

LSST Petascale Data R&D Challenges

Achieving scalability and reliability in LSST computing, storage, and network resources

The design of the DM system architecture is influenced by the technology. We expect to be available to implement it, starting with construction in 2011 – 2014 and continuing through the principal survey period until 2024. This technology includes not only more powerful components, but completely new system architectures and potentially disruptive technologies. Most computing throughput improvements will come not from increased CPU clock speeds as in the past, but from larger concentrations of CPUs/cores and advanced computing architectures. Solid state technology may change storage and the way we physically organize data. Hardware failures will be routine for the LSST data system due to the large number of CPUs and disk drives, and reliance on high-speed network connectivity. It is a challenge to create a system sufficiently robust to these failures. We need to predict the characteristics of CPU, network, storage hardware, and system software sufficiently well that our design is appropriate. Further, we need to insulate the design as much as possible from underlying platform dependencies.

Reliability and performance issues for very large databases

LSST's main data products from the 20,000 square degree survey with 2000 images over ten years per patch of sky are in the form of relational database tables. These tables are very large (50 billion rows in the Object table, 600 billion rows in the Source table). They must be extensible, and partitioned and indexed to facilitate high query performance, and replicated across multiple centers. Queries in the time domain (Source table) are likely to be of equal importance to those in the spatial domain. Since these are traditionally optimized by different database organizations, it is unclear what choices will perform best for LSST. Some intensive applications will involve n-point correlations of object attributes over all objects. All these factors suggest that database performance and reliability are risk areas.

Efficient automated data quality assessment

LSST will produce large volumes of science data. The Data Management System (DMS) produces derived products for scientific use both during observing (i.e. alerts and supporting image and source data) and in daily and periodic reprocessing. The periodic reprocessing also results in released science products. Analysis of the nightly data will also provide insight into the health of the telescope/camera system. An automated data quality assessment system must be developed, which efficiently searches for outliers in raw image data and unusual correlations. This will involve aspects of machine learning.

Operational control and monitoring of the DMS

The DMS will be a complex distributed system with enormous dataflows that operates 24/7. The DMS must be continuously monitored and controlled to ensure the proper functioning of all computing hardware, network connections, and software, including the data quality of the science pipelines. Most of the monitoring tasks, and some of the control tasks, must be highly automated, since the data volumes preclude human examination of all but a tiny fraction of the data.

Achieving acceptably low False Transient Alert Rate

The science mission places high demand on the LSST's ability to rapidly and accurately detect and classify varying and transient objects and to achieve a low false alarm rate. Given the very high data volume produced by the LSST, the corresponding large number of detections in each image (up to one million objects detected per image), as well as the likelihood of entirely new classes of transients, the LSST will not be able to rely on traditional labor-intensive validation of detections, classifications, and alerts. To achieve the levels of accuracy required, new algorithms for detection and classification must be created, as well as innovative automated techniques for alert filtering and validation.

Efficient detection and orbit determination for solar system objects

One of the LSST's science missions is to catalog the population of solar system objects, with a particular focus on potentially hazardous objects. Due to the depth of LSST's images, about 300 solar

system objects per square degree will be detected near the ecliptic. The LSST cadence on the sky is not optimized solely for tracking solar system objects, so this dense swarm of objects must be reliably tracked through considerable gaps in time. Algorithms must be developed that are robust to possible mis-associations of detections at different epochs, and have acceptable computational scalability.

Achieving required photometric accuracy and precision

The LSST Science Requirements Document (SRD) requires a level of photometric (intensity data) accuracy and precision that may be difficult to achieve over the entire sky, particularly since the LSST will be operating in a wide variety of seeing, sky brightness, and atmospheric extinction. To achieve this requires a thoroughly tested calibration procedure and associated image processing pipeline. In addition to the point-source requirements in the SRD, accurate photometric redshifts require precision photometry for spatially extended objects.

Achieving required astrometric accuracy and precision

The LSST SRD requires a level of astrometric (position on the sky) accuracy and precision that is difficult to achieve over the entire sky. Achieving this astrometric performance requires a global, whole-sky, numerical solution for all per-frame astrometric quantities that minimizes a cost function. Considerable work will be required to develop an effective cost function.

Achieving optimal object detection and shape measurement from stacks of images

Most objects that will be used for dark matter and energy science are too faint to be usefully measured in a single LSST exposure. Instead, the LSST must detect and measure the properties of objects combining information from multiple exposures of the same region of sky (image stacks). Weak lensing galaxy shape measurements are particularly vulnerable to systematic effects introduced by errors in the local point-spread function (PSF) determination, and these systematic effects must be minimized. Exposures may vary significantly in their signal-to-noise and PSF quality, and defining how to optimally combine information from all of them is a research problem. See <http://universe.ucdavis.edu/docs/MultiFit-ADASS.pdf> for more information.

Need to develop a flexible approach that enables highly reliable classification of objects

Classification of astronomical objects is important and difficult. A wide variety of information must be assessed to reliably classify an object. This includes spatial morphology in multiple colors, photometry in multiple colors, time dependent behavior, and astrometric motion. Further, the best classifications will make use of surveys in other wavelength regimes and spectral information where available, not solely information from the LSST. Experience from many surveys has shown that no single algorithm can do a good job on all objects. Rather, good algorithms tend to be specialist, limited to particular objects classes, e.g. eclipsing binaries or supernovae. A successful system must allow the development and incorporation of a wide variety of algorithms in a flexible manner.

Adaptive retuning of algorithm behavior

Several key algorithms employed in the LSST application pipelines are complex, containing many data-dependent decisions and a large number of tuning parameters that affect their behavior. As observing conditions change, an algorithm may begin to fail for a particular choice of tuning parameters. LSST's extremely large data volume makes human intervention in such cases impractical, but it is essential that the pipelines continue to function successfully.

Need to verify scientific usefulness of the LSST database schema and its implementation against realistic queries

The LSST database schema must efficiently support queries of data that have many relationships between multiple locations on the sky, epochs of observation, and filters employed. A high performance implementation of this schema has many complexities that are addressed in the peta-scale database architecture and analysis task. The ultimate test of how well these tasks have been carried out is to perform science with the database. To do this usefully, we are simulating LSST data, using data from current surveys, and engaging the LSST Science Collaborations and scientific community.