Design of the Data Management System for the protoDUNE Experiment

M. Potekhin, B. Viren, S. Fuess,
O. Gutsche, R. Illingworth, M. Mengel, A. Norman

Overview

The protoDUNE Program and Detectors

The *protoDUNE* program will help validate various DUNE technology aspects before proceeding with the construction of the principal DUNE detectors at SURF. It is designed for measurements with a test beam provided by a dedicated target and beamline system at the CERN SPS accelerator complex. It also has the potential to be an important platform for realistic LArTPC detector characterization (e.g. PID, shower response, etc.) utilizing controlled conditions of a test-beam experimental setup. The name “protoDUNE” currently applies to two full-scale LArTPC prototypes based on two different technologies. The “full-scale” designation is used to describe the fact that the prototypes contain important (and large) structural and readout elements built according to the specifications (including the size) of the eventual full detector.

The “single-phase” (SP) LArTPC functions without amplification in the medium (liquid Argon) and is in essence a very large ionization chamber equipped with a large number of readout electrodes (wires), each with its own electronics chain. In this design, the front-end electronics is situated within the cryostat in order to minimize noise (the so-called “cold electronics design”). In the “dual-phase’’ (DP) TPC ionization electrons are extracted from the liquid into the gaseous phase of Argon, and drift in Argon gas towards a specially designed 2D structure on top of the detector where they multiply according to principles of proportional chamber operation. The two designs are complementary in the sense they explore different approaches to optimization of the Liquid Argon detector characteristics.

In December of 2015 the dual-phase prototype was given the official designation as a CERN experiment “NP02”, and the single-phase was designated as “NP04”. Both are to be deployed at CERN in 2017 and scheduled to take data in 2018. The prototypes will be placed in a specially constructed large-scale extension of the existing experimental hall located in the CERN North Area. Each prototype will be provided a dedicated optical fiber network connection to the CERN central storage facilities located in the West Area campus of CERN. The nominal bandwidth of these dedicated network connections will be 20 Gbps for each experiment. The motivations for this specific choice of nominal bandwidth will be presented in the following sections.

Outlook for protoDUNE Data Characteristics

In order to provide the necessary precision for reconstruction of the ionization patterns in the LArTPC, both single-phase and dual-phase designs share the same fundamental characteristics:

- High spatial granularity of readout (e.g. the electrode pattern), and the resulting high channel count
- High digitization frequency (which is essential to ensure a precise position measurement along the drift direction)

Another common factor in both designs is the relatively slow drift velocity of electrons in Liquid Argon, which is of the order of millimeters per microsecond, depending on the drift volume voltage and other parameters. This leads to a substantial readout window (of the order of milliseconds) required to collect all of the ionization in the Liquid Argon volume due the event of interest. Even though the readout times are substantially different in the two designs, the net effect is similar. The high digitization frequency in every channel (as explained above) leads to a considerable amount of data per event. Each event is comparable in size to a high-resolution digital photograph.

As will be shown in the following sections, it is foreseen that the total amount of data to be produced by the protoDUNE detectors will be of the order of a couple of petabytes (including commissioning runs with cosmic rays). Instantaneous and average data rates in the data transmission chain are expected to be substantial. For these reasons, capturing data streams generated by the protoDUNE DAQ systems, buffering of the data, performing fast QA analysis, and transporting the data to sites external to CERN for processing (e.g. FNAL, BNL, etc.) requires significant resources and adequate planning.

Prioritization

All of the many elements in the chain of data acquisition, storage, distribution and processing are critically important to derive physics results from the data. At the same time, certain components of the data chain need to be prioritized over others in order to perform the measurements during a potentially limited time period.

The priority components are the DAQ and the Raw Data Management System, which includes capturing the data coming out of the DAQ, transporting the data to persistent mass storage and prompt Quality Assurance which is required to ensure corrective action can be taken if the detector or certain system problems are identified in the QA process. The latter can be thought of as sophisticated monitoring done in near-time, which implies a ``few minutes'' scale of processing.

A note on the DAQ Interface to the Data Handling System

The plan is to have adequate buffering capability in the DAQ for both NP02 and NP04. In this case, “adequate” indicates in part conforming to a CERN requirement that the experiment must be able to keep taking data for a least 3 days at nominal rate, even if there is an occasional problem with the data link between the experiment site and CERN storage facilities, an issue with central storage or any type of similar outage. At the time of writing, there are differences in two respective approaches:

- Buffer depth in NP02 is larger, in order to make possible some processing right in the data room of the experiment. A number of middleware options are being explored for this storage solution, in particular the BeeGFS file system.
- In NP04 the emphasis is made on a more lightweight and fault tolerant setup which satisfies the general throughput requirement. No extensive processing is foreseen on the experiment site. Among the technical options for the buffer farm file access is xrootd.

It is this “outer layer” of the data acquisition system in either experiment that will need to be interfaced with protoDUNE’s raw data management complex.

Requirements

- Transfer raw data files from both online disk buffer farms of the DP and SP prototype detectors (NP02 and NP04 respectively) to CERN EOS disk and from there to CERN tape (CASTOR), FNAL tape (Enstore) and other end-points.
- Ensure that the throughput is adequate and there are no bottlenecks for the Data Acquisition System given the expected data rates over the nominal two month running (see table)
- Record metadata about file status and outcome of file operations
- Operate at CERN and FNAL with support for initial setup and ongoing operations
- Provide monitoring of overall system health, alerts on error and support debugging of problems.
- Provide triggers to perform file operations (copy, delete) based on configurable rules
- Support “express lane” process at CERN and other institutions.

The following table gives performance measures driven by the extreme of each DP and SP detector and which the file handling system must accommodate on the assumption that they apply to both detectors.

Total raw data	2.5 PB
Total number of raw files	2 M
Sustained data rate	20Gbps
Max latency to reach EOS	10 minutes

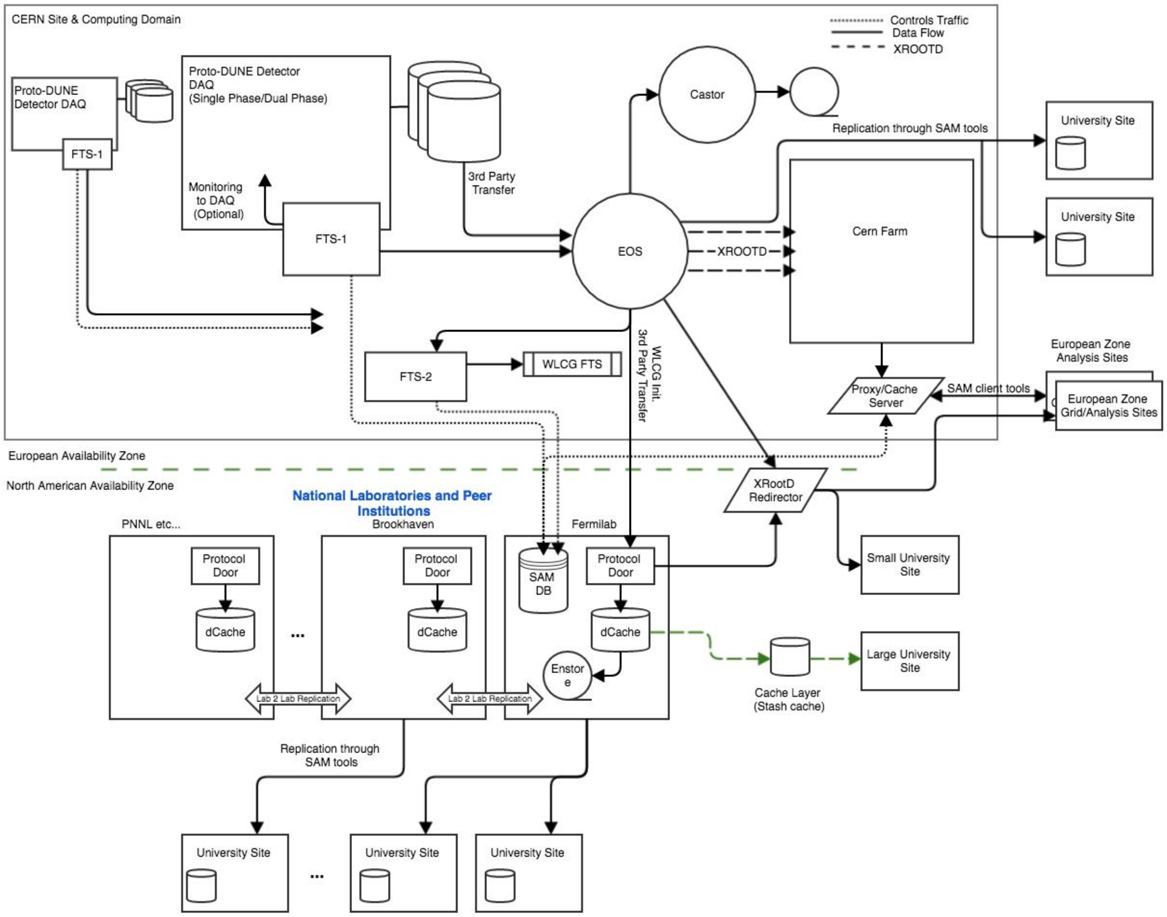
Max latency to reach express lane processing.	10 minutes

The file handling system assumes the following limits will not be exceeded

Max simultaneous files in FTS dropbox	1 M
Max simultaneous active files in transfer	50,000
Max file registrations	200,000 / day

System Topology

The baseline design for the online data handling system is shown schematically in Fig. 1. This design leverages the technology of the Fermilab File Transfer Service (F-FTS) running in two positions within the CERN computing environment. The primary F-FTS systems are homed within or straddling the border of each of the detector's DAQ domains (one for single phase and one for dual phase readout detectors) and is configured to run on a system which is able to access the DAQ's buffer disks either through a POSIX filesystem or a protocol layer (e.g. XROOTD, gridFTP or others). The primary FTS system is configured on this system to look at one or more input directories or storage locations, commonly referred to as a "dropboxes". The FTS daemon performs periodic scans of the configured dropboxes. The system will operate asynchronously and independently of the protoDUNE DAQ systems. When the protoDUNE DAQ has produced a file that it wishes to have passed off to the storage system, it will "move" (perform a filesystem/storage system atomic move operation or the equivalent) the file to the dropbox location. When a new file is located within one of the dropboxes, the system will initiate the file registration and transfer operations.

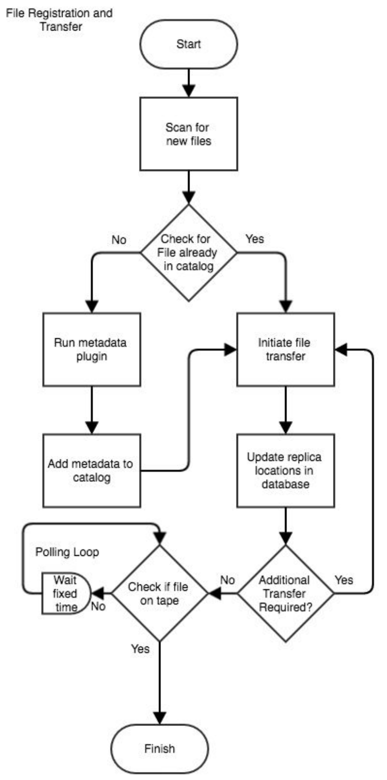


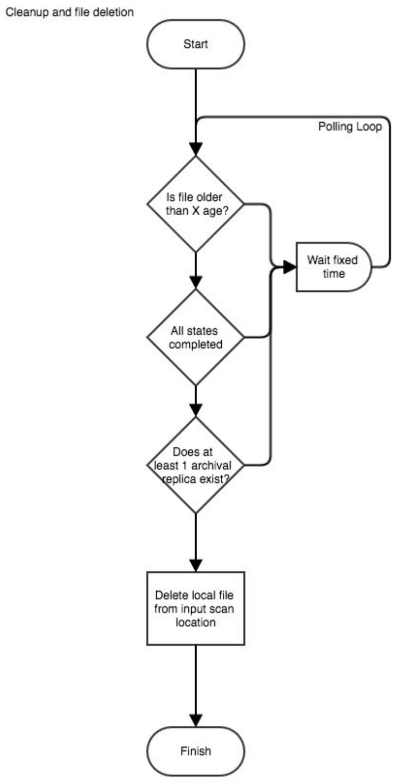
In the protoDUNE model, the primary FTS will initiate a copy operation from the online disk buffer farms into the EOS storage system. The system will use the 3rd party copy support that is provided by the XROOTD protocol to allow for optimized transfers into EOS. Upon completion of the initial copy into EOS, the FTS will initiate a “chained” copy (a copy that is dependent on the initial copy into EOS) of the data from EOS to the Castor archive system. The FTS system will register each file’s metadata records (containing both basic and physics metadata, defined below) in the SAM data handling system. Upon completion of the replication of the data to Castor, the FTS will enter a monitoring/polling state of its operations to determine when the file has been successfully written to the archival media.

The primary FTS will handle the “cleanup” of its input dropbox. The FTS performs configurable periodic cleanup passes through its current file sets. The baseline cleanup logic is shown in Fig. X. This cleanup process ensures that all files are successfully transferred and archived before they are deleted and provides an “age” criterion on files so that files can be retained within the DAQ environment for a period of time. This allows operations like online/nearline processing or log files to be examined within the DAQ environment after the actual transfers to archival storage have completed.

The secondary FTS service runs outside of the detector DAQ domains, but within the CERN computing sphere for maximum efficiency. This FTS system is configured to monitor, as its input dropboxes, the EOS locations that the primary DAQ-FTS systems are using as output endpoints. The system operates in the same manner as the primary FTS, scanning for new files and then initiating copy requests of the data to a set of one or more destination endpoints. This FTS system will initiate the copy request between the European availability zone, and the North American availability zone. At least one of the endpoint destinations for this service will be the FNAL-based dCache/Enstore system, where a second archival copy of the data will be recorded. Additional copies across the availability zones can be configured based on available resources and available transatlantic bandwidth (e.g. direct transfer from CERN to BNL's dCache or to PNNL's storage facilities). To perform the transfers, the F-FTS will interface with the WLCG FTS (as a supported transfer protocol) to schedule the actual transfer between the storage elements.

Underlying the F-FTS systems will be a fully featured data management layer (SAM) which will provide the metadata and replica catalogs. All data being handled by the transfer systems will have corresponding records in the data handling catalog so that the content, locations and provenance of the data can be fully tracked. The primary catalog systems will reside at FNAL with proxy/cache layers in the European and North American zones to ensure high speed connections between the servers and the [offline analysis] clients that will query the catalogs from these zones. Similar scalable proxy and cache layers can be instantiated for other identified zones which may require high speed or optimized access to the catalog systems.




Cleanup and file deletion
Start
Polling Loop
Is file older than X age?
Wait fixed time
All states completed
Does at least 1 archival replica exist?
Delete local file from input scan location
Finish

Replication between collaborating peer institutions in the same geographic zone (i.e. national labs and large universities with significant computing/storage resources in the North American zone) will be provided through standard “dataset” replication tools that are a part of the SAM data handling suite. Data access at smaller university sites or opportunistic computing resources (e.g. Open Science Grid (OSG) affiliated sites) is provided and optimized through support for streaming protocols such as XROOTD with redirector services and through cache layers such as the “stash cache” system employed by OSG.

Technical Requirements and Specifications

The data management and transfer systems requirements are enumerated in table XX according to the category to which they belong.

Data Throughput

The goal of the data handling system is to be performant at a level that allows for full exploitation of the underlying hardware and networking infrastructure upon which it is running. In the case of the proto-Dune experiment the FTS system will be capable of operating at a sustained data processing and transfer rate that is greater than 80% of the maximum theoretical bandwidth available between the primary DAQ storage systems and the EOS file storage system. In the current design this network bandwidth consists of two 20 Gb/s ethernet links. The RAID arrays or distributed file system architectures to which the data will be written to and read from are assumed to have similar or lower available bandwidth. The actual observed read/write bandwidth will be determined by the actual choice of hardware that is deployed to the detectors. The primary DAQ FTS system will be required to operate at a sustained effective bandwidth of the lesser of: two times 16 Gb/s (80% of theoretic network bandwidth) or 80% of the measured read access bandwidth from the storage arrays [during simultaneous write operations, if the DAQ computing models requires this mixed access mode] under the standard proto-DUNE operations.

Transaction Throughput

The goal of the data handling system is to be performant at a level that allows for full exploitation of the underlying data catalog and database infrastructure while at the same time meeting or exceeding the rate at which files or other data objects are generated by the proto-DUNE detectors. In the case of the proto-DUNE experiment that FTS systems that run on the border of the DAQ domain will be able to support a minimum transaction rate of 3600 file registration per hour (1 Hz) and an aggregated rate of 100,000 file registration operations per day. This level of performance has been demonstrated in other online and offline systems where the data handling system has been deployed.

Transaction Tracking

The goal of the online components of the data handling system are to be able to perform simultaneous end-to-end tracking of all files that are undergoing active transport through the system prior to being written/verified in the archival storage systems, without impacting or interrupting data taking. To achieve this goal the online data handling systems for proto-DUNE will be able to support a total number of “in flight” files (per FTS instance) equal to the 100,000 times the maximum supported duration (in days) of an outage that can be sustained by the DAQ system under outages in the network or storage systems downstream of the DAQ domain. In the case of proto-DUNE the FTS systems will support at least 700,000 “in flight” files at any given time, where at least 10,000 of these files are undergoing “active” registration/transport through the system at any given time (meaning at least 10k of the files are being actively managed while the remainder are “waiting” in the designated dropbox location.

Transfer Latencies

The file transfer and management systems are designed to operate in an asynchronous manner to other components of the proto-DUNE DAQ. Due to the polling models employed in this asynchronous model, many operations do not have fixed temporal relations to other events in the DAQ system, but rather will occur/complete within a well defined time window or with a certain latency. In the case of the proto-DUNE experiment, the data management systems will be capable of operating within the following parameters

In the first stage of the file transfers, the latency between the completion of the DAQ writing a file and the hand off/detection of the file by the FTS system will be controlled by a configurable delay parameter (in minutes) within the FTS which controls the intervals at which the FTS looks for “new” files in its dropbox locations. Under normal (steady state) operation, the average latency will be 1-2 polling intervals. The actual latency between when the DAQ finishes writing a given file and when the FTS picks up and begins to actively manage that file can be delayed by the number of other new or pending file currently in the system. In this case the order of handling of files will be handled via an internal queueing algorithm that efficiently handles the files but does provide any user specified prioritization.

In the second stage of file transfers, the latency between when a file is generated by the DAQ and when it is transferred to the EOS system and then written/verified to archival media (via Castor) is dependent on first stage latencies (new file detection and file registration) and then the details of the archival storage system and its characteristics. In particular the DAQ to EOS to Castor path will constitute a set of chained dependencies with an independent polling interval for each stage. Under normal operating conditions that latency for the file to be transferred to EOS and be available through the data handling system is expected to be 1-2 of these polling cycles, contingent on write congestion in the EOS storage system.   For full registration of the files in Castor and onto type, the FTS/SAM systems will support latencies in this recording of 3 hours to 30 days and will support a configurable timeout parameter to indicate failures in this transition.

The latency between when files are written by the DAQ and when the files are available on storage in the North American zone, will be determined by the second stage latency of the files appearing on the EOS system from the DAQ and then a secondary latency will be incurred based on the polling for files that require transatlantic endpoints. Similar to the other stages this is configurable through a polling interval. Once queued for handling the actual file transport will be off loaded to the WLCG FTS system which will schedule the files for transmission and will properly balance and throttle the site to site traffic and will properly conform to the CERN computing environment. Latencies at this stage are well understood based on the experience that the LHC experiments have with WLCG FTS.

Storage System Interface Specifications

The module design of the data handling system provide considerable flexibility in its ability to adapt to different requirements for both data input and output storage systems. For the proto-Dune experiment these major interface points are considered to be at the interface between the experiment's core DAQ components and at the interfaces between the mass storage and archival storage systems at both CERN and FNAL. The interface choices below detail these interfaces.

DAQ to Data Handling System Interface

The interface between the core DAQ system and the data handling system is made at the DAQ's disk buffer farm. When the final stage event builder have completed the assembly of a file, the file will be handed off to the data handling system by the DAQ by placing the file into a designated "dropbox" area on the target file system. For POSIX compliant file systems, this can be accomplished through an atomic move operation (i.e. Unix "mv" equivalent) which minimizes the IO overhead associated with the hand off. No signals need be emitted from the DAQ, nor will any other form of message be required in the direction of the DAQ to the data handling system. The F-FTS will detect new files through files becoming visible in configured dropbox locations of the storage system (i.e. a new file appearing in a directory)

The F-FTS systems running as part of the data handling system will use either a POSIX (or near POSIX) compliant file system view of the disk buffer farm, or an API that provides access to standard meta information regarding the storage system's dropbox area (directory) and contents (i.e. filenames, file size, create/modify times etc…) Additionally the FTS will require a file read API for the purpose of metadata extraction from the file and for checksum computations (if not provided through some other means). The system will require a delete API and privilege, if the automated cleanup options are desired/enabled. If the DAQ storage elements support "third party" transfer protocols, the FTS can use these to reduced overhead in performing the actual file transfer operations.

The organization of the FTS dropbox area(s) will be configured to support the volume of data that would be generated during a sustained outage (of non-DAQ systems) of no less than 3 days plus the time required to register/process the volume of data that would be generated during the outage (i.e. “the recovery time” required to clear a backlog).

Data Handling to EOS Interface

The interface between the data handling system and the EOS storage system will be made through the APIs provided by the standard protocols already supported by both the EOS system and the SAM data handling system. In particular the xrootd protocol will be the primary API used for the interaction between the system, with secondary support for the gridftp, webdav and srm protocol APIs. The EOS system will be declared, and configured in the SAM system as a standard storage element and will have files stored on it registered and mapped within SAM the replica catalog to the well defined access URIs.

Data Handling to CASTOR Interface

The interface between the data handling system and the CASTOR archival storage system for file ingest will be made through the APIs provided by the standard protocols already supported by both the CASTOR system and the SAM data handling system. In particular these system both already support the xrootd protocol, gridftp and srm. The CASTOR API to query the tape archive interfaces (to determine if a file has been successfully archived to tape and which tape it is located on) will be integrated into the SAM data handling system in a manner similar to other tape systems that SAM is already aware of (i.e. Enstore)   The CASTOR system will be declared, and configured in the SAM system as a standard tape storage system and will have files stored on it registered and mapped within SAM the replica catalog to the well defined access URIs along with additional tape location information to permit optimized retrieval. Custom modules/utilities will be developed as needed as part of the SAM data handling suite to provide additional support for the CASTOR system in performing certain operations (i.e. bulk queries or monitoring operations). These modules/utilities will be distributed as part of the core SAM distribution.

Data Handling to dCache/Enstore Interface

The interface between the SAM data handling system and the dCache/Enstore archival storage system is already fully defined and supported. The interface fully supports standard access protocols including xrootd, gridftp, webdav and srm as well as the proprietary dcap protocol and a specialized NFS implementation that provides limited basic read access for hosts with local mounts of the dCache system. The SAM system has modules that optimize data access based on Enstore tape location information. This interface is in wide scale production use across many experiments including NOvA, MicroBooNE, Dune 35t, Minos and others.

Data Handling to SAN/NAS Interface

Generalized SAN and NAS systems when acting as data sources (input) will be integrated to the data handling system through their POSIX style interface if available. When enabled as data sinks (output) for the data handling system, a front end data server(s) running standard access protocols in the form of xrootd or gridftp will be utilized. These will then be mapped into SAM as standard storage locations.

Generalized Protocol Support Interfaces

The data handling system provides a file delivery and transfer layer which acts to provide protocol abstraction to the end users (i.e. provides a consistent command interface) and provides a protocol bridge when performing transfer operations between dissimilar storage systems (i.e. cross protocol transfers). The “Intensity Frontier Data Handling” tool (IFHD) for proto-DONE will support protocol integration for:

- XROOTD -- Full support will be provided for xrootd in the Fermi-SAM data catalog, F-FTS and IFDH layers. Currently these tools support read/write access methods. Full support for additional xrootd features (listings, permissions modification, etc...) is currently under development.
- WLCG FTS -- SAM, FTS and IFDH will support WLCG-FTS as a 3rd party or proxy transfer mechanism. In this mode outbound file transfers originated from the CERN domain can offloaded to WLCG-FTS so that the data traffic across the CERN site can be properly scheduled and balanced. Support for this mode of operation will be integrated into IDFH or directly into the sam_cp layer of the F-FTS.
- GridFTP -- This protocol is fully supported for most of the components involved, but is not the preferred or documented interface for any of the CERN storage components.
- SRM -- This protocol is fully supported for most of the component systems involved, but is not the preferred or documented interface specification for the CERN storage components.
- HTTP -- Fermi-SAM uses HTTP for internal communication between Fermi-FTS and the Fermi-SAM instances, and for client programs wanting file location and metadata information.
- CVMFS -- The data handling suite includes support for distributing its client tools and homing other parts of the suite on a CVMFS read-only filesystem. This support allows the experiments to provide widespread, efficient code distribution over a HTTP based protocol while providing a file-system layer to the end user applications. This is a standard distribution method used to to distribute the DUNE and LARSOFT software, as well as Fermi-SAM client utilities.

Data Replication

The data handling systems provide for multiple replicas of files to exist across storage systems and locations. It will be necessary to replicate data files or data sets between sites or storage locations. The data handling systems for proto-DUNE will handle this replication in different ways depending on geographic proximity and bandwidth. In particular in the proto-DUNE architecture at least two “Primary Zones” exists in the form of the European zone and the North American zone. The European zone encompasses CERN and other institutions near to CERN geographically and in network bandwidth proximity, while the North American zone would encompass FNAL, BNL, the other institutions in close network proximity to the major laboratories and universities in the DUNE collaboration.

Replication within the primary zones will be handled in the following manners by the data handling systems:

- European Zone -- Replication and data transfers within the CERN domain will be handled by F-FTS instances configured with site specific endpoints rules (i.e. the EOS and CASTOR rule sets) which will provide for the full data sets to available in both EOS and CASTOR. Transfer and replication operations outside the CERN domain to institutions with registered storage elements will be performed through the SAM suite’s replication tools (sam_clone_dataset). Transfer or replications to smaller institutions or analysis sites without registered storage elements can be performed through the SAM/ifdh client tools (ifdh_fetch). The SAM client tools are which are available through CVMFS.
- North American Zone -- Initial replication of the data to FNAL, BNL and other collaborationg labs or large university sites will be automated through FTS instances at CERN and in North America which will be used to optimize the data transfer paths to each of the institutions, subject to bandwidth constraints of the host institutions. Transfers elsewhere in the North American side would be performed using the same SAM tool suites and strategy for handling registered and unregistered storage, as specified above for the European zone.
- Tertiary Institutions -- Other institutions wishing to host a set of datasets or partial data sets, will be able to register their storage elements with the SAM data catalog and will then be able to use the replication tools to clone the relevant data to their site.

File Registration

Registering files or other forms of data with the SAM catalog require a set of “meta information” about the files. This information is used to allow detailed searches to be performed to select specific data for analysis. This metadata is broken into two general types “Base Metadata” and “Physics Metadata” which can be provided at the time of file registration, or modified later. These metadata are composed of:

- Base Metadata -- For each file the the Fermi-SAM database requires a unique filename, the file size, a file type string. It also allows one or more checksums - each consisting of an algorithm type name and a value - and a description of the application used to create the file. The file can also have zero or more parents (which much be files already existing in the catalog) which creates a tree of relationships between files.

- Physics Metadata -- There are a number of predefined physics metadata fields such as the detector configuration, the run number, the data tier, the event count, the start time and end time, and the data stream. It is also possible to define arbitrary key names for other values with either integer, float (64 bit), or string types.

The experiment must provide a method for determining or generating the metadata (i.e. an external program that can be run on a file, or a python module that extracts information from a filename, etc…) and that method will be invoked by the data handling system when it encounters appropriate files. The output of the method used must conform to a suitable format supported by the F-FTS and SAM systems. The systems support JSON formated files and python dictionaries for direct upload to the data catalogs.

Data and Replica Catalogs

The data management system for proto-DUNE will leverage the DUNE experiment's SAM data catalog. The underlying resources of this catalog and it's software are capable of, and sized to support, a file inventory in excess of 150 million files (demonstrated by the DØ experiment's catalog which uses the same technology). The data catalog provides a command line toolset for both client and administrative functions, as well as a web based API for interacting with the catalog. The web tools provide additional guidance for assisting users in finding and classifying data through the indexes.

The data management system will also leverage the same DUNE instance of the SAM catalog to maintain its replica catalog. This instance will provide a unique namespace for the DUNE/proto-DUNE experiments and will contain the appropriate storage identifiers and paths to enumerate the full locations of DUNE/proto-DUNE data on all supported sites. The SAM replica catalog itself is protocol neutral, but a mapping layer within the data handling system performs the translation of locations into actual access URLs of the default, preferred or requested protocol schema for a given storage system (i.e. the system maps the location onto the method for how you retrieve the file).

Data Transfer Technology

The data management system for proto-DUNE will support, through its use the SAM tool suite, data transfers to and from CERN, FNAL and other site's storage elements using a combination of protocols supported by the specific site's storage resources. In particular the system will natively support the xrootd, gsiftp (gridftp), webdav, srm protocols for access to EOS, Castor and dCache/Enstore (as appropriate). The system can support multi-channel transfers and 3rd party transfers using these protocols to achieve high bandwidth or low overhead transfers where needed. The system also supports operations specific to local access methods on POSIX file systems.

The file transport layer uses a modular design with an abstraction layer so that new protocols or access methods can be added to the system without changes to the other interface layers. The system can also select or prioritize the replica source location and access protocol based on destination characterics (i.e. it can prefer a local cache copy of a file to a replica that is at a transatlantic location) The system will also support offloading of transfers to 3rd party transport systems (e.g. WLCG FTS) to properly balance the resources of different WAN connections and site resources.

Outline of the Offline Computing Model (Maxim)

The Offline Computing in protoDUNE includes two principal domains:

- “Express streams” to be run on the CERN cluster, for fast calibrations and QA purposes
- Production and analysis on other sites, in Europe, the United States and elsewhere

Development and management of the code to be used in these two cases will be done largely by the same group of people responsible for offline software in DUNE.

At the time of writing, only general parameters of the protoDUNE Offline Computing Model can be determined. For example, it is fairly certain that the data management system for offline computing will be based on the same infrastructure as the raw data management system as described throughout this document.

As to Workload Management (meaning assigning particular tasks or groups of tasks to Grid sites or any other suitable compute elements, and coordinating that with the data management), there are a few options currently under consideration, including the ``jobsub'' toolkit provided by FNAL, and a few of more comprehensive Workload Management Systems that need to be evaluated for the needs of protoDUNE.

Derived Data Handling

Data will be processed shortly after it is collected in order to support the physics goals of protoDUNE. The bulk of the single-phase processing is expected to take place at Fermilab and other OSG computing sites, with the processed data stored at Fermilab, indexed with SAM. Dual-phase processing will take place at CCIN2P3, at CERN, and the online cluster in EHN1. The processing is to be done with LArSoft for the single-phase prototype, and a combination of Qscan and LArSoft for the dual-phase prototype.

Three tiers of processed data are envisioned for the single-phase detector: a set of files with raw data formatted in a convenient offline form, a set of files with all of the raw and reconstructed data products in each, and a much more compact analysis tree or ntuple format.

These are described in more detail below, followed by a discussion of the optimization of the workflow, both to reduce the CPU needs and the number of copies of the raw data.

The first tier of reformatted data results from converting the raw data from the DAQ format into a version more easily processed and analyzed.   Experience with the 35-ton prototype has shown that the definition of an event as output by the DAQ may not match the expected triggered data, and that several DAQ events may need to be combined together to make one offline event, or one DAQ event may need to be split into multiple offline events. The ADC values also may arrive in a different order than expected, or be packed in a way that takes CPU and memory to unpack. It has been found useful for some analyses to preprocess the data in this way to speed up subsequent analyses.

The second tier consists of processed data files with a full set of output data products. At least for some fraction of the data, it is useful to have a copy of the raw data (ADC counts for each sample time) as well as reconstructed objects (hits, clusters, tracks, showers, lepton candidates, and other objects). Often multiple algorithms which produce the same kinds of data products are run together so that their performance can be compared. The presence of raw as well as reconstructed data in the same file allows the evaluation of the performance of the reconstruction algorithms, especially in the Monte Carlo, but also for the data. We expect to produce two passes of full reconstruction per year while the papers are in preparation.

The third tier consists of analysis ntuples or Root trees, which contain summary information about the reconstructed data products. It is not envisaged to pack raw data into analysis trees, at least not for large production jobs. These samples are expected to be small and easily distributed to collaborating institutions. They will be re-created each time the reconstruction passes are performed.

It is not necessary to duplicate the entire raw dataset in the first processing tier of reformatting the data, as reformatting can be a part of the reconstruction job. Only a subset that is required for users to scan events and develop software is needed. This data tier may be populated only for particular streams, for these purposes. It is also not necessary that all reconstruction sets contain all of the raw data. Two copies of the raw data exist in this model already for the single-phase detector -- one at CERN and one at Fermilab, and a second copy in the reconstructed sample is not necessary. It is necessary for calculating the performance of the algorithms, but the fraction of the data that needs to have a full copy of the raw data along with the reconstructed data in the reconstructed files is an optimizable number. All of the data should be processed through the ntuple stage.

Support model

Physical Hardware / System Support

The physical hardware that run the services need to be installed with an OS (ideally Scientific Linux compatible) and administered in their respective administrative domains (DAQ would administer the FTS instance in the DAQ domain, FTS2 would be easier to run on CERN-IT hardware but also would be possible to run on Fermilab or US hardware). Fermilab experts would assist protoDUNE experts in setting up the services and providing training. Maintenance is done by protoDUNE experts supported by Fermilab experts

Operations

The proposed data handling systems from the Fermilab services catalog are in use by a variety of neutrino and muon experiments in various stages of their experimental life cycle, from design to commissioning to operation. All systems are supported on a business hour basis. Communication is primarily done through tickets, Fermilab's ServiceNow system is used for operational issues while Fermilab's Redmine system is used for tracking development issues. The 8-by-5 support is agreed on with the current experiments and considered sufficient. Experiments have sufficiently decoupled data handling setups that allow to continue to take data while parts of the system are down due to various reasons. The buffer space at the detector locations are designed to be large enough so that in the case of an outage the data can be buffered and the system can catch up later. It is advisable to define the requested buffer in time of data taking, and doubling the calculated amount of needed buffer space to allow the system to catch up in a reasonable time while new data is being taken. To have sufficient support coverage, it is advisable to design the buffer space to cover a possible outage of 3 full days (i.e. a long weekend).

Experts from the collaboration will be trained by Fermilab experts in the usage of the systems and the basic operations including first line of defense debugging (start, stop, configuration change).

Documentation is available for experts and general users to use the systems and get access to data at various locations (big sites, small sites, WAN access).