

# **LSST Data Management: Prospects for Processing and Archiving Massive Astronomical Data Sets**

**Andrew Connolly University of Pittsburgh**

The LSST presents a set of unique computational challenges: the processing of streaming imaging data in real time, the robust detection and classification of a range of static and moving objects and the need for almost real time identification of rapidly varying astrophysical sources. Independent of the final design of the LSST system we can expect that we will have to have the ability to analyze data coming from a multi-Gigapixel imaging camera that reads out every 10-20s in almost real time. To accomplish this we require a data processing and archiving facility that is capable of handling a sustained data rate of approximately 300 MB/s.

The challenges that such a data rate pose are numerous. How do we process 8-16 TB of data from every 8 hours worth of observing, what algorithms are close to optimal for the detection of variable and moving sources (to the limit of the data), what are the storage requirements for such a system and how do we get the raw and processed data from the telescope to the processing center and ultimately to the community as a whole. While these issues appear daunting we can benefit substantially from the precursor work of current and planned imaging surveys (e.g. the Sloan Digital Sky Survey, the Two Micron Survey, MACHO, SuperMacho and PanSTARRS). These surveys provide the testbeds on which to develop the necessary algorithms and prototype processing facilities. In this introduction to the data management of the LSST project we utilize the existing surveys to project the hardware requirements and the potential bottlenecks in a possible processing facility.

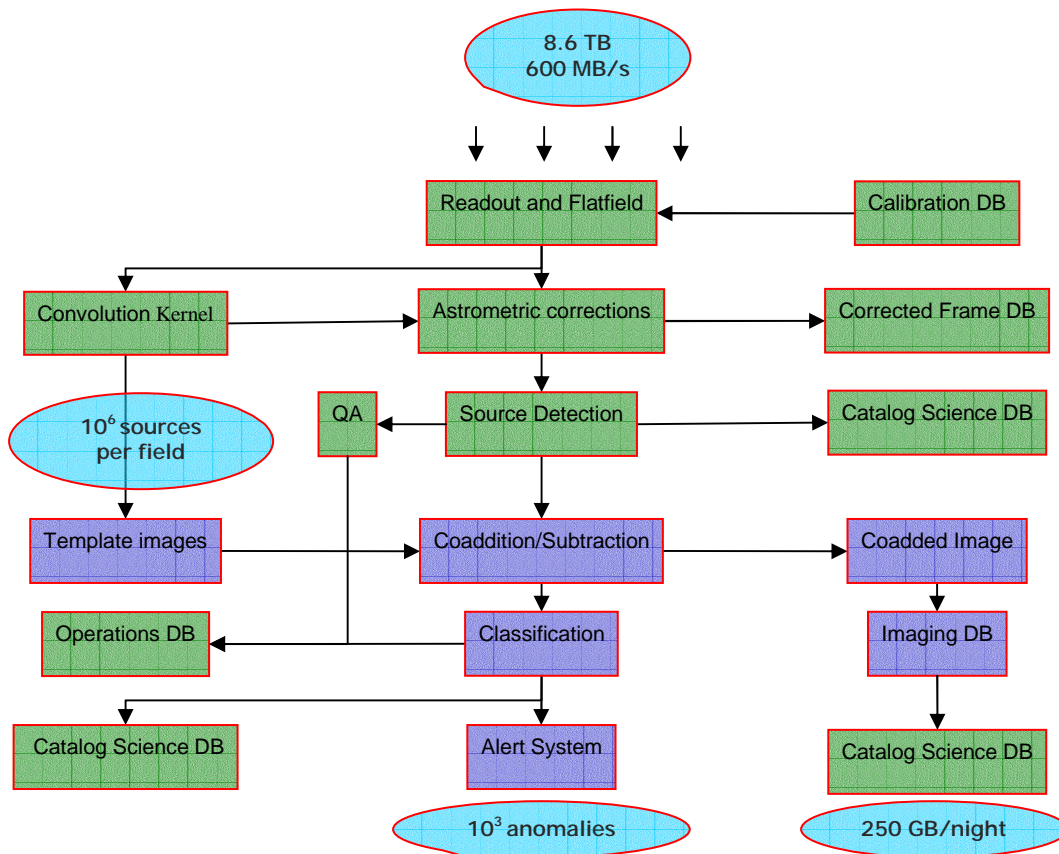
## Data Rates, Survey Volumes and Catalog Sizes

Assuming a basic design of a single 8.4m telescope with a 2.3 Gigapixel camera we can estimate the expected data flow and catalog sizes coming from the LSST. For a 10s integration time and a 7s duty cycle we expect to obtain 16 TB of data every 8 hours of observing (assuming 4 bytes per pixel). These data will come at a sustained rate of 540 MB per second. This compares to the sustained data rate obtained from the SDSS imaging camera of approximately 4.3 MB per second (i.e. science data during an imaging scan). Unlike current surveys that operate in a dual imaging and spectroscopic mode the dedicated nature of the LSST requires that the data be reduced in almost real time (in order that the data processing systems not become backlogged).

From a single 10s exposure it is expected that the LSST will reach a depth of approximately  $V=24$ . At these limits the number density of galaxies ( $6 \times 10^4$  per square degree) will dominate the source density. With an expected survey area of 14,000 square degrees, a single pass across the visible sky (which takes approximately 3 nights of observing time if two exposures are taken for each point on the sky) will produce a catalog of approximately  $8.4 \times 10^8$  sources. In a single year of observing it is expected that the LSST will be able to return to any point on the sky approximately 80 times. Over a five year period we can, therefore, expect to reach an equivalent depth of  $V>26$  in five

passbands over the full survey area. This will result in a final catalog of approximately  $3 \times 10^9$  sources.

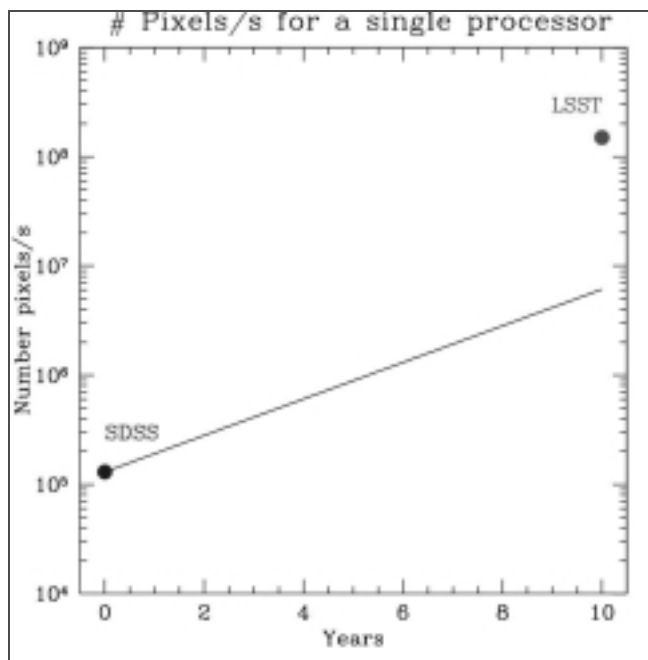
If we measure 100 parameters for each object detected in the LSST (flux, size, orientation, morphology etc) in each passband a single pass catalog of sources (in 5 bands) will amount to about 2 TB of data. The five year co-added dataset will produce around 5TB of catalog data. While this is not beyond the range of current database designs there is an additional component to the co-added data set. Each co-add can be considered independently as providing a time stream of observations (for measuring the variability of sources). It is this time domain data that present the most interesting challenge for archiving the catalog products from the LSST. The repeat observations increase the time domain catalog size to over 130 TB of data. Beyond this, archiving the reduced all-sky images (of a single pass) will require in excess of 3 Petabytes of data.



Flow chart of processing steps required to analyze LSST data

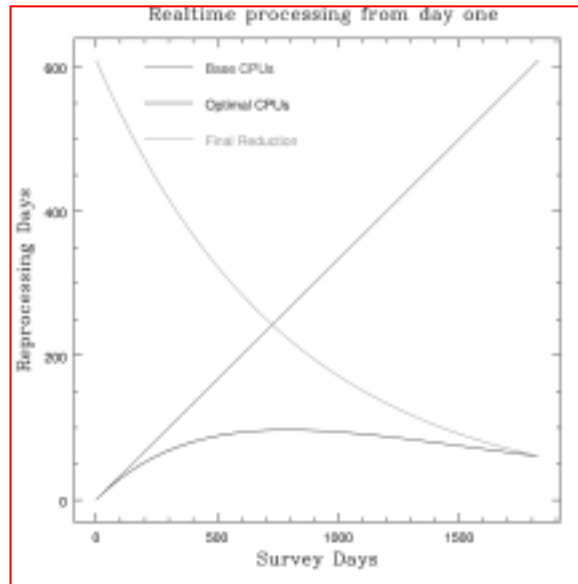
**Processing the LSST data: the challenges and a roadmap for the future**

Many of the processing steps required by the LSST reduction pipelines have already been implemented by the current generation of imaging surveys. We can, therefore, use the experience of these surveys to estimate the compute power required to process and store the LSST data. Current reduction pipelines require approximate 2-5000 operations per pixel in order to reduce the imaging data. Scaling these production systems (including the reading of the data to and from disk and caching into memory) to the size of the LSST data rate we would expect that today we would require approximately 1000 high-end workstations to keep up with the observations. Even with today's technology this is a feasible proposition. With the expected increase in processor power over the coming five to ten years we can expect that processing the LSST (allowing for four-fold redundancy in the processing power) will require on the order of 100 CPUs for real time reductions.



Processing requirements for a realtime data processing system compared to current data analysis systems

Storage of the raw and reduced images follows a similar design. Today we expect to be able to store 1 PB of for approximately 1 million dollars. Allowing for the expected increase in density of storage media and the decrease in cost (as we have witnessed throughout the last decade) the expected cost of a single Petabyte storage system should drop to approximately \$32,000 within six years. All of these factors show that while the processing of the LSST is an exciting challenge and that many new algorithms for processing and mining the data are likely to occur over the coming years there is no fundamental technological or algorithmic roadblock to prevent the success of this project.



Time required reprocessing the LSST data as a function of how long the survey has been underway. The red line shows the worst case scenario where we use the same processing system to reprocess the data as was available at the start of the survey. The blue shows time required if we take advantage of the latest processing technology when reprocessing the data. The green line shows the time to reprocess the data at the end of five years of the survey.

## Conclusions

Given the work that has been accomplished by current and planned imaging survey pipelines we can expect that the data rate for the LSST survey pipeline to be within the expected computational power of systems available by the time the LSST gets underway (i.e. systems with a few tens of nodes). In terms of algorithms prototype systems are now in place for the co-addition and subtraction of imaging data and will likely be optimized for variable seeing and photometric conditions by the SuperMacho and PannSTARRS groups prior to being implemented within an LSST framework. Algorithmically there do not, therefore, appear to be any fundamental roadblocks for processing of the data. The challenges we will face over the coming years will likely be related to how we manage the development of these software systems (controlling the features present within the software and the quality assurance of the reductions) together with how do we distribute the LSST data to these processing facilities and later on to the astronomical community as a whole.