What is Challenging About Scientific Data Sets?

Jacek Becla
SLAC National Accelerator Laboratory

NATIONAL ACCELERATOR LABORATORY

Google, Aug 21, 2012

BABAR

LSST
Large Synoptic Survey Telescope

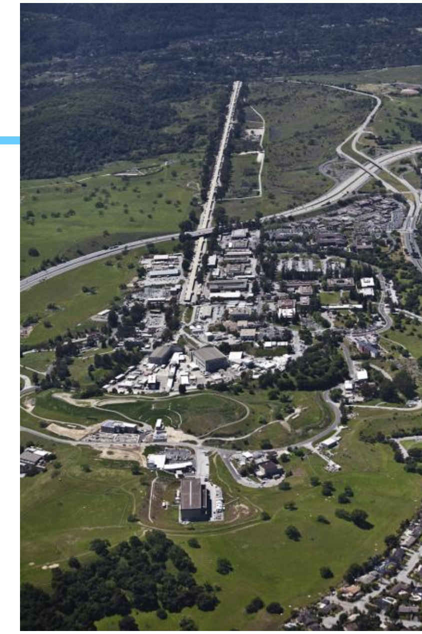
XLDB

SciDB

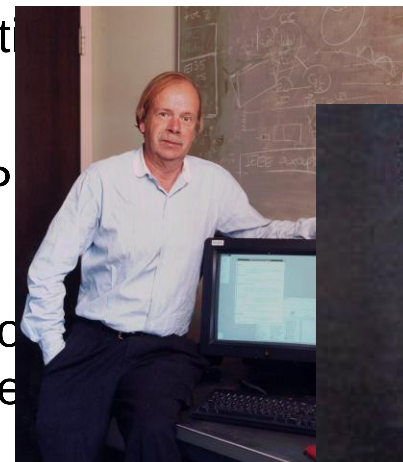
Outline

- Science & petascale
- Everything-**scientific:** data complexity, analysis, data models, architectures, HW, SW, culture
- LSST scalable database
- XLDB, SciDB
- How can **you** help?

About SLAC…



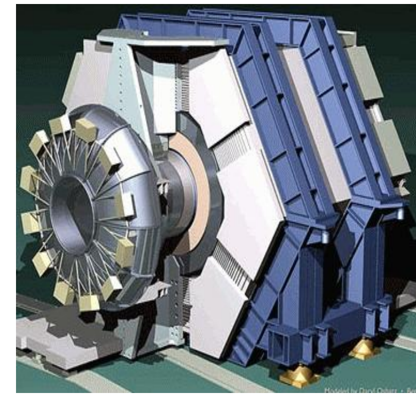
- One of 17 National Laboratories funded by the US DOE and operated by Stanford University for 50 Years
- Science-concentric mission: *no* classified research or weapons work, all research is published
- Nearly 500 acres of land and 3 MILES of tunnels
- ~1,500 staff and an equal number of visit
and researchers
- Research at SLAC has lead to 6 Nobel P
(in both chemistry and physics)
- Discoveries include the Quark, Tau Lepto
and the first direct evidence of dark matte



- First Internet Web Connection in North America (between Tim Berners-Lee at CERN and Paul Kunz at SLAC)
- First Internet Application in the World (SPIRES)

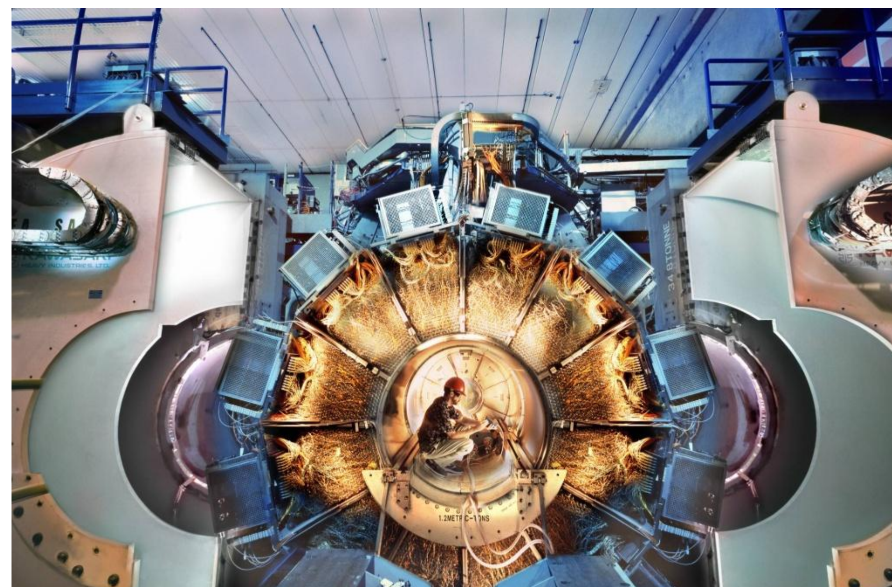
SLAC NATIONAL ACCELERATOR LABORATORY

Science & Petascale



High Energy Physics: BaBar

- 1999 – 2008
- Few TB/sec
  - Small fraction saved
- Billions of collisions
- 4 PB data set
- Petabyte database




CNN.com / SCI-TECH

SEARCH GO

MAIN PAGE
WORLD
U.S.
WEATHER
BUSINESS

Stanford researchers may have largest database

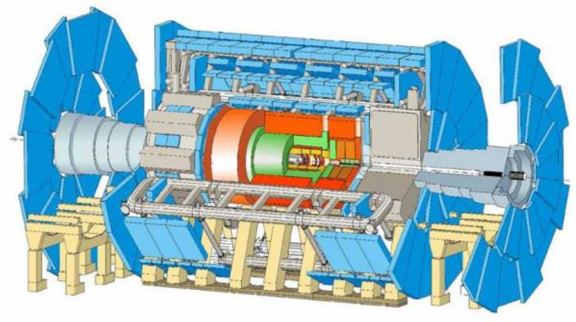
April 18, 2002 Posted: 8:15 a.m. EDT (1215 GMT)

SLAC
NATIONAL ACCELERATOR LABORATORY

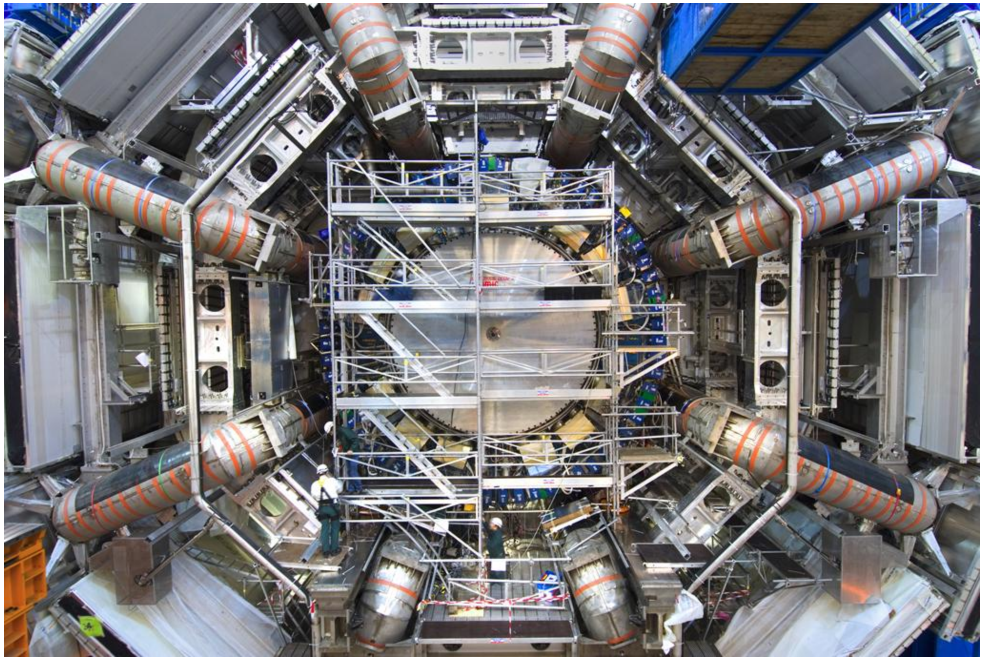
Jacek Becla

Science & Petascale

High Energy Physics: LHC



- ½ PB/sec
  - Small fraction saved
- Trillions of collisions
- 15 PB/year



Science & Petascale

NASA: Earth Observing System

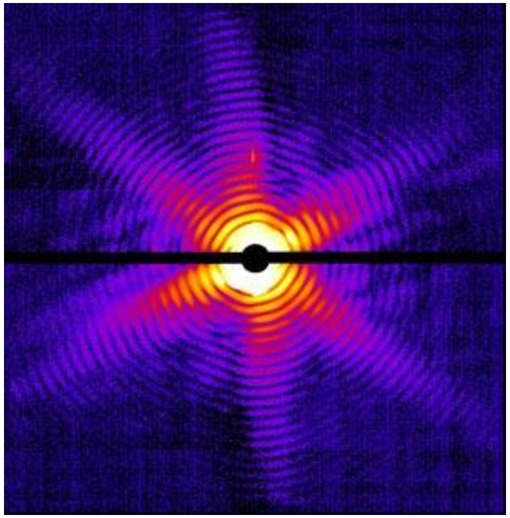
- 4 PB in 2005 (images)

Science & Petascale

Photon Science

- Huge lasers
- <100 femtosec speed
- Few MB x 120Hz
- Few PB/year

- Movies of atoms & molecules
- Portraits of viruses



X-ray diffraction pattern of a single Mimivirus particle Imaged at the LCLS. In this study, the X-ray pulse lasted a millionth of a billionth of a second and heated the virus to 100,000 degrees Celsius, but not before this image was obtained. (Image courtesy Tomas Ekeberg, Uppsala University.)

Science & Petascale

Genomics

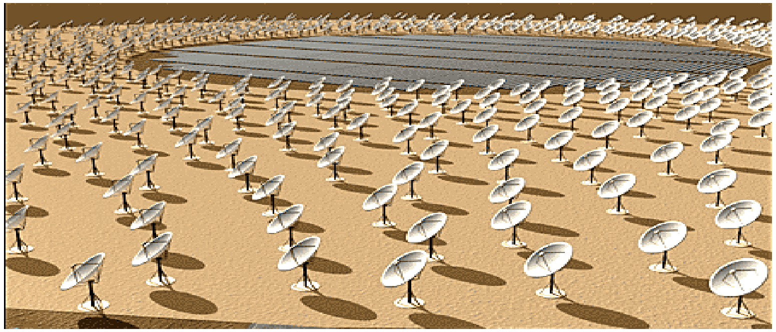
- Trying to put together database of all known DNA sequences
- Multi-petabytes

Science & Petascale

Astronomy

- Huge telescopes
- Multi-gigapixel cameras
- Thousands of dishes

- Understanding dark matter & dark energy, detecting asteroids, mapping Milky Way, …



SLAC NATIONAL ACCELERATOR LABORATORY

Petascale

- HEP – since ~2002, 15PB/year now
- Astro – PBs now, 100s PB soon, exascale planned
- Geo – now, but highly fragmented
- Bio – growth much faster than Moore's Law

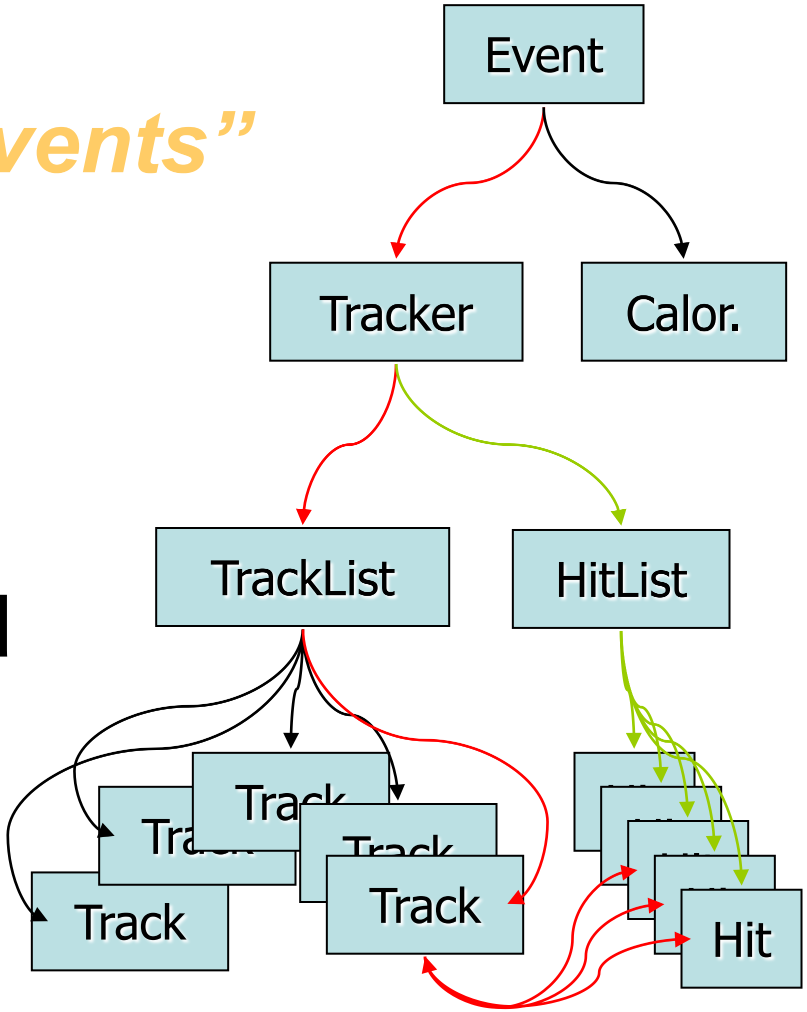
- Lots of data never saved
  - Discarded
  - Virtualized

Hunt for Higgs Boson

HEP: It's All About "Events"

- Complex hierarchical tree-like structures with many relations
- Events are uncorrelated

✓ **Needle in haystack**
✓ Spatial correlations
✓ Time series within event



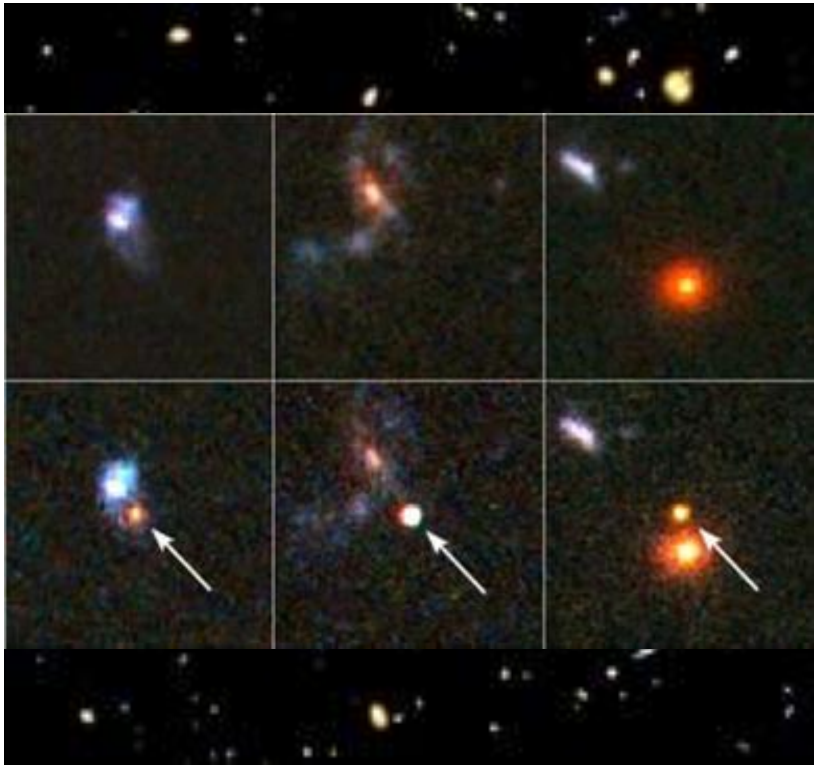
Credit: Dirk Düllmann/CERN

Untangling the Universe

Astronomy: It's All About "Astronomical Objects"

- ✓ Needle in haystack
- ✓ Spatial correlations
- ✓ Time series

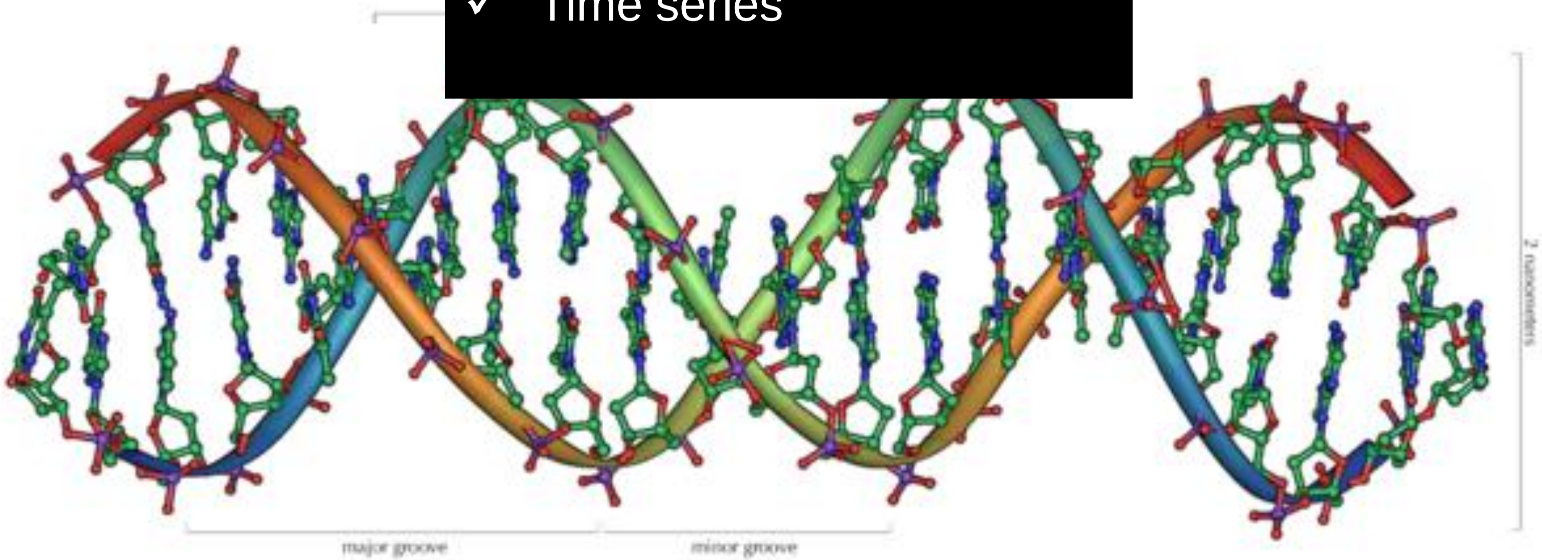
- Overlapping
- Moving
- Disappearing
- Highly correlated



Jacek Becla

Understanding Dynamics of Biological Processes

- ✓ Needle in haystack
- ✓ Correlations
- ✓ Time series



Data Complexity

- Proximity
- Adjacency
- Order
- Most data uncertain
- Multiple sources, integration and unification
  – Transform, regrid, align, calibrate
- Often distributed
- Often write-once-read-many! :-)

Queries / Analysis

Operational load

- *Still challenging @petascale*

Varying response time needs

- *Long-running – need stable environment*
- *Real-time – need speed, indexes*

Discovery-oriented

- *Complex workflows*
- *Increasingly complex*
- *Ad-hoc, unpredictable, hand-coded, sub-optimal*
- *Not just I/O, but often CPU-intensive, 100s attributes*
- *Annotate, share*
- *Repeatedly try/refine/verify*
- *More data = new ways of analyzing it!*
- *Statistical significance*
- *Avoidance of bias*

Example use cases

- *Pattern discovery*
- *Outlier detection*
- *Multi-point correlations in time and space*
  - *Time series, near neighbor*
- *Curve fitting*
- *Classification*
- *Multi-d aggregation*
- *Cross matching and anti-cross matching*
- *Dashboards / QA*

Data Models / Formats

- Relational tables rarely fit
  - Exceptions: metadata, catalogs, calibration data
- Lots of pixel data, order important
  - Fit into arrays
  - Array friendly formats (HDF5, netCDF, FITS…)
- Graphs, meshes – ocean, bio, chemistry
- Strings – bio (sequences)
- Some unstructured data
- Lots of floats (compress badly)

Architectures

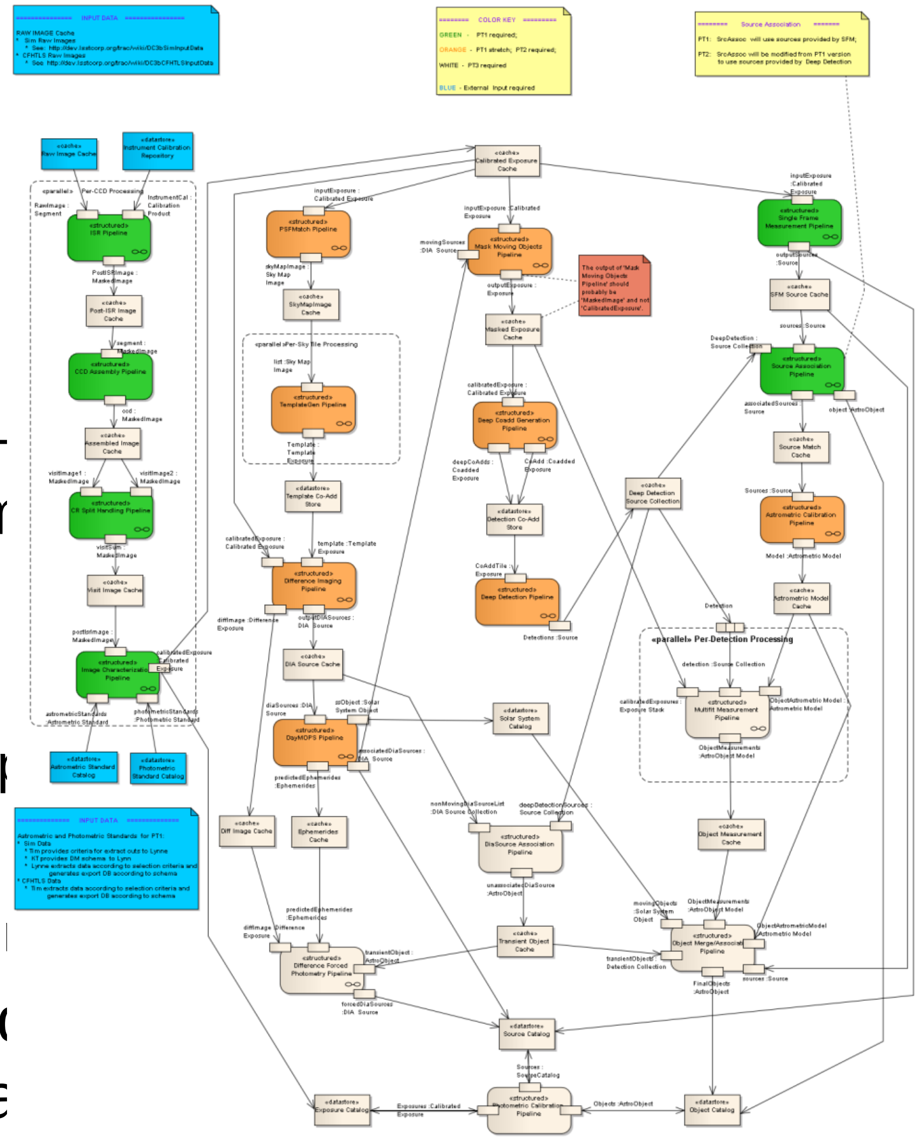
- Hierarchical data centers (tiers)
  - HEP, soon in astro
  - Geo: attempted, failed
  - Others not there yet
- Independent sites, often very different
  - Geo, bio
- Produce data, take home and analyse locally
  - Photon science
- Centralizing analysis / analytics as service
  - Requires paradigm shift
  - Must overcome desire to owning, controlling data
  - Many final (deep / specialized) analysis still on local machines

Hardware Environment

- Typically heterogeneous
- Commodity
- Moving towards shared-nothing
- Parallelization and sharing resources essential

Software

- Open source if possible
- Complex workflows
- DBMS vs files
  - Very few use real database
  - SDSS, NIF, PanSTARRS, L
- Hybrid: structured files + m
  - All HEP, NASA, bio, …
- DBMS?
  - Doesn't scale, wrong APIs, p
- M/R?
  - Complicated joins and proxi
- Tightly integrated tools and
  hamper agility / ra



Cultural Differences

Industry

Time is money
- Real time
- High availability
- Hot fail-over

Rapid change
- Agile software

"Neutrinos faster than the speed of light? Not so fast…"

Science

Severely underfunded

Multi-lab experiments
- No firm control over configuration
- Computing near funding

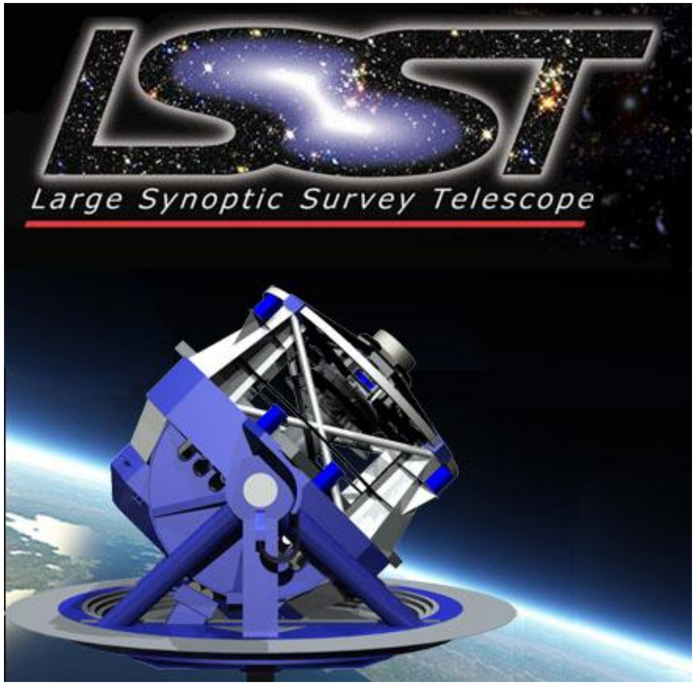
Long-term projects
- Extra isolation
- Mid-project migrations
- Unknown requirements
- Unknown hardware

Statistical errors & bias have huge impact

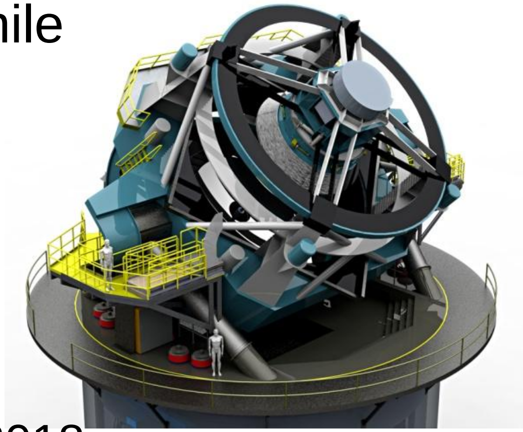

LSST

Large Synoptic Survey Telescope

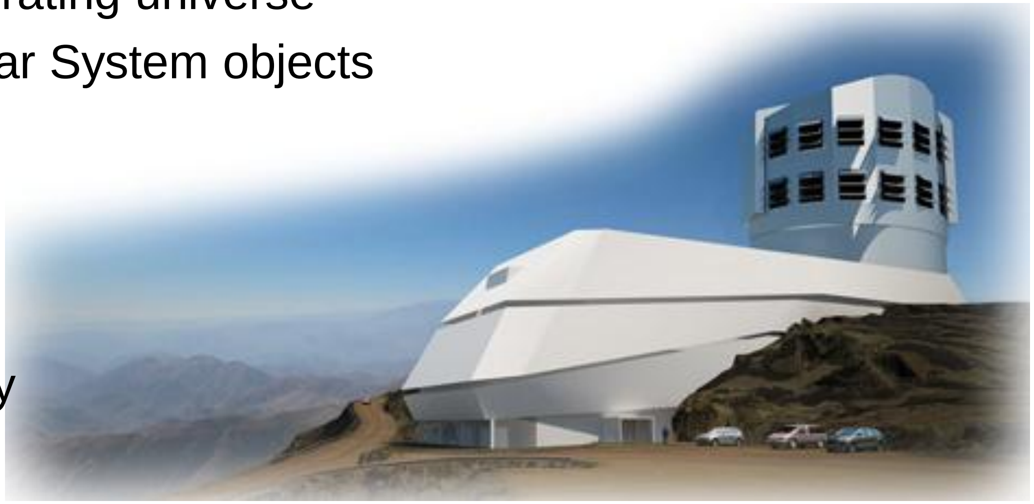
- Timeline
  - In R&D now, data challenges
  - Operations: 2022-2031
- Scale
  - *O*(100) PB
  - Plus virtual data
- Complexity
  - Time series (order)
  - Spatial correlations (adjacency)



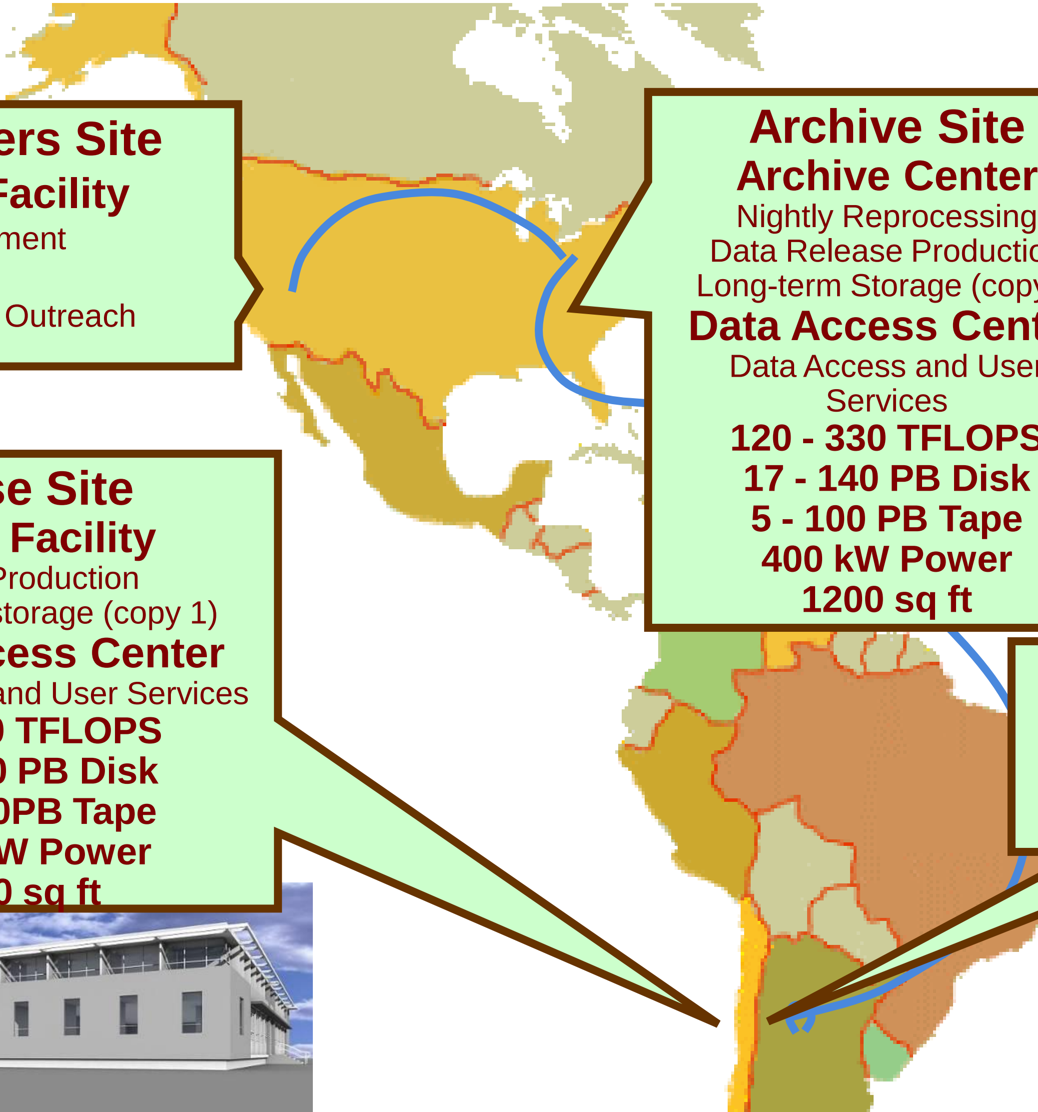
The LSST scientific instrument



- 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
  - 10-year survey begins in 2022
  - Sensitivity – per “visit”: 24.5 mag; survey: 27.5 mag
  - First computing hardware systems to be purchased in 2018
- 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

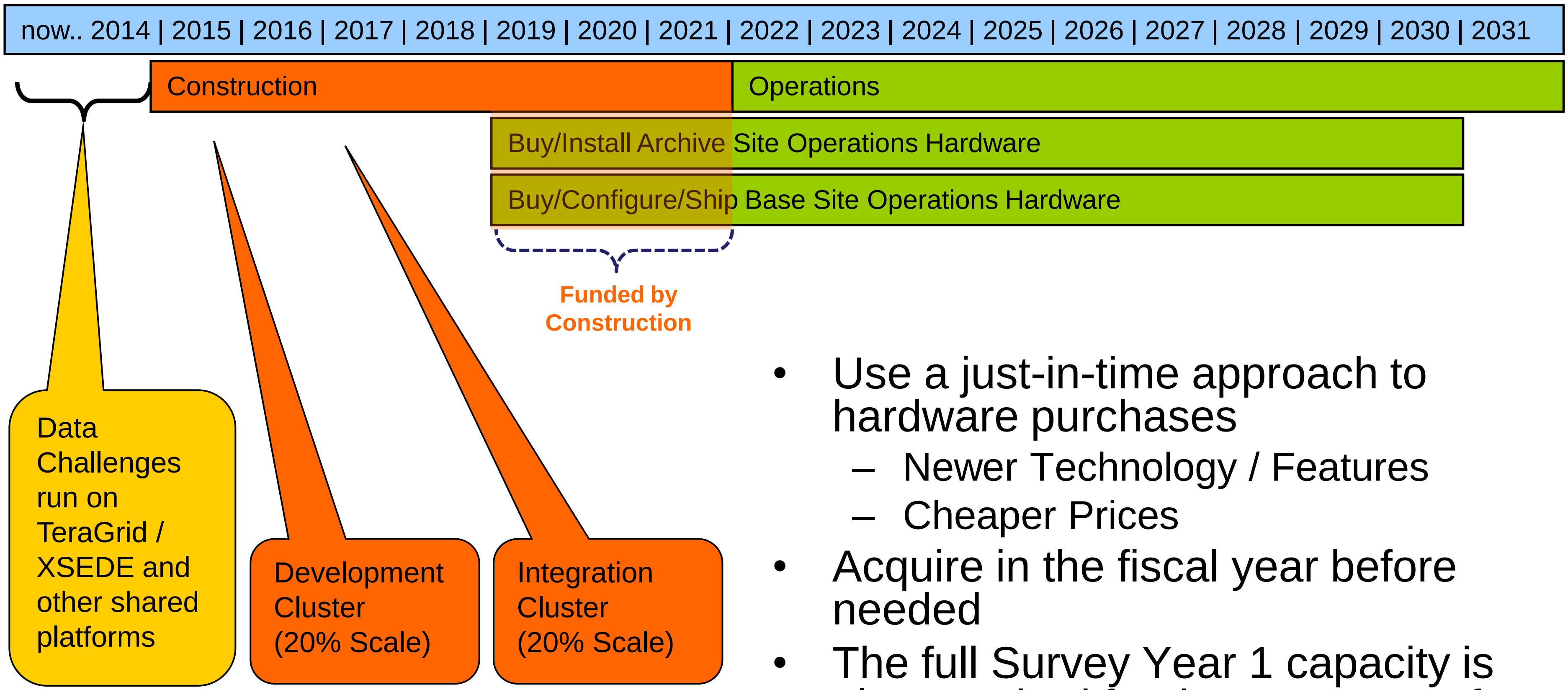
Archive Site
Archive Center
Nightly Reprocessing
Data Release Production
Long-term Storage (copy 2)
Data Access Center
Data Access and User Services
120 - 330 TFLOPS
17 - 140 PB Disk
5 - 100 PB Tape
400 kW Power
1200 sq ft

Base Site
Base Facility
Alert Production
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

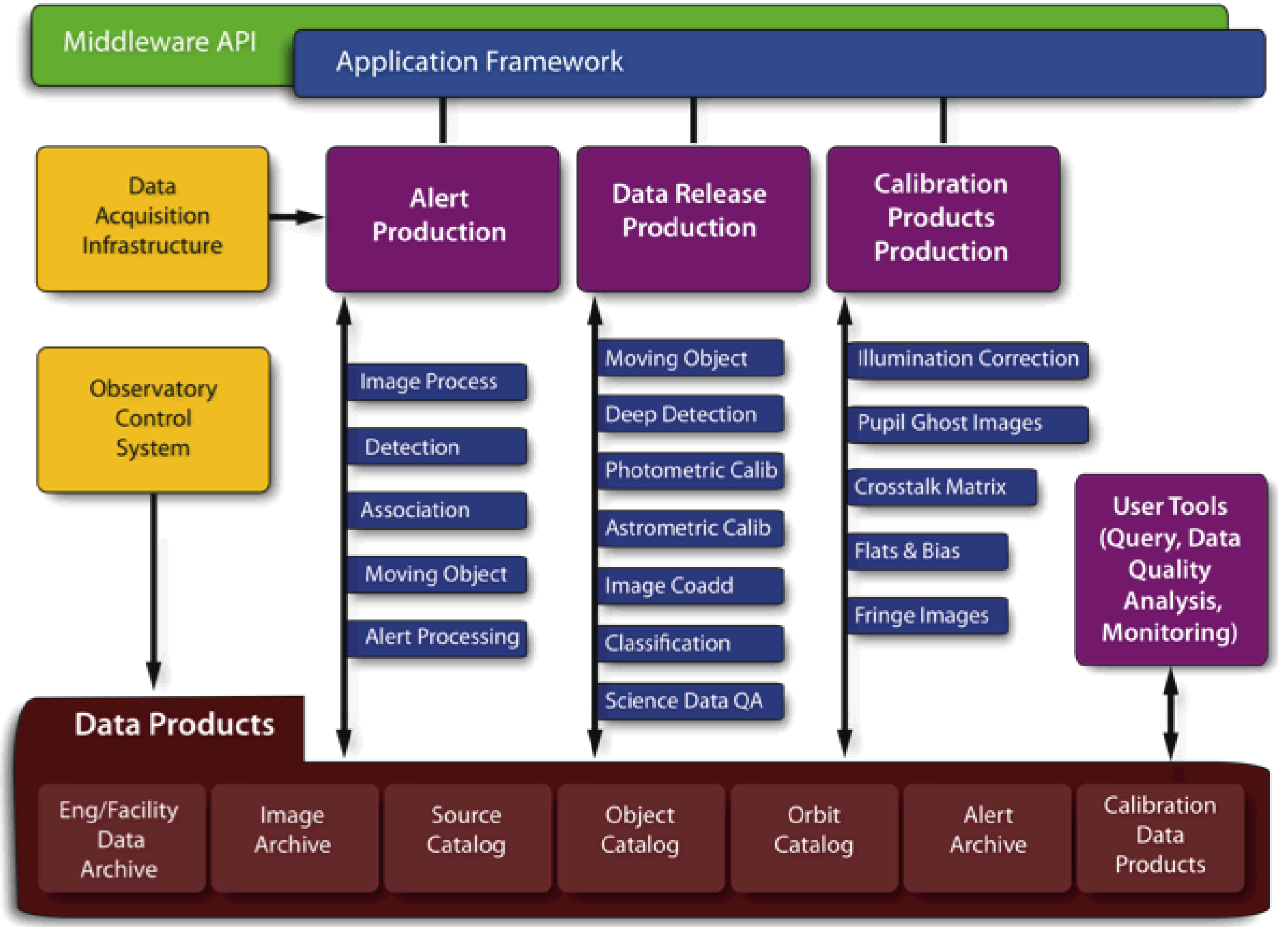
Infrastructure Acquisition Timeline



- 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 Processing



Credit: Jeff Kantor, LSST Corp

LSST Data Sets

- Images
  - Raw
  - Template
  - Difference
  - Calibrated science exposures
  - Templates
- Catalogs
  - Object
  - MovingObject
  - DiaSource
  - Source
  - ForcedSource
  - Metadata



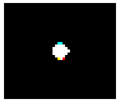
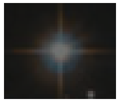
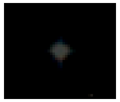




Table name	# columns	# rows
Object	500	$4x10^{10}$
Source	100	$5x10^{12}$
ForcedSource	10	$3x10^{13}$

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

Database - 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 Collaborations
- 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)

Making RDBMS Work For Us

- Offline data loading
- Real time:
  - File-based copy, partitioned, relevant columns
  - Cross match in c++
  - Localizing and minimizing updates, making non-critical
- Outside-database processing
  - partitioning, time series, 2 & 3-point auto-correlations
- Lots of “custom” features:
  - Partitioning… indexes… UDFs… synchronized scans, optimizations

Random Hard / Awkward Issues

- Cross match with external catalogs
- Object ids between data releases
- Flattening multi-d structures into tables
  – Example: 3x3 cov-matrix
- Key-value metadata

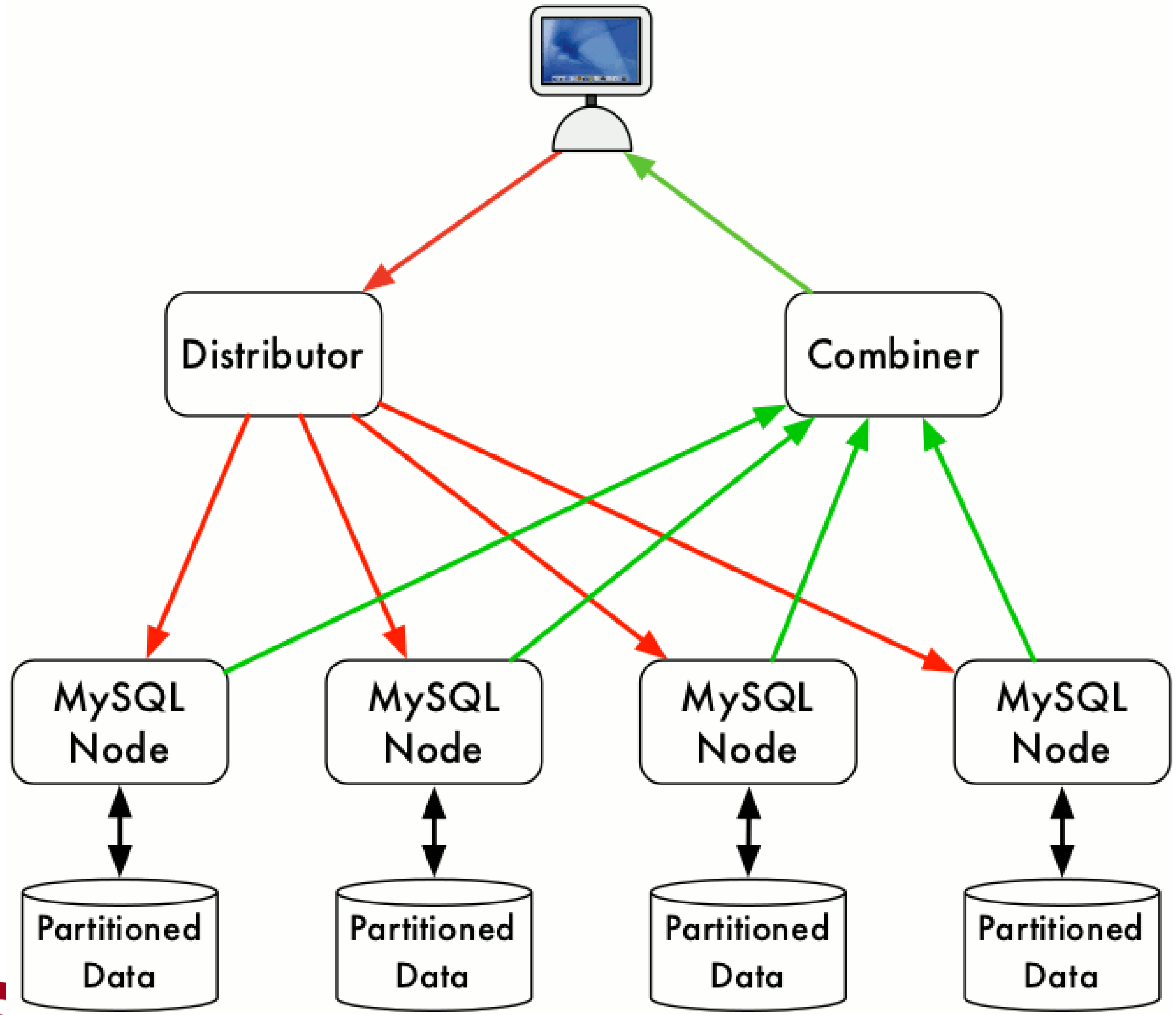
NATIONAL ACCELERATOR LABORATORY

Baseline Database Architecture

- MPP* RDBMS on shared-nothing commodity cluster, with incremental scaling, non-disruptive failure recovery
- Data clustered spatially and by time, partitioned w/overlaps
  - Data-aware two-level partitioning
  - 2^{nd} 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*

Scalable LSST DB - *Qserv*



Qserv

- Blend of RDBMS and Map/Reduce
  - Based on MySQL and xrootd
- Key features
  - Data-aware 2-level partitioning w/overlaps, 2^{nd} level materialized on the fly
  - Shared scans
  - Complexity hidden, all transparent to users
- 150-node, 30-billion row demonstration

What Is xrootd?

- A file access and data transfer *protocol*
  - Defines POSIX-style byte-level random access for
    - *Arbitrary* data organized as files of *any* type
    - Identified by a hierarchical directory-like name
- A reference *software* implementation
  - Embodied as the xrootd and cmsd daemons
    - xrootd daemon provides access to data
    - cmsd daemon clusters xrootd daemons together
- In production for 10+ years, used by many experiments
    - Antares, ALICE, ATLAS, BaBar, CMS, Compass, dchooz, EXO, Fermi, Hess, Indra, LSST, Opera, Panda, Virgo

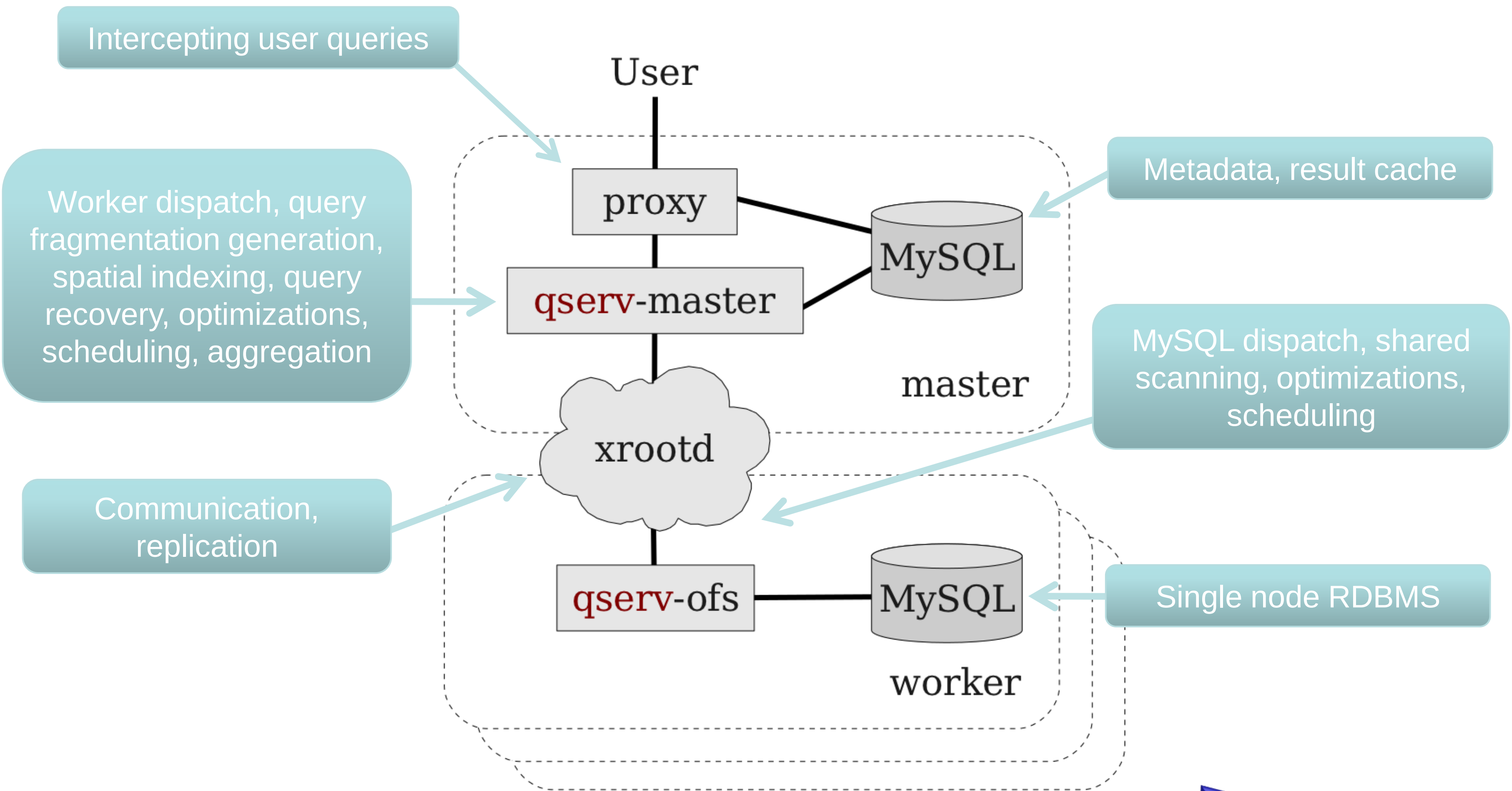
Credit: Andrew Hanushevsky, SLAC

What Makes xrootd Unusual?

- A comprehensive plug-in architecture
  - Security, storage back-ends (e.g., tape), proxies, etc
- Clusters widely disparate file systems
  - Practically any existing file system
    - Distributed (shared-everything) to JBODS (shared-nothing)
  - Unified view of disparate storage resources
    - Irrespective of physical location or makeup
- Very low support requirements
  - Hardware and human administration

Federated data sets workshop:
https://indico.in2p3.fr/conferenceDisplay.py?ovw=True&confId=6941

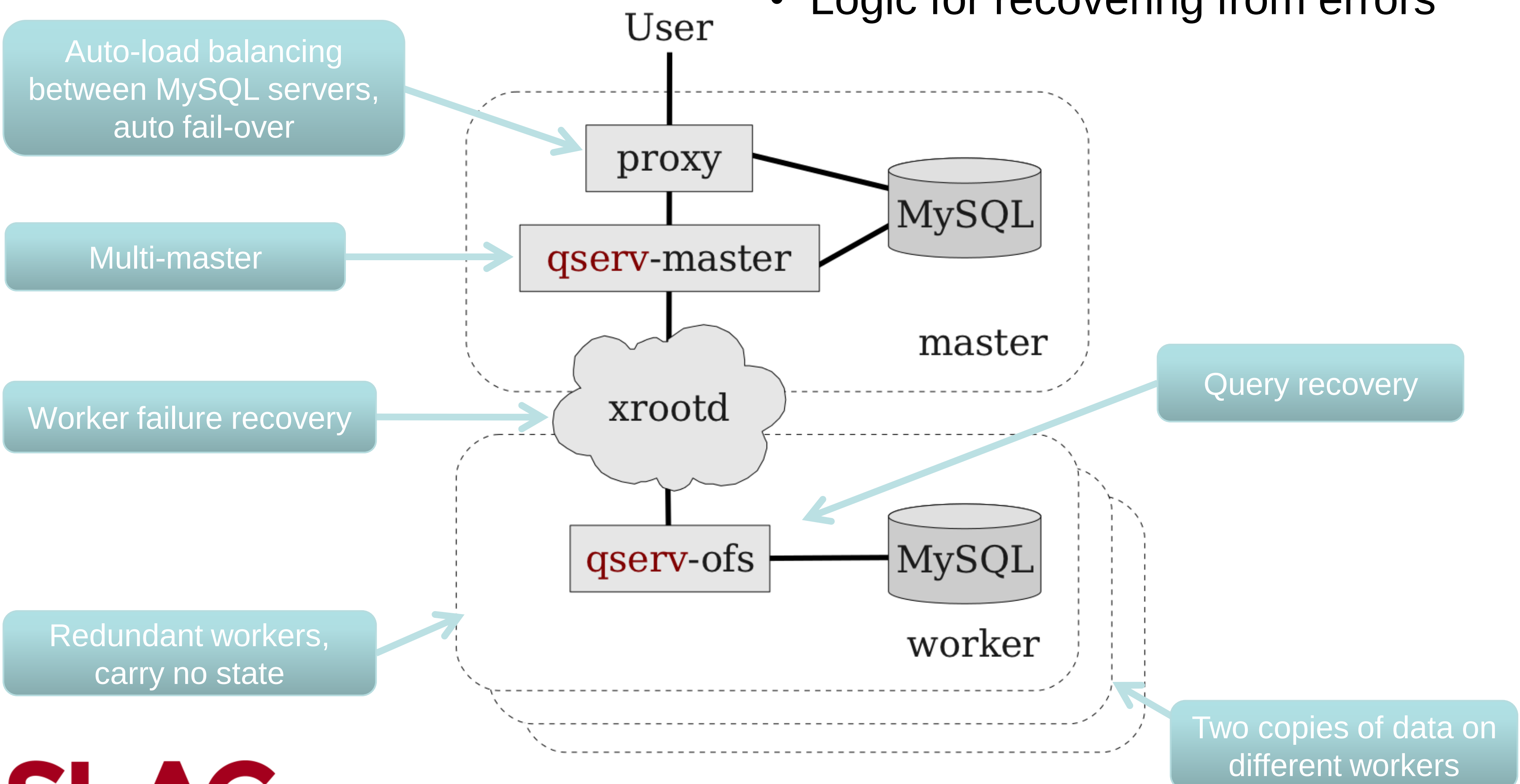
Prototype Implementation - *Qserv*



RDBMS-agnostic

Qserv Fault Tolerance

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



UDFs

sciSQL 0.1: Science Tools for MySQL

Index

- **Overview**
- **Building & Installation**
- **Spherical Geometry**
  - **UDFs**
    - scisql_angSep
    - scisql_s2CPolyHtmRanges
    - scisql_s2CPolyToBin
    - scisql_s2CircleHtmRanges
    - scisql_s2HtmId
    - scisql_s2PtInBox
    - scisql_s2PtInCPoly
    - scisql_s2PtInCircle
    - scisql_s2PtInEllipse
  - **Stored Procedures**
    - scisql_s2CPolyRegion
    - scisql_s2CircleRegion
- **Photometry**
  - scisql_abMagToDn
  - scisql_abMagToDnSigma
  - scisql_abMagToFlux
  - scisql_abMagToFluxSigma
  - scisql_dnToAbMag
  - scisql_dnToAbMagSigma

Spherical Geometry

in the ranges [350, 360) and [0, 10].

- Input values must be convertible to type DOUBLE PRECISION. If their actual types are BIGINT or DECIMAL, then the conversion can result in loss of precision and hence an inaccurate result. Loss of precision will not occur so long as the inputs are values of type DOUBLE PRECISION, FLOAT, REAL, INTEGER, SMALLINT or TINYINT.

Examples

```
SELECT objectId, ra_PS, decl_PS
    FROM Object
    WHERE scisql_s2PtInBox(ra_PS, decl_PS, -10, 10, 10, 20) = 1;
```

scisql_s2PtInCPoly

```
FUNCTION scisql_s2PtInCPoly (
        lon  DOUBLE PRECISION,        deg   Longitude angle of point to test.
        lat  DOUBLE PRECISION,        deg   Latitude angle of point to test.
        poly VARBINARY                      Binary-string representation of polygon.
) RETURNS INTEGER

FUNCTION scisql_s2PtInCPoly (
        lon   DOUBLE PRECISION,        deg   Longitude angle of point to test.
```

http://dev.lsstcorp.org/schema/sciSQL


XLDB

Extremely Large Databases

Internationally recognized conference series

- With yearly satellite events on other continents
- Started at / organized by SLAC

Philosophy

1. Identify trends, commonalities, roadblocks
2. Bridge the gap between data-intensive users and solution providers
3. Facilitate development & growth of practical technologies

☞ Focus on extreme scale & complex analytics; practical aspects

Large community

- Scientific and industrial data-intensive users
- Vendors, academic researchers

Many tangible results

- Initiated SciDB
- Collecting/publishing use cases
- Developed science benchmark
- 1000+ user community
- Blog
- Successful science-industry collaboration
- and more…

http://xldb.org

SLAC
NATIONAL ACCELERATOR LABORATORY

Jacek Becla

SciDB

- Open source, analytical DBMS
- Array data model
  - True multi-dimensional array storage
    - chunking, overlaps, non-integer dimensions
- Complex math inside database
  - window moving windows, re-grid, resampling…
- Runs on commodity H/W grid or in a cloud

So… What *is* Challenging About Scientific Data Sets?

- Scale
- Data: n-point correlations, uncertainty
- Unknown requirements
- Correctness and reproducibility
- Project and data longevity
- (Under-)funding

How Can You / How Can Google Help?

- Contribute to open source
  - Example: ProtoBuf
- Help LSST DM
  - Review
  - Advice
  - Provide access to computing resources
    - Qserv scalability tests?
    - Test of middleware software?
- Support XLDB
  - Help publicize
  - Fund

Google is one of the LSST collaborators!