LSST Database Architecture – Preparing for the Extreme Scale Analytics

Jacek Becla
SLAC National Accelerator Laboratory

ADASS XX
Boston, November 10, 2010

Outline

- LSST – intro
- Baseline architecture
- SSDs
- SciDB
- Summary

LSST - Intro

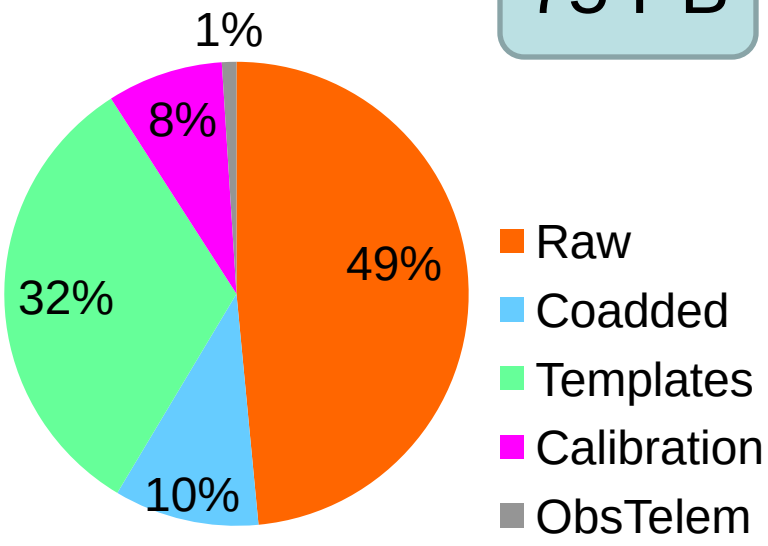
- Timeline
  - Multi-decades, in R&D now
  - Data Challenges
- Scale
  - ~75 PB images
  - ~70 PB database
- Complexity
  - Time series (order)
  - Spatial correlations (adjacency)

SLAC NATIONAL ACCELERATOR LABORATORY

Scale

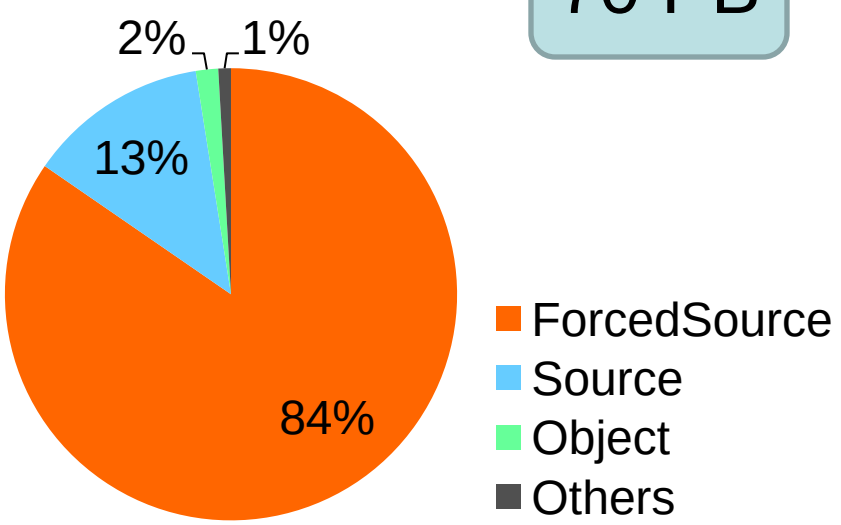
Images

75 PB



Catalogs

70 PB



The number are for all data releases, uncompressed

Database Schema (Key Tables)

←230→

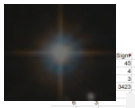
Object
0.1PB

↑ 50 B ↓

←82→

Moving Object
0.001PB

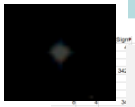
↑ 0.1 B ↓



←70→

Source
0.8PB

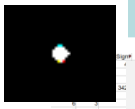
↑ 2.7 **T** ↓



←16→

Forced Source
12PB

↑ 116 **T** ↓



←51→

Dia Source
0.05PB

↑ 10 B ↓

←25→

Ccd Exposure
0.001PB

↑ 1 B ↓

Table sizes: the largest table in the last release

SLAC NATIONAL ACCELERATOR LABORATORY

DBMS or Map/Reduce

- Catalogues and placement
  - DMBS knows where relevant data resides, co-locate related sub-regions
  - M/R uses hash partitioning
- Schema
  - M/R schema in application
- Speed
  - M/R overheads ~15 sec
  - Need fast indexes
- Cost-efficiency
  - M/R checkpoints for each map and reduce
  - Spindle-level control needed to get max performance

Talking to Hadoop community

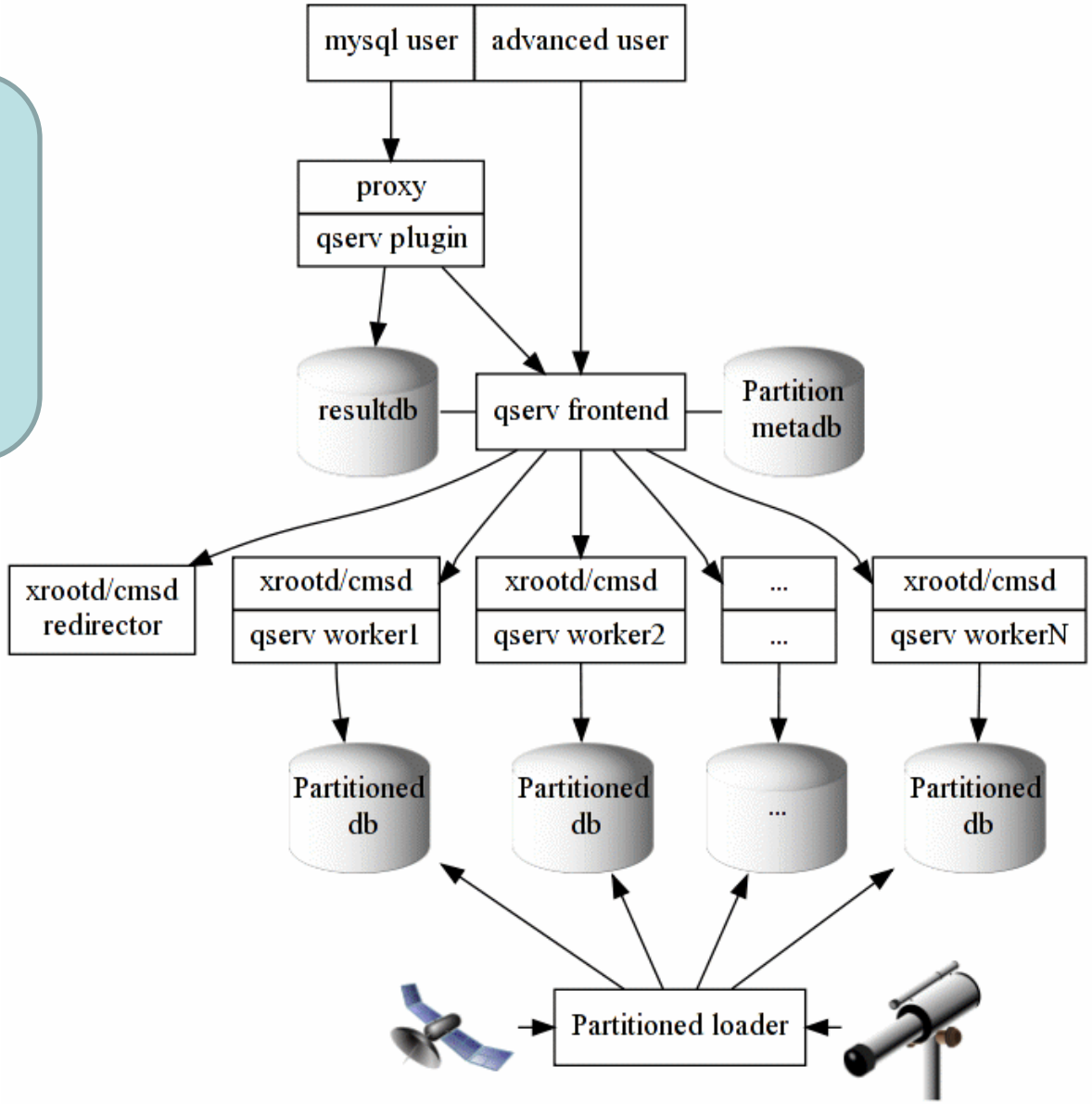
Both camps are rapidly converging

Query Service (qserv)

- Shared-nothing on top of MySQL
- Built for analyses on immutable data sets
- Optimized for spatial and temporal analyses on extreme scale data sets
- Overlapping partitioning, fixed chunks, 1st level materialized, 2nd on the fly
- Shared scans (available ~Q2'11)
- Fault tolerance
- Usable prototype in public domain

Qserv Architecture

Deploying for wide use by LSST science collaborations on multi-TB data set this year



Qserv Scaling Tests

- Tested with
  - 40 nodes (5+ years old hardware)
  - 10% of 1st LSST Data Release Object table
    - 1.5 billion 1KB rows
    - 2,000 partitions
- Demonstrated
  - <15 sec low-vol query vs LSST required 60 sec
  - <30 min high-vol query vs LSST required 60 min
- 100-node test scheduled for Nov-Dec

DC3b Tests

- Database isolated from pipelines
  - Ingest through csv during post-processing
- Multi-TB scale
  - Real and simulated data
  - Single server
- Setting up replicas for user analyses

Solid State Disks

- MySQL MyISAM on X25E SLC, X25M MLC
- Max bandwidth, random read demonstrated
- Sequential read
  - ~3-6x more expensive than mechanical disks
- Random read
  - SSD ~10-16x more IOPS/$
  - Slow mechanical disks often hide software deficiencies!
- Full table scan vs index access
- Considering SSDs for (at least) indexes

Acknowledgement: used hardware was provided by SDSC

SciDB

LSST Large Synoptic Survey Telescope

- Data Management and Analytics Software
- Array data model
  - Can be nested, ragged, unbounded, sparse
  - Native array operators (aggregate, regrid, subsample, covariance…)
  - Extensible with Postgres-style user-defined functions
  - Native support for uncertainty, provenance
- Shared nothing, distributed
  - Vertical partitioning into single-attribute arrays
  - Each single attribute array divided into overlapping chunks
- APIs
  - Array Query Language (AQL), Array Internal Language (AIL)
- V0.5 out

SLAC NATIONAL ACCELERATOR LABORATORY

Summary

- Data intensive astronomical analytics
  - Most challenges due to correlations, data adjacency
  - Require novel techniques, e.g. overlapping partitions
- LSST baseline
  - Core part implemented, deployment imminent
- SSDs
  - Rapidly becoming part of data-intensive analyses
- SciDB
  - New open source, promising solution for complex scientific analytics