1 D2.2.1 Report on data transfer experiments using DESC DC2 data

LSST: UK Phase B WP2.2.

Project Acronym	LUSC-B
Project Title	UK Involvement in the Large Synoptic Survey Telescope
Document Number	LUSC-B-05

Submission date	21/JUL/20
Version	1.0
Status	Final
Author(s) inc. institutional affiliation	Mike Read (Institute for Astronomy, University of Edinburgh)
Reviewer(s)	Richard West (Warwick), Ken Smith (QUB)

Dissemination level	
Public	

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	06/Jul/20	Initial draft	Mike R
0.2	21/JUL/20	Minor updates	George B
0.3	05/AUG/20	Change of title and minor updates following reviewers comments	Mike Read
0.4	13/AUG/20	Copyright statement date updated, reviewer names added	T. Sloan
1.0	03/SEP/20	Status changed to Final following approval at 27/AUG/20 Exec Group	T. Sloan

Table of Contents

2 Introduction

To test and gain experience with transferring significant volumes of data from a Rubin Observatory processing centre to the LSST UK DAC, as will be required to acquire each new LSST data release, we looked at transfers of DESC DC2 (Dark Energy Science Collaboration, Data Challenge 2) data from IN2P3 (French National Institute of Nuclear and Particle Physics). IN2P3 is the nearer of the two Rubin Observatory Processing Centres (the other being in the United States, provisionally at NCSA), so is a realistic source for these experiments. DESC DC2 data has been used as it is being produced by the same pipeline stack as will ultimately be used in LSST operations. Also, we initially aimed to ingest the DC2 parquet format catalogue data into Qserv (the LSST database system), but on further investigation this was determined to not be timely, and so not a sensible use of available resources.

3 The data

Colleagues at IN2P3 made available 20 Terabytes (TB) of DESC DC2 data from run DESC simulation run 1.2P. They provided a list of 1,244,118 file URLs and opened up their firewall to one of the LSST:UK test-bed machines (specifically, lsstuk4.roe.ac.uk). The bulk of data was FITS images and catalogues, with individual file sizes ranged from small ASCII configuration files to around 200 MB for the larger image files.

4 The transfer tests

We anticipated that the fastest aggregate transfer speeds would be achieved by performing multiple downloads in parallel. To this end, GNU parallel was installed, as this is readily configurable to accept a list of URLs as input and to alter the number of parallel jobs to run.

We did not have 20TB of available space attached to lsstuk4, so the first tests transferred 12 × 1TB groups of around 60,000 files each. Prior to running the transfer tests, we retrieved the byte file size of all the URLs so that the size of each group could be limited to 1TB and also to use in verifying the transfers.

The twelve chunks were then run sequentially by a script that checked the transfers (file sizes) after each group and deleted the files before the next group was run. The original directory structure was replicated as the files were copied over.

The main command in the script was:

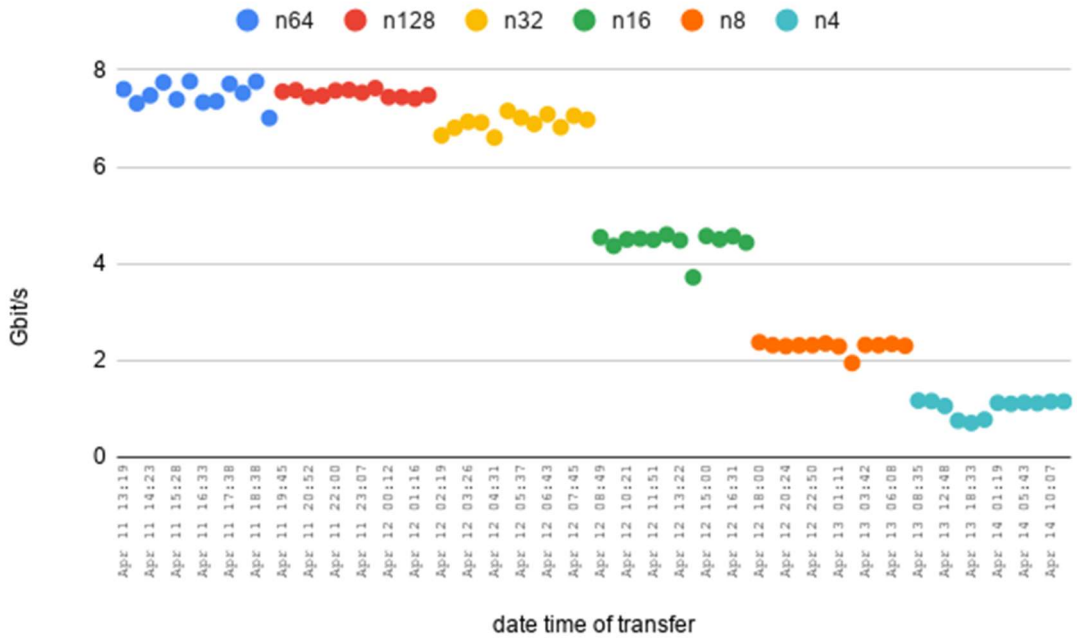
```
cat $urlFile | parallel -P $npProc wget \

  --no-check-certificate -x -q
```

—where $noProc is the number of processes running in parallel, --no-check-certificate is required as no valid certificate is available, -x creates the directory structure and -q is quiet mode.

The aggregate transfer rate for different numbers of parallel transfers is shown in the table below, with measurements for individual group runs depicted in the graph.

No of parallel processes	Average transfer rate (Gbit/s)	Median transfer rate (Gbit/s)
128	7.5	7.5
64	7.5	7.5
32	6.9	6.9
16	4.4	4.5
8	2.3	2.3
4	1.0	1.1



The highest transfer speeds achieved, **7.5 Gbit/s**, are close to saturating the bandwidth of the local, 10 Gbit/s, network switch.

After running these tests, it was noted that the files were being hosted by four servers (or virtual servers) at IN2P3:

ccnetlsst01.in2p3.fr:65010, ccnetlsst02.in2p3.fr:65010, ccnetlsst03.in2p3.fr:65010, ccnetlsst04.in2p3.fr:65010

and that, in the first test, the groups of URLs used were mainly all each from one host. A further test was carried out using chunks of 64,000 URLs split evenly across the four hosts and running 64 parallel processes. No improvement was recorded and a mean/median of 7.5 GBit/s was again achieved.

One further test was carried out, including a second test-bed machine (lsstuk1.roe.ac.uk), which IN2P3 allowed through the firewall. Transfers were carried out on lsstuk1 and lsstuk4 in parallel, with each server running 64 wget processes. The full dataset was transferred, half going to each server with multiple processes using each IN2P3 server. The 20TB were transferred in 5h 17m, equivalent to **8.5 GBit/s**, which is a modest increase on the transfer rate for lsstuk4 only.

In comparison the WFAU's VDFS transfers from CASU (University of Cambridge) to Edinburgh using a dark-fibre run are typically 40 times slower at around 15–30 MB/s.

5 Further tests

Whilst 8.5 GBit/s is a respectable transfer rate it was hoped that this could be improved upon by carrying out the tests to an IRIS system, hosted in Edinburgh University's Advanced Computing Facility (ACF). This machine would be on a faster network connection to JANET (30—40 GBit/s). Unfortunately, initial measurements from IN2P3 to the ACF were much slower than those already achieved (around **2.5 Gbit/s**). Investigations were delayed by a power outage at the ACF, which was resolved more slowly than usual due to Covid-19-related staffing restrictions.

Investigations remain on-going. Benchmark measurements on the IRIS system using the Ookla Speedtest service (www.speedtest.net) indicate a potential download speed in excess of **6Gbit/s**, though this has not been observed in practice. Attention is focused on potential overheads of virtualisation on the IRIS system (the end point for transfers is a virtual machine hosted by OpenStack). Further, PerfSonar data from the neighbouring GridPP system in the ACF indicate potential transfer speeds of around 8Gbits/sec:

http://ccperfsonar1.in2p3.fr/perfsonar-graphs/?source=193.48.99.77&dest=129.215.213.69#start=1592404556&end=1594996556&summaryWindow=86400&timeframe=2592000

We discussed implementing other transfer protocols (gridFTP, bbcp) but, at the time of writing, IN2P3 only support standard HTTP/HTTPS protocols. Furthermore, IN2P3 have measured a sustained transfer rate of 32 Gbit/s over 32 hours between themselves and NERSC (National Energy Research Scientific Computing Center) in the USA using HTTP.

We also transferred a small amount of DESC DC2 catalog files in parquet format. We had planned to ingest these into the Qserv database system running on the LSST:UK Testbed. However, on looking into the steps required we discovered that it is necessary to first convert the parquet files to CSV format. As our earlier Qserv ingest experiments using UKIDSS data [LUSC-A-07] had already employed CSV files and as the LSST DM team was actively developing the software for handling and parsing the parquet files, it was decided that the tests of the latest Qserv version would be best carried out using the UKIDSS data rather than expend resources translating DC2 data from one format to another.

6 Conclusion

A sustained aggregate transfer rate of 8.5 Gbits/s has been achieved between IN2P3 and LSST test-bed machines at Edinburgh, using standard file-transfer protocols and techniques. At this rate, we would expect to be able to transfer around 85TB of data over a 24-hour period. With an estimated size of 1.6 PB, the LSST DR1 catalogues would take around 20 days to download at this rate. Images in DR1 are estimated to total 2.9 PB, requiring some 37 days to transfer. The LSST alert stream will also require fast network transfer speeds and this will be addressed within LSST:UK Phase B Work Package 2.3.

Investigations into the performance of data transfer into the IRIS kit at the ACF will continue, with an expectation that data rates can be increased by a factor of 3 or 4 by better exploiting the theoretical 40 Gbit/s available bandwidth. Such performance would be in line with bandwidth measured by IN2P3 between them and NERSC. This report will be updated when those investigations have concluded.

The Science Requirements Document (SRD) should be updated to include more definitive estimates of the required data transfer and ingest rates as they crystallize. The risk management plan should also note the consequences of not achieving the desired rates.