Study of DUNE xrootd transactions

Heidi Schellman

July 6, 2021

**Introduction.** The SAM data access system was designed to provide granular information about file access as well as static metadata about file characteristics. The DUNE data management group is designing a SAM replacement which uses separate a MetaCat catalog and Data Dispatcher to decouple these two roles.

This note describes a study of file access using the xrootd streaming file access for DUNE compute jobs run on the grid and at Fermilab between Jan 1, 2021 and June 28, 2021.

Individual transfer action communicated to the sam system are logged in ElasticSearch and accessible through https queries such as this example.

https://fifemon-es.fnal.gov/sam-events/2021.04/\_search?q=experiment:dune%%20and%%20project\_id:%s&size=10000"%( projectID)

The ElasticSearch system is queried using code in DUNE github at data-mgmt-testing/XrootParser, the transactions are recorded in json files. This study used Utils3.py to gather the information and Analyzer.py to create summary histograms and tables.

ElasticSearch can only return 10,000 entries per query so queries are made by first selecting individual sam projects and then filtering on projectID. Sam “events” that are directly relevant for file transfers are then recorded in a json file for each project. Those events are then grouped by the name of the file transferred and sorted in time to determine the start and end of the transfer. Because of the 10,000 entry limit not all files from very long projects return information. In addition, merging file event records to form a single file processing history is quite CPU and memory intensive so SAM projects are selected by day in the data gathering phase and only merged in the analysis phase.

For this sample a total of 1,305,010 attempted file streams were studied. Of these 1,031,606 had successful final state “consumed”, 52,106 had final state “skipped” which indicate failure after file processing has started. The remaining transfers were of extremely short duration and had state “transferred” or “delivered” which, in streaming mode, means that SAM successfully communicated the file location to the process but there is no further record of a transaction. Prestaging jobs were not included in the sample.

The following two examples show the information recorded about file processing based on the ElasticSearch information (durations, locations) and general SAM information (application) about the project. The average streaming rate is calculated using the file size/duration as only the start and end times of the process are available. In the subsequent analysis, streams of duration < 100 seconds were omitted.

**First examples of single stream records**:

This is an example of a slow file read. User “drivera” ran a job at site “pp.rl.ac.uk” in the UK which took 49,013 seconds to process a 4.7 GB file from the PDSPProd4 production through application neutronana. The data were streamed from disk “fndca1.fnal.gov". The final state of the file was “consumed” which means that the process reported success in processing the file. The average transfer rate was 0.097 MB/sec which, after comparison with similar jobs running at Fermilab, indicates that the job was completely CPU bound.

{"disk": "fndca1.fnal.gov",

"user": "drivera",

"date": "2021-06-27",

"process\_id": 16059492,

"timestamp": "2021-06-27T23:57:58.047Z",

"duration": 49013.56700015068,

"file\_size": 4767154720,

"username": "drivera",

"application": "neutronana",

"version": "v09\_16\_01",

"final\_state": "consumed",

"site": "uk\_pp.rl.ac.uk",

"rate": 0.09726194218807507,

"project\_name": "drivera\_protodune-sp\_runset\_5842\_reco\_v09\_09\_01\_v0\_20210627181334",

"file\_name": "np04\_raw\_run005842\_0013\_dl3\_reco1\_14895769\_0\_20201208T225647Z.root",

"data\_tier": "full-reconstructed",

"node": "heplnc123.pp.rl.ac.uk",

"country": "uk",

"campaign": "PDSPProd4"}

**This is an example fast file read**. User “spurgeon” ran a job at site “usatlas.bnl.gov” in the US which took 226 seconds to process a 4.1 GB raw data file through application twocrtmatching. The data were streamed from disk “fndca1.fnal.gov". The final state of the file was “consumed” which means that the process reported success in processing the file. The average transfer rate was 18 MB/sec which indicates that the job was completely IO bound.

{

"disk": "fndca1.fnal.gov",

"user": "spurgeon",

"date": "2021-01-01",

"process\_id": 14503418,

"timestamp": "2021-01-01T00:46:54.253Z",

"duration": 225.9170000553131,

"file\_size": 4101329211,

"username": "spurgeon",

"application": "twocrtmatching",

"version": "v08\_40\_00",

"final\_state": "consumed",

"site": "us\_usatlas.bnl.gov",

"rate": 18.15414160951074,

"project\_name": "spurgeon\_stable\_xe\_small\_20201231183416",

"file\_name": "np04\_raw\_run011184\_0006\_dl6.root",

"data\_tier": "raw",

"node": "acas1091.usatlas.bnl.gov",

"country": "us",

"campaign": null

}

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

ROOT plots of Log10 of the streaming rate for applications “twocrtmatching” and “reco”. Internal streams are shown in red while FNAL->non-US streams are shown in blue. The average rate for “reco” differs by ~10% (see the table on the next page) from on-site to off-site while the average rate for “twocrtmatching” differs by a factor of 2.5. The average rates are summarized in Table 1.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

RStudio study of streaming rates for fast (twocrtmatching) and slow(reco) applications broken down by country. CERN and FNAL are considered to be countries. These plots were generated from a csv file produced by the ROOT analysis.



Table 1: This summarizes the average successful streaming rate from FNAL dCache for different applications measured in this data sample. Applications with no job runs outside the US are not shown. Transfers internal to FNAL are summarized on the left while transfers from FNAL to sites outside the US are summarized on the right. The ratio of external to internal is in the last column.

**Chart, histogram

Description automatically generated**

**Figure:** Streaming rate by data source location. Some raw data (suitable for reco) is stored on CERN EOS. Some processed samples (suitable for fast analysis) are stored at RAL in the UK.

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Figure:** Streaming rate by application for jobs run in the UK using data at RAL (top) and run in the UK using disk at FNAL (bottom).

It is possible to compare streaming rates within the UK with those within the FNAL site. Performance for I/O bound processing such as twocrtmatching improves when both the CPU and data source are in the UK.

**Conclusions –** ElasticSearch records, in combination with SAM project records can be mined to produce useful information about the characteristics of different DUNE applications. Within an application, the I/O rate can be reasonably well determined, both for local streaming and trans-Atlantic processing. For reconstruction, data location is reasonably unimportant as the jobs are CPU limited. For fast analysis, using the nearest data copy is important.

**Future work**. The information gathering needs to be automated so that we can collect these records overnight for further analysis as ElasticSearch information has a finite lifetime. As more recent data arrives at RAL we should be able to study local transfers within Europe. Our goal is to provide an analog to perfSonar which can monitor and describe streaming speeds between DUNE compute sites and inform job and data placement decisions.

Thanks to Sean Gilligan for the RStudio training, Steve Timm, Zac Lee, Lydia Brynmoor and Luke Penner for the ElasticSearch training and Ken Herner for identifying the mystery sites.