Designing for Peta-Scale in the LSST Database

Jeffrey P. Kantor, Tim Axelrod, Jacek Becla, Kem Cook, Jim Gray, Sergei Nikolaev, Ray Plante, Maria Nieto-Santisteban, Alex Szalay, Ani Thakar
A Quick Look at LSST

- Aperture diameter: 8.4m
- Effective aperture: 6.7m
- FOV: 3.5 deg
- Filters: u, g, r, i, z, y
- Observing mode: pairs of 15 sec exposures, separated by 5 sec slew
- Single exposure depth: 24.5
- Site: Cerro Pachon, Chile
- On sky: Late 2013
Key Requirements driving Data Management Architecture

**Key Requirements**
- 4 missions
- \((15s \exp +2s \text{ r/o})^2 + 5s\) slew cadence
- 30s transient alert latency
- 2/yr data release frequency
- 10 degree\(^2\) field of view
- 5 of 6 filters in carousel
- 0.2 arcsec optics
- 3.2 gigapixel focal plane array
- 5 milliarcsec astrometric precision
- 0.3% photometric precision

**Key Derived Data Management Requirements**
- 250 galaxies/arcmin\(^2\) at \(r=27.8\)
- 140 stars/arcmin\(^2\) at \(r=27.8\)
- 60 days cadence for deep detection
- 5 stars/arcmin\(^2\) (average)
- 70 stars/arcmin\(^2\) (include galactic center)
- 1% transient rate
- Open data access for up to 300 simultaneous users
- Open algorithm access for broader impacts

**Data Products**
- Images
- Catalogs
- Alerts

**Algorithms/Pipelines**
- Astrometric/Photometric Calibration
- Source Detection
- Source - Object Association
- Moving Object Detection/Orbit Matching
- Alert Validity
- Deep Detection/Association

**Computing Architecture**
- Multiple processing sites
- Sites specialized by role
- Layered architecture
- Automated data quality analysis
- Abstract data storage and extendable data
- Plug-and-play pipelines and abstract parallel processing machines
- Open, standard data access interfaces
Derived data and processing rates

- **Functional Requirements Document (#355)**
- **Operational Concept Document (#356)**
- **UML Use Case, Domain Models (#466)**

**DMS Sizing**
- Processing (#1193, 2116)
- Communications (#1990, 1991)

Sustained (peak) processing:
- 105 (120) TFLOPS aggregate
- Parallelism maximally exploited
- 98% availability

Sustained (peak) I/O Rates:
- 60 GB/s aggregate read
- 6 GB/s aggregate write
- 15 PB/year average growth
- 0% data loss
- 98% availability

Sustained (peak) bandwidth:
- 2.5 (10) Gb/s
- 0.1% alert publication failure
- 98% availability

- 8.5$\times 10^4$ ops/pixel,
- 6.4$\times 10^6$ ops/source,
- 9.0$\times 10^3$ ops/object,
- 4.7$\times 10^7$ ops/moving object

- process an image in 17s

- 2 - 3 Bytes/pixel transfer
- 4 Bytes/pixel process
- 4 Bytes/pixel archive,
- 300 Bytes/source,
- 1.8 KB/object, 100 KB/alert

- 6s ingest latency,
- 2.5x overhead

- 24 hrs /nightly transfer to archive,
- 300 concurrent users
- 0.5GB/10s$\times 50$ +
- 6GB/60min$\times 20$
- queries served

ADASS 2006
Tucson, AZ
October 16, 2006
LSST DMS Centers and Data Flows

**Base Facility**

- **High-Speed Storage**
  - Sky Template, Catalog Data
  - Alerts
- **Pipeline Servers**
  - Data Products
  - Raw Data, Meta Data

**Archive Center**

- **Archive Ops Servers**
- **Pipeline Servers**
- **High-Speed Storage**
  - Raw Data, Meta Data
  - Alerts

**Data Access Centers**

- **Archive Ops Servers**
- **High-Speed Storage**
- **Level 1 End User**
  - Data Products
- **Level 2 - 4 End User**
  - Data Products

**Mountain Summit**

- **Data Acquisition Interface**
  - Raw Data, Meta Data
  - Xtalk Corrected Data, Raw Data, Meta-Data
  - Alerts
- **High-Speed Storage**
  - Meta-Data DQA
  - High-Speed Storage

**Arrows Key:**
- Green: real-time, in DC1, DC2
- Red: - nightly or longer, in DC3
- Blue: - on-demand, in DC4

*LSST Camera Subsystem: Instrument Subsystem*

*LSST OCS: Observatory Control System*

ADASS
Tucson, AZ
October 16, 2006
Dark fiber capacities are in excess of 2 Tb/s today. LSST will use 2.5 Gbps protected/10 Gbps fiber optic networks (REUNA, LAUREN, and WHREN-LILA) to connect the Mountain/Base and to the U.S.

LSST will use existing NSF-funded networks (and their successors) for data transfer and distribution.

LSST traffic will drive lit fiber capacities on all LSST links to levels beyond the core LSST requirements.
Computing Requirements are within Supercomputing technology trends

Computing Requirements by Year

- Science/Operations
- Spares
- Transients
- Red. Images
- DQ Analysis
- Queries
- Deep Det.
- Routine
- Nightly
- Initial

Year

Tera_Floating Point Operations (TF)
Data Catalog Volumes and Growth

Data storage is sized to accommodate observing near/in galactic plane.

Images and Catalogs are immutable once released. Two most recent and as yet unreleased catalogs on fast disk, others on tape plus disk cache.
Primary Data Base Tables and Queries

---

**Image Metadata**
- 675 million rows*
- 1 row = metadata for 1 ccd-amplifier

**Source**
- 260 billion rows*
- 2,000 partitions*
- 306 bytes/row
- 1 row = data for 1 filter

**Object**
- 22 billion rows*
- 2,000 partitions*
- 1.8 KB/row
- 1 row = data for 6 filters

---

**Queries**

- Select all galaxies in given area
  - Object.type index
  - Object.(ra,dec) index, full index scan
  - fetch data rows

- Select transients var obj near a known galaxy
  - VarObj.(ra,dec) full index scan

- Cone-mag-color search, ra, decl-best selectivity
  - Object.(ra,dec) index
  - fetch data rows

- Cone-mag-color search, color-best selectivity
  - zMag index, full index scan
  - grColor index, full index scan
  - izColor index, full index scan
  - Object.(ra,dec) index
  - fetch data rows

- Find extremely red galaxies
  - Object.type index
  - Object.izColor index, full index scan
  - Object.xMag index, full index scan (x5colors)
  - fetch data rows

- Select time series data for given cone
  - Source.(ra, dec) index, full index scan
  - join result w/objectId index
  - join result w/tai index
  - Sort, assume in memory
  - fetch data rows for Source

---

* - as of Data Release 1, 2014
The Data Access Framework (DAF) provides ingest and query, transfer and replication services, encapsulating both distributed files and databases. The DPS and DAF ensures that all Pipelines and Data Products record metadata and provenance, so that the products can be re-generated on-demand.
The DAF supports open access via VO-compliant interfaces to all data and meta-data.

While LSST will provide a “base” set of tools and queries needed for DQA, we anticipate the VO community to provide additional tools for access, visualization, and cross-survey fusion.
Data Base Framework

Ingest Server
- ingest parallelized for scalability
- pipeline (writer)
- fault tolerance by mirroring

Cached Archived Data Server
- replicate ingested data
- split (partition) archived data
- pipeline (reader)

Data Center
- avoid competition between readers and writers
- replicate

General public
- execute query at the "best" place
- parallelize queries

Archive Center
- partition tables & indices
- cache major indexes in RAM
- combine disk arrays with virtual file systems

Professional & amateur astronomers
- load balance
Distributed File Systems in the DAF

• The DAF uses DFS for
• Staging input data for pipeline processing
• Staging output data for ingest
• Storing, replicating, and serving image files
• Current file systems under evaluation:
  • GPFS, Google File System, Lustre, IBRIX
Data Challenges validate the Design

- Data Challenge 1 July - October, 2006
  - Goal: Validate infrastructure and middleware scalability
  - Simulated data and applications running on TeraGrid
  - Simulated real-time data flows from Mountain to Base, through Nightly Pipelines, Ingest into Database, transfer to Archive, re-run Nightly Pipelines
  - Purdue cluster represents Mountain, NCSA represents Base Facility, SDSC represents Archive Center
  - Results - still tuning/improving, but results to date are:
    - 70 megabytes/second data transfers (>15% of LSST transfer rate)
    - 192 CCDs (0.1 - 1.0 gigabytes each) runs processed across 16 nodes/32 itanium CPUs with latency and throughput of approximately 75 seconds (>15% of LSST per node processing rate)
    - 4.5 megabytes/second source data ingest (>15% of LSST required ingest rate)

- Data Challenge 2 November, 2007 - Validate nightly pipeline algorithms
- Data Challenge 3 November, 2008 - Validate science pipelines, end-to-end data quality, and reliability
- Data Challenge 4 July, 2009 - Validate open interfaces and data access