Large Synoptic Survey Telescope

www.lsst.org

Data Management R&D for the LSST Project

Gregory P. Dubois-Felsmann[1], T. Axelrod[2], A. Becker[3], J. Becla[1], D. Burke[1], A. Connolly[3], R. Cutri[4], S. Dodd[4], M. Freemon[5], Z. Ivezic[3], J. Kantor[6], D. Levine[4], K. T. Lim[1], R. Lupton[7], D. Monet[8], R. Owen[3], R. Plante[5], J. A. Tyson[9], D. Wittman[9]

[1]SLAC, [2]Steward Observatory/LSSTC, [3]Univ. of Washington, [4]IPAC, [5]NCSA, [6]LSSTC, [7]Princeton Univ., [8]U.S. Naval Observatory, [9]Univ. Of California, Davis

(on behalf of the LSST Data Management team)

The Data Management system for the LSST will have to perform near-real-time calibration and analysis of acquired images, particularly for transient detection and alert generation; annual processing of the entire dataset for precision calibration, object detection and characterization, and catalog generation; and support of user data access and analysis. Images will be acquired at roughly a 17-second cadence, with alerts generated within one minute. The ten-year survey will result in tens of petabytes of image and catalog data and will require ~250 TFlops of processing to reduce.

The LSST project is carrying out a series of Data Challenges to refine the design, evaluate the scientific and computational performance of candidate algorithms, and address the challenging scaling issues that the LSST dataset will present. Algorithm development must address the dual requirements for efficient use of computational resources, including emerging computing architectures, and the accurate and reliable processing of the unprecedented combination of deep and broad data resulting from the survey. This will require substantial progress beyond the state of the art from existing surveys. We anticipate the need for novel machine-learning algorithms for data quality analysis and to enable the discovery of the unexpected.

The Data Challenges incorporate both existing astronomical images and image data resulting from a detailed photon-level simulation, from sources through the atmosphere to the LSST observatory. The simulation is used to ensure that the system can scale to the LSST field of view and 3.2 gigapixel camera scale and meet the associated image and survey quality requirements. Future Data Challenges, carried out in conjunction with the LSST Science Collaborations, are planned to deliver data products suitable for high-quality science. We report on these plans and on the progress of the Data Challenges to date.

Data Challenges and the LSST DM Design and Development Plan

The LSST DM system will break new ground in both the **scope of the data** and the **complexity of the algorithms** it must execute to support the survey's goals. From the beginning, the design and development of the system has been structured around a series of **Data Challenges**, beginning in 2006, that enable the application of the growing code base to problems of **increasing scale and scientific complexity**. We view this as essential for **risk reduction** and for a full understanding of the requirements of the project.

DC1
Demonstrate scalability of the overall processing architecture and data flows.
Completed.

We challenge all parts of the DM system: the development process, the application framework and middleware, prototype infrastructure, and science algorithms.

DC2
Prototype nightly image & alert processing and associated middleware
Completed.

Results are used for performance tests and science validation, including "blind" tests of algorithms.

DC3
Refine framework & alert processing; prototype data release processing, science database queries; demonstrate scaling to full FPA.
In progress...

DC4
Calibration pipelines; alert distribution; prototype user interfaces for science data access and analysis, further extend scaling - to 20% and beyond.
Planned for 2011...

Increasing data scale and processing complexity

Data Sources

The Data Challenges are based on a combination of data sets from existing and near-term observatories and from a highly detailed simulation of the LSST and its environment.

Precursor surveys: we are using data from the CFHT Legacy Survey (19,000+ 290 Mpix images, from both the Deep and Wide surveys), and anticipate being able to apply our code to DLS data as well.

Simulation of the LSST Data

Three simulators combine to create data sets used for the Data Challenges

- **Operations simulation**: a night-by-night, exposure-by-exposure simulation of the operational cadence of the LSST, combining a prioritized set of science goals with models of weather, seeing, downtime, and other conditions;
- **Catalog simulation**: a list of observable objects to be simulated is assembled from a set of inputs: a cosmological evolution model, for galaxies and AGN [de Lucia *et al.* 2006]; a Milky Way model incorporating main sequence and white dwarf stars, colors and dynamics from [Juric *et al.* 2008], parallax and proper motions, variability and SEDs; and a solar system model with 10^7 objects.
- **Photon-level image simulation**: a photon-by-photon simulation of the propagation of light from astronomical sources through the atmosphere and the LSST, and the consequent response of the imaging system, yielding full-frame simulated images.

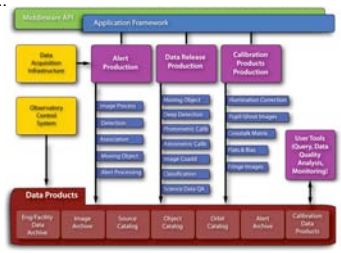
The current Data Challenge will use ~1.4M simulated CCD exposures, corresponding to the observation of ~350 deg^2 of the sky over the course of one year of the projected ten-year survey, and requiring 1.5-2 million CPU-hours to generate. Future Data Challenges will greatly increase the quantity of simulated data generated and processed.

(Additional details are available in the Krabbendam *et al.* and Connolly *et al.* posters in this session.)

LSST Data Management Tasks

Based on open-source software development...

- **Acquire and archive raw image data**
- Calibrate and analyze each image, detecting **transient and moving objects** and generating **alerts**
- Produce annual data releases, including calibrated **co-added deep images**, object detection and characterization yielding catalogs; maintain Science Database
- Support science data analysis, providing both software frameworks and computing resources, and allowing federation of user data products with the LSST catalogs



Current Data Challenge (DC3) Goals

Achievements so far:

- Complete full analysis chain for instrument signature removal and transient (alert) detection
- Refinement of a flexible processing framework (applicable to other projects as well)
- Science Data Quality Assessment framework
- Execution on existing clusters at NCSA

Phased production runs throughout 2010 at increasing scales

Final phase currently under way:

- Demonstrate full data release processing chain
- Moving-object detection and orbit determination
- Co-added image generation
- Global photometric calibration
- Astrometric model fitting
- Faint source detection and characterization
- Exercise of science database loading and querying
- Scale up to full focal plane size and to 2.5% of the volume of the first LSST data release
- Engagement of the LSST Science Collaborations in science algorithm validation

Advanced Technologies

In addition to the production- and scaling-oriented Data Challenges, the DM group is beginning R&D into several **advanced computing and storage technologies** which are expected to be available at competitive prices during LSST construction. Notably we are **porting astronomical algorithms to GPUs** and evaluating the applicability of **solid-state storage (SSDs)** to supporting high database query rates.

There is also work on porting the LSST catalog and image simulation to GPU hardware: see the poster by Juric, *et al.*, in this session.

Future Data Challenges

The final Data Challenge during design (**DC4**) is planned to demonstrate **calibration products production**, test **alert generation and distribution**, exhibit the **scaling of the database** for production and science queries, test and demonstrate **fault tolerance** in the design, and allow for the testing of **science data access and visualization tools**. DC4 is expected to demonstrate **scaling of alert production to 20%** of the full operational throughput.

DC4 is expected to demonstrate the use in production of GPUs or other novel architectures, building on the prototyping work being done during DC3.

DC4 is intended to support **publication-quality research** with the precursor datasets from CFHT-LS. It will also involve a significant increase in the LSST simulated data required over DC3.

The Data Challenge approach can then continue to be applied during the construction phase of LSST, to assist in the testing and "blind" validation of science algorithms as they are developed, and take additional steps toward the full scale of computation required in the operational system.

LSST is a public-private partnership. Design and development activity is supported in part by the National Science Foundation. Additional funding comes from private gifts, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members.