EPJ Web of Conferences **245**, 03019 (2020) https://doi.org/10.1051/epjconf/202024503019
*CHEP 2019*

# DUNE Production processing and workflow management software evaluation

*Kenneth* Herner[1,***]

[1]Fermi National Accelerator Laboratory, Batavia, IL, USA

**Abstract.** The Deep Underground Neutrino Experiment (DUNE) will be the world's foremost neutrino detector when it begins taking data in the mid-2020s. Two prototype detectors, collectively known as ProtoDUNE, have begun taking data at CERN and have accumulated over 3 PB of raw and reconstructed data since September 2018. Particle interaction within liquid argon time projection chambers are challenging to reconstruct, and the collaboration has set up a dedicated Production Processing group to perform centralized reconstruction of the large ProtoDUNE datasets as well as to generate large-scale Monte Carlo simulation. Part of the production infrastructure includes workflow management software and monitoring tools that are necessary to efficiently submit and monitor the large and diverse set of jobs needed to meet the experiment's goals. We will give a brief overview of DUNE and ProtoDUNE, describe the various types of jobs within the Production Processing group's purview, and discuss the software and workflow management strategies are currently in place to meet existing demand. We will conclude with a description of our requirements in a workflow management software solution and our planned evaluation process.

## 1 Introduction to DUNE

The Deep Underground Neutrino Experiment is an international mega-science project that aims to probe the nature of the universe through a variety of methods such as studying neutrino oscillations, searching for proton decay, and collecting a large neutrino flux from a core-collapse supernova in our galaxy. The collaboration currently has over 1200 members in over 30 countries. The full experiment will consist of a Near Detector at Fermilab, itself containing three sub-detectors, and a Far Detector at SURF in South Dakota, consisting of four 10-kt liquid argon time projection chambers utilizing various readout technologies.

In addition to the main detectors two prototype detectors are currently operational at CERN: ProtoDUNE Single-Phase (SP) and ProtoDUNE Dual-Phase (DP). These detectors utilize the same technology and design (at approximately 5% scale) intended for the Far Detector modules and serve as technology demonstrators for the larger project. The ProtoDUNE-SP detector took approximately six weeks of data with beam on a staggered schedule from September to November 2018. Another beam run is planned for 2022 with both the SP and DP detectors. In the meantime both detectors continue to take cosmic ray data and run a variety of calibration and DAQ system studies. Experience with ProtoDUNE

*for the DUNE Collaboration
**e-mail: kherner@fnal.gov

is vital to building expertise in maintaining liquid-argon TPCs over the long periods of time that DUNE will require.

### 1.1 The DUNE Production Group

The DUNE Collaboration already has an active and expanding computing effort. Ref. [1] provides an overview of the current state of DUNE computing. Within the overall Computing Consortium, the Production Group is currently responsible for running central reconstruction on ProtoDUNE raw data, large-scale Monte Carlo (MC) generation for all detectors, and works with the data management group to curate datasets. Additionally tests jobs submitted by Production group members are typically the first jobs sent to test new sites as they join the DUNE computing effort. Group members also work very closely with the data management group to test future storage solutions, and with service providers to test current and future workflow management systems. The group size has varied from two to six people over the past two years.

The main consumers of computing resources have been the ProtoDUNE reconstruction campaigns. The initial reconstruction campaign for the ProtoDUNE-SP data was from September to December 2018, and a reprocessing pass occurred between late August and early October 2019. Table 1 provides information on the CPU and storage utilization on the most recent ProtoDUNE-SP reprocessing campaign.

| | |
|---|---|
| Raw data (SP+DP) | 3329 TiB |
| Raw "physics" beam data (SP) | 786 TiB |
| Reco output size (SP "good" physics runs only) | 169 TiB |
| Wall hours (beam data+new MC) | 2.08 M |

**Table 1.** Statistics on the most recent ProtoDUNE-SP reprocessing campaign, August-October 2019. "Physics" data is defined as data taken with beam, as opposed to cosmic ray or DAQ test data. "Good" runs are those runs taken with beam when the DAQ system and detector were fully stable.

## 2 Current production workflow submission infrastructure

### 2.1 Current job submission infrastructure

Production jobs use Fermilab's Production Operations Management System (POMS) [2, 3] for submission. POMS provides both GUI and command-line options for job launches (both immediate and scheduled), recovery project setup, and provides integrated monitoring with Fermilab's Landscape project. Job submission is typically via Fermilab's Jobsub tool [4]. Jobsub in turn interfaces with the GlideinWMS workflow management system [5] for resource provisioning and matchmaking to slots at Fermilab, on dedicated DUNE resources at other sites, or opportunistic cycles on the Open Science Grid (OSG) [6]. Figure 1 illustrates the entire chain, including interaction with storage elements within the job. The architecture can also provision resources on High-Performance Computing (HPC) resources, such as Cori at the National Energy Research supercomputing Center (NERSC), within the HEPCloud [7, 8] infrastructure. The submission mechanism is unchanged whether the jobs are HTC or HPC; this seamless transition is key to efficiently utilizing available resources and also saves the job submitter significant effort by not requiring customized submission infrastructure for different resource types.

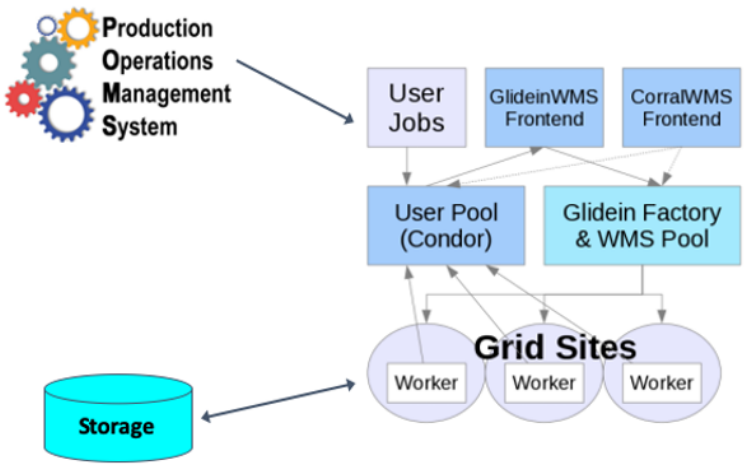


**Figure 1.** Overview of the current DUNE Production workflow setup used also by the ProtoDUNE detectors for data reconstruction and simulation. Production group members interact with POMS to submit jobs, which uses the Jobsub tool to submit jobs to a HTCondor schedd. GlideinWMS provisions worker node resources and jobs match to the available worker node slots. DUNE jobs interact with storage elements both at Fermilab and other sites both for input copy (or streaming for most production workflows) and output copyback.

Choosing a Glidein-based system at this stage of the experiment had several advantages. DUNE was able to quickly leverage the existing FIFE [2] toolset, including POMS and Jobsub, negating the need for significant effort from the experiment in getting jobs running quickly (let alone in designing a completely new system). Since the system is in use by other neutrino experiments at Fermilab, it is easy for new DUNE collaboration members coming from these experiments to begin submitting jobs quickly as they are working with a familiar system. The DUNE Production workflows were also able to leverage the existing infrastructure support teams in place to server other collaborations and consortia such as CMS and OSG. Finally, as GlideinWMS is widely used in HEP, setting up new sites becomes extremely straightforward, especially if the site is already supporting another experiment that uses GlideinWMS. Our integration times for new DUNE sites are typically less than one week and successful production jobs immediately after opening up the site are now the rule rather than the exception. This ease of setup has been a key enabler of DUNE's international expansion over the past 18 months. From January to November 2019, sites outside the United States delivered approximately 49% of the total DUNE Production wall hours as shown in Figure 2. International sites regularly run the full suite of DUNE jobs, including ProtoDUNE data reconstruction. DUNE is also considering the creation of a global GlideinWMS pool similar to the CMS Global Pool [9, 10], which would enable multiple institutions to set up their own submission hosts if they desired to do so.

### 2.2 DUNE software, input data distribution, and output file handling

DUNE builds its software suite for both Scientific Linux 6 and Scientific Linux 7, though we anticipate phasing out SL6 support in the first half of 2020. Since October 2019 DUNE jobs

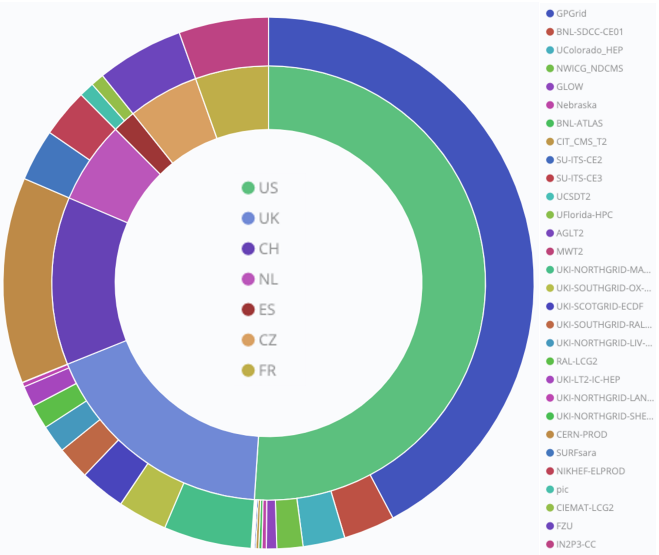


**Figure 2.** Wall time-weighted distribution of successful DUNE Production jobs January 2019 to November 2019. Inner ring: distribution by country. Outer ring: distribution by site. Sites outside the United States delivered 49% of the total wall hours for the jobs shown here.

will automatically run inside a Singularity container at supported sites via a GlideinWMS mechanism that requires no user knowledge of Singularity other than specifying the desired image, reducing the possibility of errors and guaranteeing a homogeneous environment across all sites.

For its file catalog, DUNE currently uses Fermilab's Sequential Access via Metadata (SAM) system [11], which handles input file selection and delivery for each job, and has been successfully used by numerous HEP experiments for well over a decade. In the future we expect Rucio [12] to gradually take over most of these functions. Production jobs typically stream input data files via XRootD, though in some cases they will copy a file to the worker node and directly read the local copy. The data source can be any storage system reachable from the worker node and to which DUNE has access. This is most often Fermilab dCache, but CERN EOS and other storage elements in Europe are also used (jobs run at CERN would get their inputs from CERN EOS, for example). Several DUNE workflows require one or more auxiliary input files (i.e. not detector data) such as calibration files, neutrino flux information, and other inputs necessary for analysis for MC generation. Some simulation workflows randomly choose several such input files for each job from a much larger set, so the file overlap between jobs is small. Additionally the files are typically tens of MB in size. These two attributes make these files poor candidates for placement in a standard CVMFS repository. For such files we store them in a StashCache [13] repository, accessible in a POSIX-like fashion though a CVMFS overlay. With this method there is still some level of shared caching on a worker node, but as files need to be copied in from the source (Fermilab dCache), it happens in a transparent way, meaning the user can simply access the files via a CVMFS path in the dune.osgstorage.org repository.

For output file handling nearly all workflows currently copy their outputs to Fermilab dCache, with a small minority (ProtoDUNE-DP jobs) also copying to a storage element at IN2P3 in France. We use Rucio to manage file replication to other sites. In the future DUNE

will likely move to a more distributed copyback model: at sites with local storage, a job can simply copy its output to a local location, and then we can use Rucio for replication as is already done, rather than requiring everything first go through Fermilab.

There is an exception for workflows run at NERSC; we read input auxiliary files from and copy output files out to Cori's global scratch filesystem. For job outputs, a separate process performs bulk transfer back to Fermilab from one of NERSC's dedicated data transfer nodes. We expect that workflows on other future HPC platforms will follow a similar approach, especially at places without external connectivity on the worker nodes.

## 3 Future workflow management software evaluation

The existing POMS+GlideinWMS infrastructure should be adequate to meet DUNE's needs through the second ProtoDUNE run planned for 2022. As DUNE constructs the Near and Far Detectors, it is clear that the computing challenges will be much greater than they currently are, and the current infrastructure will not suffice. The main needs are: the ability to quickly provision resources on a variety of architectures, including traditional High Throughput Computing, HPC, and GPU; the ability to schedule large blocks of jobs on HPC machines when work demands it or when such resources are foreseen to be coming available soon; and support for so-called "pipeline" workflows, which are multi-stage workflows where each stage may require a different type of resource. A provisioning system should be aware of these needs and begin provisioning the resources required for the $N+1^{th}$ stage while jobs for the $N^{th}$ are running so that the resources for the next stage are available immediately after the conclusion of the previous stage.

The POMS system will continue to evolve to meet DUNE's requirements, and DUNE will consider continuing to use the POMS system as it is upgraded. DUNE is also considering two other existing workflow management systems: PanDA [14], currently used by the ATLAS Collaboration, and DIRAC [15], currently used by LHCb and several other medium-scale experiments. At the moment DUNE has not finalized its formal requirements for functionality in a future system. It is clear, however, that the requirements should be driven by the full DUNE computing model, which takes into account limitations such as total experiment storage and network bandwidth. They should not be driven by any attempts to fit into any specific existing setup.

## 4 Summary

The DUNE Production environment builds on successful job submission and workflow management systems also used by other experiments. The Production Group is responsible for large-scale data reconstruction and simulation generation, and plays major roles in computing site commissioning and workflow management software evolution. Several workflow management systems are under consideration by DUNE for long-term adoption, and the evaluation process will be based on the requirements of the full DUNE computing model.



This research was done using resources provided by the Open Science Grid [5, 6], which is supported by the National Science Foundation award 1148698, and the U.S. Department of Energy's Office of Science.

## References

[1] H. Schellman, this conference, see https://indico.cern.ch/event/773049/contributions/3581360/
[2] K. Herner *et al.*, EPJ Web of Conf. **214**, 03059 (2019).
[3] M. Mengel *et al.*, this conference, see https://indico.cern.ch/event/773049/contributions/3473358/
[4] D. Box, J. Phys.: Conf. Ser. **513**, 032010 (2014).
[5] I. Sfiligoi *et al.*, *2009 WRI World Congress on Computer Science and Information Engineering (CSIE2009)* (IEEE, 2009) 428-432.
[6] R. Pordes *et al.*, J. Phys.: Conf. Ser. **78**, 012057 (2007).
[7] B. Holzman, L.A.T. Bauerdick, B. Bockelman *et al.*, Comput. Softw. Big Sci. **1**, 1 (2017).
[8] P. Mhashilkar *et al.*, EPJ Web of Conf. **214**, 03060 (2019).
[9] S. Belforte *et al.*, J. Phys.: Conf. Ser. **513**, 032041 (2014).
[10] J. Balcas *et al.*, J. Phys.: Conf. Ser. **664**, 062031 (2015).
[11] R. A. Illingworth, J. Phys.: Conf. Ser. **513**, 032045 (2014).
[12] M. Barisits *et al.*, Comput. Softw. Big Sci. **3**, 11 (2019).
[13] D. Weitzel *et al.*, in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)* (ACM, New York, 2019) Article 58, 1–7.
[14] P. Svirin *et al.*, EPJ Web of Conf. **214**, 03050 (2019).
[15] A. Casajus *et al.*, J. Phys.: Conf. Ser. **219**, 062049 (2010).