



LARGE SYNOPTIC SURVEY TELESCOPE

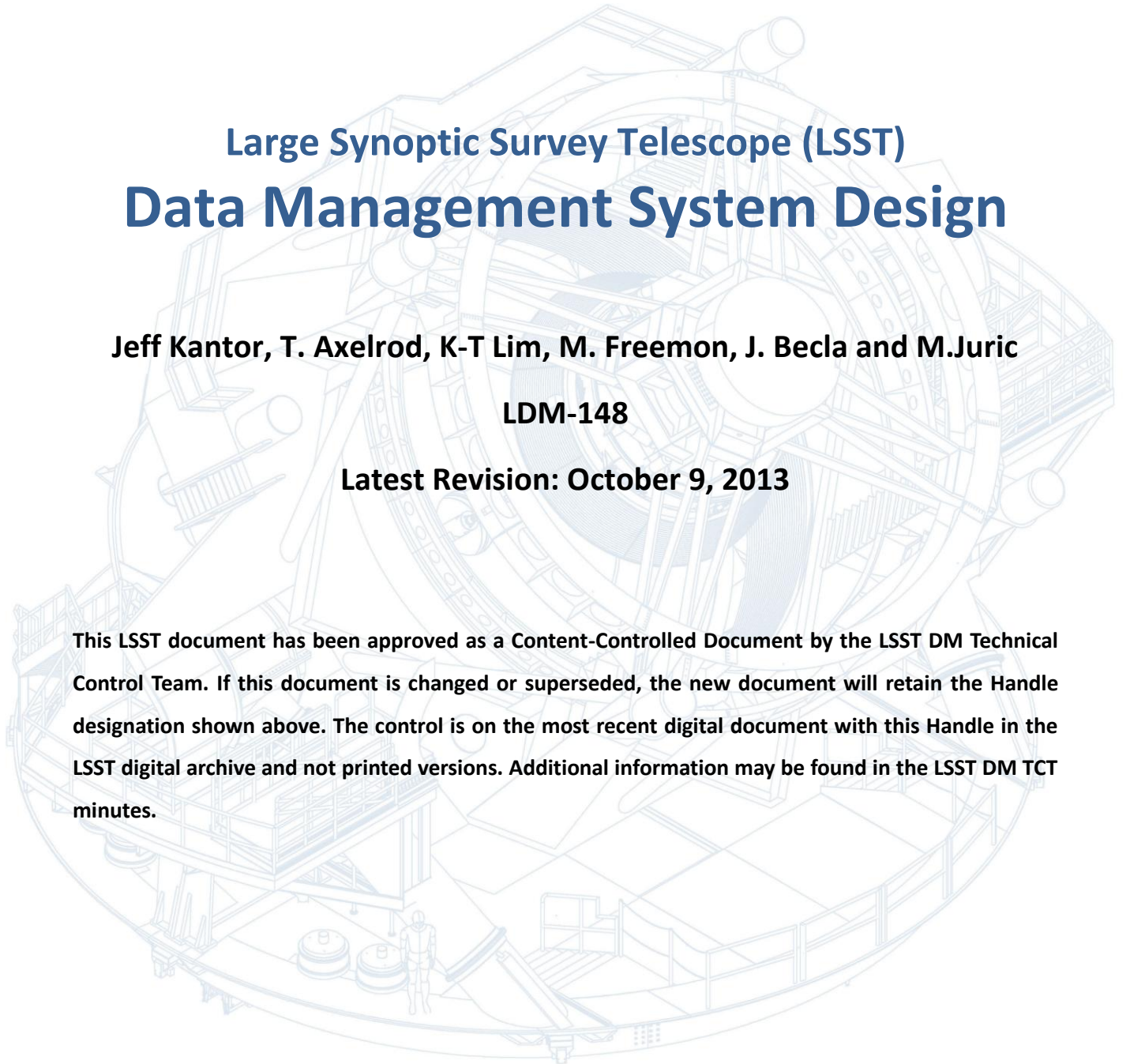
# Large Synoptic Survey Telescope (LSST) Data Management System Design

Jeff Kantor, T. Axelrod, K-T Lim, M. Freemon, J. Becla and M. Juric

LDM-148

Latest Revision: October 9, 2013

This LSST document has been approved as a Content-Controlled Document by the LSST DM Technical Control Team. If this document is changed or superseded, the new document will retain the Handle designation shown above. The control is on the most recent digital document with this Handle in the LSST digital archive and not printed versions. Additional information may be found in the LSST DM TCT minutes.



## Change Record

Version	Date	Description	Owner name
Document 2082, v1	7/24/06	Initial version	Jeffrey Kantor
Document 2082, v4	8/16/06	Updated database schema	Jeffrey Kantor
Document 2082, v9	8/22/06	Updated for review, re-organized	Jeffrey Kantor
Document 2082, v11	9/8/06	Updated query numbers and summary chart	Jeffrey Kantor
Document 2464, v3	11/2/06	Reworked into reference design detail appendix for MREFC proposal 2007	Jeffrey Kantor
Document 3859, v13	9/10/07	Reworked for Concept Design Review	Jeffrey Kantor
Document 10291, v1	11/23/10	Reworked into Data Management section of MREFC proposal 2011	Jeffrey Kantor
LDM-148 V2	8/9/11	Copied into LDM-148 handle, reformatted	Robert McKercher
LDM-148, v3	8/15/11	Updated for Preliminary Design Review	Tim Axelrod, K-T Lim, Mike Freemon, Jeffrey Kantor
V4	10/9/13	Updated for Final Design Review	Mario Juric, K-T Lim, Jeffrey Kantor



## Table of Contents

**Change Record ..... i**

**1 Introduction to the Data Management System (DMS) ..... 1**

**2 Requirements and Design Implications ..... 2**

**3 Design of the Data Management System..... 13**

**3.1 Application Layer Design .....15**

**3.2 Middleware Layer Design .....17**

**3.3 Infrastructure Layer Design .....21**

# The LSST Data Management System Design

## 1 Introduction to the Data Management System (DMS)

The data management challenge for the LSST Observatory is to provide fully calibrated public data to the user community to support the frontier science described in the LSST Science Requirements Document, while simultaneously enabling new lines of research not anticipated today. The nature, quality, and volume of LSST data will be unprecedented, so the DMS design features petascale storage, terascale computing, and gigascale communications. The computational facility and data archives of the LSST DMS will rapidly make it one of the largest and most important facilities of its kind in the world. New algorithms will have to be developed and existing approaches refined in order to take full advantage of this resource, so "plug-in" features in the DMS design and an open data-open source software approach enable both science and technology evolution over the decade-long LSST survey. To tackle these challenges and to minimize risk, LSSTC and the data management (DM) subsystem team have taken four specific steps:

- 1) The LSST has put together a highly qualified data management team. The National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign will serve as the primary LSST archive center and provide infrastructure and high-performance process control middleware; the National Optical Astronomy Observatory (NOAO) will, along with international network collaborators FIU/AmLight and REUNA, provide the long-haul networks: Princeton University, the University of Washington, and the Infrared Processing and Analysis Center (IPAC) at the California Institute of Technology will provide astronomical software, including user interfaces; and SLAC National Accelerator Laboratory will provide data access and processing services software. Google Corporation, Johns Hopkins University, and a variety of technology providers are providing advice and collaborating on DMS architecture and sizing.
- 2) The DM team has closely examined and continues to monitor relevant developments within (and in some cases recruited key members from) precursor astronomical surveys including 2MASS (Skrutskie et al. 2006), SDSS (York et al. 2000), Pan-STARRS (Kaiser et al. 2002), the Dark Energy Survey (DES, Abbott et al. 2005), the ESSENCE (Miknaitis et al. 2007), CFHT-LS (Boulade 2003, Mellier 2008) and SuperMACHO (Stubbs et al. 2002) projects, the Deep Lens Survey (DLS, Wittman et al. 2002), and the NVO/VAO.
- 3) The DM team is monitoring relevant data management developments within the physics community and has recruited members with experience in comparable high-energy physics data management systems. Specifically, members were recruited from the BaBar collaboration at SLAC (Aubert et al. 2002) who are experts in large-scale databases.
- 4) The DM team has defined and executed eight successive "Data Challenges" (DC) to validate the DMS architecture. Success with these challenges prior to the beginning of construction provides confidence that the scientific and technical performance targets will be achieved.

## 2 Requirements and Design Implications

The LSST Data Management System (DMS) will deliver:

- **Image archives:** An image data archive of over two million raw and calibrated scientific images, each image available within 24 hours of capture.
- **Alerts:** Verified alerts of transient and moving objects detected within 60 seconds of capture of a pair of images in a visit.
- **Catalogs:** Astronomical catalogs containing billions of stars and galaxies and trillions of observations and measurements of them, richly attributed in both time and space dimensions, and setting a new standard in uniformity of astrometric and photometric precision.
- **Data Access Resources:** Web portals, query and analytical toolkits, an open software framework for astronomy and parallel processing, and associated computational, storage, and communications infrastructure for executing LSST-developed and external scientific codes.
- **Quality Assurance:** A process including automated metrics collection and analysis, visualization, and documentation to ensure data integrity with full tracking of data provenance.

All data products will be accessible via direct query or for fusion with other astronomical surveys. The user community will vary widely. From students to researchers, users will be processing up to petabyte-sized sections of the entire catalog on a dedicated supercomputer cluster or across a scientific grid. The workload placed on the system by these users will be actively managed to ensure equitable access to all segments of the user community. LSST key science deliverables will be enabled by providing computing resources co-located with the raw data and catalog storage. Figure 1 shows the content of the data products and the cadence on which they are generated. The data products are organized into three groups, based largely on where and when they are produced.

Level 1 data products are intended to enable detection and follow-up of time-sensitive time-domain events. They are generated by pipeline processing of the stream of data from the camera subsystem during normal observing. Level 1 data products are therefore continuously generated and updated every observing night. This process is of necessity highly automated and must proceed with absolutely minimal human interaction. In addition to science data products, a number of Level 1 Science Data Quality Assessment (SDQA) data products are generated to assess quality and to provide feedback to the observatory control system (OCS).

Level 2 data products are generated as part of Data Releases, which are required to be performed at least yearly, and will be performed more frequently for the first year of survey data. Level 2 products include data products for which extensive computation is required, often because they combine information from many exposures. Although the steps that generate Level 2 products will in general be automated, significant human interaction will be required at key points to ensure their quality.

Scientific users create Level 3 data products from Level 1 and Level 2 data products, to support particular science goals, often requiring the combination of LSST data across significant areas on the sky. The DMS is required to facilitate the creation of Level 3 data products, for example by providing suitable software, APIs, and computing infrastructure, but will not itself create any Level 3 data product. Instead these data products will be created externally to the DMS, using software written by

researchers, e.g., science collaborations. Once created, Level 3 data products may be associated with Level 1 and Level 2 data products through database federation. The LSST Project, with the agreement of the Level 3 creators, may decide to incorporate user-provided Level 3 data product pipelines into the DMS production flow, thereby promoting them to Level 2 data products.

Level 1 and Level 2 data products that have passed quality control tests must be made promptly accessible without restriction to U.S. and Chilean communities, as well as International Partners. Additionally, the source code used to generate them will be made available under an open-source license, and LSST will provide support for builds on selected platforms. The access policies for Level 3 data products will be product- specific and source-specific. In some cases Level 3 products may be proprietary for some time.

	IMAGES CATEGORY (FILES)	CATALOG CATEGORY (DATABASE)	ALERT CATEGORY (DATABASE)
<b>NIGHTLY</b> Level 1	Raw science images Calibrated science images Subtracted science images Data quality analysis	Catalog of sources (detections) found on difference images Catalog of objects found on difference images Orbit catalog Data quality analysis	Transient alert Data quality analysis  (within 60 seconds)
<b>DATA RELEASE</b> (Annual) Level 2	Stacked science images Calibration images RGB JPEG images Data quality analysis	Source (detection) catalog from calibrated science images Object catalog from optimally measured properties Data quality analysis	Alert statistics and summaries based on annual reprocessing Data quality analysis

Figure 1 Key deliverable Data Products and their production cadence.

The system that produces and manages the LSST archive must be robust enough to keep up with the LSST's prodigious data rates and will be designed to minimize the possibility of data loss. This system will be initially constructed and subsequently refreshed using commodity hardware to ensure affordability, even as technology evolves.

The principal functions of the DMS are to:

- Process the incoming stream of images generated by the camera system during observing by archiving raw images, generating transient alerts, and updating difference source and object catalogs.
- Periodically process the accumulated survey data to provide a uniform photometric and astrometric calibration, measure the properties of fainter objects, and classify objects based on

their time-dependent behavior. The results of such a processing run form a data release (DR), which is a static, self-consistent data set for use in performing scientific analysis of LSST data and publication of the results. All data releases are archived for the entire operational life of the LSST archive.

- Periodically create new calibration data products, such as bias frames and flat fields, to be used by the other processing functions.
- Make all LSST data available through an interface that utilizes, to the maximum possible extent, community-based standards such as those being developed by the Virtual Astronomy Observatory (VAO) in collaboration with the International Virtual Observatory Alliance (IVOA). Provide enough processing, storage, and network bandwidth to enable user analysis of the data without petabyte-scale data transfers.

The Data Management Applications Design Document (LDM-151) provides a more detailed description of the data products and the processing that produces them.

All data products will be documented by a record of the full processing history (data provenance) and a rich set of metadata describing their "pedigree." A unified approach to data provenance enables a key feature of the DMS design: data storage space can be traded for processing time by recreating derived data products when they are needed instead of storing them permanently. This trade can be shifted over time as the survey proceeds to take advantage of technology trends, minimizing overall costs. All data products will also have associated data quality metrics that are produced on the same cadence as the data product, as well as metrics produced at later stages that assess the data product in the context of a larger set of data. These metrics will be made available through the observatory control system for use by observers and operators for schedule optimization and will be available for scientific analysis as well.

The latency requirements for alerts determine several aspects of the DMS design and overall cost. An alert is triggered by an unexpected excursion in brightness of a known object or the appearance of a previously undetected object such as a supernova or a GRB. The astrophysical time scale of some of these events may warrant follow-up by other telescopes on short time scales. These excursions in brightness must be recognized by the pipeline, and the resulting alert data product sent on its way, within 60 seconds. This drives the DMS design in the decision to acquire high-bandwidth/high-reliability long-haul networking from Chile to the U.S., performing cross-talk correction on the data in the data acquisition system, and parallelizing the alert processing at the amplifier and CCD levels, where possible.

Finally, and perhaps most importantly, automated data quality assessment and the constant focus on data quality leads us to focus on *key science deliverables*. No large data project in astronomy or physics has been successful without active involvement of science stakeholders in data quality assessment and DM execution. To facilitate this, metadata visualization tools will be developed or adapted from other fields and initiatives, such as the VOA, to aid the LSST operations scientists and the science collaborations in their detection and correction of systematic errors system-wide.

A fundamental question is how large the LSST data management system must be. To this end, a comprehensive analytical model has been developed driven by input from the requirements specifications. Specifications in the science and other subsystem designs, and the observing strategy, translate directly into numbers of detected sources and astronomical objects, and ultimately into

required network bandwidths and the size of storage systems. Specific science requirements of the survey determine the data quality that must be maintained in the DMS products, which in turn determine the algorithmic requirements and the computer power necessary to execute them. The relationship of the elements of this model and their flow-down from systems and DMS requirements is shown in Figure 2. Detailed sizing computations and associated explanations appear in LSST Documents listed on the Figure.

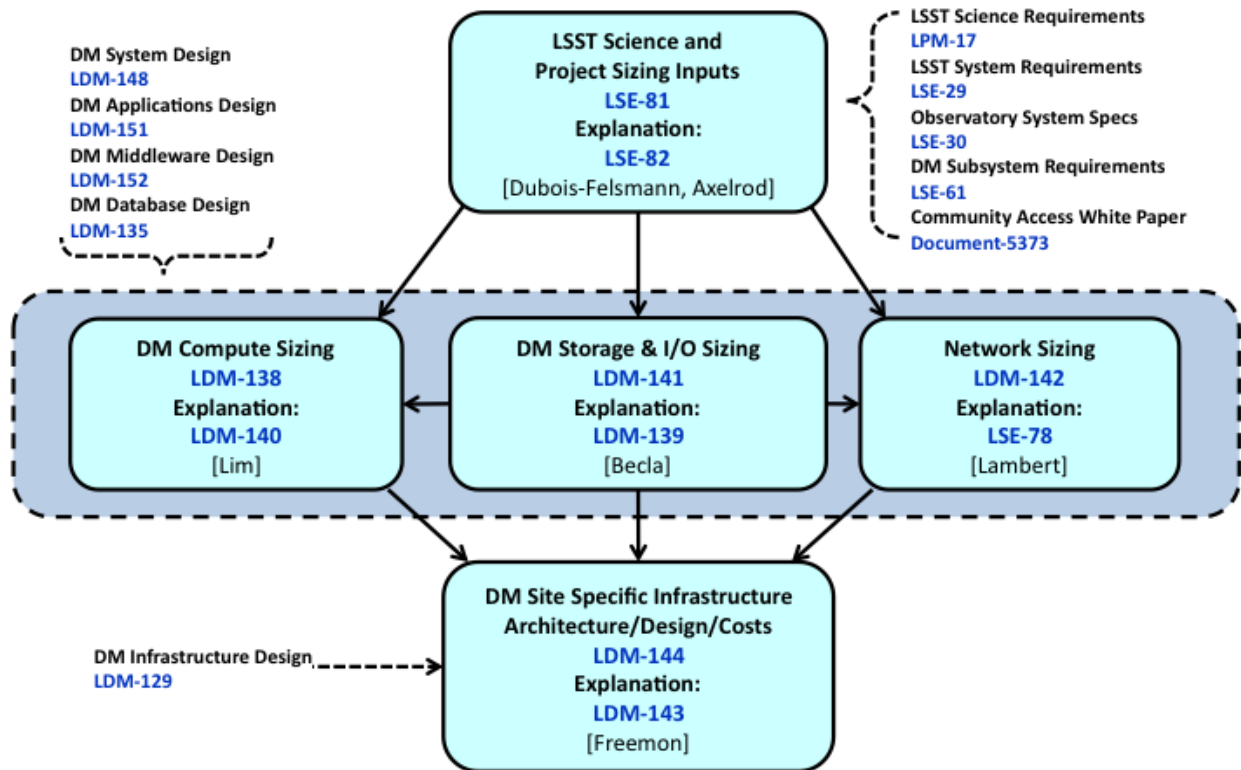


Figure 2 DMS Infrastructure sizing and estimation.

Key input parameters include camera characteristics, the expected cadence of observations, the number of observed stars and galaxies expected per band, the processing operations per data element, the data transfer rates between and within processing locations, the ingest and query rates of input and output data, the alert generation rates, and latency and throughput requirements for all data products.

Processing requirements were extrapolated from the functional model of operations, prototype pipelines and algorithms, and existing pre-cursor pipelines (SDSS, DLS, SuperMACHO, ESSENCE, and Raptor) adjusted to LSST scale. As a part of every data release, all data previously processed are reprocessed with the latest algorithms, calibration products, and parameters. This causes the processing requirements to increase with time.

Storage and input/output requirements were extrapolated from the data model of LSST data products, the DMS and precursor database schemas (SDSS, 2MASS), and existing database management system (DBMS) overhead factors in precursor surveys and experiments (SDSS, 2MASS, BaBar) adjusted to LSST scale.



Communications requirements were developed and modeled for the data transfers and user query/response load, extrapolated from existing surveys and adjusted to LSST scale.

In all of the above, industry-provided technology trends (LSST Document LDM-143) were used to extrapolate to the LSST construction and operations phases in which the technology will be acquired, configured, deployed, operated, and maintained. A just-in-time acquisition strategy is employed to leverage favorable cost/performance trends.

The results of the functional and performance flow down exercise are presented in Table 1 and discussed below for processing, storage, and communications. The resulting performance and sizing requirements show the DMS to be a supercomputing-class system and correspondingly large data input/output and network bandwidth rates. Despite this size, technology trends show this to be well within the anticipated performance of commodity-based systems during that time frame, and not even within the top 500 supercomputing systems as anticipated at that time.

In order to balance between near-real time and non-real time requirements, we have divided computing across three specialized centers: Base Center, Archive Center, and Data Access Centers. (Each center is described in detail later in this section.) As shown in Table 1 and Figure 3, the total processing requirements for LSST at the Base and Archive Center start in the commissioning period, and the Base Site grows more slowly than the Archive Site, as it is dominated by its co-located Data Access Center .

		<i>Archive Site</i>	<i>Base Site</i>
Compute	Teraflops (sustained)	200 → 1100	30 → 55
	Nodes	800 → 800 (1200 hwm)	100 → 50 (115 hwm)
	Cores	45K → 180K	7K → 10K
	Memory Bandwidth	25 → 130 TB/s	3 → 6 TB/s
Database	Teraflops (sustained)	33 → 330	30 → 310
	Database Nodes	120 → 270 (360 hwm)	100 → 250 (340 hwm)
Facilities	Floorspace	670 → 700 sq ft (875 hwm)	400 → 420 sq ft (500 hwm)
	Power	300 → 440 kW (610 hwm)	120 → 160 kW (220 hwm)
	Cooling	1.0 → 1.5 mmbtu (2.1 hwm)	0.4 → 0.5 mmbtu (0.7 hwm)

Table 1DMS Compute Infrastructure Sizing, including growth from Year 1 to Year 10.

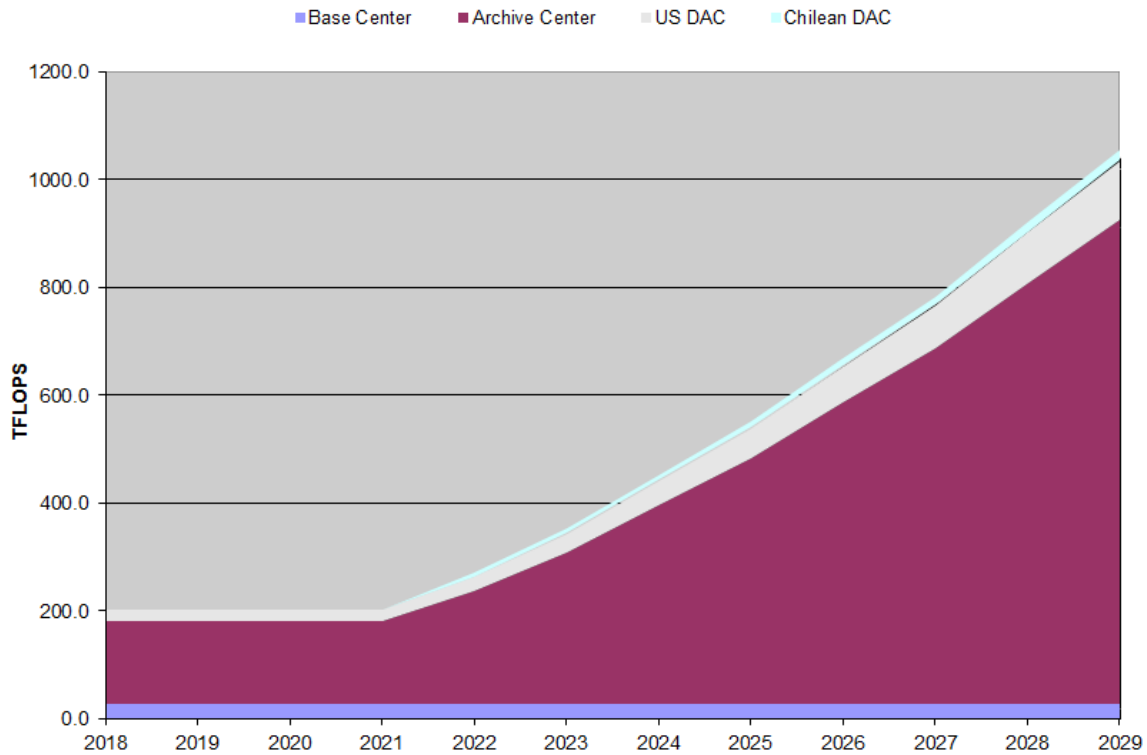


Figure 3 The computational performance requirement for the DMS grows during the life of the survey.

These requirements are technically feasible now. The growth trend is less steep than industry trends in computing system throughput, and the DMS middleware layer is designed to facilitate migration to improved architecture without a complete rewrite of the software. Though LSST processing requirements are considerable by the standards of current optical/infrared astronomical surveys, they are fairly modest when compared to the requirements of high-energy physics, radio astronomy, and other scientific programs driving the top 500 supercomputing sites (see Figure 4).



Figure 4 DMS computing requirements grow but remain readily achievable.

The sizing of the computational requirements involved a conscious trade-off between the cost of saving all computed data products at all stages of processing versus re-computing selected data products on demand. This resulted in a CPU-intensive model because raw image data and associated metadata expands from 15 TB to over 100 TB per night upon processing. Only certain data products are stored permanently along with the provenance needed to re-create all the other data products on demand.

LSST data storage requirements are substantial: LSST will reach multi-petabyte scale during the first year of operations (Table 2 and Figure 5). The raw image data will grow linearly at nearly four petabytes per year based on observing cadence and image size. The largest of the heavily used catalogs, the Source Catalog, will hold about 450 billion source measurements in the first data release (DR1) and nine trillion in the last data release (DR11). A source is defined to be a detection on a single exposure.

Once released, all data products will be permanently stored and immutable. Unreleased data collected and processed between releases will be updated nightly. This latest (Level 1) catalog data is always available for query, but it does not have the quality attributes of a Data Release, so these data will carry the caveat that analytical results based on those data may change once the data are reprocessed for inclusion in the new DR.

Row sizing for database storage estimates is derived from a provisional schema based on experience with SDSS, 2MASS, and SuperMACHO, expanded with input from the LSST Science Collaborations, and captured in the LSST Data Products Definition Document (LSE-163). For purposes of expansion, additional space has been reserved for as yet unknown attributes. A factor has been applied to accommodate "persistent overheads," i.e. database indices, database overheads, and data replication. Additional disk space is reserved for database swap space. Estimates of overheads and swap space are based on the experience of SDSS, 2MASS, and BaBar, as well as input from external database experts. These estimates and factors are described fully in the Data Management Storage Sizing and I/O Model and Explanation (LDM-141 and LDM-139).

		<i>Archive Site</i>	<i>Base Site</i>
Disk Storage for Images	Capacity	26 → 120 PB	12 → 29 PB
	Drives	1600 → 1300 (1700 hwm)	1140 → 310 (1140 hwm)
	Disk Bandwidth	190 → 580 GB/s	90 → 100 GB/s
Disk Storage for Databases	Storage Capacity	17 → 137 PB	11 → 90 PB
	Disk Drives	1700 → 2600 (3000 hwm)	1000 → 1700 (2200 hwm)
	Disk Bandwidth (sequential)	150 → 620 GB/s	100 → 310 GB/s
Tape Storage	Capacity	10 → 120 PB 10 → 120 PB (offsite)	10 → 120 PB no offsite
	Tapes	2000 → 8300 (near line) 2000 → 8300 (offsite)	2000 → 8300 (near line) no offsite
	Tape Bandwidth	22 → 52 GB/s	13 → 13 GB/s

Table 2 DMS Storage Infrastructure Sizing, including growth from Year 1 to Year 10.

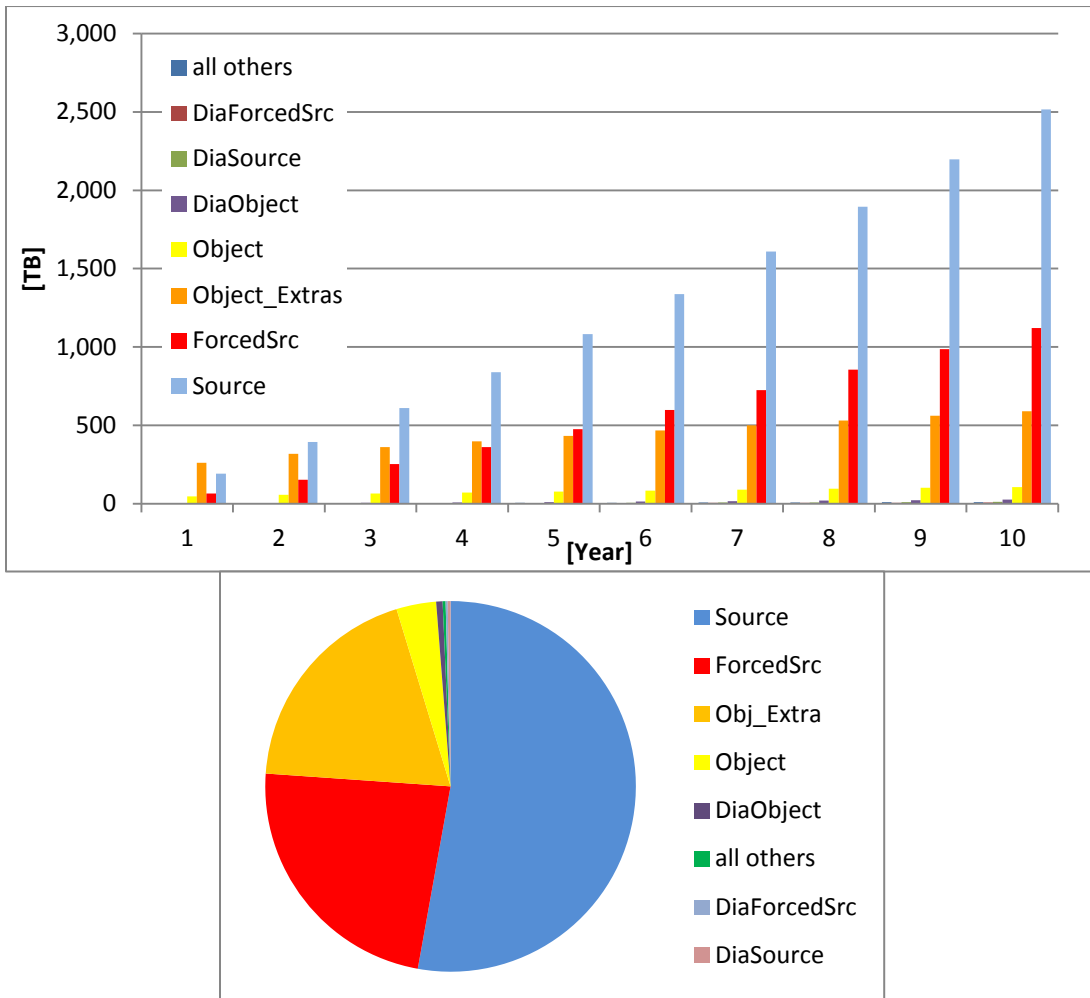


Figure 5 The DMS storage requirements are substantial, especially compared to other past and current astronomical datasets, but well within technology limits.

The two most recent data releases and the "live" Level 1 data, including their corresponding persistent overheads and database swap space, will be served from fast disks, while older releases will be served as batch downloads from more economical slower disks and tape.

The LSST disk input/output requirements are estimated based on pipeline data access and ingest rates (including re-processing), indexing rates, and a hypothetical end-user query load. The query load modeled, including both interactive and programmatic queries, is based on 65 "standard" queries developed to represent likely data access patterns and to stress the database architecture (see Data Management Database Design document LDM-135). It includes an estimate of:

- 100 simultaneous low-volume queries against up to 1 square degree in the catalog; response time 10s; resulting data set: 0.5 GB
- 20 simultaneous high-volume queries against all objects in the catalog; response time 1 hr; resulting data set: 6 GB

The system will support many times more users who may log in, but typically most users are in "think

time mode" rather than constantly executing queries.

Terabyte-sized volumes of LSST data and data products must be moved each day from the Mountain Summit to the Base Center in Chile, from the Base Center to the Archive and Data Access Centers, and finally from those centers to the users when necessary, as depicted in Figure 6.

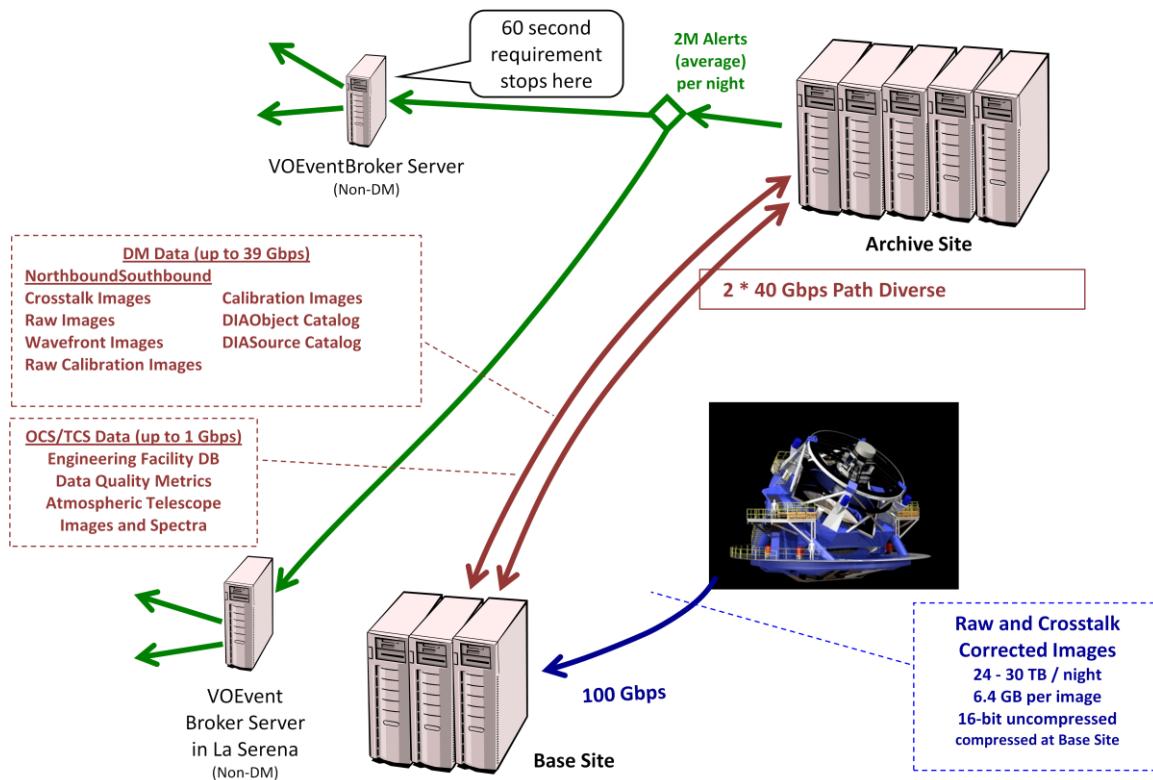


Figure 6 The DMS achieves latency and data availability requirements by transferring only time-critical information across international Chile – US fiber optic networks

The key driving requirements for the Mountain Summit to the Base Center communications are the bandwidth and reliability required to transfer the crosstalk-corrected image and the raw image data for forwarding to the Archive Center. Because this link closely follows the path of existing CTIO networks, LSST will manage and maintain this mission-critical link, providing the reliability and availability necessary to run the combined Mountain-Base infrastructure as a single component of the LSST system.

The required bandwidth from the Base Center in La Serena to the Archive Center in Illinois is 10gigabits/second to accommodate transfer of raw and crosstalk-corrected data for processing by the Alert Production. This link must also be reliable to avoid data transfer bottlenecks, and there must be spare capacity to "catch up" in the event that a failure or slow-down occurs. Commercial carriers and research networks are the only realistic source of this capacity and reliability over this distance. LSST has partnered with NOAO and CTIO to leverage and expand existing NSF-funded network initiatives that are rapidly increasing the available Chile - United States bandwidth while driving the cost downward.

Within the United States and Chile, LSST must provide open access to LSST data-not only to the scientific

community, but also to diverse segments of non-specialist users whose curiosity may be triggered by LSST educational and public outreach programs. LSST will provide portals to the public internet and Internet2 for such educational and outreach efforts.

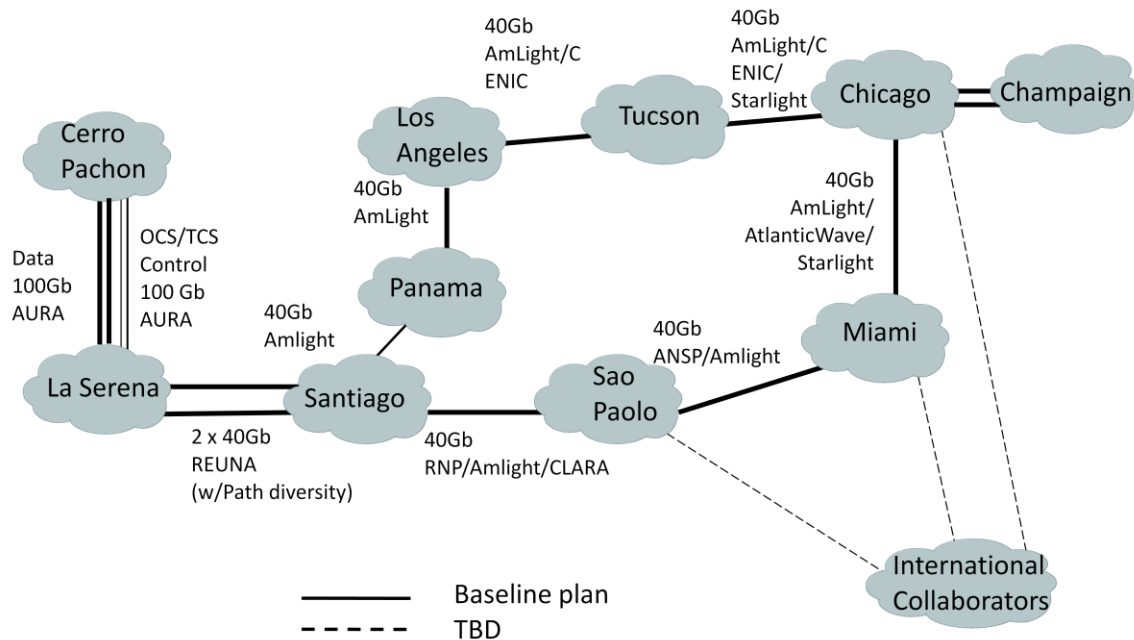


Figure 7 The DMS is driving early provisioning of the long-haul network design and leveraging NSF-funded and DOE-funded Research and Education Networks.

Based on the survey reference design, we have established an overall availability percentage for the DMS as a constraint for the design effort, because this drives the amount of spare capacity and redundancy that must be included in the design.

The requirement for overall system availability is 98%, and the justification is as follows. First, based on statistics from CTIO, CFHT, and other observatories, we anticipate losing approximately 20% of total available observing time to weather and overall operational considerations. (This lost survey efficiency is included in the operations simulations.) We have set the goal that no more than 10% of this loss will be attributable specifically to the DMS, or 2% of total available observing time, thus leading to 98% availability.

Our reliability architecture is based on two fundamental principles: *data preservation* and *keeping up with the data*.

At any given point in time, at least two copies of the data are maintained in separate geographic locations. Caches are large enough to handle any projected system downtime, and multiple recovery options exist for any failure mode. That is, the raw image data is stored (2 copies) for two nights in the Data Acquisition System on the Mountain Summit, and streamed to the Base Center, where it is permanently archived in the Chilean Data Access Center and also stored for up to 14 nights in a buffer.

As the crosstalk-corrected image data is received at the Base Center, it is streamed to the Archive Center for alert generation. At the same time, the raw images are transferred from the Mountain Summit to the Base Center and are streamed to the Archive Center over the course of the night; the transfer must be completed before observing is started for the next night.

Images are processed at the Archive Center. At the same time, data are written to tape on a nightly basis. Thus, among the Mountain Summit, Base, and Archive Centers, there are always at least two copies of every raw image. An additional archival copy is retained off-site on tape. The other established design goal is that only a small fraction (2%) of total nightly data is not processed within 24 hours. In order to achieve this, we employ a mix of redundant systems, hot spares, and cold spares to ensure a minimum number of single points of failure, while staying within our cost constraints. We have designed fault-tolerance features into the DMS in both hardware and software as defined in the Data Management Middleware Design document LDM-152, the Data Management Infrastructure Design document LDM-129, and the Data Management Automated Operations document LDM-230.

In case of storage failure in any processing center, backup copies of the data are available to perform the restore, with other Centers serving as ultimate backups. In all cases, there is sufficient spare computing, storage, and network capacity to continue nightly and data release processing while restoration is performed.

### 3 Design of the Data Management System

The collective data management experience from a wide range of institutions in astronomy and high energy physics has been used to translate the required DMS capabilities and performance specifications into a design. Experience with such systems demonstrates that these levels of performance, scalability, and reliability can be achieved through a planned, documented, systems approach to design and development, with process and specification reviews at each major milestone. A formal design process captures all design elements into specifications and models. The DMS requirements have been captured in the Data Management System Requirements (LSE-61) and a Unified Modeling Language (UML) model (LDM-133, LDM-134, and LDM-146)).

The DMS requirements flow down from the LSST Systems Engineering requirements model in SysML. Further development of detailed DMS requirements and design specifications is done in the Iconix process (Rosenberg and Stephens 2007, Rosenberg Stephens and Collins-Cope 2005). This widely used process is based on the Unified Modeling Language (OMG 2005). Cost estimates were derived from the following:

- An industry-standard software cost estimating process based on the Constructive Cost Model ("CoCoMo", Boehm 1981) and Function Point Analysis (Jones 1991).
- Vendor-provided quotes and price/performance trends for hardware and off-the-shelf software
- Comparisons with pre-cursor astronomy data management systems.

The Data Management System architecture partitions the DMS into individual centers optimized for specific responsibilities, as already described. This approach has been adopted in order to maximize the



throughput of the DMS pipelines, to meet the latency requirements, and to distribute resources to serve the data close to the users. This "vertical" partitioning is, however, only half the story. The LSST requirements and challenges for scalability, reliability, modularity, and evolutionary design are best addressed by what is commonly known as a "layered" architecture, a concept that is widely employed in information systems with similar requirements. The Open Systems Interconnect (OSI) model for networking (ISO 1994) and the Model-View-Controller (MVC) and Layered Architectures in Java give examples.

This horizontal layering separates the primarily custom application software from the physical infrastructure (i.e., hardware and system software) with "middleware" software technology. Figure 8 shows the layering structure with associated components. This layering is applied consistently in the DMS and is incorporated into the design of each of the individual DMS centers.

**Application Layer (LDM-151)**

- Scientific Layer
- Pipelines constructed from reusable, standard "parts", i.e. Application Framework
- Data Products representations standardized
- Metadata extendable without schema change
- Object-oriented, python, C++ Custom Software

**Middleware Layer (LDM-152)**

- Portability to clusters, grid, other
- Provide standard services so applications behave consistently (e.g. provenance)
- Preserve performance (<1% overhead)
- Custom Software on top of Open Source, Off-the-shelf Software

**Infrastructure Layer (LDM-129)**

- Distributed Platform
- Different sites specialized for real-time alerting vs peta-scale data access
- Off-the-shelf, Commercial Hardware & Software, Custom Integration

02C.05 Science User Interface and Analysis Tools		02C.01.02.02 - 03 SDQA and Science Pipeline Toolkits	
02C.06.01 Science Data Archive (Images, Alerts, Catalogs)		02C.01.02.01, 02C.02.01.04, 02C.03, 02C.04 Alert, SDQA, Calibration, Data Release Productions/Pipelines	
02C.03.05, 02C.04.07 Application Framework			
02C.06.02 Data Access Services		02C.07.01, 02C.06.03 Processing Middleware	
02C.07.02 Infrastructure Services (System Administration, Operations, Security)			
02C.07.04.01 Archive Site	02C.07.04.02 Base Site	02C.08.03 Long-Haul Communications	
Physical Plant (included in above)			

Figure 8 The 3-layered architecture of the Data Management System enables scalability, reliability, and evolutionary capability.

The Application layer embodies the fundamental scientific roles and responsibilities of the LSST DMS, i.e., the scientific algorithms, pipelines, and data product implementations. This layer is implemented as custom developed, open source software. The Data Management Applications Design is document LDM-151.

The Middleware Layer enables "portability," by providing an abstract and uniform interface to the hardware and system software; it also provides standard services to the applications for security, reliability, plug-in pipeline algorithms, and extendable data types. This layer is implemented by custom software and integration of off-the-shelf, open source software. The Data Management Middleware Design is document LDM-152.

The Infrastructure Layer provides all the hardware and system software in the DMS computing, storage, and network resources that are the environment in which the Application and Middleware Layers are hosted and execute. This layer is implemented as off-the-shelf commercial hardware and system software. The Data Management Infrastructure Design is document LDM-129.

The UML specifications for this design, combined with industry-standard estimating methods (CoCoMo and Function Point Analysis) provide the basis for estimating the effort and schedule for the software elements of the DMS. The effort and schedule estimates have been "calibrated" with the actual performance of the DM team over three Data Challenges and 6 software releases, totaling over 100 FTE-years of effort and 200,000 lines of C++ and python code.

In the remainder of this section, we summarize the design for each of these layers. The Applications Layer is treated first because the scientific value of the LSST resides in this layer. Next we discuss the Middleware Layer, which provides for evolution and scalability. Then we discuss the Infrastructure Layer, which hosts the applications and the middleware.

### 3.1 Application Layer Design

The Application Layer functions have been specified via UML Use Cases and Activities. A total of approximately 340 use cases and activities have been defined that cover the required capabilities. Of these, approximately 60% have been prototyped and tested for function and performance with precursor and simulated data in data challenges. At this point, the prototype software is capable of processing the entire data set from a typical current-day terabyte-scale astronomical survey in a matter of days, including point- and extended-source photometry and astrometry both on single frames and coadds. The data quality of the output data products is assessed by comparison of the input simulated data "truth" and output processed catalogs.

The Application Layer is being developed in an object-oriented fashion as custom C++ and Python class modules. C++ is used for core algorithms where computational efficiency is paramount, while Python is used for higher-level application flow, data formatting, bookkeeping, and to facilitate integration of the processing steps together by means of wrappers for the C++ classes.

The Application Framework library forms the basis for the entire layer. It is a toolkit of classes for astronomical data processing. Specific algorithms and data products are derived from the base classes of the framework. For example, there are base classes called "Image" and "PSF" (point-spread function); both can be specialized to reflect particular uses or implementations of the base class.

The Science User Interface and Analytical Tools (Figure 9) include several types of tools and user interfaces. For many of the tools, we can expect to leverage development outside the LSST project, both at IPAC and from open-source or off-the-shelf code. Some functionality will be needed prior to commissioning, all of it for science operations.

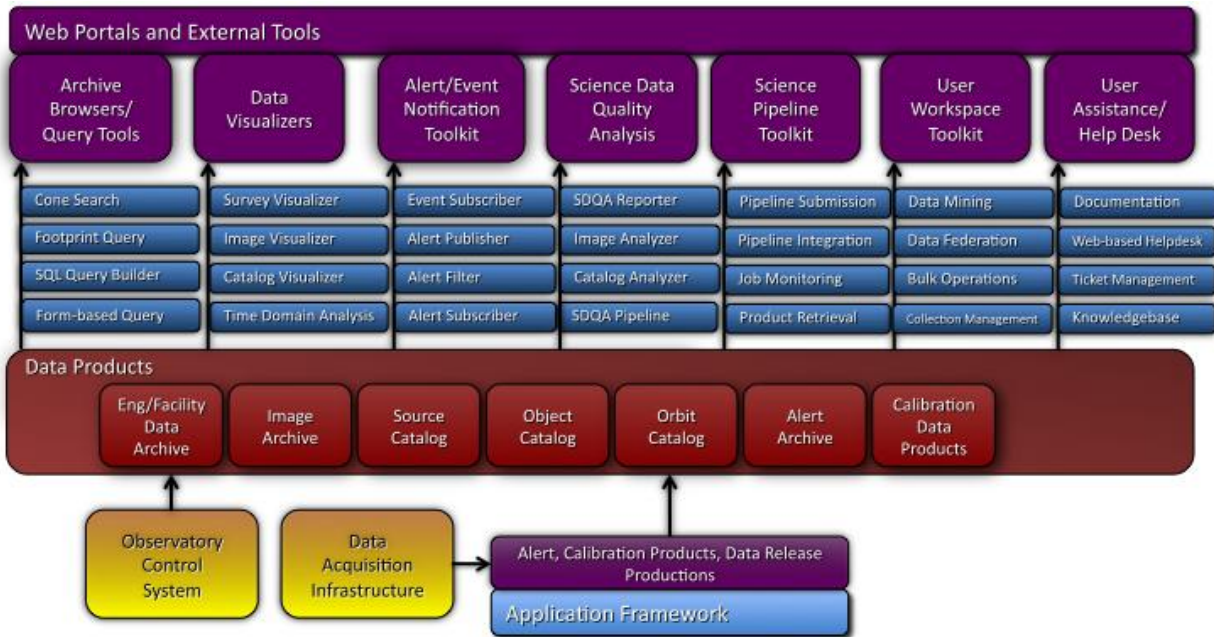


Figure 9 The Data Management System Science User Interface and Analytical Tools, SDQA Toolkit, and Science Pipeline Toolkit

The Archive Browser/Query Interface and the basic capabilities of the Data Visualization task constitute the fundamental "Science User Interface" to the LSST data. The Alert/Notification Toolkit will allow vetting and analysis of specific alerts, including retrieval of image cutouts and source/object observation history. This task essentially levies an additional set of requirements on the browse/query interface. The Data Visualizers and Analysis Tools include both basic (e.g., browse-quality image preview, and color-color and light-curve plotting) and advanced (e.g. Google-Sky-like full-sky hierarchical visualization and overlay, interactive sky movies) data visualization tools. The User Workspace Toolkit will have functionality analogous to, and will likely leverage the development of, VOSpace. The Help Desk/User Support task will be based on an existing commercial or free help desk and ticketing system.

There are two very specialized additional user interfaces provided along with the Science User Interface and Analytical Tools as shown in Figure 9. The Science Pipeline Toolkit is an interface allowing a user to construct and run a modified version of the production pipeline against LSST survey data, with the possibility of including user-created modules. The Data Quality Analyst Toolkit task will leverage the basic visualization and the archive interface, with a modest level of dedicated development for handling and visualizing the quality metadata. More advanced tools, such as database-crawling correlation seekers and synthetic images to visualize non-image quality metadata, will be treated the same way as the advanced visualization capabilities.

The Application layer is described in the remainder of this section in terms of algorithms and pipelines and the database schema and queries. The requirements for the algorithms employed in the DMS pipelines are set by several factors. The first are the data quality requirements that must be met for the LSST science missions. An example is the requirement that photometry must be consistent across the sky and between filters to 0.005 mag. An equally essential requirement derives from the fact that LSST will observe the entire visible sky under a wide variety of observing conditions i.e. seeing, sky brightness,

atmospheric extinction, and field characteristics (particularly crowding). An algorithm able to deliver the required photometric precision in excellent conditions may be unable to do so when conditions deteriorate even slightly. The enormous data rates of the LSST rule out routine human intervention to tune up an algorithm that is functioning badly. Such tuning must be performed autonomously by the DMS itself.

Algorithm development for LSST builds on the experience base and expertise already gained in large astronomical surveys (e.g., SDSS, SuperMACHO, ESSENCE, the DLS, CFHT-LS, UKIDSS) and will benefit from the advanced surveys recently started or planned for the near future, such as Pan-STARRS, the Dark Energy Survey (DES), and SkyMapper. However, while existing algorithms have been shown to meet some LSST data quality requirements, they are generally unable to meet all of them over the full range of sky and field characteristics, and do not have self-tuning or self-repair capabilities. In addition to these general areas of development, specific algorithms will require particular performance improvements. To carry out the needed algorithm development, we have undertaken early prototyping of our pipeline software and have tested the prototypes on both existing survey data and simulated LSST data.

The LSST algorithms are executed by a set of pipelines that are structured to partition the overall functionality of the DMS cleanly. Each pipeline has sole responsibility for producing one or more data products. Several pipelines integrated together into Level 1 Processing Pipelines (discussed in depth in the DM Applications Design Document; LDM-151), are run on a nightly basis at the Archive Center, producing the Nightly Data Products (Figure 1). The Camera SDS output data stream is directly fed into the Alert Production during observing. This data stream contains both unprocessed (raw) camera images that are sent directly to the Archive Center and images that have been corrected for crosstalk by the SDS on the mountain. The normal observing pattern is to take two 15-second exposures of the same field in immediate succession. These two exposures together form a Visit, which is the data unit processed by subsequent LSST pipelines.

At yearly intervals (and twice for the first year of survey data), the Data Release Production creates a new Data Release (shown in the center of Figure 10). A Data Release (DR) includes all data taken by the survey from day one to the cutoff date for the DR, and is a self-contained set of data products, all produced with the same calibration images, pipeline software, and processing parameters.

On an as-needed basis, but at least once per DR, the Calibration Products Production (shown on the right in Figure 10) is run at the Archive Center. It creates a variety of calibration data products which are required by the other two Productions.

### 3.2 Middleware Layer Design

The Data Management Middleware Layer is designed to isolate scientific applications, including the Alert Production, Data Release Production, Calibration Products Production, and Level 3 processing, from details of the underlying hardware and system software. It enables flexible reuse of the same code in multiple environments ranging from offline laptops to shared-memory multiprocessors to grid-accessed clusters, with a common communication and logging model. It ensures that key scientific and deployment parameters controlling execution can be easily modified without changing code but also with full provenance to understand what environment and parameters were used to produce any dataset. It provides flexible, high-performance, low-overhead persistence and retrieval of datasets with data repositories and formats selected by external parameters rather than hard-coding

Working in a petabyte regime will require non-conventional techniques and innovative solutions for data storage and access. A Data Access Client Framework provides applications with a high-performance abstraction layer, hiding many complexities of access to large file repositories and huge databases. Under the hood, however, a sophisticated distributed file management system and a scalable database management system are used to provide sufficient performance and capability to handle application needs. Processing queries in an optimal way is extremely important in LSST's world of multi-billion row tables. As an example, the Source table is expected to reach over 900 billion rows in the first year of data taking (Figure 11) and grow to over nine trillion rows by year 10. Triggering a single scan through this table (which can be done through a simple "SELECT \* FROM Source" query) would result in fetching many petabytes of data from disks. It is imperative that this only be done when necessary and that it be combined with other uses of the same data whenever possible. Further, having to support simultaneous multi-dimensional searches (spatial and temporal) significantly complicates the database design.

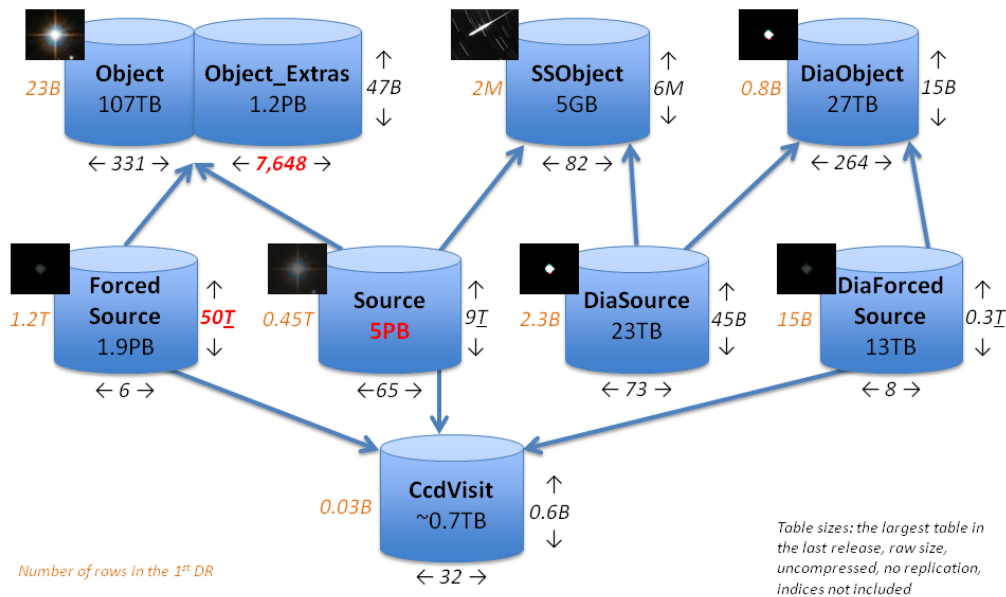


Figure 11 The Data Management System database schema features extremely large tables, some with trillions of rows

In order to achieve the performance required to effectively serve this data, a parallel query architecture has been developed as part of the middleware, as shown in Figure 12. This architecture permits users and client applications to submit standard SQL queries to the database system and handle result sets as if the system were a single large database.

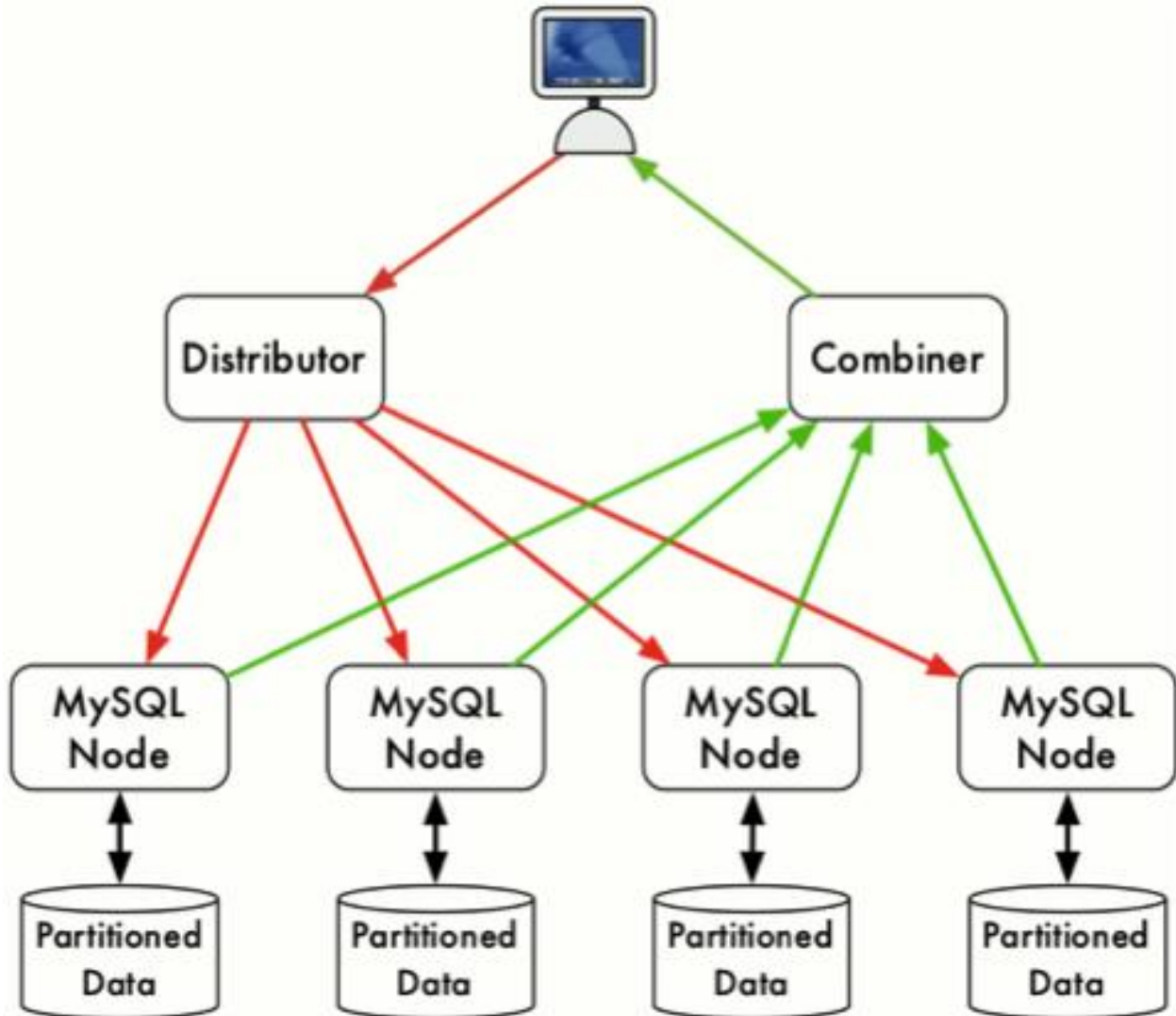


Figure 12 The Data Management System parallel query architecture supports high-performance, high-volume data access

A set of "standard" queries used for sizing and costing the system is representative enough to drive the schema design and cover both the public access as well as the professional astronomer's access. They were generated based on experience from other astronomical experiments and archives (SDSS, MACHO, IRSA) and additional input from the LSST science collaborations, carefully selected such that both spatial and temporal aspects are well covered. They are aligned with the precursor schema that was used for sizing and testing. Refer to the Data Management Database Design document LDM-135 for further details.

As shown in Figure 13, the Middleware Layer also includes services to control application pipelines and provide them with services like logging and inter-process communications. The pipeline construction toolkit supports modes ranging from debugging/interactive to batch to large-scale parallel. Orchestration manages the deployment of pipelines on the compute clusters, and the Data Management Control System oversees the execution of the application Productions. Refer to the Data

Management Middleware Design document LDM-152 for more information regarding the Data Access and Control services.

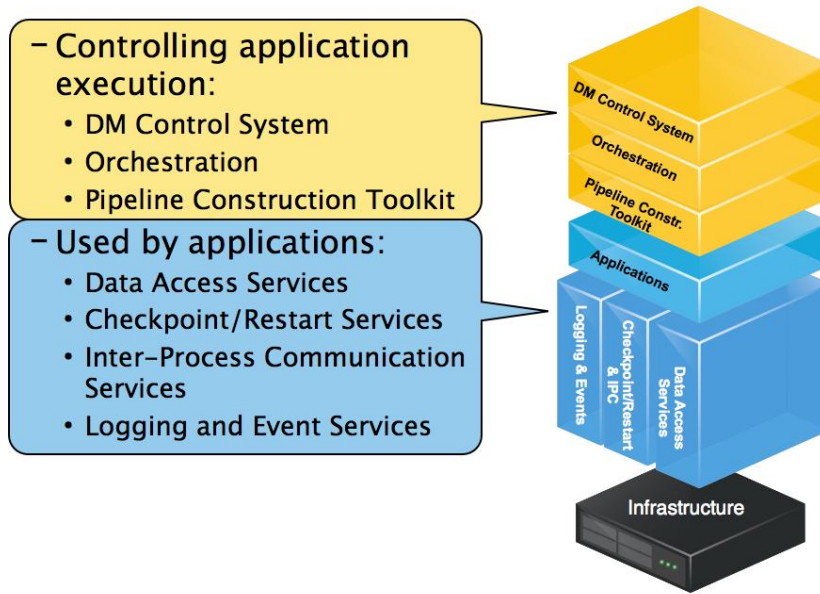


Figure 13 The Data Management System supports high-performance, high-volume data transfers and massively parallel pipelines

### 3.3 Infrastructure Layer Design

The infrastructure layer provides the total hardware and system software for DMS computing, storage, and networking. Processing and storage activities are distributed across four types of DMS centers: the Mountain Summit, Base Center, Archive Center, and Data Access Centers (Figure 14). The Data Access centers also provide data access for the Education and Public Outreach Center as well as end user sites that are external to the LSST. Connecting all the facilities are long-haul communications networks.

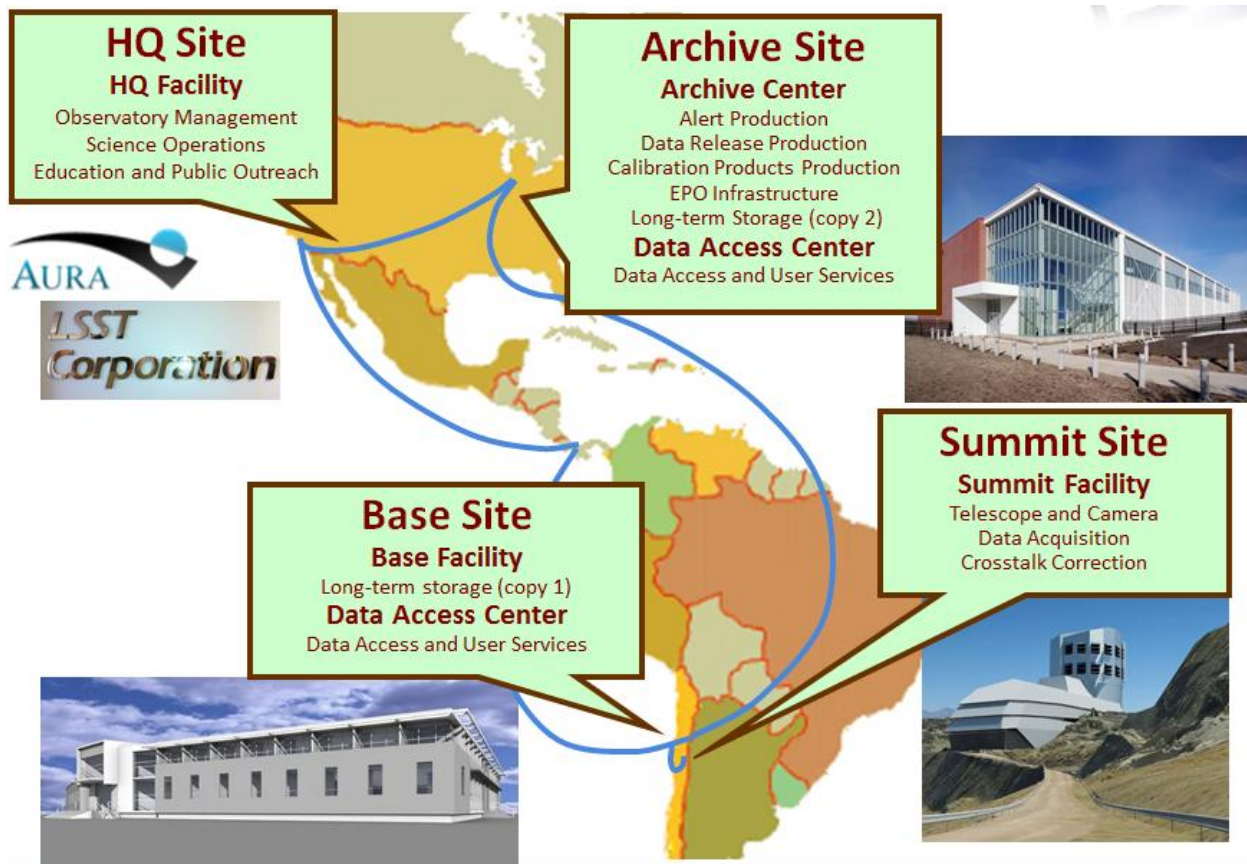


Figure 14 The Data Management System infrastructure spans multiple sites and 2 continents

Each DMS Center operates 24x7 and is staffed for both nighttime (observing) and day/swing (non-observing) shifts. Distribution of processing and storage activities among separate facilities promotes efficient performance of the system via separation of real time and non-real time resources. The design for these centers is as they would be implemented using 2013 era technology. The actual implementation will use later technology, which at then-year dollars will provide the performance needed to meet 100% of the LSST requirements. The system performance projection is based on industry-provided trends for price/performance and capacity improvements between now and 2031 (see Site Specific Infrastructure Estimation Model document LDM-144 and Explanation document LDM-143). The trends used are conservative to ensure that the estimates can be considered "not-to-exceed" costs. The remainder of this section describes each facility design and the long-haul networks design.



The Mountain Summit site is on Cerro Pachon. Its sole DMS responsibility is to capture data and move it down to the Base Center for forwarding to the Archive Center and Alert Production processing. The Data Acquisition System (part of the Camera Subsystem) and the Observatory Control System (part of the telescope subsystem) interface to the DMS, with image read-out in two seconds and immediate data transfer to the Base at 100 Gigabits/s on fiber optic lines dedicated to this traffic.

This is implemented as one 100 Gigabit/s link from the summit Ethernet switch to the base Ethernet switch. Metadata from the Engineering and Facility database in the Observatory Control System (OCS) is transferred to the Base on a dedicated 100 Gigabit/s link on a nightly basis. For details of this link refer to the LSST Network Design document LSE-78).

An analysis of running a separate fiber bundle in a different route for the backup links concluded this was not cost-effective versus the anticipated loss of alerts until a repair. Even in case of failure, the Mountain Summit store prevents loss of data within the anticipated Mean Time to Repair (MTTR). In the extremely unlikely event of simultaneous failure in the primary and backup links it is possible to transport one set of the backup data drives to the base.

This network choice is lower risk compared to other less mature high-speed communication methods and permits easy technology upgrades without impacting other systems at the Mountain Summit and Base Center. The firewall and optical switch connecting the Base Center to the Archive Center site are easily accommodated with current era technology. The Base Center and Archive Center file system design stores multiple copies of image data to further enhance data preservation.

The Base Center is at the Cerro Tololo Inter-American Observatory compound in La Serena, Chile. The Base Center's primary role is to archive a copy of the raw data and transfer it rapidly to the Archive Center for processing.

The crosstalk-corrected images from the summit will flow from the Camera data acquisition system into the memory of nodes at the Base Site and are immediately routed on to the Archive Center. This complete transfer occurs in 5 seconds, in order to permit subsequent processing 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 long-haul network connecting La Serena to Santiago and on to the United States. There are interface control documents (ICD) LSE-68 and LSE-69 that define the responsibilities between the camera system and the data management system.

Using commercially available, commodity Linux systems at the Base and Archive Centers lowers software development costs and long-term hardware support costs. These commercial Linux systems are known to be affordable, even when factoring in inevitable failed unit replacement and spare capacity.

The Base Center to Archive Center Network is a minimum guaranteed 2 x 10 Gigabits/s, anticipated baseline of 2 x 40 Gigabits/s, and up to a best case of 2 x 100 Gigabits/s burst, full duplex, path diverse, clear channel fiber optic circuit, with protocols optimized for bulk data transfer. The term path diverse means that the link is a ring with traffic possible in both directions; only a dual outage on both sides of the ring between source and destination will prevent data delivery. Clear channel means that LSST will have a dedicated portion of the network bandwidth guaranteed by an Indefeasible Right to Use (IRU). This link is used both for high priority command and alert traffic, as well as to send the raw image data and to send data products back to the Chilean Data Access Center. This traffic requires an average of 2.5 Gigabits/s. In the event of an outage in this link, once service is restored the link supports a 400% of the required average bandwidth for transmission of buffered data during the day, and much more at night and for bursts during “catch-up” periods.

The Archive Center will be located at the National Petascale Computing Facility (NPCF) at the National Center for Supercomputing Applications (NCSA) on the campus of the University of Illinois at Urbana-Champaign (Figure 15). This center will perform the Alert Production processing and publish the alerts via public alerting protocols based on the Virtual Observatory VOEvent standard (IVOA 2006). The Archive Center computing resources handle the processing needed to re-process nightly data products along with the planned annual re-processing the produces Data Releases. The Archive Center will be the source of replication operations that deliver the catalog, calibration, and sky template data to the Data Access Centers. Finally, the data releases are published from the Archive Center.

In its selection of the Archive Center site, LSST has made it possible to build its construction and operations plans on the extensive infrastructure which has been laid by NSF investments in U.S. supercomputing centers and research networks.

The LSSTC will not build or operationally staff the LSST Archive Center. The proposal calls for our partner the National Center for Supercomputer Applications at the University of Illinois Urbana-Champaign (NCSA) to be our primary archive center in the United States. NCSA is one of the five original centers in NSF’s Supercomputer Centers Program and has over 20 years of successful experience establishing specifications for, procuring, installing, deploying, and operating major resources for the national science and engineering research community.



The NCSA Building (left) is located on the campus of the University of Illinois, Urbana-Champaign; the Petascale Computing Facility (center) houses NCSA’s largest computational and storage platforms and will host the Archive Center (right).

Figure 15 Leveraging NSF Facilities at NCSA.

As depicted in Figure 16, the design includes multiple types of computing systems (pipeline and database servers) and multiple levels of storage systems (disk, disk cache, and mass storage). Disk space will be used for processing, serving data, deep storage cache, and other archive operations. The pipeline servers provide cluster nodes for running the Alert, Calibration Products, and Data Release Productions. These servers are attached to a fast storage system that supports only this processing. Other servers provide computing support for all services that run the archive, including orchestrating pipelines, managing data transfers, ingesting new data products, and running data query and access services.

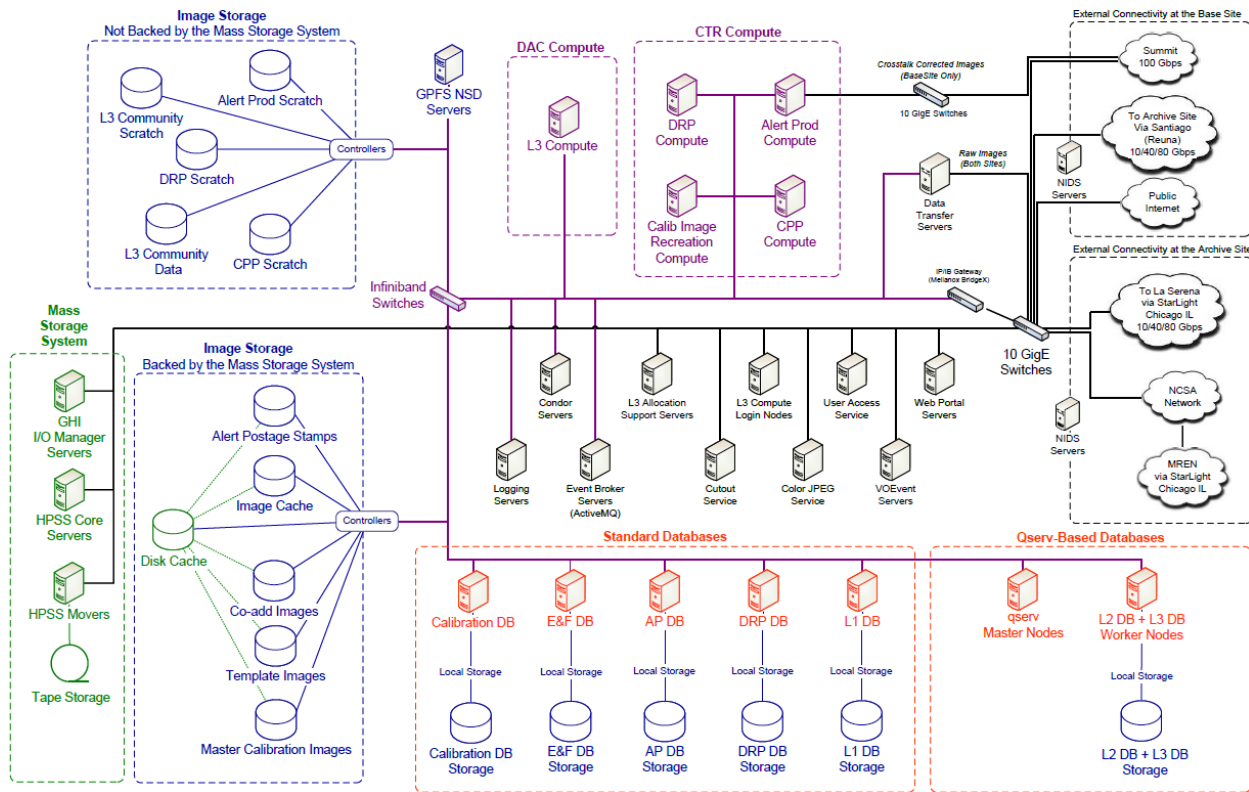


Figure 16 Computing and storage architecture is the same at the Base and Archive Centers, but scaled differently

These servers are attached to a large archive storage system, sized to hold the most recent versions of the catalogs, the latest coadded sky images, and a cache for intermediate products. Finally, a deep storage system is configured to provide archival storage and data access for older catalog versions and image files. The computing cluster is interfaced to a multi-tier storage farm housing the data archive.

The Data Access Centers provide access for external users. At present, the baseline calls for Data Access Centers co-located at the Archive Center in the U.S. and at the Base Center in Chile. This co-location allows the Data Access Centers to utilize the image archives at the Archive and Base Centers, eliminating the need for additional copies. The Data Access Centers are connected to the Base and Archive Centers via 100 Gigabits/s local area networks. No Level 1 or Level 2 data processing is performed at the Data Access Centers, the processing capacity is dedicated to running end user codes for Level 3 data production. As such, the Data Access Center design is a subset of the Archive Center design, and the details are not repeated here.

Managing access to these resources will be important for the project to achieve its mission-critical goals, both in day-to-day operations and in support of the key science programs. The concept of "service levels" will be used to manage resources effectively, and to distribute access loads over the Data Access Centers. This approach has been quite successful in the high-energy physics context where "tiers" of access and shared responsibilities are common. Service Level 1 access is relatively rare and typically requires supercomputers and storage allocations for very high-consumption applications, while Service Level 2 to 4 accesses use more limited resources but will be more common. Combining these provides

flexible access to cover the gamut of scientific users.

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.

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. Since LSST has an open data policy, theft of astronomical data is not a primary concern (although some L3 data may be subject to a proprietary period). The main considerations for LSST are data protection, data integrity, and preventing the misuse of the computing facility. See LSE-99 for additional information.