Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Computing Resources for ProtoDUNE

A. Norman, H. Schellman
Software and Computing

Questions Addressed

"Are allocated resource, provided by CERN, FNAL and other organizations, sufficient in terms of temporary, long term and archival storage to meet the proposed scope of the ProtoDUNE-SP program? Are the computing resources (CPU cycles) allocated to ProtoDUNE-SP sufficient to meet the proposed scope of the program? Are these storage/compute allocations matched to the schedule of ProtoDUNE-SP run plan? How do resource allocations evolve and match with a post beam running era?"

Summarized:

- Is there enough tape?
- Is there enough disk?
- Is there enough CPU?
- Are there enough people? (people are computing resources too!)
- What does this look like on the CY18/CY19 calendar?

Questions Addressed

"Are the resource costs associated with the reconstruction/analysis algorithms understood? How will the execution of the data processing and reconstruction be evaluated and prioritized in the context of limited computing resources?"

Will address second part (prioritization vs. other DUNE activities)

Overview of Resources and Commitments

Introduction and Organization (DUNE S&C)

Organizational Reminder:

- DUNE Software & Computing falls formerly under DUNE management
  - By design S&C is responsible for the organization and utilization of computing resources and infrastructure for all of DUNE (not just the ProtoDUNE portions)
  - Not responsible for the actual algorithms needed by physics groups
  - Not responsible for deciding which algorithms or samples are needed by a physics group
    - Physics groups determine what they need and how it should be made.
    - S&C determines how best to satisfy those requests and map them onto new or existing infrastructure
  - ProtoDUNE (DRA in particular) is treated as a physics group
  - Communication with other aspects of ProtoDUNE (DAQ, DQM, etc…) used to establish the requirements, specifications and interfaces that are needed.
- S&C was specifically re-organized (Dec. 2016) across "operational" lines to help enable this model. 5 S&C groups + monitoring (SCD experts & consulting):
  - Data Management, Central Production, Software Management
  - Database Systems, Collaboration Tools, (Monitoring Systems)
  - All Groups are fully staffed with leadership and technically skilled individuals

Resources

Will cover resource needs and commitments across:

- Archival Storage (tape)
- Durable Storage (disk)
- Computational Resources (CPU)
- Networking
  - Site specific (CERN⇒CERN)
  - Wide Area Networking (CERN⇒FNAL)
- Centralized Repositories
- Database Systems
- Monitoring Systems
- Personnel and consulting

Legend
✓ Sufficient
◆ Borderline
✕ Insufficient

Resource Planning

- Planning for ProtoDUNE resources began in Winter 2017 (Jan/Feb)
  - Iterated on during 2017 and 2018
  - Evolved with changes to run plans
  - Firm commitments from FNAL and CERN for resources being presented
    - Additional resources maybe available at each host lab
    - Each lab has procedures for requesting resources (e.g. FNAL-SCPMT reviews)
- Communication is handled through “interface committee”
  - Include Bernd Panzer-Steindel (CERN-IT) Stu Fuess (FNAL-SCD) who can commit resources at respectively
  - Computing Requests are routed through lab specific protocols

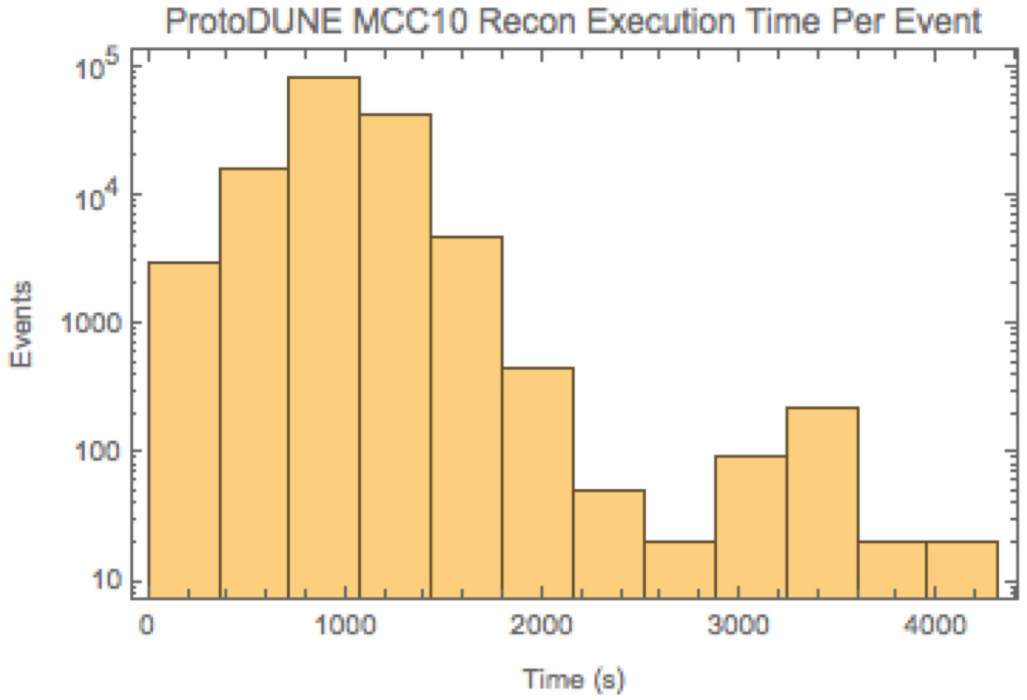
- Some resources can be provisioned quickly (Tier-0 shares), others require longer lead times and procurement (tapes, disks)
  - Personnel are a resource.
    Reallocation of personnel requires advanced planning.

Current Allocated Resources

- Archival Storage:
  - 6 PB tape (CERN), 6 PB tape (FNAL) [Shared NP02/NP04]
- Durable Storage:
  - 1.0 PB (logical) 2.0 PB (physical) EOS disk (staging + analysis)
    - Expanding to 1.5 (logical) 3.0 (physical)
  - 4.1 PB dCache (cache disk, shared)
  - 1.5 PB scratch dCache (staging, shared)
  - ~240 TB dCache dedicated write (staging) (To be allocated Summer '18)
  - 195 TB dCache (analysis disk)
- Compute:
  - ~1200 CERN Tier-0 compute nodes (0.86 Mhr/mo, 28.8 khr/day)
    - Actual allocation is 0.831% of the Tier-0
  - ~1000 FNAL Grid compute nodes (0.72 Mhr/mo, 24 khr/day)
- Network:
  - EHN1 to EOS: 40 Gb/s
  - CERN to FNAL: 20 Gb/s

Reconstruction Time

- Single event reconstruction times under current software algorithms were measured from MCC10 production campaign (commiserate with DC1.5)
- Observed event reco peaked at ~**16min/evt**
  - High side tail corresponding to reconstruction failures



- Baseline reconstruction, not advance hit finding or machine learning
- Actual data processing will require data unpacking/translation from DAQ format.
- Merging of Beam Inst. Data required for all data
- If there are other aux. data stream (i.e. non-artDAQ CRT)
  - may require other merging passes
  - Increases compute and storage requirements

Resource Assessments for Beam Operations

Resources at Baseline Scenario 1 (Uncompressed)

- Baseline Scenario for running
  - Start Aug 29th, End Nov. 11, 2018
  - 25 Hz readout rate
  - Compression factor 1 (no compression)
  - 45 beam days, 7 commissioning/cosmic days
  - Details of DAQ parameters: https://docs.google.com/spreadsheets/d/1UMJD3WAtWjnZRMam7Ltf-2BBzq25xVCnbA6QW5N5oew/edit?usp=sharing

Summary

- Average Data Rate = 1.6 GB/s (12.8 Gb/s)
- Total readout data = 3.6 PB (2.7 PB required for TDR # events)
- Total Events: 13.03 million
- Target trigger purity: 0.75

Resources at Baseline Scenario 1 (Uncompressed)

- Average Data Rate = 1.6 GB/s (12.8 Gb/s)
  - Demonstrated data transfer rates:
    - EHN1 to CERN-EOS: 33.6 Gb/s ✓
    - CERN-EOS to FNAL-dCACHE: 16 Gb/s ✓
- Total readout data (raw)= 3.6 PB (2.7 PB required for TDR # events)
  - Exceeds "fair share" of SP/DP allocations (2.5+0.5 PB/ea) ◆
    - Within allocated envelope if DP is deferred/de-scoped ◆
      - Exceeds total storage budget w/ analysis inflation factors included ✗
  - Raw data set beyond disk allocations
- Total Events: 13.03 million
  - Full Reconstruction: 3.47 MCPU hr / 7.5 weeks = 77 kCPU hr/day ◆
    - Have 52.8 kCPU hr/day dedicated from CERN+FNAL
    - Need factor of 1.45 more compute
    - Need either +1000 nodes/day or 20 days of additional scope in computing turn around
    - Ignores other DUNE compute activity in the Sept/Oct Timeframe
    - Can descope/defer full reco till post TDR

Resources at Baseline Scenario 2 (Compressed)

- Baseline Scenario for running
  - Start Aug 29th, End Nov. 11, 2018
  - 25 Hz readout rate
  - *Compression factor 5*
  - 45 beam days, 7 commissioning/cosmic days

Summary

- Average Data Rate = 0.320 GB/s (2.56 Gb/s) ✓
- Total readout data = 0.72 PB (0.54 PB required for TDR # events) ✓
  - Within tape allocation including inflation ✓
  - Permits disk resident dataset ✓
- Total Events: 13.03 million ◆
  - Requires factor 1.45 more compute for full reconstruction, same as uncompressed scenario
- Target trigger purity: 0.75

Resources at Baseline Scenario 3 (50 Hz Compressed)

- Baseline Scenario for running
  - Would follow a ramp from scenario 2
  - 50 Hz readout rate
  - *Compression factor 5*
  - Assume full run for upper limits
    (45 beam days, 7 commissioning/cosmic days)

Summary

- Average Data Rate = 0.597 GB/s (4.78 Gb/s) ✓
- Total readout data = 1.34 PB (0.54 PB required for TDR # events) ✓
  - Within tape allocation including inflation ✓
  - Permits disk resident dataset ✓
- Total Events: 24.2 million ✗
  - 6.4 MCPU hr over the run (143 kCPU hr/day)
  - Requires factor 2.72x more compute for full reconstruction
  - 122 days of processing or ~3800 more compute nodes
- Target trigger purity: 0.40

Resources at Baseline Scenario 4 (100 Hz Compressed)

- Baseline Scenario for running
  - Would follow a ramp from scenario 3
  - 100 Hz readout rate
  - *Compression factor 5*
  - Assume full run for upper limits
    (45 beam days, 7 commissioning/cosmic days)

Summary

- Average Data Rate = 1.15 GB/s (9.1 Gb/s) ✓
- Total readout data = 2.58 PB (0.54 PB required for TDR # events) ✓
  - Within tape allocation (raw) ✓
  - Exceeds w/ including inflation ◆ (Heavy filtering?)
  - Permits disk resident dataset ✗
- Total Events: 46.7 million ✗
  - 12.5 MCPU hr over the run (277 kCPU hr/day)
  - Requires factor 5.32x more compute for full reconstruction
  - 240 days of processing or ~9375 more compute nodes
- Target trigger purity: 0.21

Senario Summary

- Resource Allocations can be summarized

Scenario	Network (Local/Wide)	Tape	Disk	CPU	Databases & Other
25 Hz Uncompressed	✓/✓	◆	◆	◆	✓
25 Hz Compressed	✓/✓	✓	✓	◆	✓
50 Hz	✓/✓	✓	✓	✗	✓
100 Hz	✓/✓	◆	◆	✗	✓

Tape Summary

- Have assumed “standard” archival procedure of 2x copies of raw data
- Have assumed an “inflation” (ratio of all data to raw data) of 1.7 based on aggressive filtering and reduction schemes

Under these assumptions:

- Archival Tape allocated for FY18/FY19 is sufficient to cover the baseline beam run under two of the senarios (25 Hz and 50 Hz compressed readout) and include enough spare storage to accommodate commissioning and analysis
- Tape resources not sufficient for extended uncompressed running or prolonged high rate running.
- Tape resources not sufficient for detector operations past beam run
  - Require additional resource request to each lab
  - Request requires well defined run plan and readout strategy to set resource level

Disk Summary

- Disk resources are used for both operations (data staging) and for analysis activities.
- Analysis activities are assume to use standard computing model used by other large data volume experiments.
  - Disk is used as performant cache layer in front of slow (high capacity) tape systems.
  - Data migrates on demand between tape and disk layers
  - Cache policy (usage based) keeps cache populated with “popular data”

Under these assumptions:

- Disk resources are allocated for FY18/FY19 are sufficient to cover the baseline beam run under two of the scenarios (25 Hz and 50 Hz compressed readout) with the majority of the data and associated processed/analysis data resident within the disk systems at CERN/FNAL and without exceeding cache capacities.
- Disk resources are boarderline sufficient for extended uncompressed running or prolonged high rate running and have the potential to turn over even the large cache layers.
- Disk resources for detector operations past beam run may require additional resources
  - Request requires well defined run plan and readout strategy to set resource level

CPU Summary

- CPU resources are used for merging the raw data streams, perform calibrations, and reconstruction
  - The processing/reconstruction chain has an estimated time of 16 minutes/evt based on recent large scale reco activities.
- CPU resources are allocated as “dedicated” resources from CERN/FNAL.
  - Additional resources are available in “opportunistic” fashion from both labs

Under these assumptions:

- CPU resources are allocated for FY18/FY19 are sufficient to cover the baseline beam runs under the 25 Hz readout scenarios. These scenarios would require minimal opportunistic resources (well within normal operational capacity), or minimal extension to data processing timelines (~1 month additional running)
  - Will require balancing in context of other DUNE activities (TDR simulation needs)
  - Baseline plan would be to perform initial data processing for ProtoDUNE and prescale/defer full reconstruction for all events till after TDR simulation (i.e. defer from realtime reco to full reco campagin in Feb 2019)
- CPU needs under higher rate readout scenarios do not fit within the allocated resources without significant additional resources or scope in processing timelines
- CPU needs for post beam running require enumeration of run plan and readout strategy

Network Summary

- Networking is use for internal data transport at CERN
- Networking is used for trans-Atlantic data transport to FNAL

Under these assumptions:

- Networking resources are sufficient in all scenarios
- Networking resources are sufficient for post beam running

Central Software Repositories Summary

- Central software repositories are used to distribute the run time applications
  - DUNE/ProtoDUNE uses an established run time environment (DuneTPC) based on LarSoft which is built regularly as tagged/versioned releases.
  - All S&C managed activities use these releases
  - DUNE/ProtoDUNE uses a CVMFS based distribution scheme for deployment to computing sites

- Containerized run time environments have been prototyped and run successfully for the DUNE software stack.
  - Enables running on specific external sites (Primarily NERSC facilities)

Under these assumptions:

- Software repositories and deployment systems are sufficient to support ProtoDUNE operations
- Current schemes enable ProtoDUNE to access significant opportunistic resources
- Potential for access significant additional resources through containerized software scheme
  - Requires additional development and integration work to move to a production ready system (2-3 months)
  - Would require a resource allocation from NERSC or other facilities

Database Summary

- Beam instur. Database system and interface translate from CERN beam systems to Offline compatible systems.
- DAQ/Run database systems translate from Online to Offline compatible systems.

Under these assumptions:

- Beam Database resources are sufficient in all scenarios
- DAQ/Run databases require resources at FNAL (hosting and DB service) and development for translation software and integration with LarSoft.

Monitoring Summary

- S&C Supported Monitoring for ProtoDUNE uses standard monitoring tools provided by CERN and FNAL.
- Monitoring is aggregated under a number of “dashboard” which are supplied by FNAL-SCD
  - Modification to dashboards is through “consulting effort” and not directly supported/maintained by DUNE S&C (i.e. a request is put in for specific monitoring plots, and experts work with us to instantiate these and include them in the monitoring suite)

Under these assumptions:

- Still require development/integration effort prior to beam operations (Sept-Nov 2019)
  - Mainly integration of new plots into dashboards as experimenters identify needs
- Monitoring systems scale to support beam operations and post-beam operations
- Personnel resources may require greater allocations if significant portions of the DAQ environment are to be propagated to the central monitoring.

Personnel Summary

- All subgroups with DUNE S&C will have staffing and expertise to operate and support their associated services and systems
- Operations of ProtoDUNE-SP will require expert staffing (i.e. on-call) during beam running and during the commissioning period
- Personnel can be located at either CERN or FNAL (no technical restriction preventing remote operation of computing services)
- Presence at CERN is HIGHLY desirable
  - Based on recent Data Challenge experiences and other operational experiences

Under these assumptions:

- Personnel resources are sufficient to support beam operations (Sept-Nov 2019)
- Resources for stationing personnel at CERN during this period are not sufficient from either the collaborating University groups or FNAL. (i.e. travel budgets, teaching relief in support of extended relocation)
- Resources not sufficient for post-beam running. Requires post-beam plan and funding.

Scorecard

- Readiness and Allocation of resource

Scenario	Beam Running	Post-Beam	Notes
Archival Tape	✓/◆	◆	Highly readout scenario dependent
Disk	✓/◆	?	Highly readout scenario dependent
CPU	◆/✗	?	Highly readout scenario dependent
Networking	✓	✓	
Central Repositories	✓	✓	May enable additional resources
Databases	✓	✓	Runs Database under developement
Monitoring	✓	✓	May require additional consulting
Personnel & Consulting	◆	✗	Need support for travel and operations at CERN for U.S. collaborators

Summary

- Most resources for the beam run are in place for the beam run.
- The adequacy of the resources are HIGHLY readout/run plan dependant.
  - Baseline scenarios fit the resource windows (w/ acceptable risk, i.e. opportunistic CPU cycles or time contingency)
  - Extended scenarios do not fit the resource windows
  - Running post-beam does not have planned/allocated resources
    - Requires a run plan to establish the need level
  - Will require prioritization in context of DUNE TDR work

- Support for personnel is required
  - Both University and Lab groups
  - Presence at CERN is highly desirable
  - Extended operations in 2019 requires assessment