UCRL-TR-227567

LAWRENCE LIVERMORE NATIONAL LABORATORY

Multi-Petabyte Image Data Management Systems

Donald D. Dossa, PhD

23 January 2007

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Auspices Statement

This work was performed under the auspices of the U. S. Department of Energy (DOE) by the University of California, Lawrence Livermore National Laboratory (LLNL) under Contract No. W-7405-Eng-48. The project (06-ERD-060) was funded by the Laboratory Directed Research and Development Program at LLNL.

FY06 LDRD Final Report
Multi-Petabyte Image Data Management Systems
LDRD Project Tracking Code: 06-ERD-060
Donald D. Dossa, PhD, Principal Investigator

Abstract

This research effort is directed to determine the methods and computational infrastructure needed to save, browse, and analyze multiple petabyte databases. The data set to be generated by the Large Synoptic Survey Telescope is used as a template for this research.

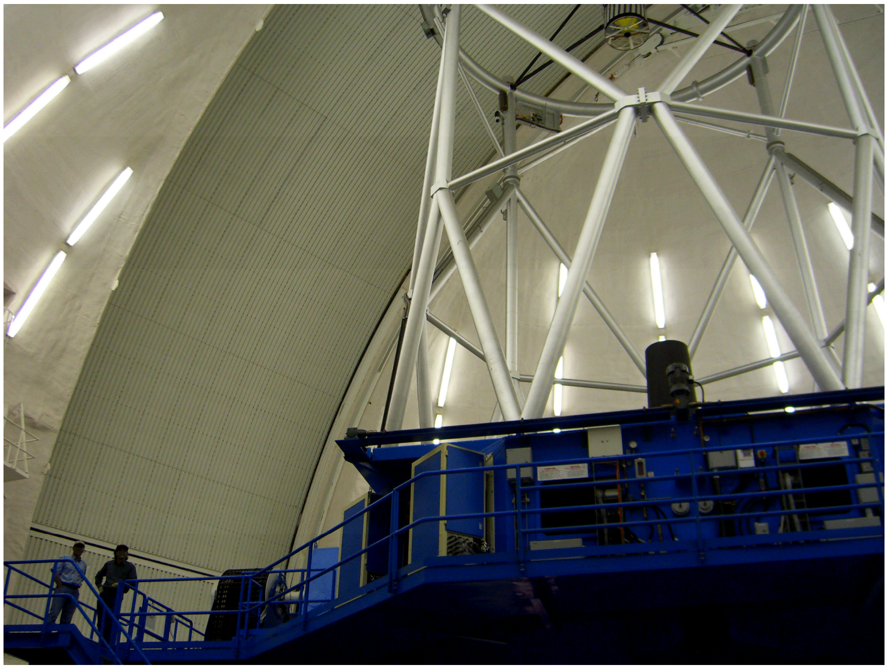
Introduction/Background

The Large Synoptic Survey Telescope (LSST) is an 8.4 meter telescope to be built on Cerro Pachon ins the Andes Mountains. It will generate a detailed map of the distribution of dark matter and dark energy in the universe. The multi-petabytes of data generated by LSST poses a significant challenge in data management, databases, and computational infrastructure. Thirty Terabytes of image data and associated meta-data will be generated every night for 10 years. The science requirements include a fast analysis of each image to determine if a significant astronomical event has occurred and generation an alert to the astronomical community. The amount of data generated every night makes alert checking impossible by human intervention. All alerts must be generated by the software with nearly100% accuracy.

Research Activities

The principal activities were in the computing infrastructure and database management.

The challenge in computing infrastructure is due to both the high data rate and large amount of data. The LSST will be located on a mountain top in the Andes Mountains in Chile. The camera system generates 15 Gbytes of data every 15 seconds over 201 data streams. The data acquisition research was a joint effort between LLNL and the camera team located at the Stanford Linear Accelerator Laboratory which is funded by DOE.



A view of the interior of the Gemini South Telescope near the site of LSST. The blue structure is a part of the mirror support. It is too large to fit into the field of view. LSST will be slightly larger. There are 2 people standing next to the mirror on the left. The PI is the person on the right.

The computing systems at the summit must acquire the image data and transmit it to the base computing system located 80 km away over Andes Mountains and desert. The science requirement is that no image data is lost. Several methods to meet this science requirement were examined. This included determining an optimal method to receive the data and a 3-way redundant data preservation system both at the mountain top and base camp capable of storing 4 nights of data.

Several advanced technologies options were examined that could possibly meet the requirement to move the image from the mountain top over 80 km of Andes Mountains and Chilean deserts to a computing base camp at La Serena, Chile. The high data rate and unfriendly environmental conditions eliminated the conventional wide area networks commonly used by telecoms and cable providers in the US. A decision was made early in the project to focus on a long-haul optical network. Calculations were made of the available data bandwidth available from the mountain to the base using laser drivers into 80 km of 1550 μ fiber optic cable. The attenuation in the fiber and sensitivity of the detectors at the base computing camp were used to insure that the data stream from the telescope was reliably received.

The view from the telescope site toward the base camp in La Serena (not visible). The data from the telescope will be transmitted over this terrain.

To meet the fast science alert times, a 60 TF computing system was designed to do the calculations using an approximately 1 Petabyte database. The astronomical analysis pipelines include functions for flat fielding, world coordinate system, swarping and source extraction. These output of these preliminary calculations are merely the first stage in the pipelines and are used for supernova and moving object detections and alerts. The LSST should detect about 1,000 nova in each image and hundreds to thousands of asteroid, and 200 million galaxies. All of these objects need to be detected and compared to the database entry for that object. There were several paper benchmarks on promising advanced computer architectures including a successor to BlueGene/L, the IBM Cell, accelerator boards, FPGAs, and commodity systems. Commodity systems had the lowest cost, ease of programming, and believable 10 year upgrade paths. A 60 Teraflop computing system would be the largest computer in South America. This poses its own challenges in terms of operation and support.

Downtown La Sera, Chile

The computing needs were combined with another requirement of preserving 4 nights of data. To have a reliable system, spare computing hardware and special keep-alive detection software, combined with job control software from NCSA, will advance LSST's capability to automatically detect hardware and software failures and reconfigure the system to continue the image processing, all without affecting the data analysis from the other 200 data streams. This method is more robust but less flexible that the usual MPI programming method in use at the national labs for writing large parallel programs. Several file systems were investigated and some benchmarks were run at NCSA. Among several file systems, Lustre appears to have the combination of highest bandwidth and reliability.

Further data analysis and long term science investigations will be carried out at the designated archive at NCSA. The computing requirement there is 3-4x higher that the computers in La Serena, Chile. A computing architecture similar to the base camp would significantly improve the ease of use and science analysis programs. The database size at NCSA would grow to several petabytes and the image data would increase by ~ 30 Terabytes per night. Unlike the base camp which needs the ability to store 4 nights of images, the archive site must maintain all of the data for at least 10 years. The combined image data will exceed 100 Petabytes, all of which must be searchable and retrievable. The high-priority DOE mission of research in dark matter and dark energy results in a need to perform 2 point correlation calculations over every pair of points across the sky and also massive weak lensing calculations involving galaxy shear. Normal methods of submitting queries to the database stored on disk would never complete during the lifetime of the astronomer. To enable the science to be done, a 3-tier storage system was designed. From the top-tier down, the storage architecture resembles the L1/L2 cache and main memory of modern microprocessors, albeit at greatly increases in size and latency. A view of the hierarchy from the bottom up resembles a web source/proxy/server architecture. Members of the team with experience in both areas made progress in defining a workable solution.

Results

Using the LSST science requirements, the research showed that for the foreseeable future, the use of industry standard blade servers and the Lustre file system is most likely optimal solution for the mountain top. A 60 TF system at the base camp is best configured based upon 3 Infiniband 4x DDR switches using commodity 4 slot 8 core processor nodes and 2 large Lustre based file systems. A similar but larger computing and database system at the archive site should be built.

Exit Plan

Continued collaboration with the DOE camera team at SLAC and with the archive site at NCSA will continue. To goal is to further improve the entire hardware and software architectures, autonomic hardware recovery, and fast database searches.

Acknowledgements

The PI acknowledges the assistance of Dr. Tony Tyson of University of California at Davis, the LSST PI, Dr Tim Axelrod of the University of Arizona, the LSST data management scientist, Dr Stuart Marshall at SLAC for his assistance on the camera output and software interfaces. Dr Ghaleb Abdulla at LLNL is the principal database expert. Dr. Chris Smith and Mr Ron Lambert at the Cerro Tololo Interamerican Observatory provided significant information on the environment at Cerro Pachon and La Serena and were wonderful hosts on the PI’s trip.