of time when full-stream data is recorded but this will an exception rather than the rule, since taking significant amount of full-stream data would require a substantially heavier infrastructure and storage footprint in the online systems, CERN, FNAL and elsewhere.
- Under conditions presented above, sustained data rate of ~200MB/s, while internal peak data rate in DAQ will be ~1GB/s (with ZS in place). This roughly corresponds to the rates in ATLAS Run 1.
- The nominal initial request for the network bandwidth was 2gbps, however in order to retain the possibility of taking a small portion of data in non-ZS mode and having a comfortable headroom for data transmission in general, it was decided to pursue a 10gbps connectivity.
- total amount of data to be recordrded is not well defined at this point, but estimates are O(100TB) which implies up to a 1PB scale.

Comparing of protoDUNE to ATLAS is of course not very straightforward. As far as the data rate is concerned, ATLAS clearly has a lot more diverse and complex hardrware, which leads to a considerabely more complex front-end electronics and trigger (including the high-level trigger). However, detailed discussion of these difference falls outside of the scope of this document and more importantly, general approach to data transfer to the CERN central services and beyond is quite similar between protoDUNE and the LHC experiments and will contain a comparable number of "moving parts".

1.3 Data Flow

Conceptual diagram of data flow in protoDUNE is presnted in Fig.1. The pattern is similar to the LHC experiments in the following aspects:

- DAQ writes its data to a disk buffer which is local to DAQ. From there, the data gets transfered to EOS (distributed mass disk storage at CERN), typically utilizing high-performance protocol such as *xrdcp*.
- Upon arrival to EOS, the data is committed to permanent tape storage at CERN (CASTOR).
- Production streams at CERN also rely on EOS-resident data. In case of protoDUNE, these would be calibration, monitoring. QA and other similar express streams.
- Data are distributed to participating data centers (e.g. in the US), and the bulk of the production is done on the Grid.
- Replication of data to remote data centers at the same rate as the data is collected from DAQ, i.e. 200MB/s (nominal). It means that the data will arrive to FNAL and other locations promptly after having been taken.

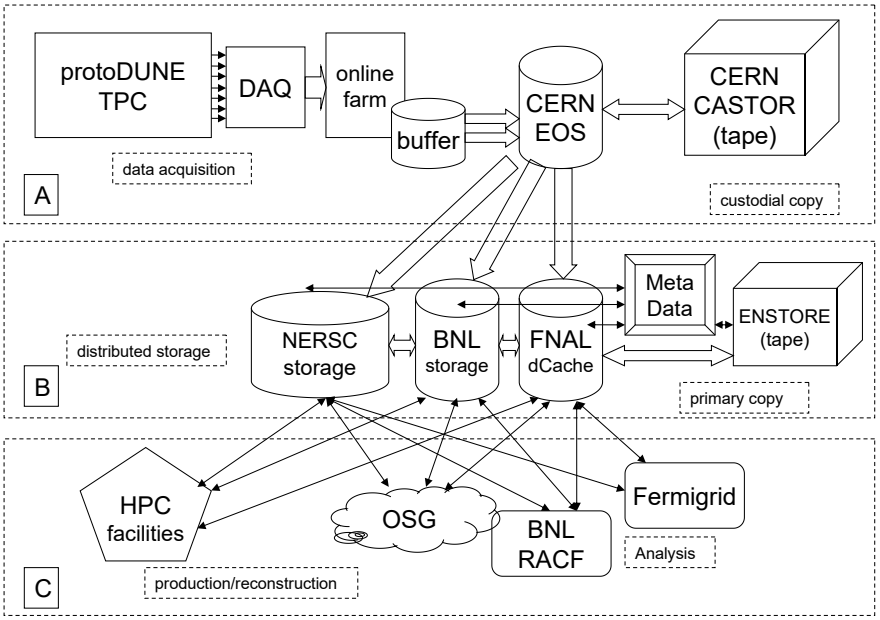

Figure 1: Data flow in protoDUNE

2 Components of the Data Handling System

2.1 Overview

Block "A" in the diagram presented in Fig.1 contains the disk buffer, which sits between the online system and EOS (distributed high-performance disk system). EOS can accessed utilizing a few different APIs, including XRootD. The latter is the main method by which the LHC experiments are transferring their data to EOS. As one example, ATLAS has the so-called "SFO" (subfarm output) system which acts as a data-aggregation buffer between DAQ and EOS [1]. The SFO nodes each open separate XRootD connections to EOS to ensure adequate bandwidth and avoid bottleneck such as throughput of an individual network card. CMS also relies on a storage cluster to absorb data coming out of multiple "filter units" and "builder units', before it get transmitted to CERN central services.' [2].

In protoDUNE, since the data in the buffer needs to be cleared once it reaches its destination (i.e. EOS in our case), this calls for a "consensus based" buffer flush, i.e. having a system in place which manages the process of data deletion based on guaranteed success of copy to EOS, and in addition to that, on completion of monitoring jobs and other processes that might be using the data in the buffer as input, for speed of access.

CASTOR serves as the keeper of the "custodial copy" of the data (i.e. it is not actively used in production). To utilize CASTOR in that role, protoDUNE will have to deploy a system for bidirectional data movement to and from CASTOR (e.g. in case where a piece of corrupted or lost raw data needs to be replaced by accessing the original copy).

Transmission of data from CERN to data centers (located primarily in the US) is a challenge since it requires 2gbps sustained rate over an extended period of time. This will likely require special arrangements to guarantee availabiliy of the requisite bandwidth. Automation, monitoring and redundant error checking and other measures need to be used to ensure robustness of the data transfer. This is graphically represented in Fig.1 by arrows connecting blocks "A" and "B".

The process of data transfer must be accompanied by generation and management of metadata, which will include both file catalog type of functionality with data characterization based on specific conditions of data taking. Beyond dealing with raw data, the metadata system will have to feed (and be fed by) the production system in which the data will be processed.

To ensure efficient sharing of data in the production stage and on to analysis, it is desirable to emply storage federation such as XRootD.

At the time of writing, issues of calibrations are in the stage of initial development and can't be included here in sufficient detail. For purposes of this document, these are considered a part of "express streams" to be run at CERN.

2.2 Itemized List of Aggregated Work Items

This is a "bird's eye" view of major items to be completed:

- DAQ output buffer farm and its interface to EOS
- Buffer management system, in particular "consensus based" buffer flush which involves interaction among a few systems such as copy to EOS, monitoring etc
- Bidirectional interface of EOS and CASTOR (to ensure the data can be restored to the disk buffer from tape is necessary)
- Metadata generation and management

- Data link between CERN and remote data centers (FNAL, BNL etc), which includes monitoring, error checking, retransmission and other aspects of automation
- Data distribution to participating data centers (e.g. in the US), and the bulk of the production done on the Grid.
- protoDUNE production system which governs transformation of data from its raw state to "Analysis Object" type of data.

General approach to implementation of these items includes early protoryping and initial testing at smaller scale with emulators and/or readily available hardware, while expanding to a full scale system with its dedicated hardware later. This will allow early idenitification of potential problems and more time for correction and reevaluation of components as needed.

3 Software reuse

Reusing the software, design and systems already proven in the LHC environment is an obvious goal. It is evident from materials such as [1], [2] etc that the online storage systems suitable for operation at rates required at the LHC (and essentially same scale as protoDUNE) can be potentially complex from the standpoing of interaction with DAQ and offline.

A collection of useful links on this subject can be found at [3]. In the past few months (mid-2015) there have been communications between DUNE and ATLAS and CMS. This was useful in the following aspects:

- Change of architecture between LHC Run 1 and Run 2 confirmed, i.e. CASTOR is no longer the endpoint for online systems. Data is first placed in EOS from where it's used in requisite production, and is sunk into tape storage (CASTOR).
- Multiple nodes utilizing XRootD connection to EOS appears to be the preferred configuration.

At the same time, based on this survey, it was found that actual scripts and tools used for moving data to mass storage at CERN in these expeiments are coupled to elements of production running at Tier-0 (i.e. CERN) which leads to considerable degree of dependency, complexity and makes it impossible to adopt a package as a ready end-to-end solution. It appears at this point that while reusing concepts and technologies from the LHC experiments, protoDUNE will still have to develop an equivalent array of utilities on its own.

In a similar fashion, software utilized to distribute data to remote data centers and subsequently manage the data is, for the most part, experiment-specific and dependent on systems supporting workflows in each respective project.

In summary, the current outlook for software reuse is as follows:

- leveraging the XRootD platform to develop data transport mechanism between protoDUNE online and EOS
- borrow elements of scripts for moving data from EOS to tape (CASTOR)
- leveraging IFDH (FNAL) or Spade (IceCube, Daya Bay, LBNL) to perform robust data movement from CERN to outside data centers

4 WBS items

In this section we present a very preliminary breakdown of work items with estimated allocation of effort, for the time period *before* the start of data taking (projected to begin in early 2018, with bulk of commissioning taking place in 2017). Dates in the "schedule" columns are very approximate and will be revised as soon as work commences on CERN systems.

Item	FTE	Duration (months)	Schedule
DAQ emulator to enable buffer integration	2	2	2016 Q1
Prototyping of the buffer farm ("subfarm")	2	3	2016 Q2
EOS interface and "consensus buffer flush"	2	2	2016 Q2
EOS-CASTOR interface (bidirectional)	1	3	2016 Q2
Provisioning DUNE software at CERN	2	2	2016 Q3
Prototyping express-streams to run at CERN, out of EOS	2	2	2016 Q3
Mock data challenge (simulated express-streams at CERN)	1	2	2016 Q3
Tech. evaluation and downselect for CERN-US data link	1	2	2016 Q4
Metadata generation and SAM interface	1	3	2016 Q3
Dress rehearsal of complete chain of data transmission (including metadata) using emulated components	3	1	2016 Q4
Procurement, configuration and testing of bandwidth out of CERN	1	2	2016 Q4
Scalability test of data transport CERN-US	1	1	2017 Q1
Data flow monitoring system	2	2	2017 Q1
Installation and configuration of the HW buffer (subfarm)	2	3	2017 Q3
protoDUNE production system (prototype)	3	4	2017 Q1
Integration with actual DAQ	2	3	2017 Q2
Data challenge with actual components in place (DAQ, subfarm, CERN streams, data link, production at FNAL and other sites)	3	4	2017 Q3
XRootD data storage federation	1	3	2017 Q4
Commissioning	3	3	2017 Q4

Table 1: WBS items for protoDUNE computing infrastructure

Not included in the above table are certain online computing items such as slow controls, conditions data base etc.

In summary, table 1 indicates effort profile commensurate with a total of ~9FTE*years, which will need to be structured roughly as 4.5FTEs over the period of two years before the start of data taking.

References

[1] "Operating the ATLAS data-flow system with the first LHC collisions", Journal of Physics: Conference Series 331 (2011) 022007

[2] "The New CMS DAQ System for Run-2 of the LHC", IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 62, NO. 3, JUNE 2015

[3] `https://dune.bnl.gov/wiki/CERN_Prototype_Data_Handling`