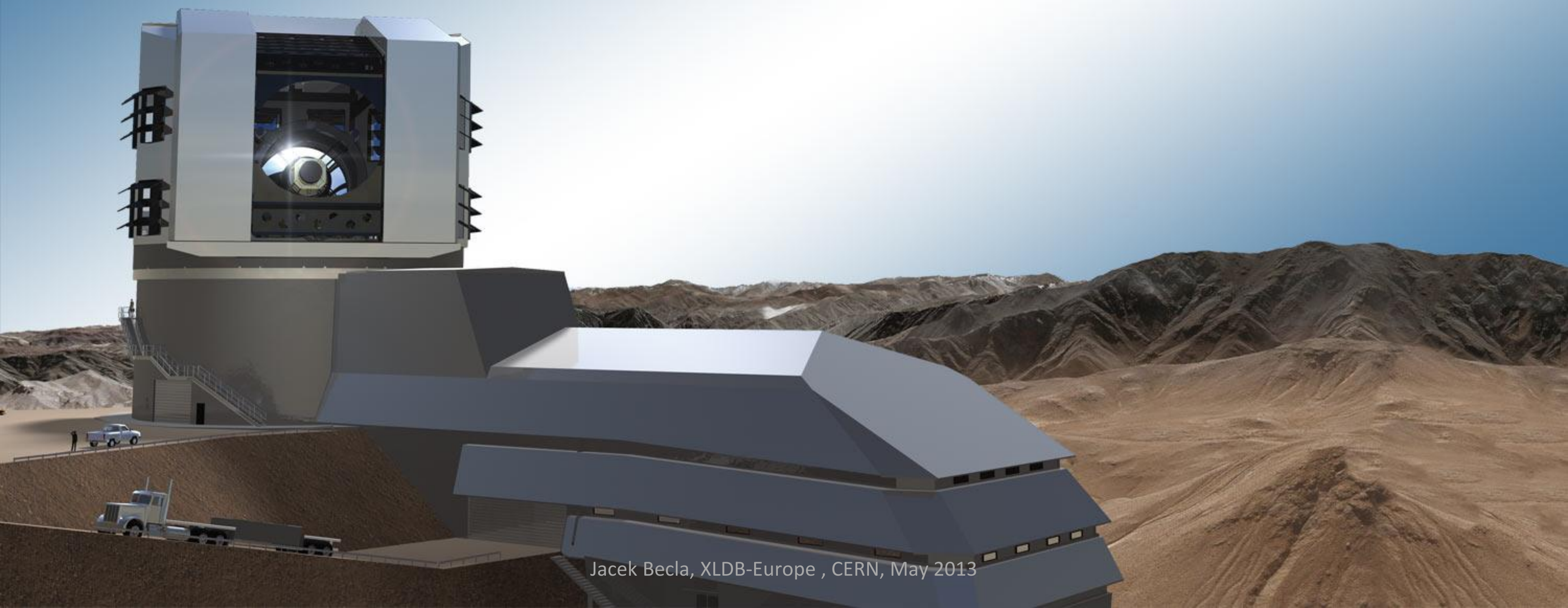
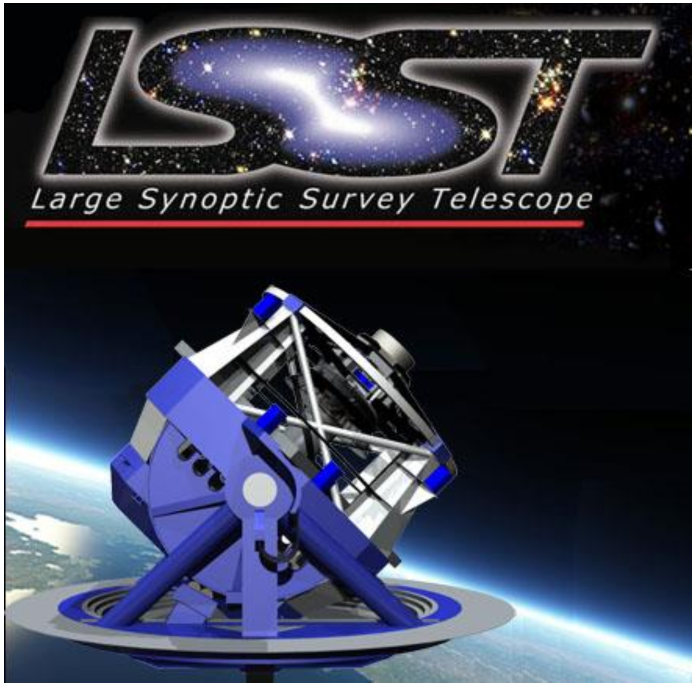
LSST Database

Jacek Becla



Large Synoptic Survey Telescope

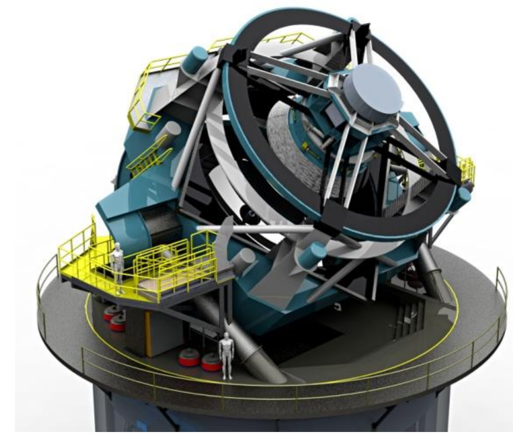
- Timeline
  - In R&D now, data challenges
  - Construction starts ~mid 2014
  - Operations: 2022-2031
- Scale
  - ~45 PB database*
  - ~65 PB images*
  - Plus virtual data
- Complexity
  - Time series (order)
  - Spatial correlations (adjacency)

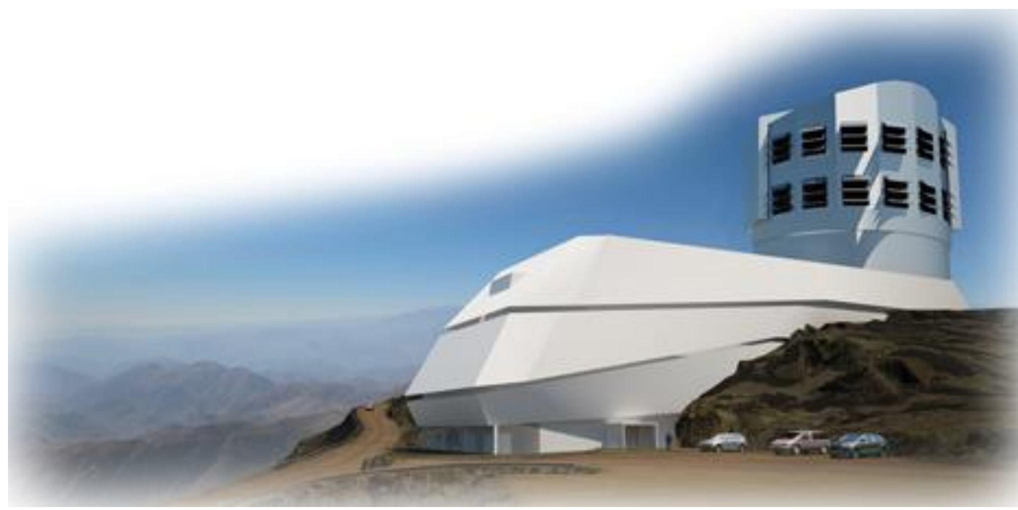


* *Compressed, single copy, includes db indexes*

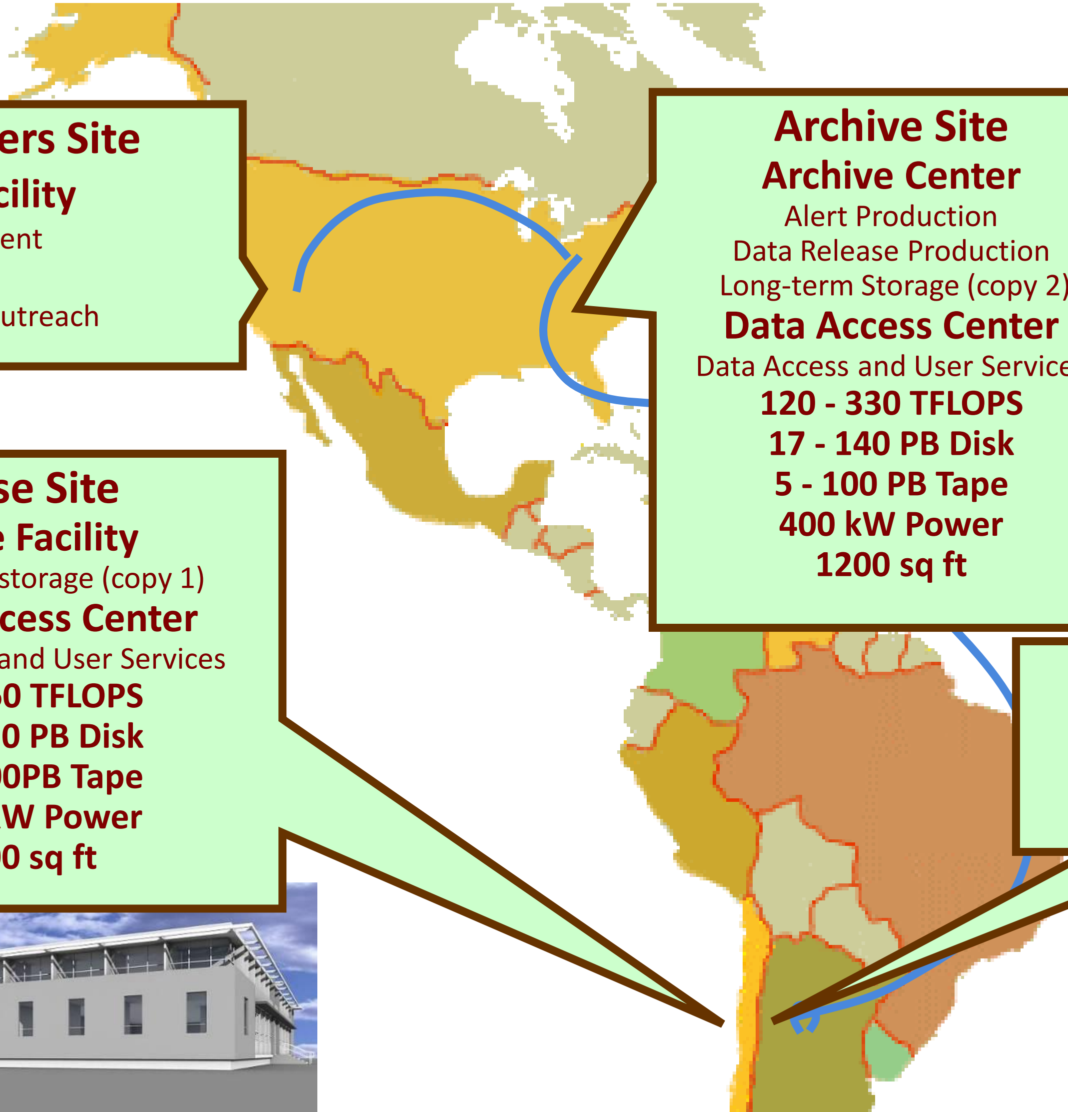
The LSST Scientific Instrument

- A new telescope to be located on Cerro Pachon in Chile
  - 8.4m dia. mirror, 10 sq. degrees FOV
  - 3.2 GPixel camera, 6 filters
  - Image available sky every 3 days
  - Sensitivity – per “visit”: 24.5 mag; survey: 27.5 mag
- Science Mission: observe the time-varying sky
  - Dark Energy and the accelerating universe
  - Comprehensive census Solar System objects
  - Study optical transients
  - Galactic Map
- Named top priority among large ground-based initiatives by NSF Astronomy Decadal Survey





LSST Data Centers



Headquarters Site

Headquarters Facility

Observatory Management
Science Operations
Education and Public Outreach

Archive Site

Archive Center

Alert Production
Data Release Production
Long-term Storage (copy 2)

Data Access Center

Data Access and User Services

120 - 330 TFLOPS
17 - 140 PB Disk
5 - 100 PB Tape
400 kW Power
1200 sq ft

Base Site

Base Facility

Long-term storage (copy 1)

Data Access Center

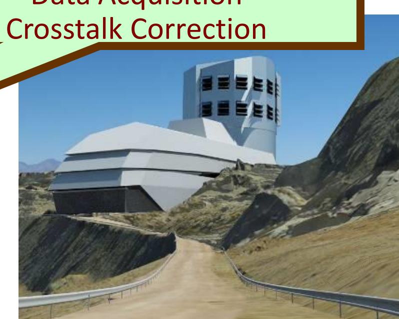
Data Access and User Services

50 - 60 TFLOPS
9 - 110 PB Disk
5 - 100PB Tape
300 kW Power
900 sq ft

Summit Site

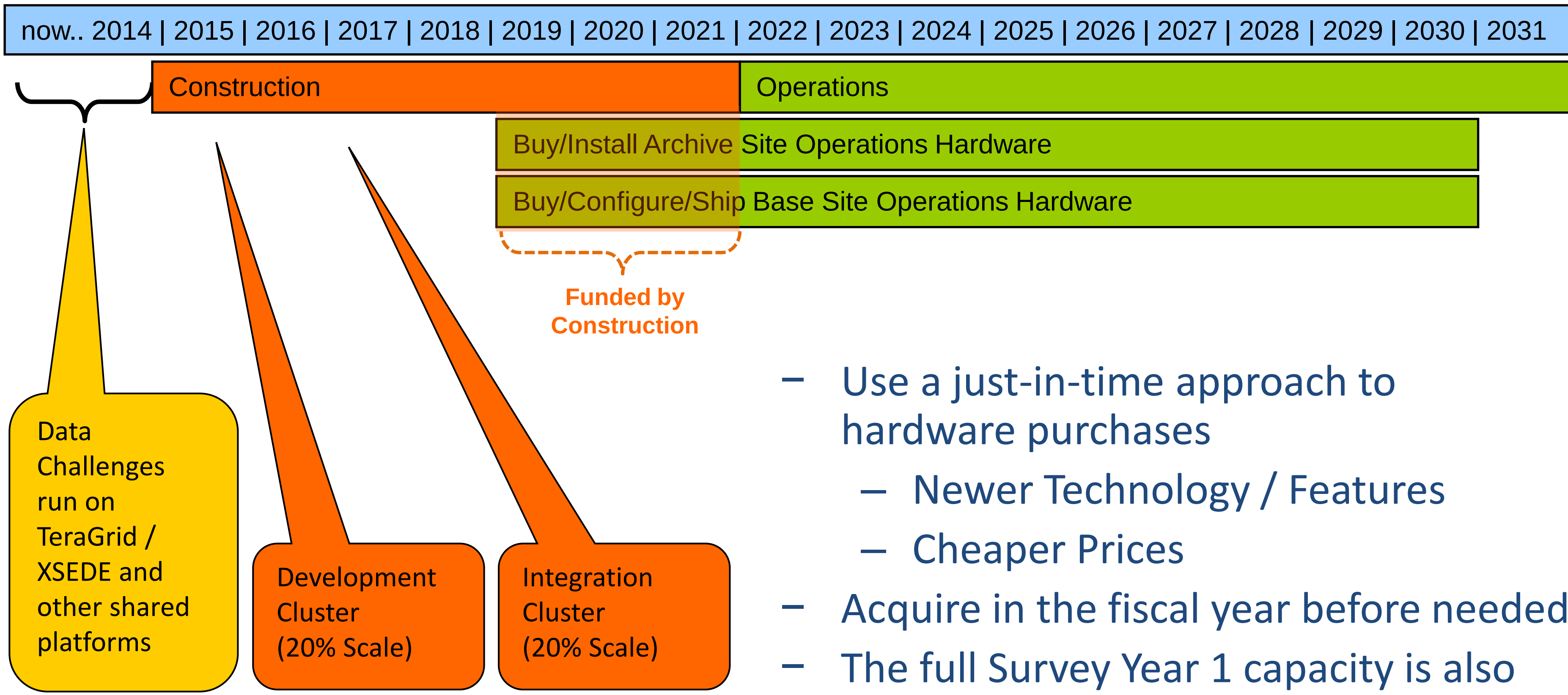
Summit Facility

Telescope and Camera
Data Acquisition
Crosstalk Correction



Credit: Jeff Kantor, LSST Corp

Infrastructure Acquisition Timeline



- Use a just-in-time approach to hardware purchases
  - Newer Technology / Features
  - Cheaper Prices
- Acquire in the fiscal year before needed
- The full Survey Year 1 capacity is also required for the two years of Commissioning

Credit: Mike Freemon, NCSA

LSST Data Sets

- Images
  - Raw
  - Template
  - Difference
  - Calibrated science exposures
  - Templates

- Catalogs
  - Object
  - MovingObject
  - DiaSource
  - Source
  - ForcedSource
  - Metadata

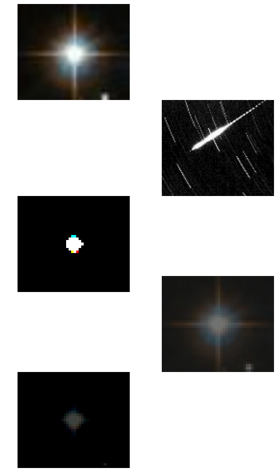


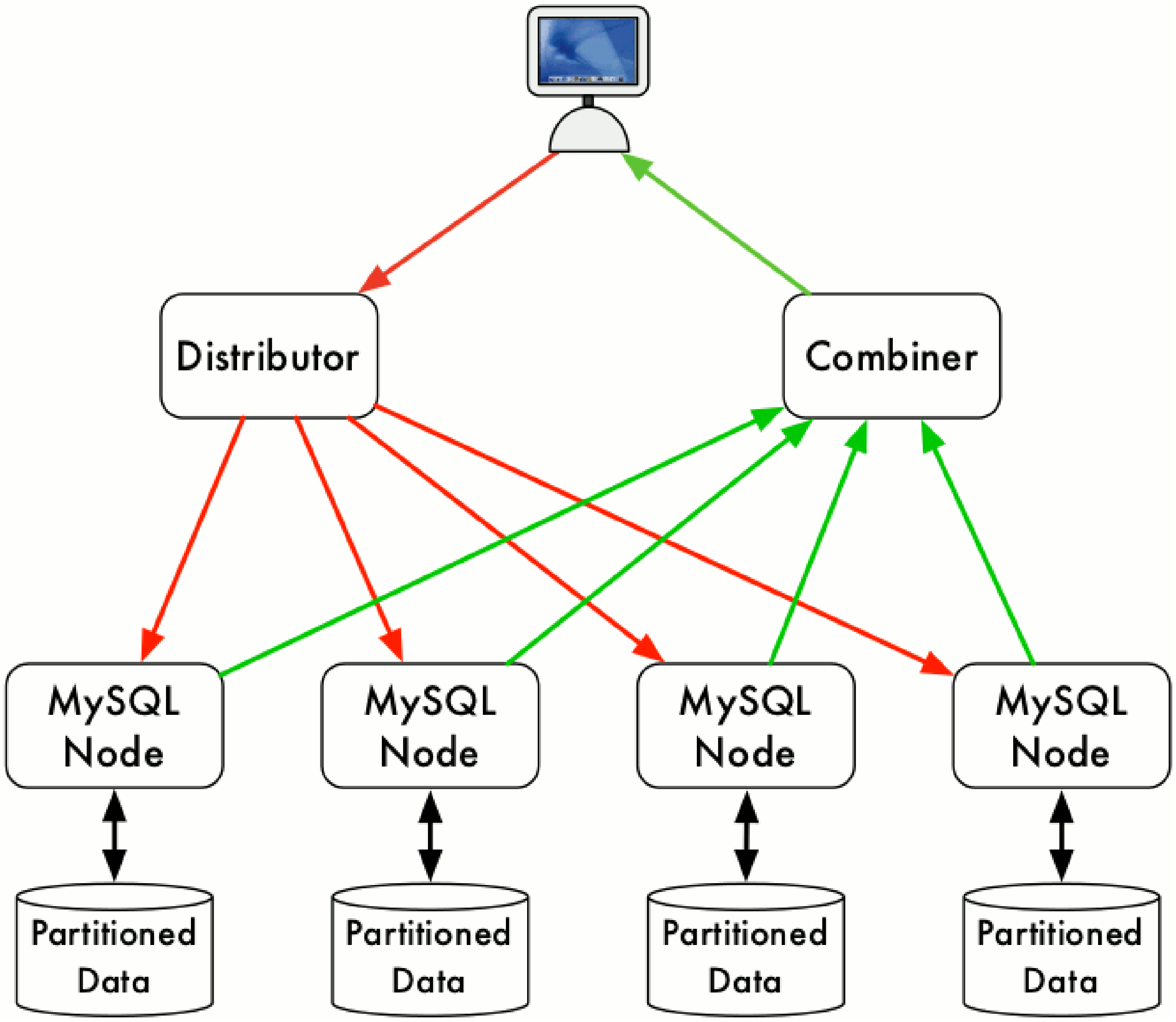
Table name	# columns	# rows
Object	500	$4x10^{10}$
Source	100	$5x10^{12}$
ForcedSource	10	$3x10^{13}$

Baseline Architecture

- MPP* RDBMS on shared-nothing commodity cluster, with incremental scaling, non-disruptive failure recovery
- Data clustered spatially and by time, partitioned w/overlaps
  - Two-level partitioning
  - 2nd level materialized on-the-fly
  - Transparent to end-users
- Selective indices to speed up interactive queries, spatial searches, joins including time series analysis
- Shared scans
- Custom software based on open source RDBMS (MySQL) + xrootd

**MPP – Massively Parallel Processing*

Baseline Architecture



Driving Requirements

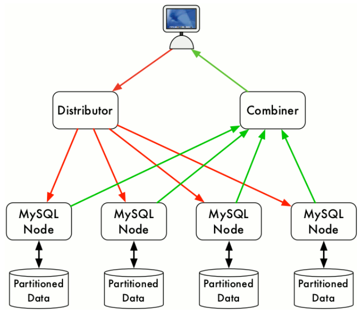
- Data volume (massively parallel, distributed system)
  - Correlations on multi-billion-row tables
  - Scans through petabytes
  - Multi-billion to multi-trillion table joins
- Access patterns
  - Interactive queries (indices)
  - Concurrent scans/aggregations/joins (shared scans)
- Query complexity
  - Spatial correlations (2-level partitioning w/overlap, indices)
  - Time series (efficient joins)
  - Unpredictable, ad-hoc analysis (shared scans)
- Multi-decade data lifetime (robust schema and catalog)
- Low-cost (commodity hardware, ideally open source)

"Standard" Scientific Questions

- ~65 "standard" questions to represent likely data access patterns and to "stress" the database
  - Based on inputs from SDSS, LSST Science Council, Science Collaboration
- Sizing and building for ~50 interactive and ~20 complex simultaneous queries
  - Interactive @<10sec
  - Object-based @<1h
  - Source-based @<24h
  - ForcedSource-based @<1 week

- In a region
  - Cone-magnitude-color search
  - For a specified patch of sky, give me the source count density of unresolved sources (star like PSF)
- Across entire sky
  - Select all variable objects of a specific type
  - Return info about extremely red objects
- Analysis of objects close to other objects
  - Find all galaxies without saturated pixels within certain distance of a given point
  - Find and store near-neighbor objects in a given region
- Analysis that require special grouping
  - Find all galaxies in dense regions
- Time series analysis
  - Find all objects that are varying with the same pattern as a given object, possibly at different times
  - Find stars that with light curves like a simulated one
- Cross match with external catalogs
  - Joining LSST main catalogs with other catalogs (cross match and anti-cross match)

Technology Choice

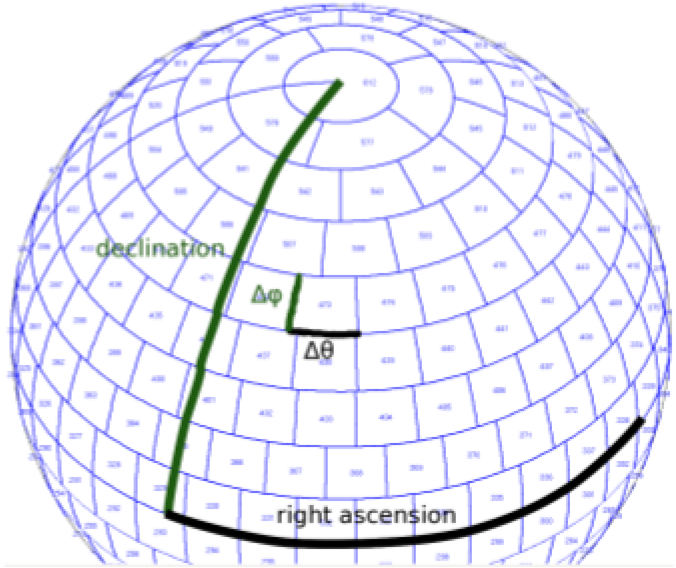


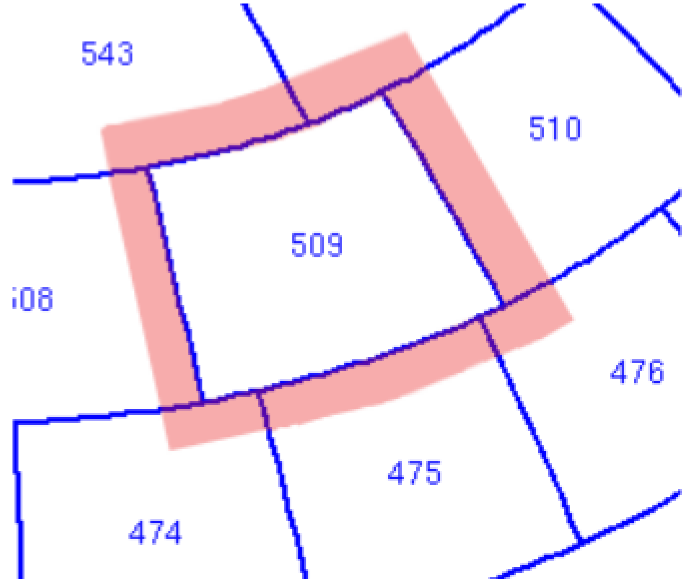
- Map/Reduce
  - 👍 Incremental scaling, fault tolerance, free
  - 👎 Indices, joins, schema and catalog, speed
    - Catching up (Hive, HBase, HadoopDB, Dremel, Tenzing)
- LSST Baseline
  - Xrootd: scalable, fault tolerance, in production, free
  - MySQL: fast, in prod., good support, big community, free
  - Custom software, including unique features: overlap partitioning, shared scans

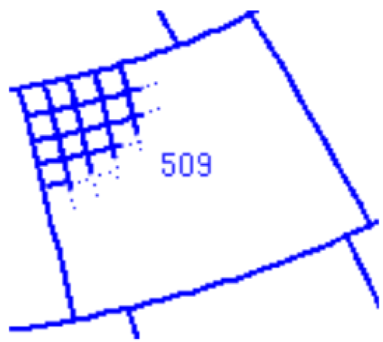
Many technologies considered

Many tests run to fine-tune and uncover bottlenecks

Partitioning



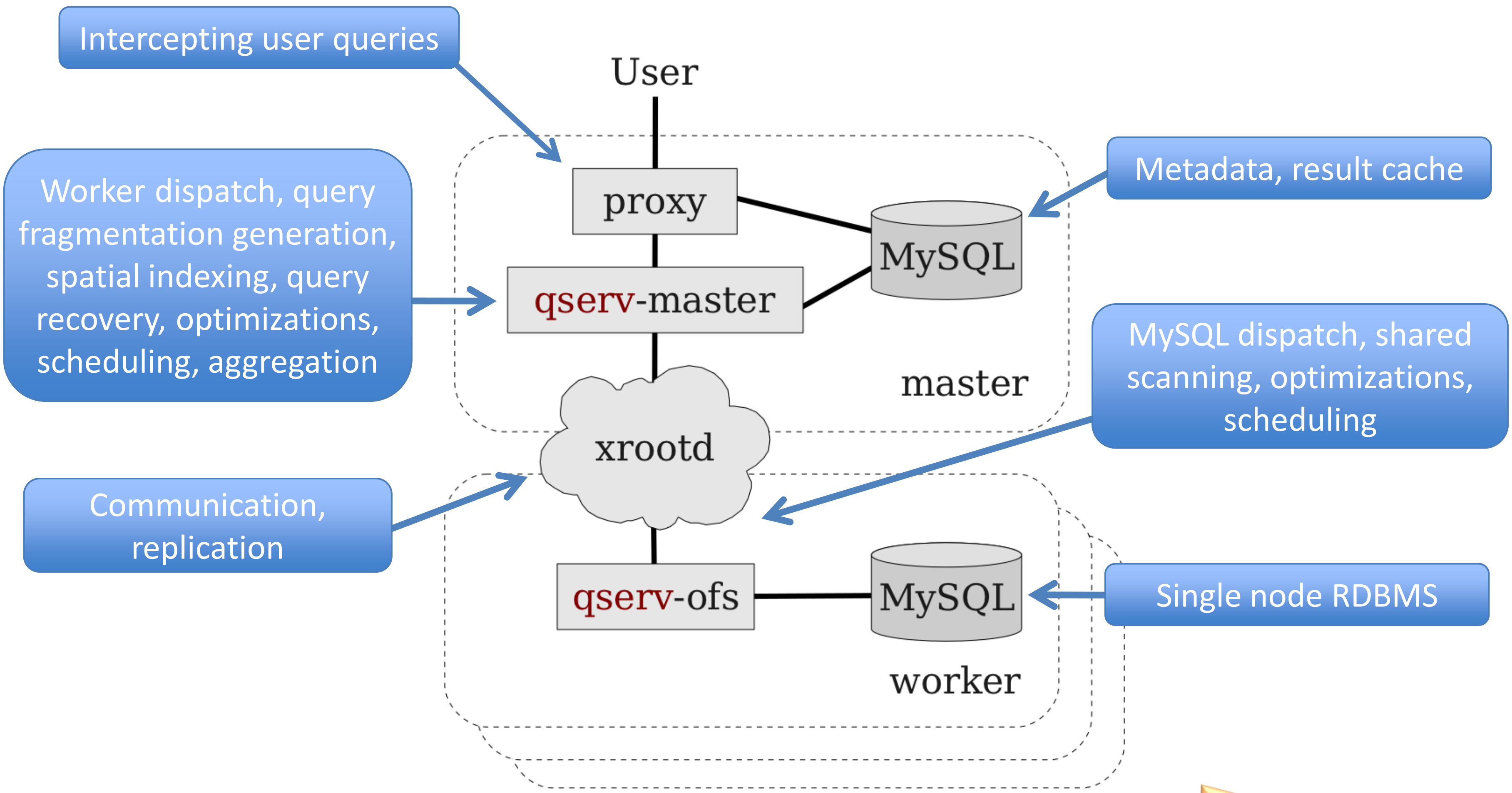




Indexing

- Nodes addressed by chunk #
- Spatial index
  - Ra-dec spec of box, ellipse, polygon, circle, …
  - Htm index
- ObjectId index
  - map objectId to chunk #
- Library of UDFs (see scisql on launchpad).
  Spatial geometry and more

Prototype Implementation - *Qserv*



RDBMS-agnostic

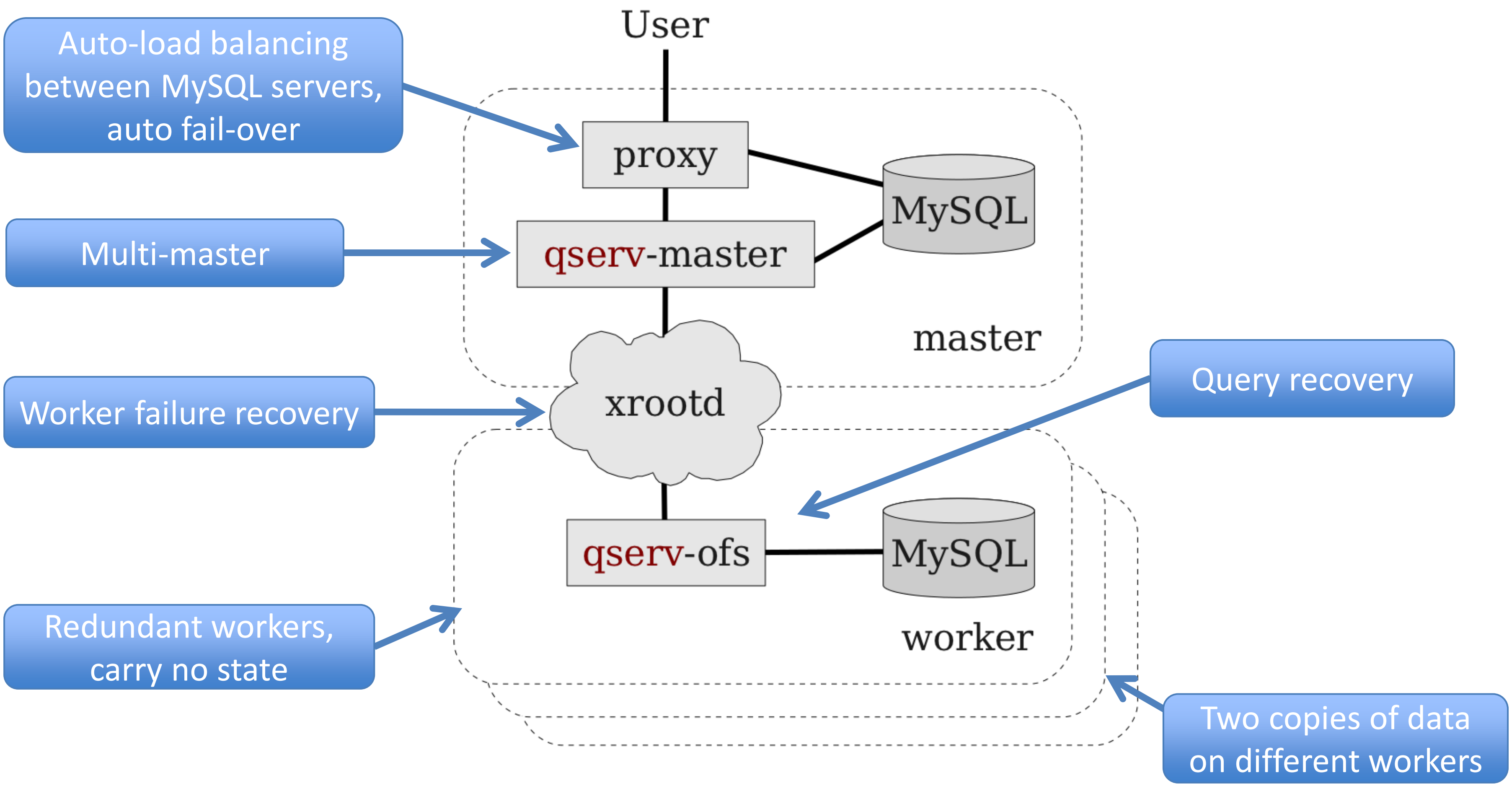
Dispatch

- Write
  - xroot://qsm@mgr:1094//query2/7505

- Result read
  - xroot://qsm@182.23.36.70:1094//result/<hash>

Qserv Fault Tolerance

- Components replicated
- Failures isolated
- Narrow interfaces
- Logic for handling errors
- Logic for recovering from errors



Status

- Working: parsing, query dispatch, 2-level partitioning, data partitioner, data loader, metadata, automated installation, …
- Tested in various configurations, including:
  - 150 nodes / 30 TB / 2B objects / 55 b sources
  - 20 nodes / 100 TB
- Working on: shared scans, query retry on failure, 300 node test, rewriting xrootd client
- Next: user table support, authentication, many others…

- Usable, stable, ~beta
- Far from production!

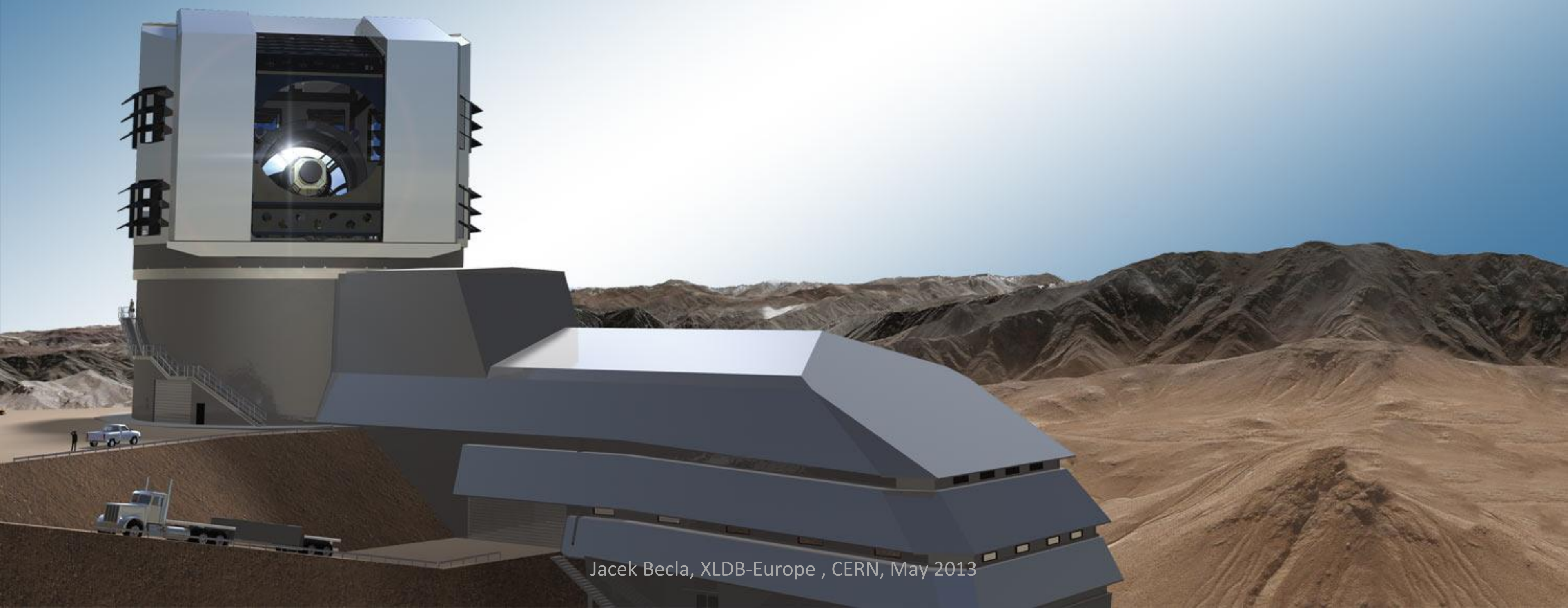
LSST & SciDB

- SciDB inspired by the needs of LSST
- LSST baseline: still qserv, not SciDB

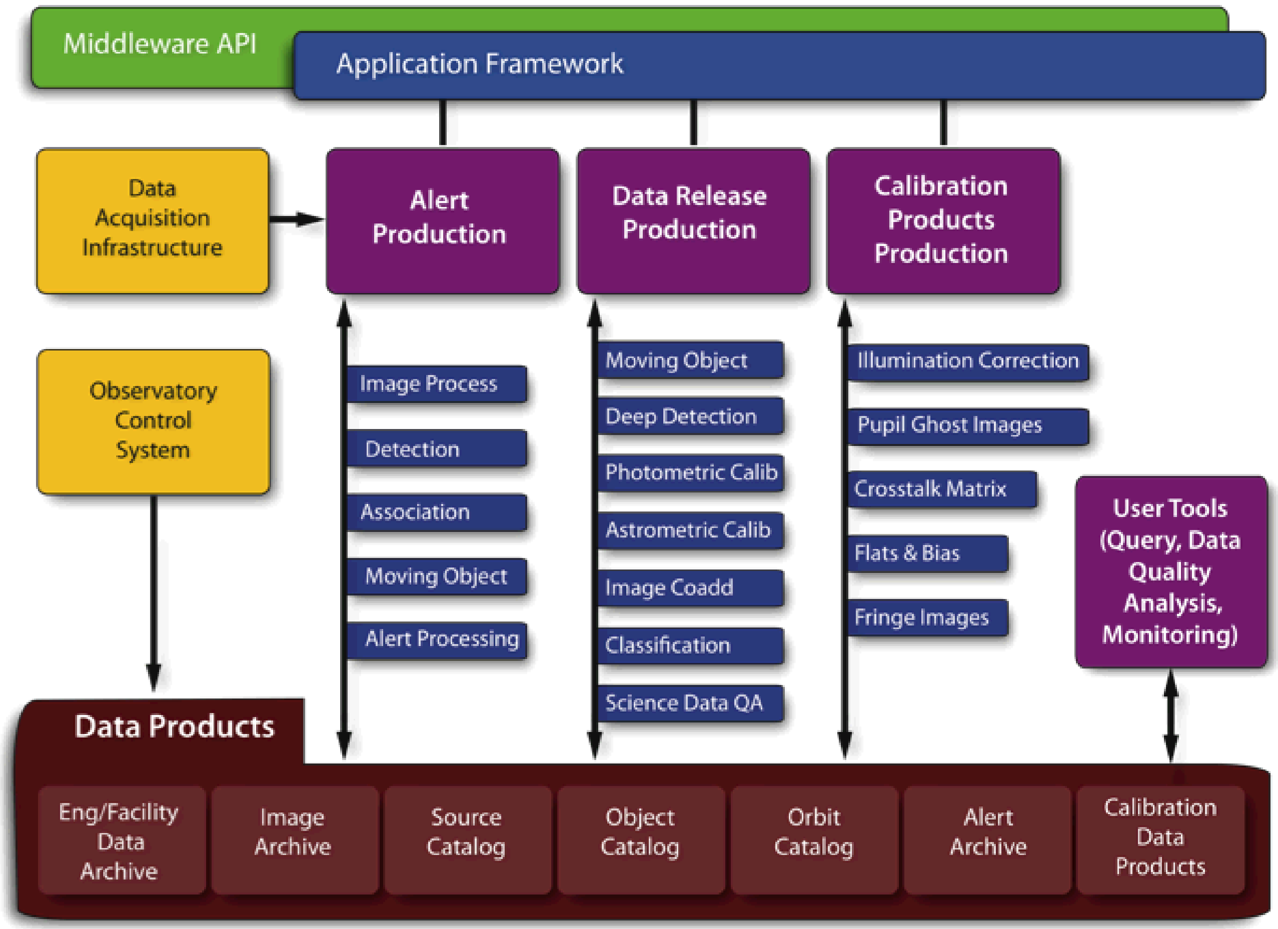
Why?

- Timescales
  - NSF/DOE reviews vs SciDB development cycle
- Domain specific legacy
  - Custom libraries perfected for decades
- Control
  - Priorities decided by stakeholder (funding issue)
  - Project longevity

Backup Slides



LSST Data Processing



Credit: Jeff Kantor, LSST Corp

How Big is the DM Archive?

Final Image Archive	345 PB	All Data Releases* Includes Virtual Data (315 PB)
Final Image Collection	75 PB	Data Release 11 (Year 10) * Includes Virtual Data (57 PB)
Final Catalog Archive	46 PB	All Data Releases*
Final Database	9 PB 32 trillion rows	Data Release 11 (Year 10) * Includes Data, Indexes, and DB Swap
Final Disk Storage	228 PB 3700 drives	Archive Site Only
Final Tape Storage	83 PB 3800 tapes	Single Copy Only
Number of Nodes	1800	Archive Site Compute and Database Nodes
Number of Alerts Generated	6 billion	Life of survey

Credit: Mike Freemon, NCSA

* *Compressed where applicable*

How much *storage* will we need?

		Archive Site	*Base Site*
Disk Storage for Images	Capacity	19 → 100 PB	12 → 23 PB
	Drives	1500 → 1100	950 → 275
	Disk Bandwidth	120 → 425 GB/s	27 → 31 GB/s
Disk Storage for Databases	Storage Capacity	10 → 128 PB	7 → 95 PB
	Disk Drives	1400 → 2600	1000 → 2000
	Disk Bandwidth (sequential)	125 → 625 GB/s	95 → 425 GB/s
Tape Storage	Capacity	8 → 83 PB	8 → 83 PB
	Tapes	1000 → 3800 (near line) 1000 → 3800 (offsite)	1000 → 3800 (near line) no offsite
	Tape Bandwidth	6 → 24 GB/s	6 → 24 GB/s
L3 Community Disk Storage	Capacity	0.7 → 0.7 PB	0.7 → 0.7 PB

Compute Nodes	1700 → 1400 nodes	300 → 60 nodes
Database Nodes	100 → 190 nodes	80 → 130 nodes

Before the right arrow is the Operations Year 1 estimate; After the arrow is the Year 10 estimate. All numbers are "on the floor"

Credit: Mike Freemon, NCSA