Large Synoptic Survey Telescope (LSST)
Data Management Infrastructure Design

Mike Freemon and Jeff Kantor

LDM-129

Latest Revision: August 15, 2011
Change Record

Version
Date
Description
Owner name

1.0
7/13/2011
Initial version as an assembled document; previous material was distributed.
Mike Freemon and Jeff Kantor
The LSST Data Management Infrastructure Design

1. Overview

The Data Management Infrastructure is composed of all computing, storage, and communications hardware and systems software, and all utility systems supporting it, that form the platform of execution and operations for the DM System. All DM System Applications and Middleware are developed, integrated, tested, deployed, and operated on the DM Infrastructure.

This document describes the design of the DM Infrastructure at the highest level of discussion. It is the “umbrella” document over many other referenced documents that elaborate on the design in greater detail.

![Diagram of 3-layered architecture of the Data Management System]

The DM System is distributed across four sites in the United States and Chile. Each site hosts one or more physical facilities, in which reside DM Centers. Each Center performs a specific role in the operational system.
The Base Center is in the Base Facility on the AURA compound in La Serena, Chile. The primary role of the Base Center is:

- Alert Production processing to meet the 60 second latency requirement
- Data Access

The Archive Center is in the National Petascale Computing Facility at NCSA in Champaign, IL. The primary role of the Archive Site is:

- Data Release Production processing
- Data Access

Both sites have copies of all the raw and released data for data access and disaster recovery purposes.
The Base and Archive Sites host the respective Base and Archive Centers, plus a co-located Data Access Center (DAC).

The Headquarters Site final location is not yet determined, but for planning and design purposes is assumed to be in Tucson Arizona. While the Base and Archive Sites provide large-scale data production and data access at supercomputing center scale, the Headquarters is only a management and supervisory control center for the DM System, and as a result is much more modest in terms of infrastructure.

2. Infrastructure Components

The Infrastructure is organized into components, which are each composed of hardware and software integrated and deployed as an assembly, to provide computing, communications, and/or storage capabilities required by the DM System. The components are named according to this role. By convention, each infrastructure component is associated with the main center (Archive or Base Center) unless it is part of the co-located DAC. The infrastructure components specific to the Data Access Center are:

- L3 Community Scratch
- L3 Community Images
- L3 Community Compute
- L3 Community Database
- Query Access Database
- Cutout Service
Both the Base Center and the Archive Center have essentially the same architecture, differing only by capacity and quantity. There are different external network interfaces depending on the site.

<table>
<thead>
<tr>
<th>Category</th>
<th>Archive Site</th>
<th>Base Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teraflops (sustained)</td>
<td>175 → 1000 (required)</td>
<td>30 → 45 (required)</td>
</tr>
<tr>
<td></td>
<td>270 → 1500 (on-the-floor)</td>
<td>50 → 30 (on-the-floor)</td>
</tr>
<tr>
<td>Nodes</td>
<td>1700 → 1400 (on-the-floor)</td>
<td>800 → 60 (on-the-floor)</td>
</tr>
<tr>
<td>Cores</td>
<td>70K → 280K (on-the-floor)</td>
<td>12K → 13K (on-the-floor)</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>25 → 130 TB/s (required)</td>
<td>4 → 6 TB/s (required)</td>
</tr>
<tr>
<td>File I/O Bandwidth</td>
<td>220 → 425 GB/s (required)</td>
<td>25 → 30 GB/s (required)</td>
</tr>
<tr>
<td>Storage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity</td>
<td>13 → 25 PB (required)</td>
<td>8 → 3 PB (required)</td>
</tr>
<tr>
<td></td>
<td>19 → 100 PB (on-the-floor)</td>
<td>12 → 23 PB (on-the-floor)</td>
</tr>
<tr>
<td>Drives</td>
<td>1500 → 1100 (on-the-floor)</td>
<td>950 → 975 (on-the-floor)</td>
</tr>
<tr>
<td>Disk Bandwidth</td>
<td>220 → 425 GB/s (required)</td>
<td>27 → 31 GB/s (required)</td>
</tr>
<tr>
<td>Mass Storage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity</td>
<td>7 → 75 PB (required)</td>
<td>7 → 75 PB (required)</td>
</tr>
<tr>
<td></td>
<td>8 → 83 PB (on-the-floor)</td>
<td>8 → 83 PB (on-the-floor)</td>
</tr>
<tr>
<td>Tapes</td>
<td>1000 → 3800 (near line)</td>
<td>1000 → 3800 (near line)</td>
</tr>
<tr>
<td></td>
<td>1000 → 3800 (offsite)</td>
<td></td>
</tr>
<tr>
<td>Database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teraflops (sustained)</td>
<td>16 → 199 PB (required)</td>
<td>12 → 128 PB (required)</td>
</tr>
<tr>
<td></td>
<td>16 → 199 PB (on-the-floor)</td>
<td>12 → 128 PB (on-the-floor)</td>
</tr>
<tr>
<td>Storage Capacity</td>
<td>4 → 50 PB (required)</td>
<td>3 → 40 PB (required)</td>
</tr>
<tr>
<td></td>
<td>10 → 125 PB (on-the-floor)</td>
<td>7 → 95 PB (on-the-floor)</td>
</tr>
<tr>
<td>Disk Drives</td>
<td>1400 → 1600 (on-the-floor)</td>
<td>1000 → 2000 (on-the-floor)</td>
</tr>
<tr>
<td>Disk Bandwidth (sequential)</td>
<td>125 → 625 GB/s (required)</td>
<td>95 → 425 GB/s (required)</td>
</tr>
<tr>
<td>Database Nodes</td>
<td>100 → 150 (on-the-floor)</td>
<td>80 → 130 (on-the-floor)</td>
</tr>
<tr>
<td>Facilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor space</td>
<td>950 → 850 sq ft (1100 high water mark)</td>
<td>460 → 350 sq ft (470 high water mark)</td>
</tr>
<tr>
<td>Power</td>
<td>720 → 640 kW (880 high water mark)</td>
<td>180 → 110 kW (210 high water mark)</td>
</tr>
<tr>
<td>Cooling</td>
<td>2.4 → 2.1 mmbtu (3.0 high water mark)</td>
<td>0.6 → 0.4 mmbtu (0.7 high water mark)</td>
</tr>
</tbody>
</table>

Table: DM Infrastructure Capacities and Quantities

The capacities and quantities are derived from the scientific, system, and operational requirements via a detailed sizing model. The complete sizing model and the process used to arrive at the infrastructure is available in the LSST Project Archive.

This design assumes that the DM System will be built using commodity parts that are not bleeding edge, but rather have been readily available on the market for one to two years. This choice is intended to lower the risk of integration problems and the time to build a working, production-level system. This also defines a certain cost class for the computing platform that can be described in terms of technology available today. We then assume that we will be able to purchase a system in 2020 in this same cost class with the same number of dollars (ignoring inflation); however, the performance of that system will be greater than the corresponding system purchased today by some performance evolution curve factor.

Note that the current baseline for power, cooling, and floor space assumes air-cooled equipment. If the sizing model or technology trends change and we find that flops-per-watt is the primary constraint in our system design, we will evaluate water-cooled systems.
Finally, note that Base Site equipment is purchased in the U.S., and delivered by the equipment vendors to the Archive Site in Champaign, IL. NCSA installs, configures, and tests the Base Site equipment before shipping to La Serena. When new Data Release data has been generated and validated, the new DRP data is loaded onto the disk storage destined for La Serena.

3. Facilities

This section describes the operational characteristics of the facilities in which the DM infrastructure resides.

1. National Petascale Computing Facility, Champaign, IL, US

The National Petascale Computing Facility (NPCF) is a new data center facility on the campus of the University of Illinois. It was built specifically to house the Blue Waters system, but will also host the LSST Data Management systems. The key characteristics of the facility are:

- 24MW of power (1/4 of campus electric usage)
- 5900 tons of CHW cooling
- F 3 tornado & Seismic resistant design
- NPCF is expected to achieve LEED Gold certification, a benchmark for the design, construction, and operation of green buildings.
- NPCF’s forecasted power usage effectiveness (PUE) rating is an impressive 1.1 to 1.2, while a typical data center rating is 1.4. PUE is determined by dividing the amount of power entering a data center by the power used to run the computer infrastructure within it, so efficiency is greater as the quotient decreases toward 1.
- Three on-site cooling towers will provide water chilled by Mother Nature about 70 percent of the year.
- Power conversion losses will be reduced by running 480 volt AC power to compute systems.
- The facility will operate continually at the high end of the American Society of Heating, Refrigerating and Air-Conditioning Engineers standards, meaning the data center will not be
overcooled. Equipment must be able to operate with a 65F inlet water temperature and a 78F inlet air temperature.

- Provides 1 or 10 gigabit high-performance Ethernet connections as required with up to 300 gigabit external network.
- There is no UPS in the PCF. LSST will install rack-based UPS systems to keep systems running during brief power outages and to automatically manage controlled shutdowns when extended power outages occur. This ensures that file system buffers are flushed to disk to prevent any data loss.

The fire suppression system at the NPCF is a double action water system. The first triggering event loads the sprinklers with water and pressurizes it in the system. The water is not released unless the second trigger occurs.

### 2. NOAO Facility, La Serena, Chile

NOAO is expanded their facility in La Serena, Chili, in order to accommodate the LSST project. Refer to the Base Site design in the Telescope and Site Subsystem for more detail.
3. Floorspace, Power, and Cooling

The following table shows the facilities usage by the LSST Data Management System over the survey period. This does not include any extra space that might be needed during the process of transitioning replacement equipment or staging of Base Site equipment at the Archive Site.
Table: Floorspace, power, and cooling estimates for the Data Management System.

<table>
<thead>
<tr>
<th></th>
<th>Archive Site</th>
<th>Base Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floorspace (sq ft)</td>
<td>950 → 850 (1100 hwm)</td>
<td>460 → 350 (470 hwm)</td>
</tr>
<tr>
<td>Power (kW)</td>
<td>720 → 640 (880 hwm)</td>
<td>180 → 110 (210 hwm)</td>
</tr>
<tr>
<td>Cooling (mmbtu)</td>
<td>2.4 → 2.1 (3.0 hwm)</td>
<td>0.6 → 0.4 (0.7 hwm)</td>
</tr>
</tbody>
</table>

Figure: Floorspace needed by the Data Management System over the survey period.

Figure: Power and cooling required by the Data Management System over the survey period.
4. Computing

This section defines the design for the computing resources at the Centers. In general, hardware is purchased in the year before it is needed in order to leverage price/performance improvements.

A special situation occurs for the Commissioning phase of Construction. In 2018, we acquire and install the computing infrastructure needed to support Commissioning, for which we use the same sizing as that for the first year of Operations.

<table>
<thead>
<tr>
<th></th>
<th>Archive Site</th>
<th>Base Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teraflops (sustained)</td>
<td>175 → 1000 (required) 270 → 1600 (on-the-floor)</td>
<td>30 → 45 (required) 50 → 80 (on-the-floor)</td>
</tr>
<tr>
<td>Nodes</td>
<td>1700 → 1400 (on-the-floor)</td>
<td>300 → 60 (on-the-floor)</td>
</tr>
<tr>
<td>Cores</td>
<td>70K → 280K (on-the-floor)</td>
<td>12K → 13K (on-the-floor)</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>25 → 130 TB/s (required)</td>
<td>4 → 6 TB/s (required)</td>
</tr>
<tr>
<td>File I/O Bandwidth</td>
<td>120 → 425 GB/s (required)</td>
<td>25 → 30 GB/s (required)</td>
</tr>
</tbody>
</table>

Table: Compute sizing for the Data Management System.
Figure: The growth of compute requirements over the survey period.
Figure. The number of compute nodes on-the-floor over the survey period.
5. Storage

Image storage will be controller-based storage in a RAID6 8+2 configuration for protection against individual disk failures. GPFS is the parallel file system.

<table>
<thead>
<tr>
<th></th>
<th>Archive Site</th>
<th>Base Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>13 → 25 PB (required) 19 → 100 PB (on-the-floor)</td>
<td>8 → 8 PB (required) 12 → 23 PB (on-the-floor)</td>
</tr>
<tr>
<td>Drives</td>
<td>1500 → 1100 (on-the-floor)</td>
<td>950 → 275 (on-the-floor)</td>
</tr>
<tr>
<td>Disk Bandwidth</td>
<td>120 → 425 GB/s (required)</td>
<td>27 → 31 GB/s (required)</td>
</tr>
</tbody>
</table>

Table: Image file storage sizing for the Data Management System.
GPFS was chosen as the baseline for the parallel filesystem implementation based upon the following considerations:

- NCSA has and will continue to have deep expertise in GPFS, HPSS, and GHI.
- NCSA is conducting extensive scaling tests with GPFS and any potential problems that emerge at high loads will be solved by the time LSST going into Operations.
- LSST gets special pricing due to the University of Illinois’ campus licensing agreement with IBM. These prices are quite favorable and even at the highest rates are lower than NCSA can currently get for equivalent Lustre service.
- NCSA provides level 1 support for all UIUC campus licenses under the site licensing agreement.
- Choice of parallel filesystem implementation is transparent to users of LSST.

Figure: Storage for LSST Image Products. The green shows the mass storage disk cache, a key element for the GPFS-HPSS integration that creates a transparent hierarchical storage environment.
6. Mass Storage

The mass storage system will be HPSS. The GPFS-HPSS Interface (GHI) is used to create a hierarchical storage system.

All client interaction (meaning both processing and people) is with the single GPFS namespace.

The mass storage system is completely transparent to all clients.

The mass storage system at the Archive Site will write data to dual tapes, with one going offsite for safe keeping. The Base Site will write a single copy.

At Year 5 during Operations, a new tape library system will be purchased to replace the existing library equipment.

The sizing for mass storage is the same at both the Archive Site and the Base Site. It starts at 7 PB for survey year 1, and grows to 75 PB by the end of the 10-year survey. There will be nearly 4000 tapes used during year 10.
7. Databases

The relational database catalogs are implemented with qserv, an approach similar to the map-reduce approach in architecture, but applied to processing sql queries. The database storage is provided via local disk drives within the database servers themselves. See Document-11625 for additional information.

![Qserv Database Infrastructure](image)

**Figure**: The Qserv Database Infrastructure

<table>
<thead>
<tr>
<th>Qserv-Based Databases</th>
<th>Archive Site</th>
<th>Base Site</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Teraflops (sustained)</td>
<td>Archive Site</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16 → 193 PB (required)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16 → 199 PB (on-the-floor)</td>
</tr>
<tr>
<td></td>
<td>Storage Capacity</td>
<td>4 → 50 PB (required)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 → 128 PB (on-the-floor)</td>
</tr>
<tr>
<td></td>
<td>Disk Drives</td>
<td>1400 → 2600 (on-the-floor)</td>
</tr>
<tr>
<td></td>
<td>Disk Bandwidth (sequential)</td>
<td>125 → 625 GB/s (required)</td>
</tr>
<tr>
<td></td>
<td>Database Nodes</td>
<td>100 → 190 (on-the-floor)</td>
</tr>
<tr>
<td></td>
<td>Disk Drives per Node</td>
<td>15 → 25 (45 hwm)</td>
</tr>
</tbody>
</table>

**Table**: Database sizing for the Data Management System.
8. Additional Support Servers

There are a number of additional support servers in the LSST DM computing environment. They include:

- User Data Access – login nodes, portals
- Pipeline Support - Condor, Activemq Event Brokers
- Inter-Site Data Transfer
- Network Security Servers (NIDS)
9. Cluster Interconnect and Local Networking

The local network technologies will be a combination of 10GigE and InfiniBand.

10GigE will be used for the external network interfaces (i.e. external to the DM site), user access servers and services (e.g. web portals, VOEvent servers), mass storage (due to technical limitations), and the Long Haul Network (see the next section). 10GigE is ubiquitous for these uses and is a familiar and known technology.

InfiniBand will be used as the cluster interconnect for intra-node communication within the compute cluster, as well as to the database servers. It will also be the storage fabric for the image data. InfiniBand provides the low-latency communication we need at the Base Site for the MPI-based alert generation processing to meet the 60-second latency requirements, as well as for the storage I/O performance we need at the Archive Site for Data Release Production. By using InfiniBand in this way, we can avoid buying, implementing, and supporting the more expensive Fibre Channel for the storage fabric.

Although consolidation of networking fabric expected, we do not yet assume this for the baseline design.
The crosstalk-corrected images from the summit will flow directly into the memory on the alert production compute nodes at the Base Site in order to meet the 60-second alert generation latency requirement. The networking to support this will consist of dedicated switching equipment to establish direct connectivity from the summit equipment to the compute cluster at La Serena. There is an interface control document (ICD) that defines the responsibilities between the camera team and the data management team.
10. Long Haul Network

The communication link between Summit and Base will be 100 Gbps.

The network between the Base Site in La Serena, and the Archive Site in Champaign, IL, will support 10 Gbps minimum, 40 Gbps during the night hours, and 80 Gbps burst capability in the event we have a service interruption and need to “catch up”.

The key features of the network plan are:

- Mountain summit – Base is only new fiber, 100 Gbps capacity
- Inter-site Long-Haul links on existing fibers, protected circuits
- LSST is leveraging and driving US - Chile long-haul network expansion
- Capacity growth supports construction and commissioning
- 1 Gb/s 2011, 3 Gb/s 2018, 10-40-80 Gb/s 2019
- Equipment is available today at budgeted cost

Additional information can be found in the Network Design Document, LSE-78.
Images crosstalk-corrected at the Summit are sent to the Base Site and fed directly into the memory of Alert Production processing. Those images are processed immediately and the resulting alerts are provided to the VOEvent network in La Serena within 60 seconds of shutter close.

Raw image data is transmitted to the Archive Site. Neither the crosstalk-corrected images, nor the products from Alert Production at the Base Site are transferred to the Archive Site. The Archive Site reprocesses raw images. The computing demands for this reprocessing at the Archive Site are far less than at the Base Site since there is no 60-second latency requirement at the Archive Site.

As noted earlier in this document, Base Site hardware is purchased at the Archive Site and shipped to the Base Site. Annual Data Release Production Products will be loaded on the new disk hardware purchased at the Archive Site before it is shipped to the Base Site. This forms a type of “sneakernet” to get the Data Release Products to the Base Site.
11. Policies

A just-in-time approach for purchasing hardware is used to leverage the fact that hardware prices get cheaper over time. This also allows for the use of the latest features of the hardware if valuable to the project.

We buy in the fiscal year before the need occurs so that the infrastructure is installed, configured, tested, and ready to go when needed. There is also a ramp up of the initial computing infrastructure for the Commissioning phase of Construction.

Shown in this section are various policies that we implement for the DM computing infrastructure. Additional supporting discussion is contained within document LDM-143.

1. Replacement Policy

- Compute Nodes: 5 Years
- Disk Drives: 3 Years
- Tape Media: 5 Years
- Tape Drives: 3 Years
- Tape Library System: Once at Year 5

2. Storage Overheads

- RAID: 20%
- Filesystem: 10%

3. Spares (hardware failures)

- Compute Nodes: 3% of nodes
- Disk Drives: 3% of drives
- Tape Media: 3% of tapes

4. Extra Capacity

- Disk: 10% of TB
- Tape: 10% of TB

5. Additional Margins

- Compute algorithms: 50% of TF
12. Disaster Recovery

Mass storage is used at both sites to ensure the safe keeping of data products. At the Archive Site, the mass storage system will write two copies of all data to different media. One set of media stays in the tape library for later recall as needed. The second copy is transported off-site. This protects against both media failures (e.g. bad tapes) and loss of the facility itself.

The Base Site will write a single copy of data to tape, which remains near-line in the tape library system.

Either Site can be the source of data for recovery of the other Site.

13. CyberSecurity

We leverage best practices to ensure a secure computing environment. This includes monitoring such as the use of intrusion detection systems, partitioning of resources such as segregating the L3 compute nodes from the core DM processing nodes, and limiting the scope of authorizations to only that which is needed.

LSST has an open data policy. The primary data deliverables of LSST Data Management are made available to the public without any proprietary period.

As a result, the central considerations are when applying security policies are not about the theft of L1 and L2 data. The main considerations are:

- Data Protection
- Data Integrity
- Misuse of Facility
- L3 Community Data

Refer to LSE-99 for additional information. This is a LSST system-wide document, not just DM, as cybersecurity reaches across to all of the LSST subsystems.