

Scale tests of the new DUNE data pipeline

Steven Timm^{1*}, For the DUNE Collaboration

¹Fermi National Accelerator Laboratory, Batavia, IL 60510, USA

Abstract. In preparation for the second runs of the ProtoDUNE detectors at CERN (NP02 and NP04)[1], DUNE has established a new data pipeline for bringing the data from the EHN-1 experimental hall at CERN to primary tape storage at Fermilab and CERN, and then spreading it out to a distributed disk data store at many locations around the world. This system includes a new Ingest Daemon and a new Declaration Daemon. The Rucio[2] replica catalog, and FTS3 transport are used to transport all files. All file metadata is declared to the new MetaCat[3] metadata service. All of these new components have been successfully tested at a scale equal to the expected output of the detector data acquisition system (~2-4 GB/s), and the expected network bandwidth out of the experimental hall. We present the procedure that was used to test and the results of the test.

1. Phase 1 Data Pipeline Test

The ProtoDUNE-II[1] detectors are located in the EHN-1 experimental hall, also known as the Neutrino Platform, at CERN, and are referred to as ProtoDUNE-IIHD Module 0 (NP04) and ProtoDUNE-IIVD Module 0 (NP02) respectively. Each is preparing for their second beam run and is capable of generating as much as 4GB/s of data running at full rate. Since the first beam run in 2018, the DUNE Data Management group has deployed a new data management system based on MetaCat[3] to store the file metadata, and Rucio[2] to be the file location catalog. We have also had to rewrite the data pipeline daemons. This requires testing the new data pipeline at scale to make sure that we can process the full expected data rate. The raw data file format has also changed in preparation for this run, to HDF5 format. This paper will describe the Phase I data pipeline test and also mention the results of the Phase II distributed processing test.

* Corresponding author timmm@fnal.gov

1.1 Description of Data Pipeline Components

Figure 1 shows a complete system diagram of the data flow for this test. Data is sourced from four data logging machines referred to as “EHN1 DAQ” in the experimental hall, each of which has several large and performant disk arrays managed under software RAID. These machines each run a XrootD server.

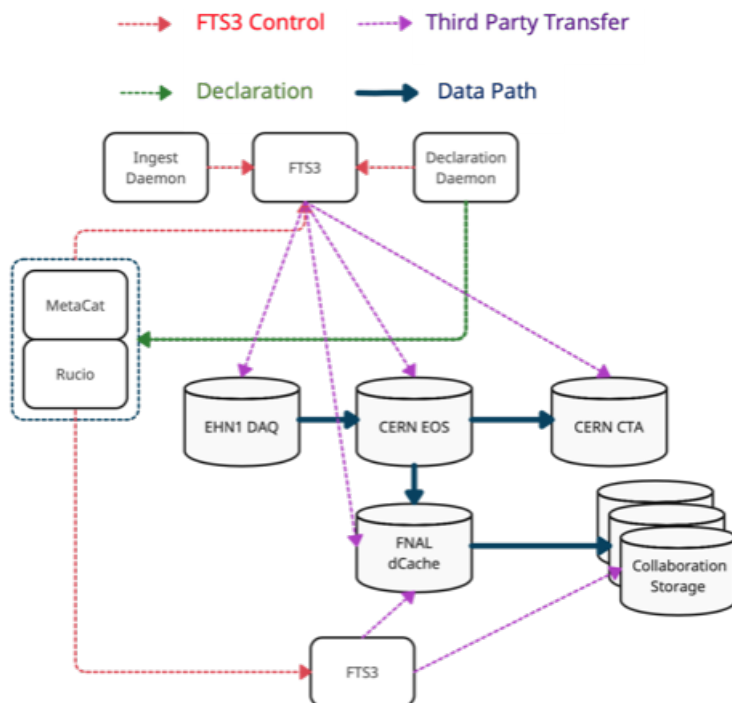


Fig. 1 System Diagram of the ProtoDUNE II Data Pipeline

There is a 40 Gbit/s data link between the DAQ system at EHN-1 and the CERN computing center. The Ingest Daemon scans the DAQ servers for new files and corresponding metadata files, and then initiates a 3rd-party transfer of the files from the DAQ server to a directory in CERN public EOS. The Declaration Daemon scans this directory and when new files arrive, it copies them to our disk Rucio Storage Element (RSE) at CERN public EOS and declares the replica both to MetaCat and to Rucio. It then creates Rucio rules which instruct the Rucio server to copy the files (using FTS3 transport) to our tape storage at CERN CTA and at Fermilab Enstore/dCache. Finally for Phase II of the data challenge we spread out five copies of the data to ten distributed RSEs.

1.2 Test Data Samples

We used three types of data in the data challenge. The first two types of data were from the Vertical Drift prototype ‘coldbox’, a test stand including a small LArTPC (Liquid Argon Time Projection Chamber) used for electronics and detector integration testing, which provided reconstructible data. This coldbox had liquid argon in it and hits from cosmic ray muons were could be reconstructed. There are two types of electronics, the “top” electronics

which were read out with the legacy NP02 DAQ readout in a binary format, and the “bottom” electronics which were read out with the updated DUNE DAQ system, which writes raw data in HDF5 file format. We also made a Monte Carlo simulation of the new ProtoDUNE-II Horizontal Drift detector data, which is Root format.

Each individual file for the “top” drift electronics, “bottom” drift electronics, and Monte Carlo simulation had sizes of 3 GB, 4 GB, and 1.4 GB, respectively. For each of these types of data, we started with a dataset corresponding to one run: approximately 250 GB for each type. To mimic the data acquisition system generating more data, we cloned these 250 GB samples every 15 minutes on each of the four DAQ servers, leading to an effective production rate of 1 TB every 15 minutes, or 4 TB/hr. The total sample size was 500 TB, which also considers minor differences in the start of data cloning and the beginning of data pipeline tests.

1.3 Phase 1 Data Pipeline Run

We officially began running the Phase 1 Pipeline at midnight UTC on July 11, 2022, and ran it until midnight UTC on July 16, 2022. Figure 2 shows the transfer rates we observed from CERN to Fermilab.

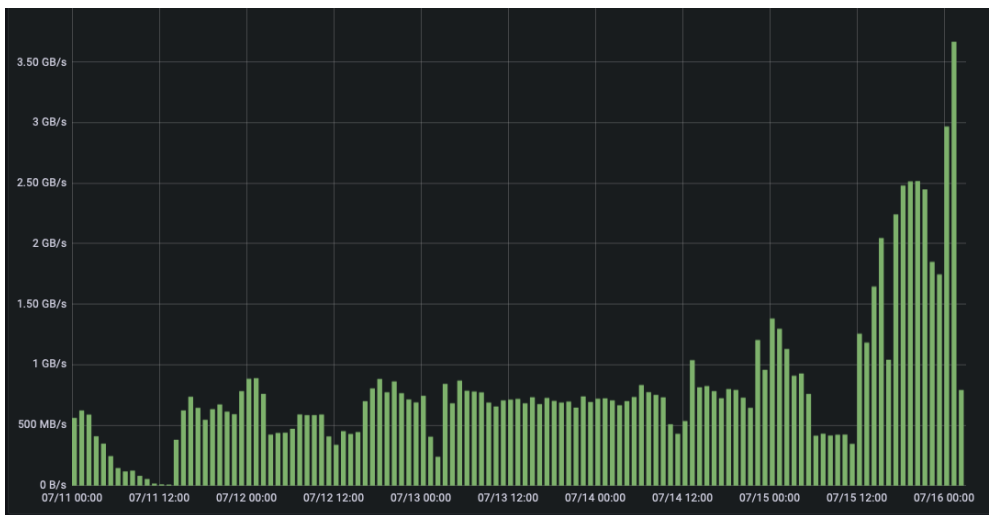


Fig. 2. Transfer Rates as monitored by FTS3 between CERN and Fermilab.

As seen in Figure 2 above, we made a number of improvements during the course of the data challenge to achieve the desired data rate. The first 12 hours saw slow rate due to the Ingest Daemon not having multithreading turned on as it should have been. We also found that the disk performance of the DAQ servers on which we were generating the copies of the data was a limiting factor at first. Each copy of the file was written once and then had to be read twice, once to calculate the checksum and once to extract the metadata. Initially the metadata extraction script was trying to compile a list of all event numbers in the file and add that to the metadata. Once we disabled that feature the rate was sufficient to keep up. After the data challenge the servers were reinstalled with better RAID parameters and the script was modified to require just one read rather than two. The dip on Tuesday 7/13 was due to an

EOS slowdown not caused by us. The key improvement was to greatly increase the size of the virtual machine on which the Declaration Daemon ran, from two cores to eight, and to run more threads. Once this was done on the morning of 7/15 the declaration daemon caught up with its backlog quickly, clearing a backlog of 55 TB and processing 45 TB more. Finally we had to adjust the rates of all the downstream daemons. The MetaCat server had to allow more incoming queue requests, and the FTS3 server had to be modified to allow up to 500 simultaneous transfers in flight between Fermilab and CERN, from the default of 120. With all these changes together, we were able to achieve a rate of 3.6 GB/s (28.8 Gbit/s) on the transatlantic link. Once we stopped generating new files at midnight of July 16, the rest of the files were cleared through the pipeline in less than 2 hours.

The network link as monitored by ESNNet is shown in Figure 3.

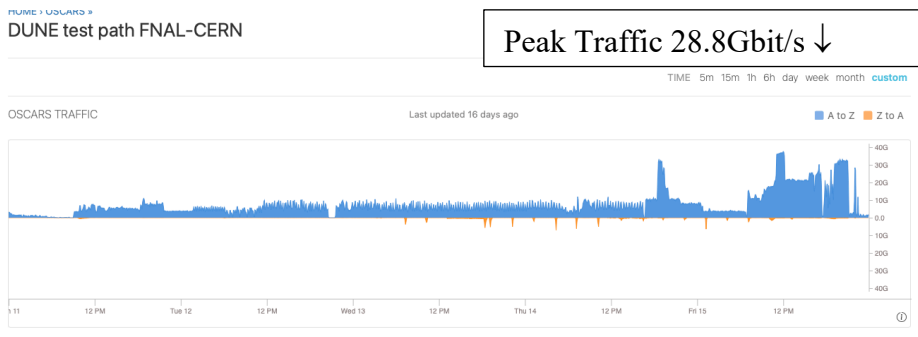


Fig. 3. OSCARS link between CERN and Fermilab

An OSCARS[5] virtual link was set up between CERN eospublic.cern.ch and Fermilab public dCache for purposes of the ProtoDUNE data traffic. The blue bars (upward pointing) in Figure 3 show CERN to Fermilab traffic and the orange bars (lower pointing) show Fermilab to CERN traffic. The peak traffic in our test here was 28.8 Gbit/s. There are a couple earlier high spikes which came from the ProtoDUNE detector also taking a small amount of data while we were doing the test.

2 Phase II Distributed Processing

We replicated the 500 TB data set to a collection of 10 disk-based RSEs. The output of the pipeline phase had already given us one full copy at Fermilab and one at CERN. We placed one full copy distributed between five data centers in the UK, another one distributed between 3 data centers in continental Europe, and one at Brookhaven National Laboratory. It was our goal for the distributed processing that we would assign the various workflows to the site which had the copies of the corresponding type of files. For this purpose the justIN[4] (just IN time) workflow system was developed. Generic jobs are submitted to all sites which then call back to the workflow manager and get both a workflow payload and a file to process which is optimal for the site at which they are executing. This was the first time we had done significant streaming of HDF5 files via XrootD by means of the POSIX plugin

The large scale tests of the justIN workflow manager were done in late December of 2022. Figure 4 shows the site distributions for two of the large workflows that processed the entire data challenge data set. For example, within the UK, the “vertical drift coldbox bottom

electronics” data was concentrated at Manchester and the “vertical drift coldbox top electronics” data was concentrated at RAL. Thus one would expect mostly jobs processing “vertical drift coldbox bottom electronics” to run at Manchester, and many fewer of the other kind. This did in fact happen. The top graph shows the distribution of vd-coldbox-top jobs. The Manchester contribution is the small two peaks at the top of the stacked line graph. The bottom graph shows the distribution of vd-coldbox-bottom jobs. The Manchester contribution there is the thick area at the top of the stacked line graph from Dec. 29 through Dec. 30. At various points more than 2000 simultaneous jobs were running there. Note also that the top graph shows a significant contribution from NL_NIKHEF which did have a vd-coldbox-top set of data. The scale of 8000 and 7000 simultaneous jobs running shows that overall the justIN system can submit and keep track of workflows at the scale needed for ProtoDUNE or full DUNE keep-up processing. The output was declared to MetaCat and Rucio (via Rucio upload) at the conclusion of each processing job.

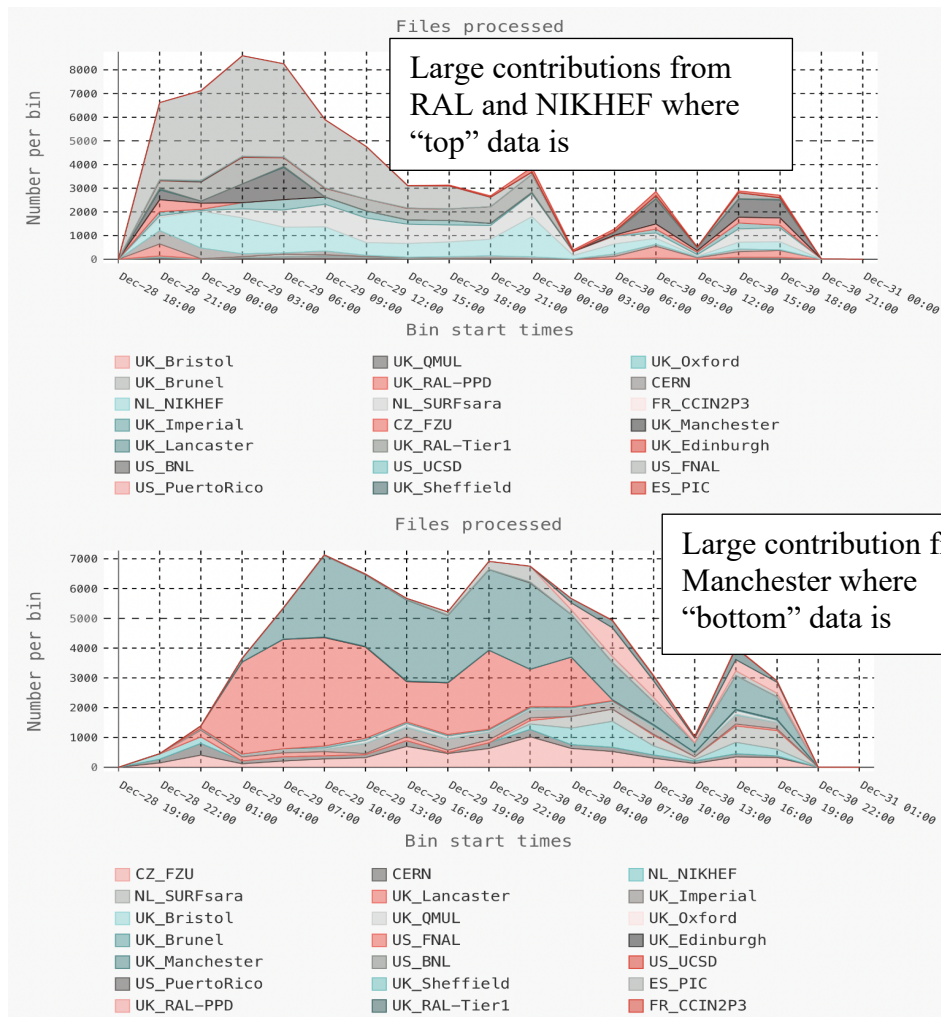


Fig. 4: Top: Site distribution of “VD coldbox top” processing. Bottom: Site distribution of “VD coldbox bottom” processing”.

3 Conclusions

Our tests have shown that the full pipeline from the experimental hall to CERN and Fermilab is functional and stable and can run at the rates expected in the forthcoming ProtoDUNE II runs, currently scheduled for 2024 (3.6 GB/s). The distributed processing phase can also operate at the scale needed for keep-up processing of that data. The software that was written and deployed for this test has now been deployed in production for use as our production data pipeline. This software has also been used to build pipelines for other DUNE-related test beams including the forthcoming near detector liquid argon cube 2x2 prototype, and has also been used to process the output of offline simulation. We expect to use a very similar software stack when data starts to flow from the DUNE far detectors at SURF.

This research is based upon work supported by the US Department of Energy, Office of Science, Office of High Energy Physics. W. Yuan receives funding from Iris project in UK. Thanks to the Scientific Storage Department at Fermilab for operating our Rucio server (Brandon White, Dennis Lee) and MetaCat servers as well as developing MetaCat and the ingest and declaration daemons (Igor Mandrichenko). Thanks to UK group (Andrew McNab, Chris Brew, Raja Nandakumar) for justIN workflow design, development, testing, and operation. Thanks to the CERN IT staff for FTS3, EOS, and CTA support. Thanks to Rucio developers. Thanks to DUNEDAQ developers particularly Kurt Biery for metadata generation script.

This document was prepared by the DUNE collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359. This work was supported by CNPq, FAPERJ, FAPEG and FAPESP, Brazil; CFI, IPP and NSERC, Canada; CERN; MŠMT, Czech Republic; ERDF, H2020-EU and MSCA, European Union; CNRS/IN2P3 and CEA, France; INFN, Italy; FCT, Portugal; NRF, South Korea; CAM, Fundación “La Caixa”, Junta de Andalucía-FEDER, MICINN, and Xunta de Galicia, Spain; SERI and SNSF, Switzerland; TÜBİTAK, Turkey; The Royal Society and UKRI/STFC, United Kingdom; DOE and NSF, United States of America.

The ProtoDUNE-SP and ProtoDUNE-DP detectors were constructed and operated on the CERN Neutrino Platform. We gratefully acknowledge the support of the CERN management, and the CERN EP, BE, TE, EN and IT Departments for NP04/ProtoDUNE-SP. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

References

1. **DUNE** Collaboration, A. A. Abud *et al.*, “Design, construction and operation of the ProtoDUNE-SP Liquid Argon TPC,” *JINST* **17** no. 01, (2022) P01005,
2. M. Baritsis and others, “Rucio: Scientific Data Management,” *Computing and Software for Big Science* **3** no. 11, (2019).

3. **DUNE** Collaboration, I. Mandrichenko, "MetaCat - metadata catalog for data management systems," *EPJ Web Conf.* **251** (2021) 02048.
4. **DUNE** Collaboration, A. A. Abud et al, "DUNE Offline Computing Conceptual Design Report" arXiv:2210.15665 [physics.data-an]
5. Chin Guok, David Robertson, Mary Thompson, Jason Lee, Brian Tierney and William Johnston, "Intra and Interdomain Circuit Provisioning Using the OSCARS Reservation System", Third International Conference on Broadband Communications Networks, and Systems, IEEE/ICST, October 1, 2006