DM sizing model and cost plan for construction and operations.

William O'Mullane, Richard Dubois, Michelle Butler, Kian Tat Lim

2021-03-22

1 Introduction

This document presents the simplified sizing model for Rubin Observatory data management in Section 6.1 based on detailed sizing presented in Section 7. Section 2 presents a very high level budget summary for DM hardware which was used for LCR-2148. More interesting now is the build up at the USDF in pre-operations which is shown in Section 3 and the full operations estimates in Section 4

This version is in agreement with SLAC on the parameters for CPU and disk price fall as well as CPU cost etc.

2 Construction budget

A high level bottom line is given in Table 1. 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 27 are applied post 2020.

Year	**2021**	**2022**	**2023**
Compute (2019 pricing)	$690,000	$0	$1,500,000
Storage (2019 pricing)	$183,054	$138,290	$1,450,989
Qserv (2019 pricing)			$280,000
Total (2019 pricing)	**$873,054**	**$138,290**	**$3,230,989**
Compute (applying price factor)	$552,000	$0	$900,000
IN2P3 (50% of compute in ops)			
UKDF (25% of compute in ops)			
Storage (applying price factor)	$155,596	$107,175	$1,015,692
Qserv (applying price factor)			$182,000
Hosting cost NCSA	$110,802	$62,802	$238,012
Total budget (using price factors)	**$818,398**	**$169,977**	**$2,335,705**

We have applied a modest cost reduction assuming that processors and disks get a little cheaper - that percentage is given in Table 27 along with many other parameters. Table 27 also contains the number of nodes we assume to need for Qserv.

Specific costs for storage are detailed in Table 28 and for compute in Table 29 the following budgets can be considered. The detailed annual purchasing based on those prices is given for storage in Table 11 and for compute in Table 10.

3 Pre-Operations budget estimate

In this section we estimate the ramp up of the USDF to be ready for start of operations. This means having compute for commissioning data as well as developer services and a small science platform in place. It is very similar to the construction needs. The summary is in Table 2.

Table 2: This table builds a ramp for build up at SLAC as USDF. These would be purchases to get initial systems in place for the first year of operations. This is based on the Rome processor price and other construction inputs.

Year (Pricing $million)	**2021**	**2022**	**2023**
Compute (2020 pricing)	$0.02	$0.04	$0.43
Qserv (2020 pricing)			$0.28
Storage (2020 pricing)	$0.58	$0.14	$1.45
Total (2020 pricing)	**$0.60**	**$0.18**	**$2.16**
Applying price factor (CPU)	$0.02	$0.03	$0.31
Qserv (applying factor)	0	0	$0.21
Applying price factor (Storage)	$0.54	$0.12	$1.15
Hosting Overhead SLAC	$0.13	$0.11	$0.18
Total budget (using price factors)	**$0.68**	**$0.26**	**$1.86**
Total Pre Ops hardware to 2023	**$2.80**	**million**	

Currently we assume exactly the construction profile for storage plus space for datasets currently held at NCSA. This is presented in Table 4.

The compute is a little different and uses Rome which is captured in Table 3.

Table 3: Preoperations compute build up at USDF, cores we need to purchase per year.

Year	**2021**	**2022**	**2023**
DRP cores (from construction)	0	1,836	2,837
Alerts cores (from construction)			1188
Dev Cores	100	440	
K8S (science platform)	100		924
Total cores	**200**	**440**	**4,950**
Number of Large Rome USDF	2	4	43
Compute (2020 pricing)	$0.02	$0.04	$0.43

Table 4: Preoperations storage to be purchaed each year

Year	**2021**	**2022**	**2023**
Fast Storage (TB)	12	12	26
Normal Storage (TB)	3680	68	5494
Latent Storage (TB)	319	557	4090
High Latency (TB)	2910	3218	10605

Fast (2020 pricing)	$4,736.84	$4,736.84	$10,428.00
Normal (2020 pricing)	$489,405.06	$9,063.48	$730,646.97
Latent (2020 pricing)	$42,363.20	$74,135.60	$543,948.46
High Latency (2020 pricing)	$45,549.15	$50,354.27	$165,965.39
Total Storage PreOps (2020 pricing)	**$582,054.25**	**$138,290.19**	**$1,450,988.82**

4 Operations budget estimate

Based on the needs in Table 25 and the costs in Table 18 and Table 29 we get the estimate presented in Table 5. In Table 5 we should note that IN2P3 do 50% and UKDF do 25% of the processing so we reduce the processing cost by three quarters. This does not reduce the storage cost.

Table 5: This table adds the USDF (see Table 7) and Table 8 to give total operations hardware costs.

Year (all prices Million$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
USDF hardware	$7.48	$8.74	$8.77	$8.26	$8.75	$11.80	$11.16	$11.01	$10.28	$9.76
Chile hardware	$1.07	$1.17	$1.17	$0.96	$0.88	$1.52	$1.43	$1.41	$1.22	$1.13
Total budget (using price factors)	**$8.55**	**$9.90**	**$9.94**	**$9.22**	**$9.63**	**$13.32**	**$12.58**	**$12.42**	**$11.51**	**$10.90**
Total Operations hardware to 2033	**$107.97**	**million**	**to 2035**	**$116.28**						

Again in Table 5 we assume IN2P3 do 50% of processing (see Table 7). We have applied a compounded modest cost reduction assuming that processors and disks get a little cheaper - that percentage is given in Table 27.

It must be noted that the price of disk and tape have a profound effect over 10 years. We have been fairly conservative on the base prices in Table 28. An even bigger effect is in the compounding of the presumed fall in storage cost. Here we have used an extremely conservative 5% per year (Table 27) - changing this to 15% halves the cumulative ops estimate, setting it to 10% brings the total down by about 30%.

4.1 Cloud costs

In addition there are some cloud costs. We run certain jobs and host websites on Amazon and Google. In operations the validation team may also wish to run simulations on cloud resources. This estimate is in Table 6.

Table 6: We have on going cloud costs and assume some other activities may be on cloud in the future - we make an estimate of those costs here.

Year (all prices Million$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**	**2034**
Jira Cloud	$80,000	$80,000	$80,000	$80,000	$80,000	$80,000	$80,000	$80,000	$80,000	$80,000	$80,000

Current actuals	$20,000	$20,000	$20,000	$20,000	$20,000	$20,000	$20,000	$20,000	$20,000	$20,000	$20,000
RPF sims, V&V	$10,000	$10,000	$10,000	$10,000	$10,000	$10,000	$10,000	$10,000	$10,000	$10,000	0
Total	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$110,000**	**$100,000**

More details on the inputs are in Section 5.1

4.2 US and Chile

While Table 5 present the total ops cost for Rubin Observatory a fraction of this is in Chile and would potentially remain an NSF cost in operations. Table 7 presents just the US Data Facility budget and Table 8 presents the Chile budget.

Table 7: This table pulls together all the information in a high level summary for USDF operations - in this table Rome pricing(see Table 14). Price factors, defined in Table 27 are applied in all cases - other input values come from Table 25, Table 19.

Year (all prices Million$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Compute (2020 pricing)	$0.65	$0.69	$1.14	$1.43	$1.47	$1.66	$1.55	$1.55	$1.66	$1.55
Qserv (2020 pricing)	$1.62	$2.42	$1.86	$2.40	$2.74	$3.60	$2.10	$2.18	$2.78	$3.12
Storage (2020 pricing)	$7.92	$9.73	$11.22	$10.94	$12.62	$18.97	$20.83	$22.33	$22.05	$22.47
Total (2019 pricing)	**$10.19**	**$12.84**	**$14.22**	**$14.77**	**$16.83**	**$24.23**	**$24.48**	**$26.06**	**$26.49**	**$27.14**
Applying price factor (CPU)	$0.43	$0.41	$0.61	$0.68	$0.63	$0.64	$0.54	$0.49	$0.47	$0.39
IN2P3 (50% of compute)	-$0.21	-$0.20	-$0.30	-$0.34	-$0.32	-$0.32	-$0.27	-$0.24	-$0.23	-$0.20
UKDF (25% of compute)	-$0.11	-$0.10	-$0.15	-$0.17	-$0.16	-$0.16	-$0.14	-$0.12	-$0.12	-$0.10
Qserv (applying factor)	$1.12	$1.53	$1.07	$1.26	$1.32	$1.58	$0.84	$0.80	$0.93	$0.95
Applying price factor (Storage)	$5.80	$6.59	$7.03	$6.34	$6.76	$9.41	$9.55	$9.47	$8.65	$8.16
Hosting Overhead SLAC	0.5	0.5	0.5	0.5	0.5	0.7	0.6	0.6	0.6	0.6
Total budget (using price factors)	**$7.48**	**$8.74**	**$8.77**	**$8.26**	**$8.75**	**$11.80**	**$11.16**	**$11.01**	**$10.28**	**$9.76**
Total Operations hardware to 2033	**$96.01**	**million**	**to 2035**	**$103.46**						

Note that in Table 8 a fraction of the USDF storage is considered, big enough for Raws and a cache of products. It also does not include a Qserv. See also Table 26

Table 8: "This table pulls together all the information in a high level summary for Chile operations - in this table Rome pricing(see Table 17) is used Price factors, defined in Table 27 are applied in all cases - other input values come from Table 25, Table 20.

Year (all prices Million$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Compute (2020 pricing)	$0.04	$0.04	$0.04	$0.08	$0.08	$0.08	$0.08	$0.08	$0.08	$0.08
Storage (2020 pricing)	$1.21	$1.50	$1.71	$1.51	$1.50	$2.71	$3.00	$3.21	$3.01	$3.00
Total (2020 pricing)	**$1.25**	**$1.54**	**$1.75**	**$1.59**	**$1.58**	**$2.79**	**$3.08**	**$3.29**	**$3.09**	**$3.08**
Applying price factor (CPU)	$0.03	$0.02	$0.02	$0.04	$0.03	$0.03	$0.02	$0.02	$0.02	$0.02
Applying price factor (Storage)	$0.89	$1.02	$1.07	$0.88	$0.80	$1.35	$1.38	$1.36	$1.18	$1.09
Overhead hosting Chile	$0.15	$0.13	$0.08	$0.05	$0.05	$0.14	$0.03	$0.03	$0.02	$0.03
Total budget (using price factors)	**$1.07**	**$1.17**	**$1.17**	**$0.96**	**$0.88**	**$1.52**	**$1.43**	**$1.41**	**$1.22**	**$1.13**
Total Chile hardware to 2033	**$11.97**	**million**	**to 2035**	**$12.82**						

5 Cost details

The summary table (Table 1) uses Xeon pricing for compute as shown in Table 9.

Table 9: Implementation with Intel Xeon

Year	2021	2022	2023	2024
Number of Xeon	69	0	150	283
Approximate cost	$690,000.00	$0.00	$1,500,000.00	$2,830,000.00

An alternative architecture would be Rome - SLAC have chosen this for the Ops pricing, Table 10 gives the price of compute based on Rome -small and large. Rome large are used in the operations calculations.

Table 10: Implementation with AMD Rome (we have no good proce for these reallly)

Year	2021	2022	2023	2024
number of small rome	49	0	75	201
Approximate cost of small rome	$637,000.00	$0.00	$975,000.00	$2,613,000.00
number of large rome	16	0	25	66
Approximate cost of large rome	$208,000.00	$0.00	$325,000.00	$858,000.00

Table 11 gives the price of storage using all types that we need. This would be needed regardless of the compute chosen.

Table 11: Total storage cost estimate

Year	2021	2022	2023	2024
Fast Storage	$4,736.84	$4,736.84	$10,428.00	$62,568.00
Fast Storage Chile	$0.00	$0.00	$0.00	$62,568.00
Normal Storage	$90,405.06	$9,063.48	$730,646.97	$3,955,645.91
Latent Storage	$42,363.20	$74,135.60	$543,948.46	$3,177,069.33
Latent Storage Chile	$0.00	$0.00	$0.00	$3,837,516.59
High Latency Storage	$45,549.15	$50,354.27	$165,965.39	$727,912.22
Total	**$183,054.25**	**$138,290.19**	**$1,450,988.82**	**$11,823,280.05**

Table 12 gives the annual cost of hosting compute at NCSA for construction. This includes purchasing racks to house new nodes etc.

Table 12: Overheads(NCSA) per year based on number of cores in Table 27 and costs in Table 30 assuming Xeon density from Table 29.

Year	2021	2022	2023	2024
Total Incremental cores (USA)	**1,836**	**0**	**4,026**	**7,521**
Total owned cores (USA)	**3,528**	**3,528**	**7,554**	**15,075**
Total owned nodes	**111**	**111**	**251**	**567**
Cost for hosting nodes	$62,802	$62,802	$142,012	$320,801
Total new nodes	**58**	**0**	**140**	**317**
Total new racks	**2**	**0**	**4**	**9**
Rack install cost	$48,000.00	$0.00	$96,000.00	$216,000.00
Total Overhead (NCSA)	**$110,802.27**	**$62,802.27**	**$238,012.35**	**$536,800.80**

5.1 Ops Cost details

Table 13 gives the price of compute based on Xeons. This is broken down further for US in Table 15 and Chile in Table 16. However the Ops costing for SLAC was done using Table 14.

Table 18 gives the price of storage using all types that we need. This is broken down further for US in Table 19 and Chile in Table 20 This would be needed regardless of the compute chosen.

Table 13: Implementation with Intel Xeon for full Rubin Observatory

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Number of Xeon	283	298	491	619	634	716	672	672	716	672
Approximate cost (2020 Mdollars)	$2.83	$2.98	$4.91	$6.19	$6.34	$7.16	$6.72	$6.72	$7.16	$6.72

Table 14: Implementation with Rome for USDF

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Number of Large Rome USDF	65	69	114	143	147	166	155	155	166	155
Approximate cost (2020 Mdollars)	$0.65	$0.69	$1.14	$1.43	$1.47	$1.66	$1.55	$1.55	$1.66	$1.55

Table 15: Implementation with Intel Xeon for USDF

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Number of Xeon USDF	279	295	487	612	628	710	666	666	710	666
Approximate cost (2020 Mdollars)	$2.79	$2.95	$4.87	$6.12	$6.28	$7.10	$6.66	$6.66	$7.10	$6.66

Table 16: Implementation with Intel Xeon for Chile Compute

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Number of Xeon Chile	4	4	4	8	7	7	7	7	7	7
Approximate cost (2020 Mdollars)	$0.04	$0.04	$0.04	$0.08	$0.07	$0.07	$0.07	$0.07	$0.07	$0.07

Table 17: Implementation with AMD Rome for Chile Compute

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Number of Large Rome Chile	3	3	3	6	6	6	6	6	6	6
Approximate cost (2020 Mdollars)	$0.04	$0.04	$0.04	$0.08	$0.08	$0.08	$0.08	$0.08	$0.08	$0.08

Table 18: Total storage cost estimate for operations of Rubin Observatory USDF and CHile

Year (all in M$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Fast Storage	$0.13	$0.16	$0.19	$0.13	$0.12	$0.23	$0.27	$0.30	$0.23	$0.22
Normal Storage	$3.96	$3.86	$4.20	$4.19	$4.89	$8.13	$8.05	$8.41	$8.42	$8.41
Latent Storage	$4.33	$6.09	$6.99	$6.13	$6.68	$10.46	$12.23	$13.12	$12.27	$12.27
High Latency Storage	$0.73	$1.13	$1.56	$2.00	$2.43	$2.86	$3.28	$3.72	$4.15	$4.58
Total (M$)	**$9.14**	**$11.23**	**$12.93**	**$12.45**	**$14.12**	**$21.68**	**$23.83**	**$25.55**	**$25.07**	**$25.47**

Table 19: Total storage cost estimate for operations at USDF

Year (all in M$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Fast Storage USDF	$0.06	$0.07	$0.09	$0.03	$0.04	$0.09	$0.09	$0.11	$0.06	$0.05
Normal Storage USDF	$3.96	$3.86	$4.20	$4.19	$4.89	$8.13	$8.05	$8.41	$8.42	$8.41

Latent Storage USDF	$3.18	$4.69	$5.37	$4.72	$5.26	$7.90	$9.40	$10.09	$9.44	$9.44
High Latency Storage USDF	$0.73	$1.13	$1.56	$2.00	$2.43	$2.86	$3.28	$3.72	$4.15	$4.58
Total (M$)	**$7.92**	**$9.73**	**$11.22**	**$10.94**	**$12.62**	**$18.97**	**$20.83**	**$22.33**	**$22.05**	**$22.47**

Table 20: Total storage cost estimate for operations in Chile **Note** the latent storage here is 0.3 of the USDF latent Storage

Year (all in M$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Fast Storage Chile	$0.06	$0.09	$0.10	$0.10	$0.08	$0.15	$0.18	$0.19	$0.18	$0.17
Latent Storage Chile	$1.15	$1.41	$1.61	$1.42	$1.42	$2.57	$2.82	$3.03	$2.83	$2.83
Total (M$)	**$1.21**	**$1.50**	**$1.71**	**$1.51**	**$1.50**	**$2.71**	**$3.00**	**$3.21**	**$3.01**	**$3.00**

Table 21 gives the annual cost of hosting compute in NCSA. This includes purchasing racks to house new nodes etc.

Table 21: Overheads(NCSA) per year based on number of cores in Table 25 and costs in Table 30 assuming Xeon density from Table 29.

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Total Incremental cores (USA)	**7,521**	**7,958**	**8,979**	**8,979**	**8,979**	**8,979**	**8,979**	**8,979**	**8,979**	**8,979**
Total owned cores (USA)	**15,075**	**23,033**	**32,012**	**40,990**	**49,969**	**58,947**	**67,926**	**76,904**	**85,883**	**94,861**
Total owned nodes	**567**	**936**	**1,310**	**1,629**	**1,926**	**2,294**	**2,559**	**2,812**	**3,051**	**3,383**
Cost for hosting nodes	$320,801	$529,576	$741,180	$921,666	$1,089,704	$1,297,914	$1,447,847	$1,590,991	$1,726,214	$1,914,055
Total new nodes	**317**	**370**	**374**	**401**	**418**	**461**	**386**	**390**	**420**	**437**
Total new racks	**9**	**11**	**11**	**12**	**12**	**13**	**11**	**11**	**12**	**13**
Rack install cost	$216,000	$264,000	$264,000	$288,000	$288,000	$312,000	$264,000	$264,000	$288,000	$312,000
Total Ops Overhead (NCSA)	**$536,801**	**$793,576**	**$1,005,180**	**$1,209,666**	**$1,377,704**	**$1,609,914**	**$1,711,847**	**$1,854,991**	**$2,014,214**	**$2,226,055**

Note: rack costs are for new racks so only paid for the added racks each year, hence some zeros appear when we do not intend to add racks.

For Chile Table 22 gives the cost of hosting in Chile (las Serena).

Table 22: Overheads(Chile) per year based on number of cores in Table 25 and costs in Table 30 assuming Xeon density from Table 29.

Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Total Incremental cores (Chile)	**103**	**83**	**93**	**93**	**93**	**93**	**93**	**93**	**93**	**93**
Total owned cores (Chile)	**103**	**187**	**280**	**373**	**466**	**560**	**653**	**746**	**840**	**933**
Compute nodes	4	6	9	12	15	18	21	24	27	30
Qserv nodes	95	216	309	348	364	451	436	408	367	418
Total Nodes	**99**	**222**	**318**	**360**	**379**	**469**	**457**	**432**	**394**	**448**
Total Compute Racks	**3**	**7**	**9**	**10**	**11**	**14**	**13**	**12**	**11**	**13**
Total Storage	**8,751**	**10,786**	**12,430**	**10,992**	**11,007**	**19,751**	**21,649**	**23,172**	**21,655**	**21,704**
Toital Storage Racks	2	2	2	2	2	2	3	3	3	3
Cooling Power Kw	3.81	6.86	8.38	9.14	9.91	12.95	12.19	11.43	10.67	12.19
Computing Power kW	73	131.4	160.6	175.2	189.8	248.2	233.6	219	204.4	233.6
Power Cost	$8,467	$15,240	$18,627	$20,320	$22,014	$28,787	$27,094	$25,401	$23,707	$27,094
Compute rack install costs	$85,902.00	$114,536.00	$57,268.00	$28,634.00	$28,634.00	$85,902.00	$0.00	$0.00	$0.00	$0.00
Storage rack install costs	$57,268.00	$0.00	$0.00	$0.00	$0.00	$28,634.00	$0.00	$0.00	$0.00	$0.00
Total Ops Overhead Chile (USD)	**$151,637**	**$129,776**	**$75,895**	**$48,954**	**$50,648**	**$143,323**	**$27,094**	**$25,401**	**$23,707**	**$27,094**

For Chile the rack costs are outlined in Table 23.

Table 23: This table details the cost per rack which is added in Table 22.

2020 Rack Component	Unit Cost	Spine Port		Total	
2 x Leaf	$7,000.00	$2,187.00	$9,187.00	$18,374.00	Cisco Nexus 93108TC-EX, + overhead of Spine port
2 x PDU	$2,980.00			$5,960.00	Raritan PX3 5085U-N2
1 x IPMI	$1,800.00			$1,800.00	Cisco Catalyst 2960-X / Cisco 9200
1 x Rack	$2,500.00			$2,500.00	APC AR3357
Total				**$28,634.00**	

For Chile the power costs are outlined in Table 24.

Table 24: Cost is estimated to increase 5-10% every 2-3 years

Year	2021	2024	2027	2030	2033
Power Cost	100.21	110.231	121.2541	133.37951	146.717461

Various other inputs to ops costing are given in Table 25.

Table 25: Various inputs for deriving costs in operations - these drive the costs in Table 5. This is based on Table 13, Table 18

Year	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033
Core-hours Needed Total (DRP)	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
Core-hours Annual Increase	3.40E+07	3.6E+07	4.1E+07	4.1E+07	4.1E+07	4.1E+07	4.1E+07	4.1E+07	4.1E+07	4.1E+07
Time to Process days	200	200	200	200	200	200	200	200	200	200
Time to Process hours	4,800	4,800	4,800	4,800	4,800	4,800	4,800	4,800	4,800	4,800
Cores (DRP) Annual increase	7,093	7,594	8,512	8,512	8,512	8,512	8,512	8,512	8,512	8,512
Cores (DRP) Annual refresh			2,837	7,093	7,594	8,512	8,512	8,512	8,512	8,512
Cores (DRP) Annual purchase	7,093	7,594	11,349	15,605	16,106	17,024	17,024	17,024	17,024	17,024
Cores (Alerts)	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188
Cores (Alerts) Annual refresh			1,188			1,188			1,188	
Cores (US DAC/ Staff)	568	933	1,399	1,866	2,332	2,798	3,265	3,731	4,198	4,664
Cores (US DAC/ Staff) Annual increase	428	364	466	466	466	466	466	466	466	466
Cores (US DAC/ Staff) Annual refresh			141	428	364	466	466	466	466	466
Cores (US DAC/ Staff) Annual purchase	428	364	607	894	831	933	933	933	933	933
Cores (Chilean DAC)	103	187	280	373	466	560	653	746	840	933
Cores (Chilean DAC) Annual increase	103	83	93	93	93	93	93	93	93	93
Cores (Chilean DAC) Annual refresh			0	103	83	93	93	93	93	93
Cores (Chilean DAC) Annual purchase	103	83	93	197	177	187	187	187	187	187
Qserv nodes (US DAC/ Staff)	95	216	309	348	364	451	436	408	367	418
Qserv nodes (US DAC/ Staff) Annual Increase	81	121	93	120	137	180	105	109	139	156
Qserv nodes (Chilean DAC)	95	216	309	348	364	451	436	408	367	418
Qserv nodes (Chilean DAC) Annual Increase	95	121	93	134	137	180	119	109	139	170
Total Cores Annual Increase	**7,624**	**8,042**	**13,238**	**16,696**	**17,113**	**19,332**	**18,144**	**18,144**	**19,332**	**18,144**
Fast Storage (TB)	206	371	586	667	735	798	859	918	974	1029
Annual Increase (Fast)	156	164	215	81	68	63	60	59	57	55
Annual Refresh (Fast)					26	156	164	215	81	68
Annual Purchase (Fast)	156	164	215	81	94	220	225	275	138	123
Normal Storage (TB)	38,983	67982	99544	131031	162327	193733	225294	256997	288794	320737
Annual Increase (Normal)	29,742	28999	31563	31487	31296	31406	31560	31703	31797	31943
Annual Refresh (Normal)					5,494	29,742	28,999	31,563	31,487	31,296
Annual Purchase (Normal)	29,742	28,999	31,563	31,487	36,790	61,148	60,559	63,266	63,284	63,239
Latent Storage (TB)	28,854	64,086	104,491	139,969	175,447	210,925	246,403	281,881	317,359	352,837
Annual Increase (Latent)	23,888	35,232	40,405	35,478	35,478	35,478	35,478	35,478	35,478	35,478
Annual Refresh (Latent)					4,090	23,888	35,232	40,405	35,478	35,478
Annual Purchase (Latent)	23,888	35,232	40,405	35,478	39,568	59,366	70,710	75,884	70,956	70,956
High Latency (TB)	63,245	135,135	234,943	362,508	517,497	699,929	909,829	1,147,221	1,412,122	1,704,554
Annual Increase (High Latency)	46,512	71,890	99,809	127,565	154,989	182,432	209,900	237,392	264,900	292,432
Chilean DAC Fast Storage (TB)	335	570	825	1,064	1,274	1,484	1,694	1,904	2,114	2,324

Annual Increase (Fast Chilean DAC)	156	235	255	239	210	210	210	210	210	210
Annual Refresh (Fast Chilean DAC)						156	235	255	239	210
Annual Purchase (Fast Chilean DAC)	156	235	255	239	210	366	445	465	449	420
Chilean DAC Latent Storage (TB)	8,656	19,226	31,347	41,991	52,634	63,278	73,921	84,564	95,208	105,851
Annual Increase (Latent Chilean DAC)	8,656	10,570	12,121	10,644	10,643	10,644	10,643	10,643	10,644	10,643
Annual Refresh (Latent Chilean DAC)						8,656	10,570	12,121	10,644	10,643
Annual Purchase (Latent Chilean DAC)	8,656	10,570	12,121	10,644	10,643	19,300	21,213	22,764	21,288	21,286

5.1.1 Alternative costing for Chile

The alternative cost for Chile including Xeons and Qserv but with a similar storage model is given in Table 26.

Table 26: This table pulls together all the information in a high level summary for Chile operations - in this table Xeon pricing(see Table 16) is used since that is the more expensive but better known option. Price factors, defined in Table 27 are applied in all cases - other input values come from Table 25, Table 20.

Year (all prices Million$)	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Compute (2020 pricing)	$0.04	$0.04	$0.04	$0.08	$0.07	$0.07	$0.07	$0.07	$0.07	$0.07
Qserv (2020 pricing)	$1.90	$2.42	$1.86	$2.68	$2.74	$3.60	$2.38	$2.18	$2.78	$3.40
Storage (2020 pricing)	$1.21	$1.50	$1.71	$1.51	$1.50	$2.71	$3.00	$3.21	$3.01	$3.00
Total (2020 pricing)	**$3.15**	**$3.96**	**$3.61**	**$4.27**	**$4.31**	**$6.38**	**$5.45**	**$5.46**	**$5.86**	**$6.47**
Applying price factor (CPU)	$0.03	$0.02	$0.02	$0.04	$0.03	$0.03	$0.02	$0.02	$0.02	$0.02
Qserv (applying factor)	$1.32	$1.53	$1.07	$1.41	$1.32	$1.58	$0.95	$0.80	$0.93	$1.03
Applying price factor (Storage)	$0.89	$1.02	$1.07	$0.88	$0.80	$1.35	$1.38	$1.36	$1.18	$1.09
Overhead hosting Chile	$0.15	$0.13	$0.08	$0.05	$0.05	$0.14	$0.03	$0.03	$0.02	$0.03
Total budget (using price factors)	**$2.38**	**$2.70**	**$2.24**	**$2.37**	**$2.20**	**$3.09**	**$2.38**	**$2.21**	**$2.15**	**$2.17**
Total Operations hardware to 2033	**$23.90**	**million**	**to 2035**	**$25.00**						

6 Models

6.1 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 27 gives the annual requirements for the next few years.

Table 27: Various inputs for deriving costs - 2019 represents current holdings.

Year	**2019**	**2021**	**2022**	**2023**	**2024**
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) Total		1,836	1,836	4,673	9,430

Instantaneous cores (DRP) Annual increase	1152	1,836	0	2,837	7,093
Instantaneous cores (Alerts)		0	0	1188	1188
Cores (Alerts) Annual increase		0	0	1188	0
Instantaneous cores (US DAC/ Staff)	540	540	540	141	568
Cores (US DAC/ Staff) Annual increase		0	0	0	428
Instantaneous cores (Chilean DAC)		0	0	0	103
Cores (Chilean DAC) Annual increase		0	0	0	103
Qserv nodes (US DAC/ Staff)				14	95
Qserv nodes (US DAC/ Staff) Annual Increase				14	81
Qserv nodes (Chilean DAC)				0	95
Qserv nodes (Chilean DAC) Annual Increase				0	95
Total Cores Annual Increase		**1,836**	**0**	**4,026**	**7,624**
Fast Storage (TB)		12	24	50	206
Annual Increase (Fast)		12	12	26	156
Normal Storage (TB)	3000	3680	3748	9241	38983
Annual Increase (Normal)		680	68	5494	29742
Latent Storage (TB)		319	876	4966	28854
Annual Increase (Latent)		319	557	4090	23888
High Latency (TB)		2910	6128	16733	63245
Annual Increase (High Latency)		2910	3218	10605	46512
Chilean DAC Fast Storage (TB)					156
Annual Increase (Fast Chilean DAC)					156
Chilean DAC Latent Storage (TB)					28854
Annual Increase (Latent Chilean DAC)					28854
Annual price decrease CPU		10%			
Annual price decrease Storage		8%			
Annual price decrease Qserv		9%			
Chile peso rate		720			

6.2 Compute and storage

We which to base our budget on reasonable well know machines for which we have well know prices. Table 29 gives an outline of a few standard machines we use and a price. This table also gives a FLOP estimate for those machines. Table 28 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 28: Storage types and costs used as inputs used for calculations

Storage type	**Estimate(SLAC)**	**Estimate(NCSA)**
fast – NVMe (50GB/ s each) / TB	$400.00	$1,000.00
normal - SATA GPFS file systems/ TB	$133.00	$135.00
latency – slower but on disk	$133.00	$45.00
high latency – very slow – on tape	$15.65	$25.00

In Table 28 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 29: 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	27	$10,000.00	current K8 node
Qserv	12	128	12	$20,000.00	current qserv node
small rome	64	256	38	$13,000.00	https:/ / www.microway.com/ product/ navion-1u-amd-epyc-gpu-server/
large rome	128	512	116	$13,000.00	from Richard
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 30.

Table 30: 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* chilled water per MBTU* hours/ year * 1KW in (MBTU)	$210
** Need to add annually: maintenance for nodes – 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
Power per Rack (for Chile) Watts	14600
Approx PB per Storage Rack	8
Compute node lifetime (years)	3
Storage lifetime (years)	5
Chile Power CLP / KW-hr	105.65
Cooling Power Kw	0.7620164127

7 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 33 and Table 34 while the compute is given in Table 36 and Table 37.

7.1 Processing Plan

This model assumes the following processing:

- Precursor data (HSC RC2 and a similarly-sized DESC DC2 subset) is reprocessed each month during the Construction period using the Data Release Production (DRP).
- A large precursor reprocessing of HSC PDR2 (or equivalent) is completed twice a year. Products from one of these reprocessings will be released as Data Preview 0. One or more of these processings during Commissioning could be devoted instead to ComCam science data for Data Preview 1 or LSSTCam science data in preparation for Data Preview 2.
- Alert Production (AP) processing happens continuously as LSSTCam science images are obtained. AP hardware is purchased in FY22 to support this, presumably on a limited basis during that year and then to whatever extent possible during the first year of the survey.
- Commissioning processing of LSSTCam science images for Data Preview 2 is assumed to happen towards the end of FY22 as a single execution of the DRP. The hardware for this can be purchased early in that fiscal year.
- Annual DRP execution starts at the beginning of LSST Operations Year 2 with the processing for DR2. The hardware for each year's processing must be purchased and ready for use at the beginning of the year, so it is allocated in the tables to the prior fiscal year, when the images for that processing were taken.
- DR1 processing begins after the first 6 months of the survey; the hardware for this can be part of the DR2 purchase during FY23.

Some storage for raw data needs to be in place at the beginning of the fiscal year, but it can be ramped up over the course of the year. As a simplification it is allocated to the fiscal year in which it will be used.

7.2 Storage Model

Table 31: Inputs used to calculate storage needs

Parameters	unit	FY2020	FY2021	FY2022	FY2023	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
Number of coadd data products	number	2				deep and good-seeing
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	

7.2.1 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 presence or absence of ”heavy footprints” is assumed to be correct. The coadd storage is doubled to account for an additional ”good-seeing” coadd along with the existing ”deep” coadd.
- 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. Object storage is expected to be cheaper and more scalable for read-only data products; filesystem storage is used for data that is being generated or modified.
- Raw images and coadd images are only temporarily stored on filesystem disk and are then rapidly moved to object storage, where they are retained.
- 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.

7.2.2 Parameters

The key parameters in Table 31 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.

Two complete all-sky coadds are assumed, one for "good seeing" and one deep.

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, taken on test stands, is estimated as a ramp up to the full science cadence. The numbers of engineering (unprocessed) 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.

7.2.3 Data Product Sizing

Images and the results of processing them are the dominant factor controlling tthe storage sizing which is outlined in Table 32. 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) for the precursor data where we do not have object count estimates. For LSSTCam, we use the catalog row estimates to derive Qserv table sizes, but the Parquet file sizes are scaled based on HSC, as they may differ from the Qserv schema.

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 32: Inputs on dataset sizes used to calculate storage needs

Dataset Sizing	unit	FY2020	FY2021	FY2022	FY2023	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
Object store datasets:						
LSSTCam Raw Images	TB	319	557	1290	4816	compressed, immediate object store
LSSTCam Output Coadd Images	TB			909	7727	lossless-compressed, immediate object store
Normal disk datasets:						
Precursor Input Images	TB	549	549	549	549	HSC RC2, HSC PDR2, DC2
Precursor Output Images	TB	916	916	916	916	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 Parquet	TB			1329	7973	moves to object store
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	761	814	2003	8684	20% of total

7.2.4 Storage Sizing

Finally, storage is allocated to specific types as shown in Table 33. 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 the datasets labeled as such, including output images (initially), output catalogs, and scratch. 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 raw images, lossy-compressed output images, lossless-compressed coadd images, and output catalogs. It also accumulates from year to year. Tape is used for long-term archiving of all the raw images, all the data products (both filesystem and object store), and filesystem backups. Again, this accumulates from year to year.

Note that no replication is assumed in the object store.

Table 33: On floor LDF storage estimates based on Table 32 and Table 31

Storage Sizing (on the floor)	unit	FY2020	FY2021	FY2022	FY2023	Notes
APDB	TB	12	24	24	24	
Qserv Czar/ Object	TB			26	182	accumulates with time
Total Fast	**TB**	**12**	**24**	**50**	**206**	**SSD, sum of previous two rows**
Total Normal	**TB**	**3680**	**3748**	**9241**	**38983**	**enterprise-grade SATA**
Total Qserv Storage	**TB**	**1088**	**1088**	**585**	**4094**	**local consumer-grade SATA, accumulates with time**
LSSTCam Raw Images	TB	319	876	2166	6982	accumulates with time
LSSTCam Output Images	TB			562	3933	lossy-compressed, accumulates with time
LSSTCam Output Coadd Images	TB			909	8636	accumulates with time
LSSTCam Output Parquet	TB			1329	9302	accumulates with time
Total Object Store	**TB**	**319**	**876**	**4966**	**28854**	**consumer-grade SATA, sum of previous four rows**
LSSTCam Raw Images	TB	319	876	2166	6982	accumulates with time
All Data Products/ Backup	TB	2592	5252	13658	47626	normal storage minus Qserv/ scratch, accumulates with time
All Object Store-Only Products	TB	0	0	909	8636	accumulates with time
Total Tape	**TB**	**2910**	**6128**	**16733**	**63245**	**sum of previous three rows**

An additional table (Table 34) 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 34: On floor Chile storage estimates for Base Data Center

Chile Storage (on the floor)	unit	FY2020	FY2021	FY2022	FY2023	Notes
Fast	TB				156	SSD
Normal	TB				0	Enterprise-grade SATA
Qserv Storage	TB				4094	Local consumer-grade SATA
Object Store	TB				28854	
Tape	TB				0	

7.3 Compute Model

Table 35: 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	166				measured 83 on DECam 2Kx4K CCD
Additional AP steps	factor	1.25				DCR, real_bogus, etc.
AP compute per visit	core-hours/ visit	1.1E+01				
Qserv data/ node	TB	43.2				1 GB/ sec for 12 hours

7.3.1 Overview

This simplified computing model (Table 35) 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). For example, if a pipeline running on precursor data took an average of one hour on a 32-core nominal CPU, 32 core-hours would be used as its compute requirement. 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.

Scaling to other CPUs of the same architecture is based on the ratios of nominal GHz clock rates and core counts. For different architectures (e.g. Rome), the scaling is based on the ratio of industry-reported achievable FLOPS for the two architectures.

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 node 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.

7.3.2 Parameters

The key parameters in Table 35 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. These CCDs are from DECam, which is half the size of an LSST CCD, so the total time is doubled. 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.)

7.3.3 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 36. 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 36: Compute needs for DRP and AP

Data Release Production	**units**	**FY20**	**FY21**	**FY22**	**FY23**	**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	FY20	FY21	FY22	FY23	Notes
AP cores	cores			1,188	1,188	minimum necessary to keep up

7.3.4 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 36. 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.

7.3.5 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 37. 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 37.

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 37: Compute needs for the Science Platform instances

US DAC	**units**	**FY20**	**FY21**	**FY22**	**FY23**	**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	FY20	FY21	FY22	FY23	Notes
LSP cores	cores			26	103	20% of US DAC
Qserv nodes	nodes			14	95	
Staff LSP	units	FY20	FY21	FY22	FY23	Notes
LSP cores	cores			13	52	10% of US DAC

7.3.6 DES Comparison

As another check on the model, core-hour figures for Dark Energy Survey (DES) processing were obtained. These are given in Table 38. 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 38: 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					

7.4 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.

7.4.1 Storage in Operations

The Object, Source, and ForcedSource numbers in Table 39 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 40 are calculated using the same formulas and proportionality constants as in Table 32.

The on-the-floor storage estimates in Table 41 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 written immediately to object storage, as are Parquet-format catalogs. PVIs are lossy-compressed and placed in object storage. The complete set of raw images is available, whereas the catalogs from only the last two Data Releases and the one in preparation are kept, and the PVIs from only the last Data Release and the one in preparation are online.

Table 39: Inputs used to calculate storage needs during Operations

Parameters	unit	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10
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
Number of coadd data products	number	2	2	2	2	2	2	2	2	2	2
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 40: Dataset sizes used to calculate storage needs during Operations

Dataset Sizing	unit	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10
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
Object store datasets:											
Incremental LSSTCam Raw Images	TB	4816	4816	4816	4816	4816	4816	4816	4816	4816	4816
LSSTCam Output Coadd Images	TB	7727	7727	7727	7727	7727	7727	7727	7727	7727	7727
Normal disk datasets:											
LSSTCam Output Images	TB	13485	26970	40456	53941	67426	80911	94397	107882	121367	134852
LSSTCam Output Parquet	TB	7973	15946	23919	31893	39866	47839	55812	63785	71758	79731
Sims output	TB	5	5	5	5	5	5	5	5	5	5
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	8209	13425	18685	23933	29149	34383	39643	44927	50227	55551

Table 41: On floor LDF storage estimates during Operations

LDF Storage (on the floor)	**unit**	**LOY1**	**LOY2**	**LOY3**	**LOY4**	**LOY5**	**LOY6**	**LOY7**	**LOY8**	**LOY9**	**LOY10**
APDB	TB	24	24	24	24	24	24	24	24	24	24
Qserv Czar/ Object	TB	182	347	562	643	711	774	835	894	951	1006
Total Fast	**TB**	**206**	**371**	**586**	**667**	**735**	**798**	**859**	**918**	**974**	**1029**
Normal	TB	38983	67982	99544	131031	162327	193733	225294	256997	288794	320737
Qserv Storage	TB	4094	9257	17275	24144	31277	38734	46555	54716	63206	72017
LSSTCam Raw Images	TB	6982	11798	16614	21430	26246	31062	35878	40694	45510	50326
LSSTCam Output Images	TB	3933	10676	16857	23599	30342	37084	43827	50570	57312	64055
LSSTCam Output Coadd Images	TB	8636	16364	23182	23182	23182	23182	23182	23182	23182	23182
LSSTCam Output Parquet	TB	9302	25248	47839	71758	95678	119597	143516	167436	191355	215275
Object Store	TB	28854	64086	104491	139969	175447	210925	246403	281881	317359	352837
LSSTCam Raw Images	TB	6982	11798	16614	21430	26246	31062	35878	40694	45510	50326
All Data Products/ Backup	TB	47626	106973	194238	309260	451705	621594	818951	1043800	1296157	1576046
All Object Store-Only Products	TB	8636	16364	24091	31818	39545	47273	55000	62727	70455	78182
Tape	TB	63245	135135	234943	362508	517497	699929	909829	1147221	1412122	1704554

Table 42: On floor Chile storage estimates during Operations

Chile Storage (on the floor)	**unit**	**LOY1**	**LOY2**	**LOY3**	**LOY4**	**LOY5**	**LOY6**	**LOY7**	**LOY8**	**LOY9**	**LOY10**	
Fast	TB	335	570	825	1064	1274	1484	1694	1904	2114	2324	
Qserv Storage	TB	4094	9257	17275	24144	31277	38734	46555	54716	63206	72017	
Object Store	TB	8656	19226	31347	41991	52634	63278	73921	84564	95208	105851	

Table 43: Compute needs during Operations

Data Release Production	**units**	**LOY1**	**LOY2**	**LOY3**	**LOY4**	**LOY5**	**LOY6**	**LOY7**	**LOY8**	**LOY9**	**LOY10**
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	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10
AP cores	cores	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188
US DAC	units	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10
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	216	309	348	364	451	436	408	367	418
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	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10
LSP cores	cores	103	187	280	373	466	560	653	746	840	933
Qserv nodes	nodes	95	216	309	348	364	451	436	408	367	418
Staff LSP	units	LOY1	LOY2	LOY3	LOY4	LOY5	LOY6	LOY7	LOY8	LOY9	LOY10
LSP cores	cores	52	93	140	187	233	280	326	373	420	466

All data products and new raw images for each Data Release are copied to tape, but scratch space and the Qserv-schema catalogs are not.

Table 42 extracts the Qserv and object store sizing needed to populate the Chilean DAC with a copy of the data products and raw images.

7.4.2 Compute in Operations

The DRP compute sizing in Table 43 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 http://dmtn-113.lsst.io

[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-091.lsst.io`, LSST Data Management Technical Note

B Acronyms

Acronym	Description
AMD	Advanced Micro Devices
AP	Alert Production
APDB	Alert Production DataBase
CCD	Charge-Coupled Device
CLP	Chilean Peso
CPU	Central Processing Unit
ComCam	The commissioning camera is a single-raft, 9-CCD camera that will be installed in LSST during commissioning, before the final camera is ready.
DAC	Data Access Center
DC2	Data Challenge 2 (DESC)
DCR	Differential Chromatic Refraction
DDN	Data Delivery Network
DES	Dark Energy Survey
DESC	Dark Energy Science Collaboration
DM	Data Management
DMTN	DM Technical Note
DP	Data Production
DPDD	Data Product Definition Document
DR1	Data Release 1
DRP	Data Release Production
FLOP	FLoating point Operation
FLOPS	FLoating point Operation per Second
FY21	Financial Year 21
GB	Gigabyte
GFLOPS	Giga FLOP per Second
GPFS	General Parallel File System (now IBM Spectrum Scale)
HSC	Hyper Suprime-Cam

IN2P3	Institut National de Physique Nucléaire et de Physique des Particules
KW	Kilowatt
LATISS	LSST Atmospheric Transmission Imager and Slitless Spectrograph
LCR	LSST Change Request
LDF	LSST Data Facility
LDM	LSST Data Management (Document Handle)
LSE	LSST Systems Engineering (Document Handle)
LSP	LSST Science Platform (now Rubin Science Platform)
LSST	Legacy Survey of Space and Time (formerly Large Synoptic Survey Telescope)
MB	MegaByte
MBTU	Mega British Thermal Unit
MOF	Multi-Object Multi-Band Fitting
NCSA	National Center for Supercomputing Applications
NSF	National Science Foundation
NVMe	Non Volatile Memory Express
PB	PetaByte
PCI	Peripheral Component Interconnect
PDR	Preliminary Design Review
PDR1	Public Data Release 1 (HSC)
PDR2	Public Data Release 2 (HSC)
QA	Quality Assurance
RAM	Random Access Memory
RFC	Request For Comment
S3	(Amazon) Simple Storage Service
SATA	Serial Advanced Technology Attachment
SLAC	SLAC National Accelerator Laboratory
SOF	Single-Object Fitting
SQuaSH	Science Quality Analysis Harness
SSD	Solid-State Disk
SSP	Solar System Processing
TB	TeraByte
UKDF	United Kingdom Data Facility
US	United States

USDF	United States Data Facility
VM	Virtual Machine
deg	degree; unit of angle