DUNE Computing Plan
In Support of the ProtoDUNE Program
Executive Summary

A. Norman* and T. Junk
Fermi National Accelerator Lab
P.O. 500 Batavia, IL 60510
(Dated: May 9, 2017)

This document details the computing plans and models that are to be executed for the ProtoDUNE test beam projects. This summary also details the assumptions that the model has been developed under and the resulting impact that the plan is expected to have on resource allocations at FNAL, CERN and other collaboratoring institutions.

I. INTRODUCTION

The DUNE and ProtoDUNE projects will require significant computing resources to acquire, record, process and analyze their data. The flow, format and resources used to work with this data through the various stages of its lifecycle constitute the computing model used for the experiment.

This computing model is the sole responsibility of the DUNE computing organization and will be executed by the DUNE computing organization and its subgroups. Official data processing, simulation and other tasks that require significant resources will fall under this model or will be coordinated to interface with this model. Requests by the collaboration for significant computing resources are expected to be compatible with this model.

Requests for resources or services are handled through requests to the DUNE computing organization, and the leadership of the organization will work to evaluate and fulfill the requests in a manner that is commensurate with available effort, physical resources and other constraints.

This document outlines the baseline computing plan with a focus on the ProtoDUNE detector programs.

II. BASELINE DATA ACQUISITION AND PROCESSING

The baseline DUNE computing plan assumes a linear flow and processing of data as shown in Fig. 2 and a migration of data between the CERN and FNAL facilities as shown in Fig. 1. Data are initially acquired for each ProtoDUNE detector's data acquisition system and written to one or more files on a storage device residing within the DAQ environment. The raw-data format is defined by the DAQ teams, with review from the software and computing organization and the teams responsible for reconstruction. The data, still in its raw format, is then transmitted to the EOS storage facility operated by CERN-IT. From the EOS storage facility, a copy of the data is made to the CASTOR tape archive system. This copy constitutes the first of two archival replicas of the data and is retained indefinitely. A copy originating from the EOS storage facility is also made to the dCache storage facility at FNAL. This copy is then used to propagate the data to the Enstore tape archive system at FNAL. The copy of the raw data that resides on the Enstore system is considered to be the second of two archival copies of the raw data and is retained indefinitely.

The copies of raw data that reside on both the EOS and dCache storage systems are used as data sources for data processing, reconstruction or other activities that occur on the computing resources local to each of the labs (i.e. EOS is the data source for activities on the Tier 0 at CERN, and dCache is the data source for activities on the FNAL grid or Tier 1 facilities.) These replicas are managed through a combination of cache policies specific to the storage systems and through the DUNE data management system that uses a centralized data and replica catalog. In particular the EOS copies of raw ProtoDUNE data are intended to have short lifetimes, consistent with time-sensitive data processing that is targeted for execution on the Tier 0 facility.

The size of the storage requested at each lab is shown in Table I. Each of the resources has either been requested to match the FY17/FY18 need profiles for ProtoDUNE, or has been requested and allocated for use by the DUNE collaboration. The breakdown of the needs of the two ProtoDUNE experiments is shown in Table II.

III. DATA MOVEMENT

Detector data will be acquired by dedicated DAQ computers running Linux. These DAQ machines will be connected via private networks to the front-end electronics of each ProtoDUNE detector. The data will first be stored on disk buffers local to the EHN1 counting houses in a fully assembled and compressed form. The DAQ software, in addition to assembling the event data, processes will supply descriptive "metadata" relating to the raw data files [e.g. filenames, file sizes, number of events, run begin and end times, run and detector configuration strings]. The DUNE software and computing group will

* anorman@fnal.gov

Storage	Type	Data Type	Status	Total(TB)
CERN				
Castor	Tape	Raw	Requested (FY18)	6000
EOS	Disk (Durable)	Raw	Allocated	700
EOS	Disk (Durable)	General	Allocated	3000
FNAL				
Enstore	Tape	Raw	Requested (FY18)	6000
Enstore	Tape	General	Allocated	1000
dCache	Write Cache	Raw	Allocated	40
dCache	Read Cache	General	Allocated	2000
dCache	Temp Shared	General	Allocated	1000
dCache	Disk (Durable)	Raw	Allocated	200

TABLE I. Archival and durable storage by laboratory.

be responsible for providing a data management system that can handle these files and their associated metadata (the system may also handle DAQ logfiles or other types of filebase information that the DAQ systems need to catalog and retain).

The baseline data migration path, shown in FIG. 1, allows for the data to initially exit the DAQ/Detector environments of EHN1 and be transported to the CERN EOS storage system and the associated Castor tape archive system. From there the data is registered and replicated to the FNAL dCache and Enstore systems.

Data integrity is preserved through the data migrations via file level checksums which are computed by a combination of the mass storage systems that the data is housed in, and the data management system which catalogs the data (i.e. The file checksums will be computed during the transfer to EOS using the `xrootd` implementation of the `adler32` checksum algorithm, which is also used by SAM and dCache.) Details of the data integrity checks provided by the systems are documented in the SAM and Fermi File Transfer System (F-FTS) documentation[1]. The checksum scheme planned for ProtoDUNE provides integrity betweens all points in the data movement chain except the first copy from the DAQ disk buffer into EOS[2].

Two full, redundant copies of the ProtoDUNE raw data are planned to be kept for archival purposes. One copy will be kept on the Castor tape system at CERN. The other copy will be kept on the Enstore tape system at Fermilab. In the event of a media loss at one of the sites, the files will be recopied to that site from the other lab's archive.

Both the EOS and dCache systems are high performance, distributed disk systems that are designed to be accessed from computing facilities at the labs (e.g. the Tier 0 at CERN or the Tier 1 and grid facilities at FNAL). The data migration path takes advantage of theses systems and their caching capabilities. Both the CERN EOS and FNAL dCache systems will have an active (temporary) cache copy of the data after the initial data replication out of the DAQ systems. These cached copies of the data permit monitoring programs or analysis jobs running on compute resources local to a given lab to access the data.

In the baseline plan, for ProtoDUNE-SP, no monitoring or processing of data is planned or intended to use the EHN1 disk buffers as their input source. Low latency (order of seconds) DAQ "online monitoring" is performed using the in memory data sharing model provided by the *artdaq* tool suite.

Lower latency (minutes to hours) "nearline" monitoring will utilize the CERN EOS copy of the data as its input source. The compute resources for this processing will come from the DUNE experiment's allocation on the Tier 0. The data to be processed will be pre-scaled to match the allocated resources and other priorities of the DUNE experiment.

Routine "keep-up" processing, which involves the first stages of data processing and reconstruction, will be run on FNAL compute resources and will use the dCache copy of the ProtoDUNE files for input. This output from this stage of processing will be retained as an intermediate record of the processing. Data reduction schemes will be applied to in the processing chain after this initial translation stage.

IV. PROTODUNE TEST BEAM RUNS

The ProtoDUNE detectors will both collect physics data in beam during the second half of 2018. The beam run is expected to last three months and will terminate with the beginning of CERN's Long Shutdown 2 (LS2), which is scheduled to start in October 2018. Resource estimates for both the ProtoDUNE-SP and ProtoDUNE-DP computing are based on this run plan.

The ProtoDUNE detectors will acquire some cosmic-ray data prior to the beam run, and may take additional cosmic-ray data after the beam run. The exact time profile and amount of cosmic-ray data that may be acquired after the initial beam run has not been determined. Computing resources relating to this additional running after the conclusion of beam running have not been requested, but can be extrapolated in a linear fashion based on the knowledge of the event sizes and processing time per trigger. The DUNE computing group will make these requests for these resources when post beam run plans are established.

For the planning purposes, it is assumed that both detectors will take data simultaneously both before and during the beam running.

Paramater	Value
Dual Phase	
Event Size (compressed)	15.9 MB
Beam Event Triggers	100 M
Cosmic Ray Event Triggers	8.4 M
Projected Total Data Size	1.61 PB
Requested Total Data Size	2.5 PB
Compression Factor	10
Single Phase	
Event Size (uncompressed)	230.4 MB
Compression Factor	4

TABLE II. Raw data volumes predicted for ProtoDUNE-DP and ProtoDUNE-SP.

A. ProtoDUNE-SP

The ProtoDUNE-SP beam data-taking run is expected to last for 120 days. The detector program has a goal of acquiring approximately 15 million physics triggers, split between negative and positive beam polarities. To achieve this physics goal, and based on assumptions for trigger efficiencies, purities and live times described in the ProtoDUNE-SP TDR, it is estimated that approximately 50 M beam triggers will need to be collected. These triggers will constitute a raw data volume of approximately 3 PB. The actual data volume will be contingent on the compression ratio that is achieved and the noise rates that the detector experiences.

For planning purposes, ProtoDUNE-SP is assumed to require the processing of 50 M beam triggers through the initial reconstruction, but that the dataset will be reduced to approximately 15 M for the later stages of event processing and reconstruction (a $>3\times$ data reduction factor).

B. ProtoDUNE-DP

ProtoDUNE-DP has requested 120 days of beam running, and expects to take data with 50% efficiency, collecting a total of 100M triggers. The ProtoDUNE-DP event size is expected to be 15.9 MB/event resulting in a total data volume to be acquired of 1.48 PB. Cosmic-ray counters mounted on the vertical sides of the cryostat will trigger horizontal cosmic-ray events at a rate of the order of 1 Hz, which is expected to contribute a small fraction of the total data volume.

For planning purposes, ProtoDUNE-DP is assumed to require the processing of 100 M beam triggers through the initial stages of reconstruction and that this data set will be carried through to final reconstruction.

V. DATA PROCESSING MODEL

The data are to be processed using the LArSoft toolkit which is based on the *art* event-processing framework.

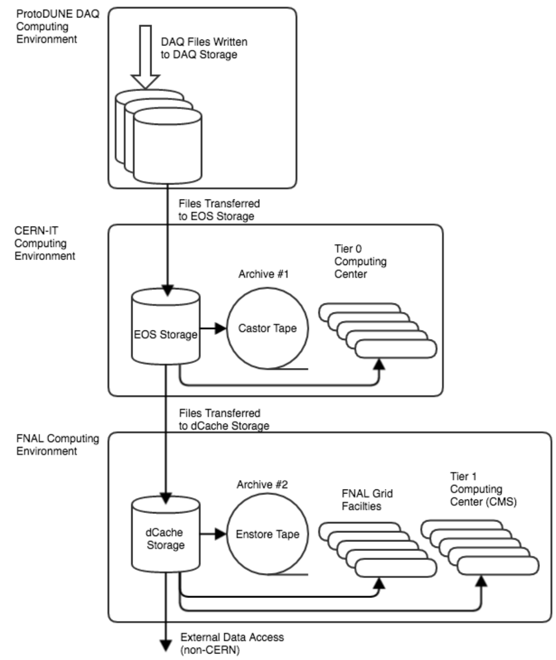

FIG. 1. Baseline data movement path for ProtoDUNE raw data.

The baseline reconstruction chain requires approximately 10 minutes per event to properly reconstruct. These event processing times, combined with the expected trigger sets to be taken, lead to computing requirements that can not be fulfilled by a single Lab. Data are to be processed for both ProtoDUNE-SP and ProtoDUNE-DP at a mixture of resources at Fermilab, at CERN, and on the Open Science Grid (OSG). Additional resources, such as allocations on high performance computing (like NERSC), commercial clouds and resources from collaborating institutions are being pursued.

Before a large production run of data processing is launched, calibration jobs are run first. These calibrations will determine the values of pedestals, gains, identify dead channels, and approximate the electron lifetime and space charge distributions. The results of these calibrations will be recorded in a database visible to subsequent processing jobs, and hosted at Fermilab, with caches as appropriate at computing centers providing resources for production.

For ProtoDUNE-SP, the data production jobs will mitigate stuck bits read out by the ADC's, subtract pedestals, apply calibrations, filter noise, deconvolve the wire and electronics response, and find hits, clusters, tracks, and showers. Derived data, including the hits, clusters, tracks, and showers will be written to recon-

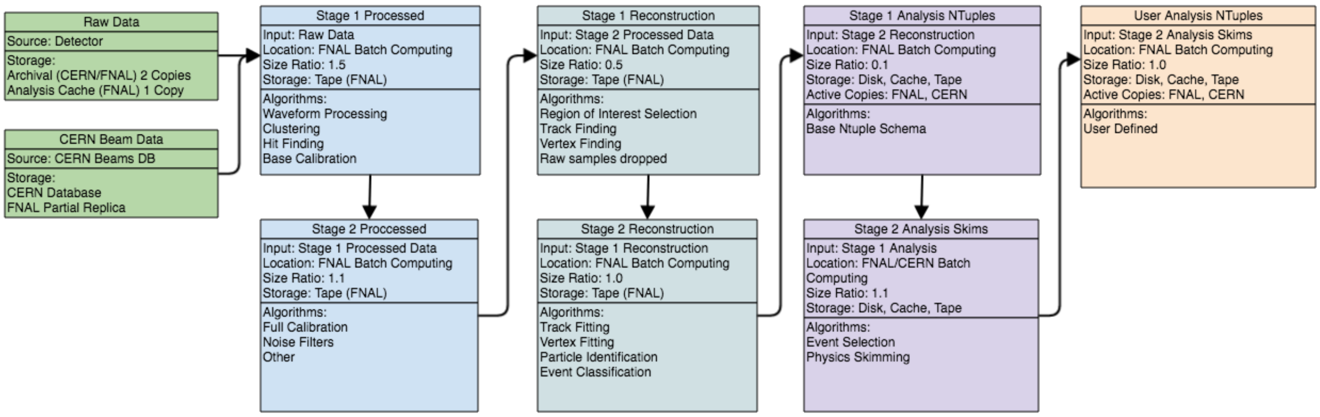

FIG. 2. Baseline data processing chain for ProtoDUNE raw data analysis.

structed data files. A more compact ROOT tree will also be produced for analysis convenience. A reduced version of the data, including the filtering and deconvolution, but also including region-of-interest sampling and downsampling the ADC readout times, will be produced along with the reconstructed output files. The reduced data may be used as input for subsequent processing steps, as well as to assist user code development and analysis. Because it is lossy, however, the un-reduced data-flow methods will be preserved.

Cosmic-ray rejection is intended to be performed as early as possible, in order to reduce the processing time and potentially the data storage volume.

Beam-particle association can happen at one of two steps. It can be done early in the reconstruction chain, to guide the selection of hits in order to purify the measurement of a particular kind of beam particle, assuming the performance of the beam instrumentation. Or, the association of beam instrumentation information with reconstructed particles can be performed after all reconstruction has been done, in order to cross-calibrate the efficiencies of both the TPC reconstruction and the beam instrumentation.

Figure 3 outlines the steps envisaged for the reconstruction of the raw data.

A small fraction, to be determined, of processed data files will be produced with copies of the raw data and filtered, deconvoluted signals. These are intended for calibration and performance evaluation, but are not intended for processing the entire sample for physics, as they are larger than the original raw data files.

ProtoDUNE-DP data processing will be similar to the single-phase processing. The larger signal-to-noise ratio, the longer drift lengths, and the fact that there are two wire directions instead of three require some adaptation of data preparation and reconstruction algorithms, though it is expected that much of the software can be shared. The step of disambiguating hits is not needed in ProtoDUNE-DP due to the lack of wrapped wires.

VI. RESOURCE ALLOCATION

In the DUNE computing model, collaborators are allowed to use the resources allocated to DUNE for DUNE-related work including simulation, algorithm development, data processing, analysis, and document preparation, and other activities needed for the collaboration. In some cases these activities may represent relatively large requests for resources. The DUNE Software and Computing Coordinators take responsibility for balancing the relative priorities of work within the DUNE computing sphere. They can adjust priorities of individuals and official campaigns. This is intended to balance the production and user usage of the resources during and after ProtoDUNE operations, but takes a much smaller role before the operations start.

Requests for additional resources from the Laboratories (CERN or FNAL) are submitted to the DUNE Software and Computing Coordinators and are presented to the labs and/or detector groups through the DUNE/Fermilab/CERN interface group.

For Fermilab resource requests, an annual procedure has been established by the Scientific Computing Division. Each year, usually in January, resource requests are made for the next year. Typically, equipment purchases are made towards the end of the fiscal year, which ends at the end of September, and the equipment arrives and is commissioned the following fiscal year. Requests are made to the Scientific Computing Division's Portfolio Mangement Team (SCPMT) for the three years following the request date, including the current fiscal year, in order to balance the needs and resource availability, and to give time to propagate the requests. Temporary reallocation of laboratory resources between experiments may be done under exceptional circumstances; a weekly meeting is held by the SCPMT for these purposes. Large-scale ProtoDUNE computing needs are expected to be identified well in advance of the actual need. Resources requests for DUNE/ProtoDUNE were presented at the

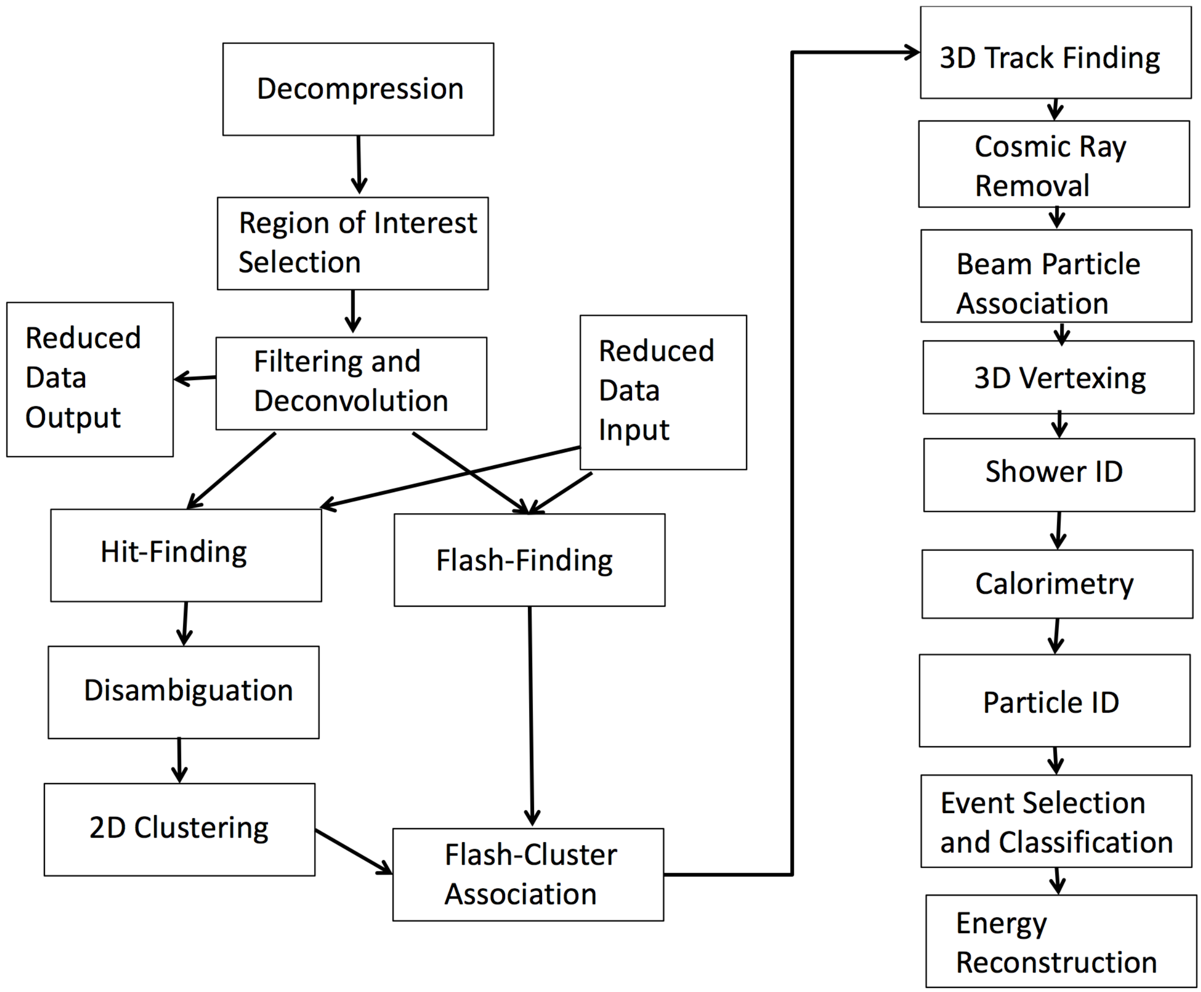

FIG. 3. Baseline data processing chain for ProtoDUNE-SP raw data analysis.

2017 SCPMT review.

Requests were made in a similar manner to CERN-IT for computing time on the Tier 0. The DUNE experiments has been granted an allocation of 1500 compute slots on the Tier-0 batch along with the requisite storage to match the system, as well as 3 PB of disk space and the networking needed to transfer data have been allocated. The tape space allocated within CASTOR is 6 PB. These Tier-0 batch slots and the EOS disk will be fully available in August 2017, and as of February 2017, 0.6 PB of EOS space, 500 Tier-0 cores, and 3 PB of CASTOR tape space are available for use. The DUNE Software and Computing Organization is tasked with providing tools that make collaboration use of these resources convenient, and to set up priorities so that they serve ProtoDUNE operations, data production, and analysis during the run.

Resources that are not directly controlled or administered by the cloud may be compatible with the DUNE software and its distribution. These resources will be used by ProtoDUNE as appropriate.

[1] https://cdcvs.fnal.gov/redmine/projects/filetransferservice https://cdcvs.fnal.gov/redmine/projects/fts-light https://cdcvs.fnal.gov/redmine/projects/sam

[2] Checksums can be calculated for this copy but are not planned due to the IO limits of the DAQ disks