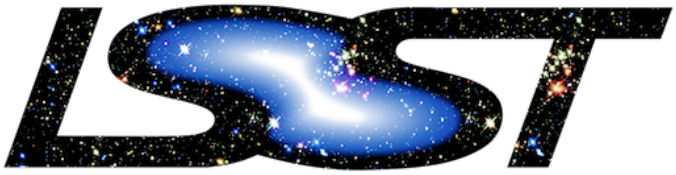

LARGE SYNOPTIC SURVEY TELESCOPE

Large Synoptic Survey Telescope (LSST)

Cloud technical assesment

William O'Mullane and John Swinbank

DMTN-072

Latest Revision: 2018-02-09

DRAFT

1 Introduction

The complexity of LSST's sizing model, and the intricacies of planning spreadsheets developed by NCSA, make it hard to understand and assess the overall compute budget for Data Management, especially vis-a-vis potential alternatives.

To address this issue, I develop a simplified hardware cost model[1], and apply it to both physical hardware procurement (the current baseline) and a representative "cloud-based" computing service (Amazon AWS). This demonstrates that LSST's requirements would theoretically be addressed by commodity cloud computing infrastructure if the I/O costs could be mitigated.

2 A simplified computing cost model

2.1 Preamble

One of the most notoriously difficult things to assess in any astronomy project is how much compute power is needed and how much it will cost to get it.

An alternative model I have used in other projects it to simply estimate the number of floating point operations (FLOPs) and storage needed at a point in time and use those figures to estimate an instantaneous cost. The model can then be scaled by assuming prices continue to evolve as they have done previously.

The FLOP itself is a notorious unit, so it is worth clarifying its definition. In this document we use FLOP, plural FLOPs, to mean FLoating point OPeration: a unit describing the total number of calculations required to complete some calculation. FLOP/s, by contrast, refers to floating point operations *per second*, a measure of the instantaneous compute power needed or available from some particular processor. For an example of the former use, refer to the first row of Table 1; for the latter, line 4 of the same table.

[1] A simplified model has been discussed for some years.

2.2 Total compute requirements

Table 1: Various inputs for deriving costs

Year	2017	2018	2019	2020	2021	2022
FLOPs Needed Total (no Alerts)	9.48261E+19	1.00E+19	1.00E+19	9.48261E+19	1.00E+19	4.74131E+20
Time to Process days	252.0	365.0	365.0	252.0	365.0	252.0
Time to Process seconds	21772800.0	31536000.0	31536000.0	21772800.0	31536000.0	21772800.0
Instantaneous GFLOP/ s	4355.255691	3.17E+02	3.17E+02	4355.255691	3.17E+02	21776.27846
Instantaneous GFLOP/ s (inc Alerts)	4355.255691	3.17E+02	3.17E+02	30025.25569	2.60E+04	21776.27846
Disk Space TB	1000	1000	1000	10000	20000	30000
I/ O for year TB	10	100	3000	30000	60000	90000
Base numbers	Ecyc	FLOP	GFLOP			
LDM-138 DR1,2 Data Rel sheet row 1	155.17	4.26718E+20	426717500000			
LDM-138 DR3 Data Rel sheet row 2	348.76	9.5909E+20	959090000000			
LDM-138 Alert Instananeous	0.00023434	25670000000000	25670			
Alert Total, assuming 275k visits/ year	64.4435	1.7722E+20	177219625000			
Total Yr1 (inc DAC)		**4.74131E+20**	**474130555556**			
	Optimistic	Pessimistic				
Moore Factor Proc	0.7	0.9				
Kryder Factor Disk	0.8	0.9				

The DM Sizing Model (LDM-138, LDM-144) contains estimates of the scale of computing which LSST must undertake; important values summarized in Table 1 for convenience. In particular, note that the first year of operations is expected to require around 4.7×10^{20} FLOPs for data release processing, with an additional sustained $25,670$ GFLOP/s for alert processing. During the first data release, this (naïvely) averages to $15,034$ GFLOP/s continuously (ie, assuming that all compute systems are kept busy continuously).

During the Construction period, we will deploy approximately 20% of the Data Release 1 capacity; we reflect this in Table 1 for 2017, and again for an assumed hardware refresh in 2021. This three year refresh cadence is typical across the industry, and has the convenient effect of being directly comparable with Amazon's three year pricing on their cloud compute offering.

2.3 Physical hardware estimates

Table 2: Estimates for physical hardware; large range of possible prices

Year	2017	2018	2019	2020	2021	2022	2023
Optimisitic price per GFLOP	$4.40	$3.08	$2.16	$1.51	$1.06	$0.74	$0.52
Likely price per GFLOP	$11.76	$8.23	$5.76	$4.03	$2.82	$1.98	$1.38
Pessimistic price per GFLOP	$53.00	$47.70	$42.93	$38.63	$34.77	$31.29	$28.16
Optm. total compute	$19,184.21	$977.74	$684.42	$45,363.97	$27,484.02	$16,121.45	$14,079.18
Pess. total compute	$230,811.45	$15,124.45	$13,612.01	$1,159,999.87	$903,590.21	$681,459.27	$765,169.40
Optm. price per TB	$42.66	$29.86	$20.90	$14.63	$10.24	$7.17	$5.02
Likely Price TB	$309.84	$216.89	$151.82	$106.28	$74.39	$52.08	$36.45
Pess. price per TB	$3,015.54	$2,110.88	$1,477.61	$1,034.33	$724.03	$506.82	$354.78
Optm. total disk	$42,660.00	$29,862.00	$20,903.40	$146,323.80	$204,853.32	$215,095.99	$150,567.19
Likely total disk	$309,841.42	$216,888.99	$151,822.29	$1,062,756.06	$1,487,858.49	$1,562,251.41	$1,093,575.99
Pess. total disk	$3,015,540.00	$2,110,878.00	$1,477,614.60	$10,343,302.20	$14,480,623.08	$15,204,654.23	$10,643,257.96
Optm. total cost	$61,844.21	$30,839.74	$21,587.82	$191,687.77	$232,337.34	$231,217.44	$164,646.37
Likely total cost	$540,652.87	$232,013.44	$165,434.30	$2,222,755.93	$2,391,448.70	$2,243,710.69	$1,858,745.39
Pess. total cost	$3,246,351.45	$2,126,002.45	$1,491,226.61	$11,503,302.07	$15,384,213.29	$15,886,113.51	$11,408,427.36
Cost Estimate (opt+4*likely+pess)/ 6	$911,801.19	$514,149.33	$362,425.27	$3,431,002.26	$4,197,057.57	$4,182,028.95	$3,168,009.22
Total construction (to 2022)		**$13,598,464.57**					

We convert the compute requirements developed in the previous section into dollar values by combining:

- LINPACK and Flops[2] as benchmarks of compute hardware;
- Costings based on experience from Gaia and other projects.

the input figures are provided in Table 3.

Table 3: Various inputs for physical hardware mainly from Josh Hoblit running flops code

Xeon	Cost	Price per GFLOP
2 CPUs, 16 core full install and power	$14,736.00	
Optm. GFLOP/ s	3,345.41	$4.40
Likely		$11.76
Pess. GFLOP/ s (*inefficiency)	1,668.35	$53.00
Storage	Cost/ TB	
Optm. (cheap NAS RAID0)	$42.66	
Likely (Spinning)	$309.84	
Pess. (NVMe)	$3,015.54	

Based on these figures, we expect a wide range of possible costings: this reflects the variety of hardware available (from cheap commodity desktop PCs to Cray supercomputers). We combine these values using formula based on PERT: given multiple estimates, we assume the most plausible true value is $(\text{optimistic} + 4 \times \text{likely} + \text{pessimistic})/6$. Throughout, we use twice the optimistic price as the likely value for physical hardware, and 1.5 times for cloud systems.

[2] https://github.com/Mysticial/Flops

We follow a similar procedure to estimate storage costs. A range of possible prices exist from ultra-cheap build-your-own systems around $40/TB[3] to full solutions like NetApp ($300 - $1K/TB) or Non-Volatile Memory($3K/TB) : these give good optimistic, likely and pessimistic prices to work with. For 2017 we need to purchase about 1PB (1000 TB) so the optimistic and pessimistic prices are approximately $40K and $3M allowing.

The result of these estimates is shown in Table 2. Note that the headline cost is around $900k in 2017; we present estimates for several other years throughout Construction and Operations, and provide a summary of the total cost through the construction period. Year 2023 is DR3 and the first non construction year.

An optimistic and pessimistic price scaling is also applied these rates are shown in the end of Table 1. In fact the price of machines usually does not fall but for the same price a more power full machine is usually available for our purposes the distinction does not matter. Likewise there is a lot of licensing, networking interconnect and racks which average over a number of machines if we buy physical hardware - this is simply bundled in the unit GFLOP price.[4]

[3] Estimate from Szalay, JHU, private communication.
[4] I have not gone back to work this number out for 2017 purchases.

2.4 Cloud computing estimates

Table 4: Estimates for Amazon; dominated by I/O

Year	2017	2018	2019	2020	2021	2022	2023
Optimistic price per GFLOP	$8.86	$6.20	$4.34	$3.04	$2.13	$1.49	$1.04
Likely price per GLOP	$10.02	$7.01	$4.91	$3.44	$2.41	$1.68	$1.18
Pessimistic price per GFLOP	$322.20	$289.98	$260.98	$234.88	$211.40	$190.26	$171.23
Optm. total compute	$38,583.23	$1,966.42	$1,376.49	$91,235.89	$55,275.77	$32,423.41	$28,316.00
Likely total compute	$43,630.04	$2,223.63	$1,556.54	$103,169.85	$62,506.03	$36,664.51	$32,019.83
Pess. total compute	$1,403,263.38	$91,952.05	$82,756.85	$7,052,446.15	$5,493,553.48	$4,143,064.98	$4,651,997.07
Optm. total I/ O	$250.00	$2,500.00	$75,000.00	$750,000.00	$1,500,000.00	$2,250,000.00	$2,250,000.00
Likely total I/ O	$500.00	$5,000.00	$150,000.00	$1,500,000.00	$3,000,000.00	$4,500,000.00	$4,500,000.00
Pess. total I/ O	$900.00	$9,000.00	$270,000.00	$2,700,000.00	$5,400,000.00	$8,100,000.00	$8,100,000.00
I/ O cost estimate (opt+4*likely+pess/ 6)	$441.67	$4,416.67	$132,500.00	$1,325,000.00	$2,650,000.00	$3,975,000.00	$3,975,000.00
Optm. Price TB	$84.00	$58.80	$41.16	$28.81	$20.17	$14.12	$9.88
Likely Price TB	$252.00	$176.40	$123.48	$86.44	$60.51	$42.35	$29.65
Pess. Price TB	$1,440.00	$1,008.00	$705.60	$493.92	$345.74	$242.02	$169.41
Optm. total disk	$84,000.00	$117,600.00	$123,480.00	$288,120.00	$403,368.00	$423,536.40	$296,475.48
Likely total disk	$252,000.00	$176,400.00	$123,480.00	$864,360.00	$1,210,104.00	$1,270,609.20	$889,426.44
Pess. total disk	$1,440,000.00	$2,016,000.00	$2,116,800.00	$4,939,200.00	$6,914,880.00	$7,260,624.00	$5,082,436.80
Optm. total cost	$122,833.23	$122,066.42	$199,856.49	$1,129,355.89	$1,958,643.77	$2,705,959.81	$2,574,791.48
Likely total cost	$296,130.04	$183,623.63	$275,036.54	$2,467,529.85	$4,272,610.03	$5,807,273.71	$5,421,446.27
Pess. total cost	$2,844,163.38	$2,116,952.05	$2,469,556.85	$14,691,646.15	$17,808,433.48	$19,503,688.98	$17,834,433.87
Cost Estimate (opt+4*likely+pess)/ 6	$691,919.46	$495,585.50	$628,259.92	$4,281,853.58	$6,142,919.56	$7,573,123.94	$7,015,835.07
Cost Estimate without IO	$691,477.80	$491,168.84	$495,759.92	$2,956,853.58	$3,492,919.56	$3,598,123.94	$3,040,835.07
Total construction (to 2022)		**$19,813,661.96**					
Total construction excl. I/ O		**$11,726,303.63**					

To properly size Amazon we should run some pipeline code and benchmark it. Here a paper [3] was used which provided LINPACK numbers for a specific type of Amazon machine. The peak and average FLOPs are used to make the optimistic and pessimistic prices per GFLOP. We use the three year leasing price, which makes these estimates comparable to the lifetime of directly purchased hardware. DM code is unlikely be reach the same level of efficiency as the LINPACK benchmark, so we build in a further inefficiency factor to produce the pessimistic price. In addition since we are using LINPACK and our code may not be efficient on these machines a further inefficiency factor is used to arrive at the pessimistic price. The results of these considerations are presented in Table 4.

Note that a major component of the expense when running on cloud systems is for egress bandwidth: the cost of transferring data out of the cloud system itself. For convenience, Table 4 presents total costs both inclusive and exclusive of this I/O cost This issue is discussed in detail in Section 3.2

Table 5: Various inputs used to cost Amazon

Amazon	3 yr price	1 year price	Price GFLOP
Optimistic price GFLOP (c5.18xlarge)	$29,637.00	$9,879.00	$8.86
Likely price GFLOP (c4.8xlarge)	$16,329.00	$5,443.00	$10.02
Pessemistic price GFLOP		$9,487.00	$322.20
GFLOP/ s (see arxiv paper c4.8xlarge)	530		
GFLOP/ s peak (see arxiv paper c4.8xlarge)	1630		

GFLOP/ s c5.18xlarge (Hoblitt)	3345.41		
Inefficiency factor	6		
Machine lifetime	3		
Cost per TB for machine lifetime	monthly/ GB	Yearl/ TB	
Optm. (cheaper than S3 Glacier?)	$0.007	$84.00	
Likely (AWS S3)	$0.021	$252.00	
Pess. (GPFS on AWS some SSD)	$0.120	$1,440.00	
I/ O pricing	GB	TB	
Optm. (Min listed price)	$0.05	$50.00	
Pess. (S3 out max price)	$0.09	$90.00	

Prices assumed for the Amazon systems are shown in Table 5. These are drawn from figures published on the Amazon website[5].

3 Potential ways forward

It is clear the cloud model is now well established and here to stay. It is incumbent upon us to consider whether and how we can best take advantage of it for LSST. In this section, we discuss a number of potential migration scenarios, and discuss further avenues for investigation.

3.1 Migration to the cloud

Both Amazon AWS and Microsoft Azure now support the Kubernetes deployment and management system used by DM. 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-physical infrastructure.

Moving to a cloud-based infrastructure would also likely 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.

The numbers presented in Table 4 assume a wholesale migration of all DM functionality to the cloud. Unfortunately, this is impractical: for example, we are committed to providing the Chilean DAC in Chile[6], and some physical hardware must remain on the mountain and in the Commissioning Cluster. However, there are potentially a number of opportunities to migrate a subset of DM services to the cloud.

[5] https://aws.amazon.com/ec2/pricing/reserved-instances/pricing/, https://aws.amazon.com/ec2/pricing/on-demand/

[6] Though one could discuss the new Amazon AWS offering in Chile with the Chileans.

3.1.1 Developer Services

DM already uses cloud-based systems for continuous integration (CI) testing[7] and for standing up JupyterHub instances for workshops and demos. These are relatively easy to set up[8] and have proven reliable.

A move in this direction would be popular with developers: it enables us to provide them with greater flexibility over the systems in use, and the ability to self-manage their own development environment when applicable. This would address numerous points of contention with our current infrastructure.

We do note, though, that it is important to maintain some developer infrastructure in large facilities to make sure we understand our deployment environment and to have access to large-scale extra-cloud data storage.

3.1.2 Cloud based Science Platform

The Science Platform 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.

To date, the prototype DAC (PDAC) has been somewhat successful in the NCSA data facility. However, delays in procurement and other work, have made it difficult to capitalize on this success. A cloud based system would enable us to sidestep many of these concerns.

It is worth noting that the EPO subsystem immediately plans to deploy their systems to cloud infrastructure. As such, they would act as a "proving ground" for the bigger DM project. However, note that DM's much larger data volumes make it more at risk of data storage and ultimately I/O problems as discussed in Section 3.2.

The Qserv database system has not yet been tested in a could based environment. However, we note that it is now deployable with Kubernetes, and no longer requires special hardware: physically attached storage is needed, but this is available on the Google and Amazon cloud offerings. Proper testing would be needed to understand how Qserv performs in this environment before committing to it.

A key benefit of a cloud-based Science Platform would be scalability: when user demands

[7] Due to the recent (Jan 2018) problems with Nebula in NCSA the test data was moved to Amazon.
[8] `https://github.com/lsst-sqre/jupyterlabdemo`

exceed the 10% of the compute budget dedicated to serving them, they would seamlessly be able to purchase more capacity from the cloud provider. There is no analogue to this in terms of physical infrastructure.

3.1.3 Consider cloud based data release processing

Use of cloud based systems for producing annual data releases would avoid I/O issues: the result would only need to be collected once. Furthermore, if combined with a cloud-based Science Platform and backup service (e.g. Amazon Glacier), almost all I/O could be avoided altogether.

3.1.4 Consider cloud based prompt processing

Prompt processing is the the only questionable part of processing on the cloud. To meet the one minute goal for processing money has been put in 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. The fast networks already deployed for LSST should be applicable, but further analysis is required.

3.2 I/O: the cloud's Achilles heel

The main expense with the cloud is neither the storage nor compute, but rather than of transferring data out ("egress"): see the summary lines in Table 4. If most science is done in the cloud this is not a problem (§§3.1.2 & 3.1.3). Alternatively, it may be possible to develop a partnership with an organization willing to give us a preferential rate on these services. If we went down this route we would probably want to stage the output catalogs in some place after bulk transfer i.e. only do large transfer out of the cloud once. The providers say "call us" to discuss large transfers: we should, at least, start that conversation.

We also note that total bandwidth may be an issue, but there are ESNET endpoints to the cloud at 10Gbps[9].

[9] http://fasterdata.es.net/performance-testing/DTNs/

4 Conclusion

The model and scenarios presented in this document demonstrate that further consideration of a cloud-based infrastructure is merited. Future investigations should consider:

- Profiling the performance of the DM stack on cloud infrastructure;
- Discussions with contacts at Google and Amazon;
- Migrating some developer services and environments to the cloud.

The last point is particularly important: if we can develop and test well on the cloud it would tell us a lot about the capabilities and limitations of the main vendors.

A References

[1] **[LDM-144]**, Freemon, M., Pietrowicz, S., Alt, J., 2016, *Site Specific Infrastructure Estimation Model*, LDM-144, URL `https://ls.st/LDM-144`

[2] **[LDM-138]**, Kantor, J., Axelrod, T., Lim, K.T., 2013, *Data Management Compute Sizing Model*, LDM-138, URL `https://ls.st/LDM-138`

[3] Mohammadi, M., Bazhirov, T., 2017, ArXiv e-prints (`arXiv:1702.02968`), ADS Link

B Acronyms

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

Acronym	Description
AURA	Association of Universities for Research in Astronomy
AWS	Amazon Web Services
CI	Continuous Integration
DAC	Data Access Center

DM	Data Management
DMLT	DM Leadership Team
DOE	Department of Energy
EPO	Education and Public Outreach
ESNET	Energy Sciences Network
EVMS	Earned Value Management System
FLOP	FLoating-point OPeration
FTE	Full-Time Equivalent
GFLOP	Giga FLOatingpoint OPeration
GPFS	General Parallel File System
JHU	Johns Hopkins University
LINPACK	a software library for performing numerical linear algebra on digital computers
LSST	Large Synoptic Survey Telescope
NAS	Network Attached Storage
NCSA	National Center for Supercomputing Applications
NSF	National Science Foundation
PC	Personal Computer
PDAC	Prototype Data Access Center
PERT	Project Evaluation and Review Techniques
PO	Purchase Order
SSD	Solid-State Disk
TB	TeraByte