Simulating the LSST System

A.J. Connolly[*a], John Peterson[b], J. Garrett Jernigan[c], Bob Abel[d], Justin Bankert[b], Chihway Chang[e], Charles F. Claver[f], Robert Gibson[a], David K. Gilmore[e], Emily Grace[b], R. Lynne Jones[a], Zeljko Ivezic[a], James Jee[g], Mario Juric[h], Steven M. Kahn[e], Victor L. Krabbendam[f], Simon Krughoff[a] Suzanne Lorenz[b], James Pizagno[a], Andrew Rasmussen[e], Nathan Todd[b], J. Anthony Tyson[g], Mallory Young[b]

[a]Department of Astronomy, University of Washington, Seattle, WA, 98195, [b]Department of Physics, Purdue University,West Lafayette, IN, 47907, [c] Space Sciences Laboratory, University of California, Berkeley, CA 94720, [d]Olympic College, Bremerton, WA 98337, [e]SLAC National Accelerator Laboratory, Menlo Park, CA, 94025, [f]National Optical Astronomical Observatory, Tucson, AZ, 85719, [g]Department of Physics, University of California, Davis, CA 95616, [h]Harvard-Smithsonian, Center for Astrophysics, Cambridge, MA 02138

ABSTRACT

Extracting science from the LSST data stream requires a detailed knowledge of the properties of the LSST catalogs and images (from their detection limits to the accuracy of the calibration to how well galaxy shapes can be characterized). These properties will depend on many of the LSST components including the design of the telescope, the conditions under which the data are taken and the overall survey strategy. To understand how these components impact the nature of the LSST data the simulations group is developing a framework for high fidelity simulations that scale to the volume of data expected from the LSST. This framework comprises galaxy, stellar and solar system catalogs designed to match the depths and properties of the LSST (to r=28), transient and moving sources, and image simulations that ray-trace the photons from above the atmosphere through the optics and to the camera. We describe here the state of the current simulation framework and its computational challenges.

Keywords: LSST, simulations, image, catalogs, telescope, atmosphere, camera

1. INTRODUCTION

The Large Synoptic Survey Telescope (LSST) will generate tens of terabytes of images and detect hundreds of millions of sources every night. The need to analyze these data in real-time, in order to identify moving sources (e.g. potentially hazardous asteroids) or transient objects (e.g. supernovae), requires that we develop new approaches for analyzing astronomical data streams; ones that can scale to the volume and complexity of the LSST. To enable the development of these techniques the LSST has undertaken a program to generate high fidelity simulations of the LSST data flow (comprising images and catalogs). The results of this work will be used in designing and testing algorithms for use by the data management groups, evaluating the capabilities and scalability of the reduction and analysis pipelines, testing and optimizing the scientific returns of the LSST survey and providing realistic LSST data to the science collaborations. In the following sections we outline the LSST simulation framework, describe the types of astronomical sources we simulate, the physics of the image simulations, how we access and process the data, and the computational challenges of distributing this workload across thousands of processors.

2. A SIMULATION FRAMEWORK

The framework underlying the LSST simulations is designed to be extensible and scalable (i.e. it is capable of being run on a single processor or across many-thousand core compute clusters). It comprises three primary components: databases of astrophysical catalogs, a system for generating observations based on the pointing of the telescope and a series of algorithms for simulating LSST images. Computationally intensive routines are written in C/C++ with the overall framework and database interactions using Python. The purpose of this design is to enable the generation of a wide

[*] ajc@astro.washington.edu

range of data products for use by the collaboration; from all-sky catalogs used in simulations of the LSST calibration pipeline, to studies of the impact of survey cadence on recovering variability, to simulated images of a single LSST focal plane.

In Figure 1 we show the overall flow of information within this framework. In the first subsystem (referred to as the *base catalog*) data are derived from cosmological N-body simulations, models for Galactic structure, simulations of Solar System sources, and other user generated catalogs. For all sources within the base catalog their spectral, photometric, astrometric and morphological properties (including proper motions and variability) are stored within a SQL database. These data provide a parameterized view of the universe outside of our atmosphere.

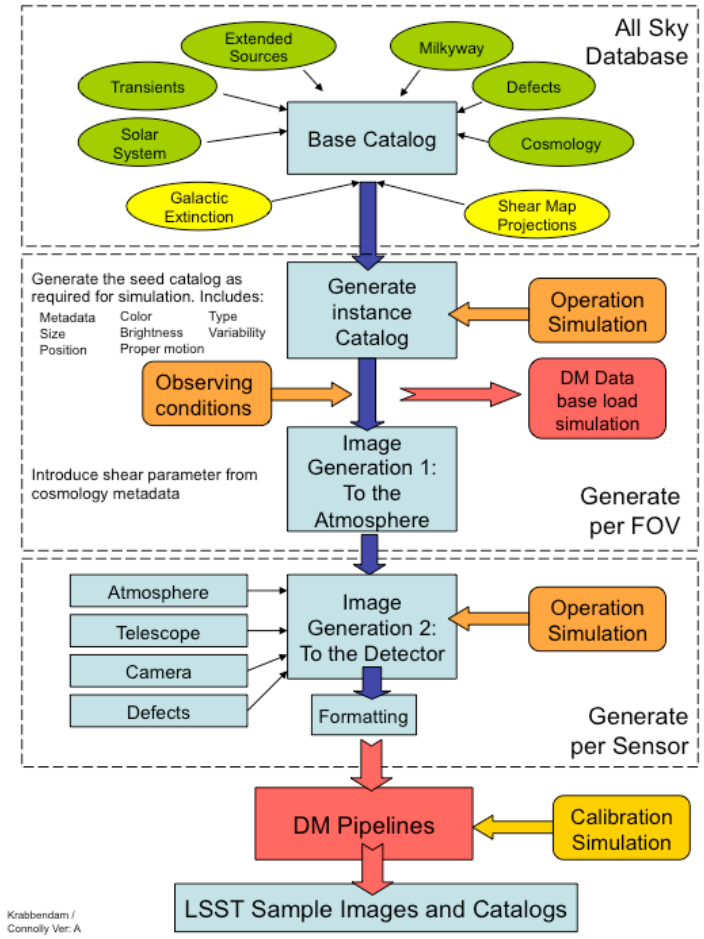

Figure 1. Data flow for the LSST simulation framework. The base catalog contains the underlying astronomical catalogs that are stored in SQL databases. These catalogs are queried based on the pointing of the LSST to generate instance catalogs that are either formatted for output for users or used as input to the image simulator. Images are simulated from the atmosphere to the camera (using fast ray-trace algorithms) and formatted to run through the LSST analysis pipelines.

This base catalog is queried using sequences of observations derived from the Operations Simulator. Each simulated pointing provides a position and time of the observation together with the appropriate sky conditions (e.g. seeing, moon phase and angle, and sky brightness). Positions of sources are propagated to the time of observation (from the proper motion information for stars and orbits for Solar System sources). Magnitudes and source counts are derived using the atmospheric and filter response functions appropriate for the airmass of the observation and after applying corrections for source variability. The resulting catalogs (*instance catalogs*) are then formatted for either output to users such as the science collaborations or fed into an image simulator.

In the third sub-system images are generated by ray-tracing individual photons through the atmosphere, telescope and camera systems. Photons are drawn from the spectral energy distributions that define the simulated data and ray-traced

through the optical system before conversion to electrons by simulating the camera physics. Images are read-out using a simulation of the camera electronics and amplifier layout and formatted for ingestion into the LSST data management system. All observing conditions, defined by the Operations Simulator, are propagated through the catalog and image generation to preserve fidelity and consistency between the derived catalogs and images.

3. CATALOG GENERATION

The current version of the LSST simulation framework (as used in Data Challenge DC3b) incorporates galaxies derived from an N-body simulation of a Λ-CDM cosmology, stars that match the observed stellar distributions within our Galaxy, asteroids generated from simulations of our Solar System, and a 3-D model for Galactic extinction.

Stellar sources are based on the Galactic structure models of Juric et al. (2008)[1] and include thin-disk, thick-disk, and halo star components. The distribution and colors of the stars match those observed by the Sloan Digital Sky Survey (SDSS, York et al. 2000[2]). Each star in the simulation is matched to a template spectral energy distribution (SED). Kurucz (1993)[3] model spectra are used to represent main-sequence F, G, and K stars as well as RGB stars, blue horizontal branch stars, and RR Lyrae. SEDs for white dwarf stars are taken from Bergeron et al. (1995)[4]. SEDs for M, L, and T dwarfs are generated from a combination of spectral models and by stacking spectra from the SDSS (e.g., Cushing et al. 2005, Bochanski et al. 2007, Burrows et al. 2006, Petterson & Hawley 1989, Kowalski et al. 2010[5,6,7,8,9]). The SED libraries were initially generated using a discretized grid of physical parameters. To provide smoother coverage of the stellar color-space, the output SEDs are generated by interpolating between the grid points. Proper motions for each star are based on the kinematic survey of Bond et al. (2010)[10]. Light curves are randomly assigned to a subset of the stellar population so that variability may also be simulated.

For Galactic reddening, a value of E(B-V) is assigned to each star using the three-dimensional Galactic model of Amores & Lepine (2005)[11]. To provide consistency with extragalactic observations the dust model in the Milky Way is re-normalized to match the Schlegel et al. (1998)[12] dust maps at a fiducial distance of 100 kpc. Figure 2 shows the effect of the extinction model on the derived colors (in this case the SDSS g-r colors). The left panel shows the density of stars within a region close to the galactic plane. The right panel shows the result of this extinction on the g-r colors of the stars (reddening by over 2 magnitudes).

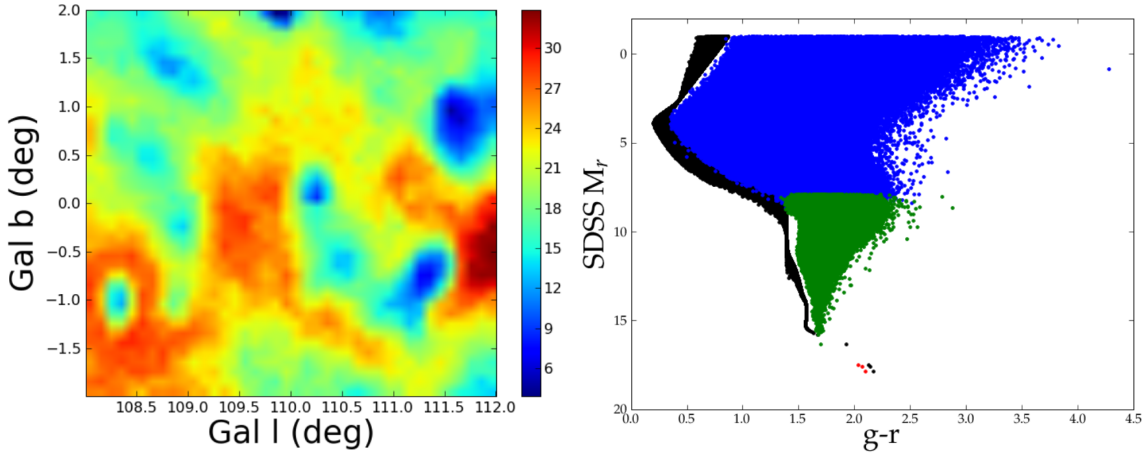

Figure 2. The left panel shows the density of stars for a region in the Galactic plane. The variation in density reflects the 3D distribution of dust in the model. The right panel shows the resulting g-r color magnitude diagram for the stars when the reddening by the dust is taken into account (note the absolute magnitudes have been corrected for dust extinction while the colors have not).

Galaxy catalogs are derived from the Millennium simulations of de Lucia et al.[13]. These models extend the dark matter N-body simulations to include gas cooling, star formation, supernovae and AGN and are designed to reproduce the observed colors, luminosities, and clustering of galaxies as a function of redshift. To generate the LSST simulated catalogs, a light-cone, covering redshifts $0<z<6$, was constructed from 58 $500h^{-1}$Mpc simulation snapshots. This light-

cone extends to a depth of approximately r=28 and covers a 4.5x4.5 degree footprint on the sky. Replicating this catalog across the sky simulates the full LSST footprint.

As with the stellar catalog, an SED is fit to the colors of each source using Bruzual and Charlot spectral synthesis models[14]. These fits are undertaken separately for the bulge and disk components and, for the disk, include inclination dependent reddening. Morphologies are modeled using two Sersic profiles. The bulge-to-disk ratio and disk scale lengths are taken from de Lucia et al. Half-light radii for bulges are estimated using the empirical absolute-magnitude vs half-light radius relation given by Gonzalez et al[15].

Comparisons between the redshift and number-magnitude distributions of the simulated catalogs with those derived from deep imaging and spectroscopic surveys showed that the de Lucia models under-predict the density of sources at faint magnitudes and high redshifts. To correct for these effects, sources are cloned in magnitude and redshift space until their densities reflect the average observed properties. The left panel of Figure 3 shows the r-band number-magnitude relation, after cloning, compared to a compilation of deep imaging surveys[16]. The right panel compares the redshift distributions, for a bright and faint sample, with those from the DEEP2 survey.

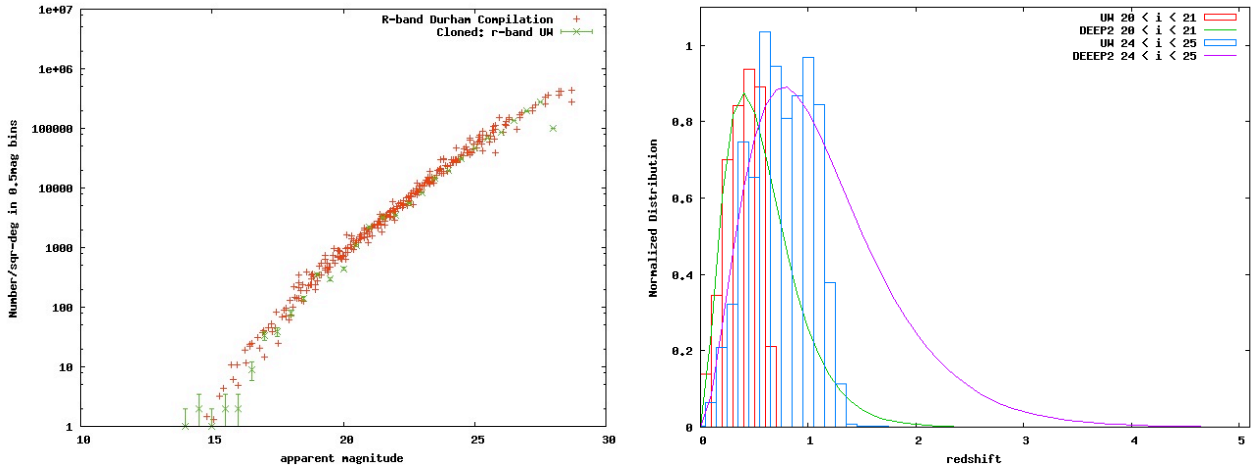

Figure 3. A comparison of the simulated galaxy number counts (left panel) and redshift distributions (right panel) with those observed from deep photometric and spectroscopic surveys. The number counts show data from a compilation of r-band imaging surveys. The lines in right panel reflect the redshift distributions from the DEEP2 redshift survey. Even with the cloning of the survey data the simulations do not generate as many high redshift galaxies as observed.

Asteroids are simulated using the Solar System models of Grav et al. (2007)[17]. They include: Near Earth Objects (NEOs), Main Belt Asteroids, the Trojans of Mars, Jupiter, Saturn, Uranus, and Neptune, Trans Neptunian Objects, and Centaurs. Spectral energy distributions are assigned using the C and S type asteroids of DeMeo et al (2009)[18]. Positions for the 11 million asteroids in the simulation are stored within the base catalog (sampled once per night for the ten year duration of the LSST survey). Querying the base catalog returns all sources within a given LSST pointing for which we generate accurate ephemerides using the PyOrb software package[19]. With typically 8000 sources per LSST plane, this procedure was developed to decrease the computational resources required to simulate asteroid ephemerides. The left panel of Figure 4 shows the distribution of eccentricities as a function of semi-major axis for the simulated populations. The right panel shows the positions of these sources within the field-of-view of the LSST after the orbits have been propagated and accurate ephemerides generated.

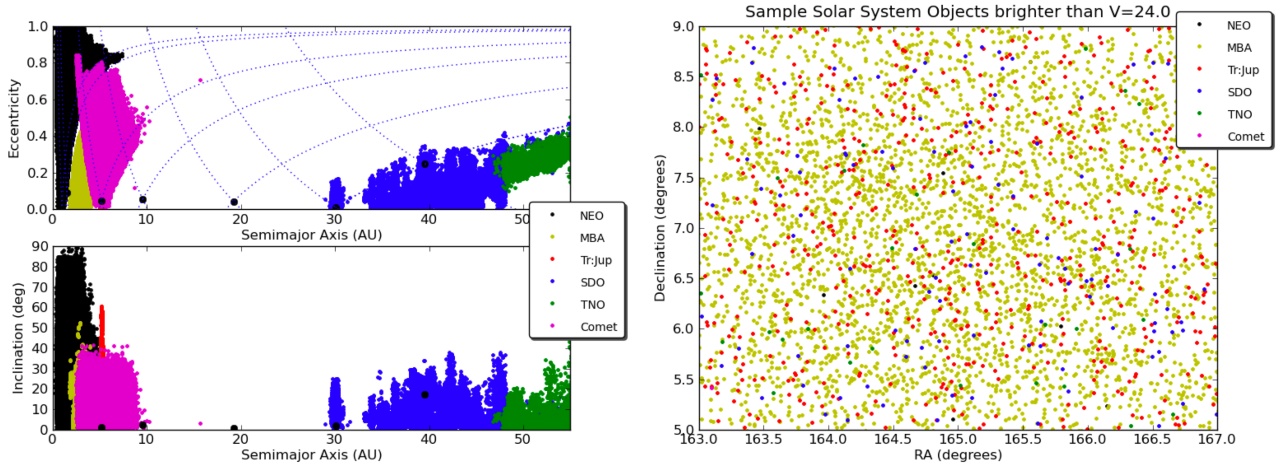

Figure 4. The distribution of asteroids in the LSST simulations is shown as a function of eccentricity and semi-major axis (left panel). By storing the position of these 11 million asteroids within a database (at a resolution of one position per night) we can efficiently prune those asteroids within the LSST focal plane and calculate their ephemerides (right panel).

4. IMAGE SIMULATIONS

The framework described above provides a parameterized view of the sky above the atmosphere. To generate images, photons are drawn from the spectral energy distribution of each source (scaled to the appropriate flux density based on the apparent magnitude of a source and accounting for the spatial distribution of light for extended sources). Each photon is ray-traced through the atmosphere, telescope and camera to generate a CCD image. The atmosphere is modeled using a Taylor frozen screen approximation (with the atmosphere described by six layers). The density fluctuations within these screens are described by a Kolmogorov spectrum with an outer scale (typically, 10m to 200m). All screens move during an exposure with velocities derived from NOAA measurements of the wind velocities above the LSST site in Chile. Typical velocities are on the order of 20 m/s and are found to have a seasonable dependence that is modeled when generating the screens. Each photon's trajectory is altered due to refraction as it passed through each screen. The left panel of Figure 5 shows an example of a single atmospheric screen with the color of the image corresponding to the density. The lower left panels of Figure 6 show the impact of the screens and the wind on the PSF and its homogenization.

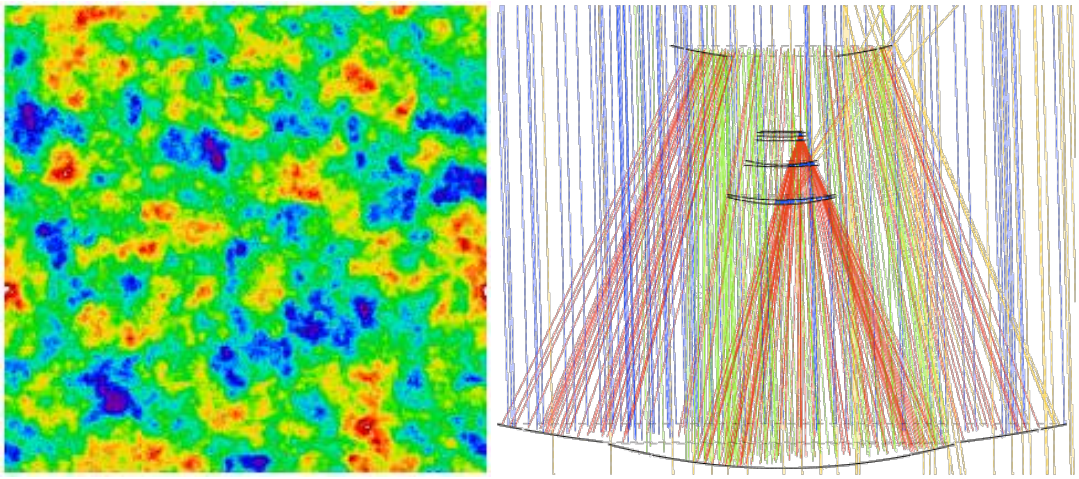

Figure 5. Images are simulated using a series of fast ray-trace algorithms. The left panel shows a single atmospheric screen (with the variations in density due to Kolmogorov turbulence). Six screens are used to simulate the atmosphere. The right panel shows the ray trace of photons through the telescope (including those photons that miss the optical surfaces). Geometric optical techniques are used to make this process efficient in order to simulate the 10^{11} photons in a single LSST focal plane.

After the atmospheric refraction, photons are reflected and refracted by the optical surfaces within the telescope (see Figure 5) and camera. The mirrors and lenses are simulated using geometric optics techniques in a fast ray-tracing algorithm and all optical surfaces include a spectrum of perturbations based on design tolerances. Each optic moves according to its six degrees of freedom within tolerances specified by the LSST system. Fast techniques for finding intercepts on the aspheric surface and altering the trajectory of a photon by reflection or wavelength-dependent refraction have been implemented to optimize the efficiency of the simulated images. Wavelength and angle-dependent transmission functions are incorporated within each of these techniques including simulation of the telescope spider.

Ray tracing of the photons continues into the silicon of the detector. Conversion probability and refraction (a function of wavelength and temperature) and charge diffusion within the silicon are modeled for all photons. Photons are pixelated and the readout process simulated including blooming, charge saturation, charge transfer inefficiency, gain and offsets, hot pixels and columns, and QE variations. Figure 6 shows the effect of these individual components within the image simulator.

The sky background is added as a post-processing step with the sky background generated, including Rayleigh scattering of the moon's light, based on SEDs for the full moon and the dark sky. The background is vignetted according to the results of ray-trace simulations. The simulator generates ~300,000 photons per second on an average workstation. In Figure 7 we show the full LSST focal plane (left panel). The variation in density of this image corresponds to the change in gain for each amplifier. In the right panel of Figure 7 we zoom in on a single amplifier region and show the resulting distribution of stars and galaxies (including the simulated diffraction spikes and bleed trails).

Figure 6. The image simulation framework is flexible enough to switch off and on different optical components. From left to right and top to bottom we show the resulting PSF as we progressively add more components to the optical path (including perturbations in the optical surfaces and a six-layer atmosphere).

5. SCALABLE SIMULATIONS

Catalogs and images generated for the LSST Data Challenge 3b will amount to approximately 1.5 million CCD images or 47TB of imaging data (two nights of LSST operations). To produce this volume of data requires approximately 1-2 million CPU hours and, therefore, necessitates the use of large compute clusters. There are a number of different approaches for distributed computing, including batch scripts (e.g. PBS), a Condor scheduler and general cloud computing models. All of these approaches have been implemented for the LSST simulations. In the following section we will focus on one of these techniques; simulating images in a Map-Reduce framework using Hadoop[20].

Commodity cloud computing has been shown to provide scalable solutions to a number of different computational challenges in the business community. In principle it abstracts the optimization of a problem, as the user often knows nothing of the underlying configuration of the compute system. The map-reduce model, implemented by Google, breaks a job into a series of serial tasks that are distributed across multiple processors (the mapper stage) and then the results of

these serial tasks are combined (the reducer stage). The user writes both the mapper and the reducer. We implement such a framework for the LSST simulations using Hadoop as deployed on the IBM/Google Clue cluster.

In the simulation framework the ray-trace code spawns photons from astronomically motivated source distributions above a model atmosphere. Since the buildup of the electrons in the detector from these simulated photons is, essentially, a linear process, the individual photon events can be assumed to be independent and treated as a sequence of serial operations. In the map-reduce framework we assign each map task a set of photons to trace and then a reduce task aggregates the resulting images into a final image containing all the co-added electrons. A threshold is applied to simulate saturation of the pixels and the image is stored in the shared filesystem. In the case of the CluE cluster this is the Hadoop distributed file system (HDFS).

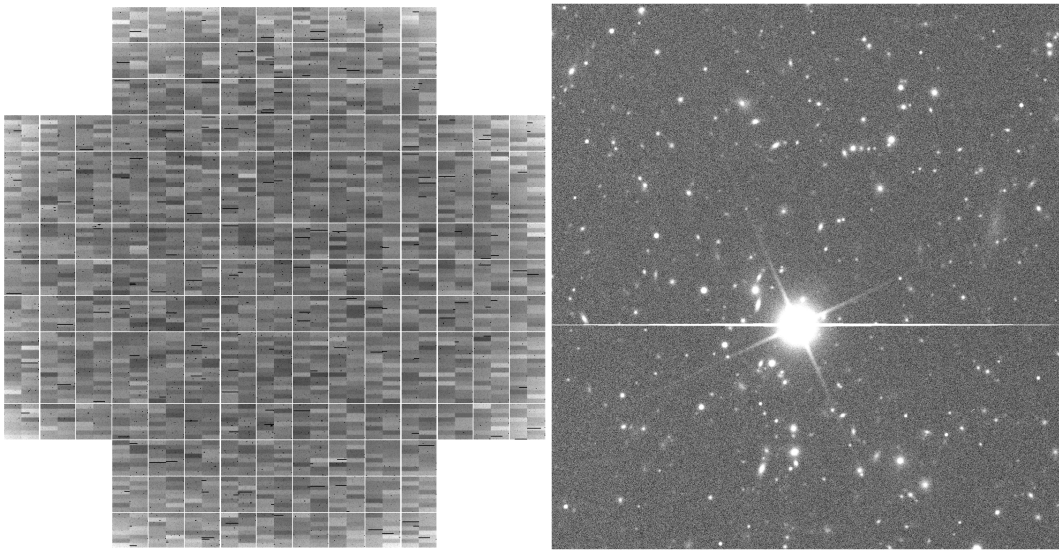

Figure 7. The left panel shows a full focal plane simulated for the LSST. The patchwork structure reflects the variation in the gain of the individual amplifiers. The right panel shows an image from part of a single amplifier. The distribution of galaxies, and the diffraction spikes and bleed trails of bright stars are visible within the image.

In our application, the interesting aspect is how to package up the sets of photons for each map task to simulate. In astronomy, we naturally partition by objects (i.e. stars galaxies, quasars, etc.). However, in a single frame the sources span ~5 orders of magnitude in flux, meaning that a single object can dominate the runtime of a single mapper. This suggests that careful consideration should be given to how best to partition the workload. The first cut at balancing the load is the most simplistic approach; that is to give each mapper the same number of objects to process. As pointed out earlier, this is unlikely to give the optimal performance, but is a starting point for determining the behavior of the system as a whole.

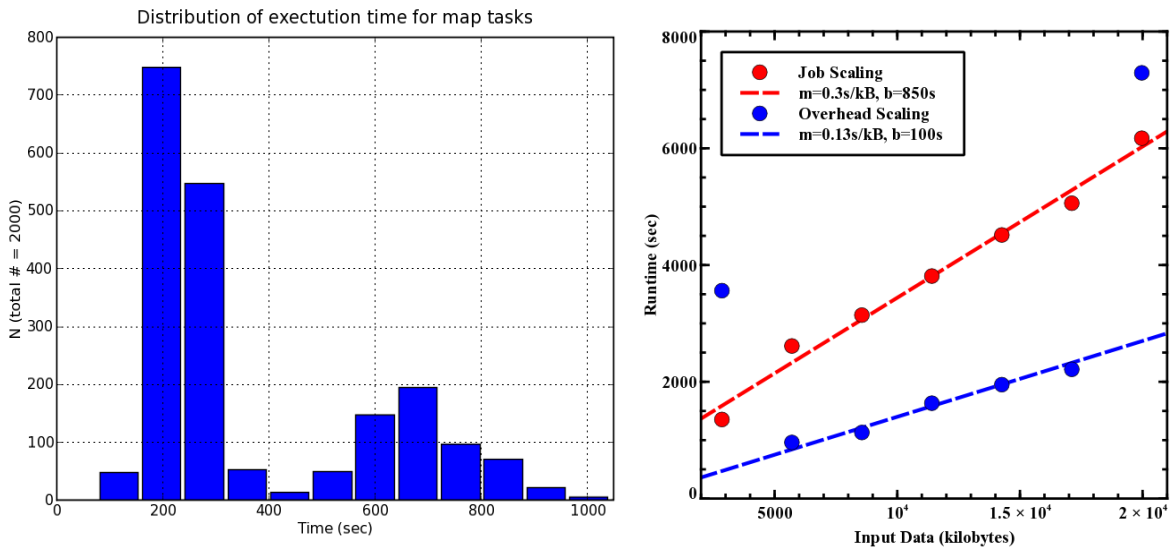

Figure 8. Simulating LSST volumes of data requires large computational resources. One such approach uses the map-reduce paradigm common in cloud computing to break down the image simulation into a set of serial tasks that are mapped to a large number of processors. The left panel shows the distribution of execution times for each task when we break the simulation on an object-by object basis. Bright sources dominate the run time. The right panel shows the scaling of the mapper and reducer as a function of data volume.

The left pane of Figure 8 shows the distribution of runtimes for 2000 map tasks for the simulation of a single CCD image. The bimodal distribution indicates that the runtime is not balanced over all map tasks. The time to simulate a single object is, clearly, proportional to the flux of the object. Thus, bright objects tend to dominate runtime. The simple approach to splitting the objects among mappers is not sufficient to balance the map task runtimes as the density of bright sources on the sky is not uniform.

The distribution of runtimes of the map tasks is but one consideration when implementing the simulation using cloud-computing resources. The overhead of job scheduling and I/O impact the scaling of runtimes in non-intuitive ways. In testing the scaling of our problem in the map-reduce framework, we first scale the problem size while maintaining the number of mappers at 100 and the number of reducers at one. The right pane of Figure 8 shows scaling of the runtime with input size. The total runtime was broken into the overhead portion (blue points) and job portion (red points). The job time is the time from the first mapper starting to the reducer finishing. The overhead associated with the task is mostly due to writing work units to HDFS and starting the jobs. This indicates that there is a minimum unit of work before the overhead dominates (in this case a few hundred sources). For runs larger than this minimum, both the overhead and job runtimes scale close to linearly with the job size. Since we have no control over how long it takes the scheduler to start jobs, improvements are needed in optimizing the file transfer. The outlier blue points are due to errors in writing to HDFS. The fact that the red line, which is the overall runtime scaling, has a positive slope indicates that the reduce phase is the dominant component of the runtime. This reveals another important synchronization point. With only one reducer, the amount of work in the serial phase can easily outpace that of the parallel mappers.

A less naïve approach is to split all objects into a superposition of several objects so that the total number of photons per object is more constant across the catalog. The benefit is that, by reordering the catalog, an approximately constant number of photons can be given to each mapper to simulate. The limitation on this approach is that some preprocessing must be done to the input catalogs and significant domain specific knowledge is required to do the processing (e.g. when simulating bright stars that saturate we optimize the computational load by increasing the weight of an individual photons and simulating fewer of them). A third approach is to have each mapper take a set of objects, but to emit only a random subset of the photons generated by the catalog. This maximizes load balancing while confining the domain specific knowledge to the code base and input catalog generation. It also has the benefit that the reducer can aggregate electron events rather than images. Since the images will, in general, have many zero pixel values, this saves significantly on data transfer overhead.

6. CONCLUSIONS

The LSST simulations group is creating a framework to simulate catalogs and images that reflect the expected properties of the LSST survey. This framework is designed to be extensible and scalable, enabling users to add their own data and to simulate of data products ranging from all-sky catalogs of sources to images of an individual CCD. The initial work has focused on generating images for the LSST data management group and to support the LSST data challenges. Future work will validate the performance of the simulations using existing telescopes and data sets, and increase the fidelity of the input catalogs. This will include: incorporating new cosmological simulations that better match the observed properties of distant galaxies, including gravitational lensing within the generated light cones, using deep multi-band HST data sets to constrain the morphological properties of galaxies at high redshift, adding AGNs, supernovae, and RR Lyrae stars, introducing artifacts into the optical system such as glints and ghosts, and extending the framework to enable users to ingest their own data.

This material is based upon work supported through the National Science Foundation grants AST-0551161 (to LSST collaboration for design and development activity) and IIS-0844580 (as part of the Cluster Exploratory, CLuE, program).

REFERENCES

[1] Juric, M., et al., "The Milky Way Tomography with SDSS. I. Stellar Number Density Distribution", Astrophysical Journal, 673, 864-914 (2008)
[2] York, D., et al., "The Sloan Digital Sky Survey: Technical Summary", Astrophysical Journal, 120, 1579-1587 (2000)
[3] Kurucz, R.L., "CD-ROM No.13", Cambridge, Mass.,Smithsonian Astrophysical Observatory, (1993)
[4] Bergeron, P., Wesemael, F., & Beauchamp, A., Wesemael, F., "Photometric Calibration of Hydrogen- and Helium-Rich White Dwarf Models", Publications of the Astronomical Society of the Pacific, 107, 1047-1054 (1995)
[5] Cushing, M.C., Rayner, J.T., & Vacca, W.D., "An Infrared Spectroscopic Sequence of M, L, and T Dwarfs", Astrophysical Journal, 623, 1115-1140 (2005)
[6] Bochanski, J.J., West, A.A., Hawley, S.L., & Covey, K.R., "Low-Mass Dwarf Template Spectra from the Sloan Digital Sky Survey", Astronomical Journal, 133, 531-544 (2007)
[7] Burrows, D., Sudarsky, D., & Hubeny, I., "L and T Dwarf Models and the L to T Transition", Astrophysical Journal, 640, 1063-1077 (2006)
[8] Pettersen, B.R., & Hawley, S.L., "A spectroscopic survey of red dwarf flare stars", Astronomy & Astrophysics, 217, 187-200 (1989)
[9] Kowalski, A., Hawley, S.L., Holtzman, J.A., Wisniewski, J.P., Hilton, E.J., "A White Light Megaflare on the dM4.5e Star YZ CMi", Astrophysical Journal, 714, L98 (2010)
[10] Bond, N., Ivezic, Z., Sesar, B., Juric, & Munn, J., "The Milky Way Tomography with SDSS: III. Stellar Kinematics", arXiv:0909.0013 (2009)
[11] Amôres, E.B., & Lépine, J.R.D., "Models for Interstellar Extinction in Galaxy", Astronomical Journal, 130, 659-673 (2005)
[12] Schlegel, D.J., Finkbeiner, D.P., & Davis, M., "Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds", Astrophysical Journal, 500, 525-553 (1998)
[13] De Lucia, G., Springel, V., White, S.D.M., Croton, D., & Kauffmann, G., "The Formation History of Elliptical Galaxies", Monthly Notices of the Royal Astronomical Society, 366, 499-509 (2006)
[14] Bruzual, G. & Charlot, S., "Stellar population synthesis at the resolution of 2003", Monthly Notices of the Royal Astronomical Society, 344, 1000-1028 (2003)
[15] González, J.E., Lacey, C. G., Baugh, C. M., Frenk, C. S., & Benson, A. J., "Testing model predictions of the cold dark matter cosmology for the sizes, colours, morphologies and luminosities of galaxies with the SDSS", Monthly Notices of the Royal Astronomical Society, 397, 1254-1274 (2009)
[16] http://astro.dur.ac.uk/~nm/pubhtml/counts/counts.html
[17] Grav, T., Jedicke, R., Denneau, L., Holman, M. J., & Spahr, T., "The Pan-STARRS Synthetic Solar System Model and its Applications", BAAS, 211, 4721 (2007)

[18] DeMeo, F.E., Binzel, R.P., Slivan, S.M., & Bus, S.J., "An Extension of the Bus Asteroid Taxonomy into the Near-Infrared", Icarus, 202, 160-180 (2009)
[19] Granvik, M., Virtanen, J., Oszkiewicz, D., & Muinonen, K., "OpenOrb: Open-Source Asteroid Orbit Computation Software Including Statistical Ranging", Meteoritics & Planetary Science, 44, 1853-1861 (2009)
[20] http://hadoop.apache.org/