LARGE SYNOPTIC SURVEY TELESCOPE

Large Synoptic Survey Telescope (LSST)

Potential proofs of concept for Google Cloud

William O'Mullane and John Swinbank and K.T. Lim and Margaret Gelemann and Xiuqin Wu and Fritz Mueller

DMTN-078

Latest Revision: 2018-04-24

DRAFT

1 Introduction

In this brief note we wish to discuss some potential routes to collaboration with Google. We will revisit some of the potential ways forward to the cloud from DMTN-072 and try to quantify what a Proof of Concept might look like. There is also scope for technical collaboration on infrastructure.

If we go ahead with one specific POC we should define more clearly the goal and duration of it and how that might feed into future operations.

Margaret/Michelle - should we have - expected results/metrics could also be the next round and can you look at th technical collaboration section.

2 Technical collaboration

We are already heavy users of Kubernetes both internally and via Google cloud. One assumes setting up of private clouds is something enterprises may want to do and we are pushing the envelope on that. We have specific technical issues running on our own K8S enabled hardware such as start up times[1] and security of the GPFS filesystem through K8S. Obviously help would be appreciated in this area and we may provide a large test bed outside Google for feedback on this.

3 Cloud proof of concept

Much of LSST Data Management (DM) is Kubernetes deployable - all items mentioned here are. This provides us with a lot of flexibility to port our system across service offerings, and would enable us to easily adopt a hybrid cloud plus on-premises infrastructure.

Moving to a cloud-based infrastructure could potentially save on personnel, as no hands-on hardware maintenance would be required. Although this is equivalent to a relatively small fraction of the construction budget, it would represent a substantial sum dedicated to non-core-business during operations.

We can probably not move wholescale to the cloud: we are committed to providing the Chilean DAC in Chile, and some physical hardware must remain on the mountain and in the

[1] Solved for now I believe.

Commissioning Cluster. However, there are potentially a number of opportunities to migrate a subset of DM services to the cloud if we could see a sensible way forward.

3.1 Qserv - LSST in house database from SLAC

The Qserv database system [LDM-135] has not yet been tested in a cloud based environment. However it is now deployable with Kubernetes, and no longer requires special hardware: physically attached storage is needed, but this is available on cloud offerings.

This is one major component of the Science Platform (Section 3.2). Proper testing would be needed to understand how Qserv performs in the cloud environment. As a proof of concept this would be interesting: well-bounded and well-understood in terms of performance. Without the full science platform it may not be very useful in the long run beyond that.

3.1.1 Potential needs

Fritz To set up Qserv we would need at least 40 large nodes with physically attached storage in to order of 2 terabytes per node. To run a set of convincing tests we would need that up for order two months.

This would be a demonstration only - the final catalogs (2032) will be order 15PB and the number of nodes and attached storage eventually have to scale to that size.

3.2 Cloud based Science Platform

The Science Platform LSE-319 is intrinsically a cloud-oriented solution to the data transfer problem: it envisions user code being collocated with the data on which it is running.

The prototype DAC (PDAC) is deployed in the NCSA data facility with potential access to 2 PB of storage. There are 3 aspects: a visualization Portal, the Jupyter Notebooks, and Web services.

The web services are an interface to the images on disk as well as the Qserv system (Section 3.1).

A key benefit of a cloud-based Science Platform would be scalability: when user demands exceed the 10% of the compute budget dedicated to serving them more capacity would at least be available even if it had to be purchased on demand. There is no analogue to this in terms of on-premises infrastructure as cloud bursting from our internal cloud infrastructure

to a commercial provider would require transferring potentially large amounts of data.

3.2.1 Aside:Public Data Releases

LSST data becomes public after 2 years. However there is no budget allocated to serve this public data - one could envision a public version of the Science Platform serving the old data as something potentially interesting for some foundations/companies e.g. to enable science in underdeveloped countries.

3.2.2 Potential needs

All the science platform components are deployable with Kubernetes. The Qserv database component is a fixed size resource as discussed in Section 3.1.1. In addition one or preferably two servers should be provisioned for the web services.

Alongside that one needs to have the JupyterHub environment [2]; depending on the assumed load, this is relatively modest as it requires only ~ 2 servers to set up, and it is recommended to have 2 CPUs per simultaneous user. For a proof of concept let's assume we would go with 20 simultaneous users to 40 CPUs or 10 nodes depending on the type of node. Each user should also have around 4GB of RAM. Theoretically we would also have a batch system and more near the data compute resources but perhaps for a proof of concept this may be treated as a desirable.

Xiuqin Firefly also requires at least a pair of server - theses should be 32 cores with 128 GB memory. In addition these should have a shared disk volume order 500GBi, preferably SSD.

Finally there is a filesystem to store the image data. Our current code assumes a Posix filesystem, but we have made some modifications towards supporting a back-end object store. Additional investment in this direction is unlikely to happen before Fall 2018. LSST will produce over its lifetime around 60 PB of raw image data, the final data volume including the processed images is estimated to be around 0.5 Exabytes. For the POC 1 PB would be sufficient to see some performance and management of a large disk volume.

For the proof of concept we could leave out the Prompt Database (this is a conventional e.g. Oracle database).

[2] see https://github.com/lsst-sqre/jupyterlabdemo

3.3 Cloud based prompt processing

Prompt processing [LDM-151] is a questionable part of processing on the cloud - hence also a very interesting use case, because the answer is not obvious. It is also the part of the system which runs in a *bursty manner* e.g. it runs at night as images are taken and in a more limited form during only part of the day. To meet the one minute goal for processing money has been put into building a rapid transfer of files to NCSA. We would have to assess if we could transfer files into the cloud fast enough to make the alert processing work to schedule and its requirements to produce alerts in one minute.. The fast networks already deployed for LSST should be applicable, but further analysis would be required. We have a preliminary Alert Distribution framework [3] - which was already run on AWS so this is easy to do. We have not yet deployed a large-scale alert Production which is specified to run on 2 clusters of one CPU per CCD on the camera i.e. 189 CPUs in each cluster. This would be an interesting real-time test i.e. this system is to process the images from the telescope in 1 minute.

3.3.1 Potential needs

John ? Assume we did a single chain instead of 2 that would be 189 CPUs. One night's data is 20TB; with processing and staging this system would need about 60TB of disk - for a POC though a couple of TB should be sufficient to run a few images though the system.

The Alert distribution would need 8 nodes with 4CPUs/30GB and 200GB storage.

4 Conclusion

A number of potential POCs are discussed above with approximates sizes/needs. We should pick one or more to develop further.

A References

[1] **[LDM-135]**, Becla, J., Wang, D., Monkewitz, S., et al., 2013, *Database Design*, LDM-135, URL `https://ls.st/LDM-135`

[3] https://dmtn-028.lsst.io/

[2] **[LSE-319]**, Jurić, M., Ciardi, D., Dubois-Felsmann, G., 2017, *LSST Science Platform Vision Document*, LSE-319, URL https://ls.st/LSE-319

[3] **[DMTN-072]**, O'Mullane, W., Swinbank, J., 2018, *Cloud technical assesment*, DMTN-072, URL https://dmtn-072.lsst.io,
LSST Data Management Technical Note

[4] **[LDM-151]**, Swinbank, J.D., et al., 2017, *Data Management Science Pipelines Design*, LDM-151, URL https://ls.st/LDM-151

B Acronyms

The following is a complete list of acronyms used in this document.

Acronym	Description
CCD	Charge-Coupled Device
CPU	Central Processing Unit
DAC	Data Access Center
DM	Data Management
DMLT	DM Leadership Team
DMTN	DM Technical Note
GB	GigaByte
GB	Giant Branch (star)
GPFS	General Parallel File System
K	Kelvin; SI unit of temperature
LDM	Light Data Management
LSE	LSST Systems Engineering (Document Handle)
LSST	Large Synoptic Survey Telescope
NCSA	National Center for Supercomputing Applications
PB	PetaByte
POC	Proof Of Concept
RAM	Random Access Memory
SLAC	Stanford Linear Accelerator Center
SSD	Solid-State Disk
TB	TeraByte
TN	Technical Note

s	second; SI unit of time