DM sizing model and purchase plan for the remainder of construction.

Michelle Butler, Kian Tat Lim, William O'Mullane

2019-12-16

1 Introduction

2 Proposed budget

Based on the needs in Table 6 and the costs in Table 7 and Table 8 the following budgets can be considered.

A high level bottom line is given in Table 1. Since Xeon is more expensive that is the number used for the budget calculation, should we get Rome and it really performs we may save a little. The remainder of the document is all the details that went into that.

Table 1: This table pulls together all the information in a high level summary - in this table Xeon pricing is used since that is the more expensive but better known option. Price factors, defined in Table 6 are applied post 2020.

Year	2020	2021	2022	2023
Compute (2019 pricing)	$690,000	$0	$1,286,151	$2,826,893
Applying price factor (CPU)		$0	$1,028,921	$1,978,825
IN2P3 (50% of compute)				-$989,412
Qserv (2019 pricing)			$560,000	$3,791,195
Qserv (applying factor)			$476,000	$2,938,176
Storage (2019 pricing)	$333,952	$333,863	$1,813,698	$11,316,420
Applying price factor (Storage)	$333,952	$317,170	$1,632,328	$9,618,957
Hosting Overhead NCSA	$119,855	$71,855	$237,447	$537,242
Total budget (using price factors)	**$453,807**	**$389,025**	**$3,209,104**	**$14,083,788**

In Table 1 we should note that IN2P3 do 50% of processing so we reduce the processing cost by half. This does not reduce the storage cost. We have applied a modes cost reduction assuming that processors and disks get a little cheaper - that percentage is given in Table 6.

3 Potential scope option

In the 2019 JSR we discussed the possibility of delaying purchasing DR1 hardware in DM. Table 2 defines what this would be worth using the cost/sizing model in this document.

Table 2: Considering a scope option of delaying the purchase of LOY1 processing hardware and only purchasing what is needed for commissioning we would only purchase up to and including 2022 hardware of Table 1. If we consider that amount and the current remaining construction budget for hardware the potential worth of such a scope option is given in this table.

Budget for commissioning (to 2022)	**$4,051,936**
DM construciton budet remaining	$14,000,000
Total potential to delay to ops	**$9,948,064**

3.1 Buy Xeon for compute

Table 3 gives the price of compute based on Xeons.

Table 3: Implementation with Intel Xeon

Year	2020	2021	2022	2023
Number of Xeon	69.00	0.00	128.62	282.69
Approximate cost	$690,000.00	$0.00	$1,286,151.12	$2,826,892.68

3.2 Buy Rome for compute

Table 4 gives the price of compute based on Rome -small and large.

Table 4: Implementation with AMD Rome (we have no good proce for these reallly)

Year	2020	2021	2022	2023
number of small rome	50.00	0.00	77.00	205.00
Approximate cost of small rome	$650,000.00	$0.00	$1,001,000.00	$2,665,000.00
number of large rome	16.00	0.00	25.00	66.00
Approximate cost of large rome	$379,200.00	$0.00	$592,500.00	$1,564,200.00

3.3 Storage

Table 5 gives the price of storage using all 3 types that we need. i This would be needed regardless of the compute chosen.

Table 5: Total storage cost estimate

Year	2020	2021	2022	2023
Fast Storage	$11,842.11	$11,842.11	$26,070.00	$312,840.00
Normal Storage	$52,828.02	$9,199.77	$790,725.05	$4,383,311.12
Latent Storage	$31,852.03	$55,741.05	$318,074.72	$3,637,761.12
High Latency Storage	$237,429.69	$257,080.04	$678,828.48	$2,982,507.91
Total	**$333,951.85**	**$333,862.97**	**$1,813,698.24**	**$11,316,420.15**

4 Models

4.1 Sizing model

An exhaustive and detailed mode is provided in [LDM-138; LDM-144] - here we concentrate on the needs for the final years of construction. We explore the compute and storage needed to get us through commissioning and suggest a 2023 purchase for DR1,2 processing which could be pushed to operations.

Table 6 gives the annual requirements for the next few years.

Table 6: Various inputs for deriving costs - 2019 represents currentl holdings.

Year	2019	2020	2021	2022	2023
Core-hours Needed Total (DRP)		4.41E+06	4.41E+06	1.12E+07	4.53E+07
Annual Increase		4.41E+06	0.00E+00	6.81E+06	3.40E+07
Time to Process days		100.0	100.0	100.0	200
Time to Process hours		2,400	2,400	2,400	4,800
Instantaneous cores (DRP) Annual increase	1152	1,836	0	2,837	7,093
Instantaneous cores (Alerts)		0	0	594	594
Cores (Alerts) Annual increase		0	0	594	0
Instantaneous cores (US DAC/ Staff)	538	538	538	141	568
Cores (US DAC/ Staff) Annual increase		0	0	0	428
Instantaneous cores (Chilean DAC)		0	0	26	103
Cores (Chilean DAC) Annual increase		0	0	26	78
Qserv nodes (US DAC/ Staff)				14	95
Qserv nodes (US DAC/ Staff) Annual Increase				14	81
Qserv nodes (Chilean DAC)				14	95
Qserv nodes (Chilean DAC) Annual Increase				14	81
Total Annual Increase		**1,836**	**0**	**3,457**	**7,599**
Fast Storage (TB)		12	24	50	206
Annual Increase (Fast)		12	12	26	156
Normal Storage (TB)	3000	3391	3459	9317	41786
Annual Increase (Normal)		391	68	5857	32469
Latent Storage (TB)		319	876	4057	20217
Annual Increase (Latent)		319	557	3181	16160
High Latency (TB)		3710	7727	18333	64935
Annual Increase (High Latency)		3710	4017	10607	46602
Chilean DAC Fast Storage (TB)					156
Annual Increase (Fast Chilean DAC)					156
Chilean DAC Latent Storage (TB)					20217
Annual Increase (Latent Chilean DAC)					20217
Annual price decrease CPU		10%			
Annual price decrease Storage		5%			
Annual price decrease Qserv		8%			

4.2 Compute and storage

We which to base our budget on reasonable well know machines for which we have well know prices. Table 8 gives an outline of a few standard machines we use and a price. This table also gives a FLOP estimate for those machines. Table 7 gives costs for different types of storage - we will require various latency for different tasks and those have varying costs. These tables

are used as look ups for the cost models in Section 2

Table 7: Storage types and costs used as inputs used for calculations

Storage type	cost
fast – NVME (50GB/ s each) / TB	$1,000.00
normal - SATA GPFS file systems/ TB	$135.00
latency – slower but on disk	$100.00
high latency – very slow – on tape	$64.00

In Table 7 we should consider for NVME for each TB with file system servers two DDN NVME box with GPFS servers. The price is based on the TOP performer with best price . The Normal price is for each TB with file system disks and servers locally attached to production resources.

In the latency and high latency prices are only at NCSA: for each TB with file systems and all people/services. The complete service not usually attached. S3 bucket type. Can be mounted if needed but not for production worthy speeds. The complete service with data flowing to tape using policies.

Table 8: Machine types and costs used as inputs used for calculations

Type of machine	Cores	Memory(GB)	Eff cores/ node	Cost	purpose/ use
xeon	32	192	26.88	$10,000.00	current K8 node
qserv	12	128	12	$20,000.00	current qserv node
small rome	64	256	37.1	$13,000.00	https:/ / www.microway.com/ product/ navion-1u-amd-epyc-gpu-server/
large rome	128	512	115.2	$23,700.00	
current compute node	24	128	24	$9,000.00	current compute node

There is also an associated running cost for machines included in the total cost of ownership. These overheads are listed in Table 9.

Table 9: Overhead costs per rack

Item	Number/ Cost
Compute nodes in a rack	36
Rack initial cost has power, networking switches, networking cables, ready for machine installation– switches last 5 years. Will need to refresh, but rack should last entire project.	$24,000.00
** need to add annually: floor space for rack for 1 years. need to renew after new nodes are racked/ stacked	$300
** Need to add annually: power for 1 node for 1 yr - kw * rate * hours/ year *	$348
** need to add annually: cooling for 1 node for 5 years kw* chillded water per MTBU* hours/ year * 1KW in (MTBU)	$210

** Need to add annually: maintenance for node s – can't purchase more than what the contract has in time left. could be included in the price of the machine, and might not be added in here.	$1,500
Cost for each machine for 1 year in a rack.	$566
**** need to add in at an annual basis. software maintenance (oracle and other software not associated with specific node annually) Oracle license, VM licensing.	$35,000

5 Sizing inputs

The following simplified sizing was used to give the input sizes for the cost model in Section 2. The storage sizes are given in Table 12 and Table 13 while the compute is given in Table 15 and Table 16.

5.1 Storage Model

Table 10: Inputs used to calculate storage needs

Parameters	unit	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
Objects	number			4.58E+09	2.75E+10	from LSE-81, scaled to 2 months for 2022, ComCam ignored
Sources	number			1.50E+11	9.01E+11	from LSE-81, scaled to 2 months for 2022, ComCam ignored
ForcedSources	number			4.85E+11	2.91E+12	from LSE-81, scaled to 2 months for 2022, ComCam ignored
Science users	users	50	100	5000	5000	"Stack Club" to 2021, DP users thereafter
Storage per science user	TB	0.1	0.2	0.2	0.4	ramp to LSE-81 number; includes oversubscription
LSSTCam image size	TB	0.0152				uncompressed, 32 bit, with overscan and corner rafts
Raw image compression	factor	0.42				lossless-compressed divided by uncompressed for raws
Lossy image compression	factor	0.250				lossy-compressed divided by lossless-compressed for PVIs
Observing nights per year	nights	300				maximum
Visits per night	visits	1000				maximum
Images per visit	images	2				
Calibration images per day	images	500				
LSSTCam Science images	images			100000	600000	test images until 2 months of science in 2022
LSSTCam Test images	images	25000	50000	50000		ramp to science images
LSSTCam Engineering images	images	12500	12500	15000	6000	decreasing ramp
LSSTCam Calibration images	images	12500	25000	37500	150000	estimates based on science and test images; actual for 2023
Object table row size	bytes			1896	1896	from LDM-141
Object_Extra tables row size	bytes			21005	21005	from LDM-141
Source table row size	bytes			467	467	from LDM-141
ForcedSource table row size	bytes			41	41	from LDM-141
Qserv replication factor	factor	3.0	3.0	3.0	3.0	

5.1.1 Overview

This simplified storage model eliminates many details in the previous storage model [LDM-141] that end up being insignificant. There are relatively few data products that require significant amounts of fast SSD or slower disk or tape storage; the others complicate the model without giving much insight. In addition, it is assumed that bandwidth is not a significant constraint, other than the distinction between SSD and spinning disk. With the advent of highly-parallel shared and object storage, having large numbers of spindles solely to achieve high bandwidth for certain operations is not thought to be necessary.

Values are computed for the amount of storage expected to be ”on the floor” at the beginning of each fiscal year from FY2020 through FY2023 (which is LSST Operations Year 1). Not included is any storage already present at the end of FY2019 holding past data.

Key scientific and algorithmic assumptions made include:

- All significant intermediates and data products generated by Data Release Production processing need to be kept on filesystem disk until the DRP is complete. Some scratch space is provided to hold small, temporary intermediates. If some intermediates could be removed during DRP when it is known they will no longer be needed, some space savings could be realized.
- HSC RC2 processing is representative of the outputs that DRP will generate. In particular, the number of coadds and the presence or absence of ”heavy footprints” are assumed to be correct.
- Processed visit images (PVIs) and catalogs in Parquet format start on ”normal” filesystem disk but then move to object storage at the completion of the DRP, with lossy compression of the PVIs at that time. This is in accordance with RFC-325, although the relevant LCR has not yet been approved.
- Raw images are only temporarily stored on filesystem disk and are then rapidly moved to object storage, where they are retained.
- Coadd images are generated and kept on filesystem disk.
- Intermediates like warped images for coaddition are not survey data products and do not need to be kept beyond the end of the DRP and subsequent QA.

All data is backed up to tape permanently, including annual snapshots of filesystems. Any incremental backups are assumed to be reusable or otherwise purged and hence not significant.

5.1.2 Parameters

The key parameters in Table 10 are described below.

The numbers of Objects, Sources, and ForcedSources are taken from LSE-81, with the FY2022 numbers reduced by a factor of 2/12 to account for the anticipated 2 months of on-sky science validation time for LSSTCam before the survey begins. These numbers are ultimately based on models for stars in the galaxy and galaxies in the universe that are dependent on the limiting magnitude achieved in each year.

The numbers of science users are estimates, using "Stack Club" users and Commissioning users for FY2020 and 2021, followed by US science users in FY2022 and FY2023 for Data Preview data. The bulk of US science users are not expected to arrive until after Data Release 1 at the beginning of FY2024.

Storage per science user is estimated based on today's usage at NCSA, scaled up as users become more active, and approaching the number given in LSE-81 as Operations begins. Note that it is expected that there will be a wide distribution of usage by user, with some using almost none and some using much more than their proportional share.

The LSSTCam image size is uncompressed and includes overscan, 4 bytes of raw data per pixel, and both science and corner rafts.

The raw image compression factor was measured on simulated LSST images. The lossy image compression factor for processed visit images is the ratio between the lossy-compressed file size (estimated at 1/6 of uncompressed) and the lossless-compressed file size (estimated at 66% of uncompressed). Note that PVIs do not compress losslessly as well as raw images due to their floating point planes.

The number of observing nights per year and the number of visits per night are maximal estimates. 2 images per visit is still the baseline and a possibility that must be accounted for. The number of calibration images per day was derived from the calibration plan.

As stated above, the number of LSSTCam science images is scaled by 2/12 for FY2022 given the length of science validation time. The number of test images is estimated as a ramp up to the full science cadence. The numbers of engineering and calibration images are estimated as ramping-down fractions of the number of science and test images, with calibration images ending at the number per day given previously.

Sizes of rows in various data product tables are taken from LDM-141, which was in turn derived from the DPDD.

Qserv replicates its data for fault tolerance; a typical replication factor is selected here.

5.1.3 Data Product Sizing

Images and the results of processing them are the dominant factor controlling tthe storage sizing which is outlined in Table 11. Precursor survey and LSSTCam images are the largest; ComCam, at less than 5% of the size of LSSTCam and with little on-sky science time is negligible, as is LATISS, which is less than 1% of the size of LSSTCam, though it has considerable on-sky time.

The sizing of the Alert Production Database (APDB) is based on experiments in Salnikov (DMTN-113) which found that 57,000 visits took 4.5 TB including indexes. A simple linear scaling to a full year's visits was performed, with half that purchased in 2020 for large (but not full) scale testing.

HyperSuprime-Cam (HSC) RC2 is a relatively small dataset used for monthly processing tests, but it is highly representative of the currently-known DRP work and so is used as the basis for scaling. The size of the input images was taken from Wood-Vasey et al. (DMTN-091); the size of the outputs (image and Parquet/other non-image files) was measured from the latest execution. A similar size dataset based on DESC DC2 is assumed to be being used for an additional monthly processing test. Note that this is a very small subset of the full DESC DC2, which is expected to cover 300 square degrees to 10-year LSST depth (approximately 1000 epochs per point on the sky). The full DESC DC2 is not currently scheduled to be reprocessed by the construction team. Instead, twice-a-year processings of the full HSC SSP PDR2 dataset (including PDR1) are assumed to occur. The size of this dataset was measured on disk; it is 2,564,358 CCD images, each at 18.2 MB (approximately three times the size of PDR1 alone).

Output sizes are assumed to scale linearly with input size, and by the same factor for each instrument, except for coadds which scale by the sky area processed. While the Object catalog ought to be proportional to sky area as well, its size is expected to be dominated by Source and ForcedSource, so we conservatively make them all proportional to input size (visits).

Scratch space is set at 10% of the output image storage for LSSTCam processing; it is assumed to be already present for precursor processing.

Qserv Czar fast (SSD) storage is assumed to be used for the primary Object table; additional space for the so-called ”secondary index” mapping object identifiers to spatial chunks is negligible in comparison.

The main Qserv database storage is based on the Parquet file sizing for precursor data and on the estimated numbers of Objects, Sources, and ForcedSources for LSSTCam data.

Note that no space is explicitly reserved for Qserv query result storage.

An additional 20% disk and tape storage is added to account for all other needs.

Table 11: Inputs on dataset sizes used to calculate storage needs

Dataset Sizing	unit	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
HSC RC2 Area	deg2	3.0	3.0	3.0	3.0	
HSC SSP PDR2 Area	deg2	300	300	300	300	
DESC DC2 Area	deg2	300	300	300	300	
LSSTCam Area	deg2			2000	17000	
APDB	TB	12	24	24	24	4.5/ 57K TB per visit; 1 year retention; 6 months in 2020
HSC RC2 Input Images	TB	0.8	0.8	0.8	0.8	428 visits * 104 CCDs * 18.2 MB uncompressed
HSC RC2 Output Images	TB	2.4	2.4	2.4	2.4	lossless-compressed, not including warps
HSC RC2 Output Coadd Images	TB	0.7	0.7	0.7	0.7	lossless-compressed
HSC RC2 Output Catalogs	TB	1.4	1.4	1.4	1.4	
HSC SSP PDR2 Input Images	TB	93.3	93.3	93.3	93.3	2564358 CCDs * 18.2 MB uncompressed (3 * PDR1)
DESC DC2 Input Images	TB	455	455	455	455	300 sq deg, 10 year depth
LSSTCam Raw Images	TB	319	557	1290	4816	compressed, moves to object store
Precursor Output Images	TB	763	763	763	763	monthly RC2 and DC2 subset plus biannual PDR
Precursor Output Parquet	TB	361	361	361	361	
LSSTCam Output Images	TB			2248	13485	lossless-compressed, moves to object store
LSSTCam Output Coadd Images	TB			455	3864	
LSSTCam Output Parquet	TB			1329	7973	
Scratch	TB			225	1349	10% of output images
Qserv Czar/ Object	TB			26	156	based on row sizes and counts
Qserv Database	TB	1088	1088	585	3510	based on Parquet for preliminary; based on row sizes and counts
Science User Home	TB	5	20	1000	2000	
Other/ Misc	TB	620	673	1772	7771	20% of total

5.1.4 Storage Sizing

Finally, storage is allocated to specific types as shown in Table 12. Fast storage (SSD) is used for the APDB and Qserv Czar, which accumulates data from year to year until Data Releases are retired. Normal storage is used for inputs, scratch, and output images (initially). Local Qserv storage is used for Qserv catalogs. It is assumed that precursor data will be removed from Qserv once LSST data is available, but the LSST data accumulates from year to year. Object storage is used for output tables each year and output images after one year. Lossy compression is applied at this time. Since only one year of operational processing is in the model, nothing is removed from the object store; it accumulates from year to year. Tape is used for long-term archiving and filesystem backup. Again, this accumulates from year to year.

Note that no replication is assumed in the object store.

Table 12: On floor LDF storage estimates based on Table 11 and Table 10

Storage Sizing (on the floor)	unit	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
Fast	TB	12	24	50	206	SSD
Normal	TB	3391	3459	9317	41786	Enterprise-grade SATA
Qserv Storage	TB	1088	1088	585	4094	Local consumer-grade SATA
Object Store	TB	319	876	4057	20217	
Tape	TB	3710	7727	18333	64935	

An additional table (Table 13) gives the storage needs in the Chilean Data Access Center (DAC). This comprises Qserv fast and local storage plus the data products in object storage. Since no DRP computation occurs in Chile, no ”normal” filesystem disk is required. Chilean user home directories are assumed to be negligible at this level.

Table 13: On floor Chile storage estimates for Base Data Center

Chile Storage (on the floor)	unit	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
Fast	TB				156	SSD
Normal	TB				0	Enterprise-grade SATA
Qserv Storage	TB				4094	Local consumer-grade SATA
Object Store	TB				20217	
Tape	TB				0	

5.2 Compute Model

Table 14: Inputs used to calculate compute needs

Parameters	units					Notes
HSC PDR1 Input Images	TB	13.7				7238 visits of 104 CCDs
HSC PDR1 small-memory compute	core-hours	64392				measured on E5-2680 v3 @ 2.50GHz
HSC PDR1 high-memory compute	core-hours	78523				measured on E5-2680 v3 @ 2.50GHz

Small-memory DRP algorithm ratio	factor	1.5				image differencing, etc.
High-memory DRP algorithm ratio	factor	2.5				stackfit, etc.
DRP compute per TB	core-hours/ TB	2.1E+04				
Percent DRP on high-memory	factor	67%				
ap_pipe single-core sec/ CCD	core-sec/ CCD	83				measured
Additional AP steps	factor	1.25				DCR, real_bogus, etc.
AP compute per visit	core-hours/ visit	5.4E+00				
Qserv data/ node	TB	43.2				1 GB/ sec for 12 hours

5.2.1 Overview

This simplified computing model (Table 14) divides computation into three classes: Data Release Production (DRP), Alert Production, and LSST Science Platform (for the US DAC, Chilean DAC, and LSST staff internal use). Calibration Products Production is assumed to be negligible.

The pipelines have advanced considerably in terms of fidelity and science performance since the previous computing model [LDM-138] was developed. Scaling compute needs based on an execution of the nascent DRP pipeline on HSC PDR1 data and nightly executions of the nascent `ap_pipe` pipeline on HiTS2015 data is thus appropriate, but the fact that several steps are still missing from these pipelines must be taken into account.

Elapsed times are measured on existing hardware and converted into core-hours on a nominal CPU (Intel Xeon E5-2680v3 at 2.50 GHz). This estimation methodology incorporates all I/O, memory bandwidth, cache miss, and other overheads into the core-hour measurement, simplifying calculations. Note that the nominal CPU does not evolve with time; if future CPUs do more work per core, the actual core-hours may be less than estimated here.

Key scientific and algorithmic assumptions are:

- DRP compute time is proportional to the input data size (or, equivalently, the number of visits). While certain tasks are undoubtedly proportional to sky area or number of Objects, overall the pipeline elapsed times are a better fit to the number of visits. Some of this may be because the Object density increases as the number of visits to the same sky patch increases.
- HSC PDR1 processing is generally representative of the final DRP, with an allocation for future additional steps as described below.
- Qserv nnode counts should remain proportional to the size of data loaded into the

database in order to maintain sufficient disk bandwidth and query processing capability, but the proportionality constant changes with time as new generations of system bus with greater bandwidth become available.

- The US DAC LSP is sized at 10% of the DRP compute budget in core-hours, readjusted to be spread over an entire year. The Chilean DAC LSP is sized at 20% of the US DAC (as in LDM-138). The LSST staff LSP is sized at 10% of the US DAC.

5.2.2 Parameters

The key parameters in Table 14 are described below.

HSC PDR1 was executed on the NCSA verification cluster, which uses the nominal CPU. The Alert Production executes on Kubernetes nodes, which are a bit slower; to be conservative, this is neglected.

The most recent run of DRP on HSC PDR1 data is described at `https://confluence.lsstcorp.org/x/WpBiB`. The input data size is measured; note that the input data files are lossless-compressed. Most jobs (but not most of the time) could run on relatively small-memory machines with 24 cores and 5 GB RAM per core. The largest and longest-running jobs, however, required up to 4 times as much memory, using half or a quarter of the cores. To be conservative, we assume that half the cores were used for the large-memory jobs. The percentage of DRP core-hours that will need to execute on large-memory nodes is estimated.

Since the HSC PDR1 processing did not include several steps from the Science Pipelines Design document [LDM-151] such as image differencing and full multi-epoch characterization, the core-hours used are scaled up to the expected pipeline consumption. Note that these algorithmic adjustments are multiplicative.

The SQuaSH system reports the execution time of `ap_pipe` in seconds per CCD. A mean was taken over all processed CCDs, and it was assumed that each CCD is processed on a single core. A factor is added to account for additional steps like differential chromatic refraction compensation and false positive detection that are not well-represented in the current pipeline. Multiplying by the number of LSSTCam science CCDs gives the total number of core-hours per visit.

The amount of Qserv data that can be handled by one node is estimated based on the amount of disk that can be scanned in 12 hours at an aggregate rate of 1 GB per second. (Since the Qserv data replicas are not all anticipated to be accessed at the same rate, this is a conservative estimate.)

5.2.3 Data Release Production

The number of nominal core-hours per TB of input data is multiplied by the precursor (HSC RC2 and DESC DC2 subset for 12 months and HSC PDR2 twice a year) and LSSTCam input data sizes (with lossless compression) to determine the total number of core-hours needed in each year. This is shown in Table 15. Approximately one-third of these core-hours need to be provided by small-memory (4-5 GB/core) machines; the other two-thirds need to come from large-memory (8-20 GB/core) machines.

Table 15: Compute needs for DRP and AP

Data Release Production	**units**	**FY2020**	**FY2021**	**FY2022**	**FY2023/ LOY1**	**Notes**
Precursor input size	TB	206	206	206	206	
LSSTCam visit input size	TB			319	1911	raw images / images/ visit, lossless-compressed
Precursor compute	core-hours	4.4E+06	4.4E+06	4.4E+06	4.4E+06	
LSSTCam compute	core-hours			6.8E+06	4.1E+07	
Total DRP compute	**core-hours**	**4.4E+06**	**4.4E+06**	**1.1E+07**	**4.5E+07**	
Alert Production	units	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
AP cores	cores			594	594	minimum necessary to keep up

5.2.4 Alert Production

The core-hours per visit are divided by the minimum visit length (30 sec plus 1 sec shutter motion plus 2 sec readout) to give the minimum number of cores needed to keep up with image taking. This is shown in Table 15. These cores are expected to be provided over multiple ”strings” of nodes. Note that the current AP design is not readily able to take advantage of more than one core per CCD.

5.2.5 LSST Science Platform

LSST Science Platform needs for US DAC science users are derived as 10% of the DRP core-hour requirement and are shown in Table 16. The LSP core-hours are assumed to be spread over a year, giving the total number of nominal cores needed in the DAC. Peak loads are expected to be handled by ”borrowing” elastically from the DRP compute pool.

As a reasonableness check, the number of cores per science user is computed, but it must be noted that an oversubscription factor needs to be taken into account since not all users are expected to be simultaneously active.

Similar computations for the Chilean DAC (at 20% of the US DAC) and the LSST staff LSP (at 10% of the US DAC) are also in Table 16.

The number of Qserv nodes needed is computed from the storage devoted to it and the storage per node number. Note that staff use of Qserv is taken into account by loading the Data Release products into an internal-only Qserv instance and then making that instance part of the DAC at Data Release, so the compute sizing is part of the US DAC.

Table 16: Compute needs for the Science Platform instances

US DAC	**units**	**FY2020**	**FY2021**	**FY2022**	**FY2023/ LOY1**	**Notes**
LSP cores	cores			128	517	10% of DRP, over a year
Qserv nodes	nodes			14	95	
LSP cores/ science user	cores/ user			0.03	0.10	includes oversubscription
Chilean DAC	units	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
LSP cores	cores			26	103	20% of US DAC
Qserv nodes	nodes			14	95	
Staff LSP	units	FY2020	FY2021	FY2022	FY2023/ LOY1	Notes
LSP cores	cores			13	52	10% of US DAC

5.2.6 DES Comparison

As another check on the model, core-hour figures for Dark Energy Survey (DES) processing were obtained. These are given in Table 17. The CPUs used for single-frame and coadd processing had slightly slower clock rates but better bandwidths and expected instructions per clock performance, so they were considered equivalent to our nominal core. The CPUs used for Multi-Object Multi-Band Fitting and Single-Object Fitting (MOF/SOF) included a large contribution from the Blue Waters machine at NCSA. Those CPUs (AMD 6276) are somewhat older and were estimated at 0.245 nominal cores.

The single-frame processing measured number of 5.2 core-hours per visit compares well with the 5.4 core-hour per visit parameter used in our sizing model. Similarly, the overall DES compute figure of 21,000 core-hours per terabyte is virtually identical to our estimate (including the factors for additional steps).

Table 17: Comparison with DES compute

DES Comparison	**units**					**Notes**
Input data size	TB	50				

Single-frame data size	TB	0.001				
Single-frame processing	core-hours/ visit	5.2				Xeon E5-2680 v4 2.4GHz
Coadd processing	core-hours/ deg2	34.7				Xeon E5-2680 v4 2.4GHz
MOF/ SOF measurement	core-hours/ deg2	108.0				AMD 6276 (313 GFLOPS/ 32 scheduled cores) and Xeon E5-2680 v4
Sky area	deg2	5707				
DES compute per TB	core-hours/ TB	2.1E+04				

5.3 Operations Sizing

Five tables use some of the parameters from the above model to project LSST storage and compute needs throughout the 10 years of Operations.

5.3.1 Storage in Operations

The Object, Source, and ForcedSource numbers in Table 18 are taken from LSE-81, as before. The number of science users and storage per user is ramped up. Note that the number of images needing storage and processing grows linearly with time. Table row sizes are taken from LDM-141; they include growth over time as columns are added.

The dataset sizes in Table 19 are calculated using the same formulas and proportionality constants as in Table 11.

The on-the-floor storage estimates in Table 20 include fast (SSD) storage for the APDB and Qserv Czar, with the latter being sized for three Data Releases (two being served and one being prepared).

”Normal” filesystem storage holds raw images, data products, scratch space, Qserv data prior to loading, science user workspace, and a 20% allocation for everything else.

Qserv local storage holds catalogs for three Data Releases.

Raw images (lossless-compressed) are copied to object storage. Lossy-compressed PVIs, and catalogs in Parquet format are also moved there, with sizing for three Data Releases.

All data products and new raw images for each Data Release are copied to tape.

Table 21 extracts the Qserv and object store sizing needed to populate the Chilean DAC with

Table 18: Inputs used to calculate storage needs during Operations

Parameters	unit	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32
Objects	number	2.75E+10	3.25E+10	3.57E+10	3.82E+10	4.03E+10	4.22E+10	4.38E+10	4.53E+10	4.64E+10	4.74E+10
Sources	number	9.01E+11	1.80E+12	2.70E+12	3.60E+12	4.51E+12	5.41E+12	6.31E+12	7.21E+12	8.11E+12	9.01E+12
ForcedSources	number	2.91E+12	6.87E+12	1.13E+13	1.61E+13	2.13E+13	2.67E+13	3.24E+13	3.83E+13	4.41E+13	5.01E+13
Science users	users	5000	6000	7000	7500	7500	7500	7500	7500	7500	7500
Storage per science user	TB	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3
LSSTCam image size	TB	0.0152	0.0152	0.0152	0.0152	0.0152	0.0152	0.0152	0.0152	0.0152	0.0152
Raw image compression	factor	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
Lossy image compression	factor	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
Observing nights per year	nights	300	300	300	300	300	300	300	300	300	300
Visits per night	visits	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Images per visit	images	2	2	2	2	2	2	2	2	2	2
Calibration images per day	images	500	500	500	500	500	500	500	500	500	500
LSSTCam Science images	images	600000	1200000	1800000	2400000	3000000	3600000	4200000	4800000	5400000	6000000
LSSTCam Engineering images	images	6000	12000	18000	24000	30000	36000	42000	48000	54000	60000
LSSTCam Calibration images	images	150000	300000	450000	600000	750000	900000	1050000	1200000	1350000	1500000
Object table row size	bytes	1896	1953	2012	2073	2136	2201	2268	2337	2408	2481
Object_Extra tables row size	bytes	21005	21636	22286	22955	23644	24354	25085	25838	26614	27413
Source table row size	bytes	467	482	497	512	528	544	561	578	596	614
ForcedSource table row size	bytes	41	41	41	41	41	41	41	41	41	41
Qserv replication factor	factor	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0

Table 19: Dataset sizes used to calculate storage needs during Operations

Dataset Sizing	unit	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32	
LSSTCam Area	deg2□	17000	17000	17000	17000	17000	17000	17000	17000	17000	17000	
APDB	TB	24	24	24	24	24	24	24	24	24	24	
LSSTCam Raw Images	TB	4816	9632	14448	19264	24080	28896	33712	38528	43344	48160	
LSSTCam Output Images	TB	13485	26970	40456	53941	67426	80911	94397	107882	121367	134852	
LSSTCam Output Coadd Images	TB	3864	3864	3864	3864	3864	3864	3864	3864	3864	3864	
LSSTCam Output Parquet	TB	7973	15946	23919	31893	39866	47839	55812	63785	71758	79731	
Scratch	TB	1349	2697	4046	5394	6743	8091	9440	10788	12137	13485	
Qserv Czar/ Object	TB	156	190	215	238	258	279	298	318	335	353	
Qserv Database	TB	3510	5748	8018	10378	12881	15475	18199	21042	23965	27010	
Science User Home	TB	2000	3000	4200	5250	6000	6750	7500	8250	9000	9750	
Other/ Misc	TB	7435	13614	19838	26049	32228	38426	44649	50896	57159	63446	

Table 20: On floor LDF storage estimates during Operations

LDF Storage (on the floor)	unit	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32	
Fast	TB	206	371	586	667	735	798	859	918	974	1029	
Normal	TB	36970	72030	104556	137006	169265	201634	234158	266824	299584	332491	
Qserv Storage	TB	4094	9257	17275	24144	31277	38734	46555	54716	63206	72017	
Object Store	TB	16160	43665	82515	121364	160213	199063	237912	276761	315611	354460	
Tape	TB	64935	117041	222152	358342	525422	723502	952739	1213273	1505202	1828670	

Table 21: On floor Chile storage estimates during Operations

Chile Storage (on the floor)	unit	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32	
Qserv Storage	TB	4094	9257	17275	24144	31277	38734	46555	54716	63206	72017	
Object Store	TB	16160	43665	82515	121364	160213	199063	237912	276761	315611	354460	

Table 22: Compute needs during Operations

Data Release Production	units	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32
LSSTCam visit input size	TB	1911	3822	5733	7644	9556	11467	13378	15289	17200	19111
DRP compute	core-hours	4.5E+07	8.2E+07	1.2E+08	1.6E+08	2.0E+08	2.5E+08	2.9E+08	3.3E+08	3.7E+08	4.1E+08
Alert Production	units	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32
AP cores	cores	594	594	594	594	594	594	594	594	594	594
US DAC	units	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32
LSP cores	cores	517	933	1,399	1,866	2,332	2,798	3,265	3,731	4,198	4,664
Qserv data per node	TB/ node	43	43	86	86	86	86	173	173	173	173
Qserv nodes	nodes	95	214	307	346	362	448	434	406	366	417
LSP cores/ science user	cores/ user	0.1	0.2	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.6
Chilean DAC	units	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32
LSP cores	cores	103	187	280	373	466	560	653	746	840	933
Qserv nodes	nodes	95	214	226	213	269	328	284	227	260	295
Staff LSP	units	LOY1/ FY23	LOY2/ FY24	LOY3/ FY25	LOY4/ FY26	LOY5/ FY27	LOY6/ FY28	LOY7/ FY29	LOY8/ FY30	LOY9/ FY31	LOY10/ FY32
LSP cores	cores	52	93	140	187	233	280	326	373	420	466

a copy of the data products and raw images.

5.3.2 Compute in Operations

The DRP compute sizing in Table 22 follows directly from the size of the input data to be processed. The number of cores for Alert Production does not change with time. The DAC and staff LSP instances are sized based on the assumed percentages of DRP compute. The amount of Qserv data that can be handled by a node is assumed to grow with time, doubling every four years (PCI Express has gone from 1.0 GB/sec to 16 GB/sec between 2003 and 2019). The number of Qserv nodes is calculated by dividing each Data Release's storage by the storage-per-node figure for its year; older nodes are assumed to be retired.

A References

References

[LDM-141], Becla, J., Lim, K.T., 2013, *Data Management Storage Sizing and I/O Model*, LDM-141, URL https://ls.st/LDM-141

[LSE-81], Dubois-Felsmann, G., 2013, *LSST Science and Project Sizing Inputs*, LSE-81, URL https://ls.st/LSE-81

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

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

[DMTN-113], Salnikov, A., 2019, *Performance of RDBMS-based PPDB implementation*, DMTN-113, URL https://dmtn-113.lsst.io, LSST Data Management Technical Note

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

[DMTN-091], Wood-Vasey, M., Bellm, E., Bosch, J., et al., 2019, *Test Datasets for Scientific Performance Monitoring*, DMTN-091, URL https://dmtn-007.lsst.io/v/DM-15448/index.html, LSST Data Management Technical Note

B Acronyms

Acronym	Description
DB	DataBase
DBB	Data Back Bone
DDN	Data Delivery Network
DM	Data Management
DMTN	DM Technical Note
FLOP	FLoating point Operation
FLOPS	FLoating point Operation per Second
GFLOP	Giga FLOP
GPFS	General Parallel File System (now IBM Spectrum Scale)
LDF	LSST Data Facility
LDM	LSST Data Management (Document Handle)
LSP	LSST Science Platform
NCSA	National Center for Supercomputing Applications
NVME	Non Volatile Memory Express.”DM IT”
Qserv	LSST’s distributed parallel database. This database system is used for collecting, storing, and serving LSST Data Release Catalogs and Project metadata, and is part of the Software Stack
SATA	Serial Advanced Technology Attachment
TB	TeraByte