

# Cosmic Peta-Scale Data Analysis at IN2P3

**Fabrice Jammes**

Scalable Data Systems Expert  
LSST Database and Data Access Software Developer

**Yvan Calas**

Senior research engineer  
LSST deputy project leader at CC-IN2P3

**Fabio Hernandez**

Senior research engineer  
LSST project leader at CC-IN2P3

**Jacek Becla**

SLAC Technology Officer for Scientific Databases  
Project Manager for LSST Data Management

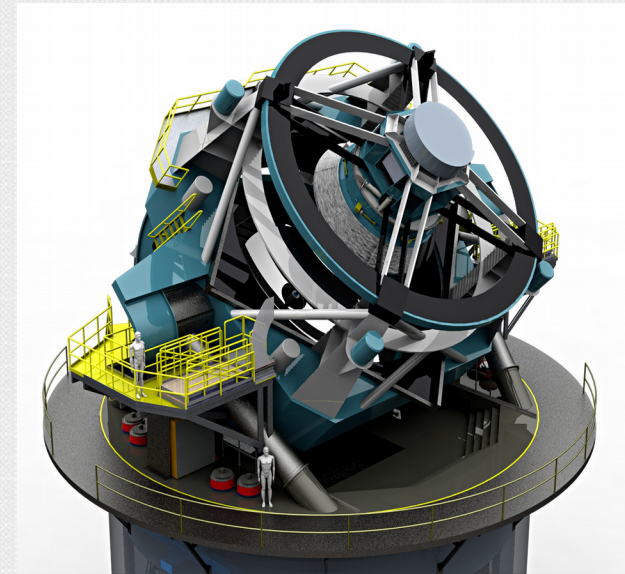




# LSST in short



- 8.4 m telescope
- Cerro Pachon (Chile)
- (Very) wide-field astronomy
- All visible sky in 6 bands  $\sim 20000 \square$
- 15 s exposure, 1 visit / 3 days
- During 10 years !
- **60 Pbytes of raw data**





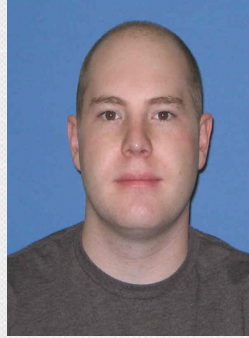
# Who We Are



Andrew  
Hanushevsky  
0.4



Andrei  
Salnikov  
0.5



Brian Van  
Klaveren  
0.4



Jacek  
Becla



John  
Gates  
1



Fabrice  
Jammes  
0.3 (+0.5)



Fritz  
Mueller  
1



Nate  
Pease  
1



Vaikunth  
Thukral  
(1)



Igor  
Gaponenko  
(1)

???  
1



# Who We Are: French Operation Team



Yvan  
Calas



Fabio  
Hernandez

And others experts:  
Fabien Wernli (Monitoring)  
Loïc Tortay (GPFS),  
Mathieu Puel (System administration)


# What We Do



- **Data Access and Database**
- Data and metadata
- Images and databases
- Persisting and querying
- For pipelines and users
- Real time Alert Prod and annual Data Release Prod
- For Archive Center and all Data Access Centers
- For USA, France and international partners
- Persisted and virtual data
- Estimating, designing, prototyping, building, and productizing



# Database Schema

**LSST Database Schema Browser** alpha

Schema versions available for browsing: [baseline](#) | [DC3a](#) | [PT1\\_1](#) | [PT1\\_2](#) | [lmSim](#) | [S12\\_sdss](#) | [S12\\_lsstsim](#) (underlined showed)

User defined functions documentation: version 0.1, version 0.2, version 0.3 (default on lsst10)

Table List	Details for table <i>Object</i>																																																																																																																		
AAA_Version_3_2_4 ApertureBins CodVisit CodVisitMetadata DiaForcedSource DiaObject DiaObject_To_Object_Match DiaSource ForcedSource LeapSeconds <b>Object</b> Object_APMean Object_Extra Object_NonPeriodic Object_Periodic prv_Amp prv_Ccd prv_cnf_Amp prv_cnf_Ccd prv_cnf_Filter prv_cnf_Fpa prv_cnf_InputDataSet prv_cnf_Node prv_cnf_Raft prv_cnf_Run prv_cnf_Task prv_cnf_Task2TaskExecution prv_cnf_Task2TaskGraph prv_cnf_TaskExecution prv_cnf_TaskGraph prv_cnf_TaskGraph2Run prv_Filter prv_Fpa prv_InputDataSet prv_Node prv_ProcHistory prv_Raft prv_Run prv_Snapshot prv_Task prv_Task2TaskExecution	The Object table contains descriptions of the multi-epoch static astronomical objects, in particular their astrophysical properties as derived from analysis of the Sources that are associated with them. Note that fast moving objects are kept in the MovingObject tables. Note that less-frequently used columns are stored in a separate table called Object_Extra. <table border="1"> <thead> <tr> <th>name</th> <th>type</th> <th>not null</th> <th>unit</th> <th>ucd</th> <th>description</th> </tr> </thead> <tbody> <tr> <td>objectId</td> <td>BIGINT</td> <td>y</td> <td></td> <td>meta.id;src</td> <td>Unique id.</td> </tr> <tr> <td>parentObjectId</td> <td>BIGINT</td> <td></td> <td></td> <td></td> <td>Id of the parent object this object has been deblended from, if any.</td> </tr> <tr> <td>procHistoryId</td> <td>BIGINT</td> <td>y</td> <td></td> <td></td> <td>Pointer to Processing-History table.</td> </tr> <tr> <td>psRa</td> <td>DOUBLE</td> <td></td> <td>deg</td> <td>pos.eq.ra</td> <td>RA-coordinate of the center of the object for the Point Source model at time 'psEpoch'.</td> </tr> <tr> <td>psRaSigma</td> <td>FLOAT</td> <td></td> <td>deg</td> <td>stat.error;pos.eq.ra</td> <td>Uncertainty of psRa.</td> </tr> <tr> <td>psDecl</td> <td>DOUBLE</td> <td></td> <td>deg</td> <td>pos.eq.dec</td> <td>Decl-coordinate of the center of the object for the Point Source model at time 'psEpoch'.</td> </tr> <tr> <td>psDeclSigma</td> <td>FLOAT</td> <td></td> <td>deg</td> <td>stat.error;pos.eq.dec</td> <td>Uncertainty of psDecl.</td> </tr> <tr> <td>psMuRa</td> <td>FLOAT</td> <td></td> <td>mas/yr</td> <td>pos.pm</td> <td>Proper motion (ra) for the Point Source model.</td> </tr> <tr> <td>psMuRaSigma</td> <td>FLOAT</td> <td></td> <td>mas/yr</td> <td>stat.error;pos.pm</td> <td>Uncertainty of psMuRa.</td> </tr> <tr> <td>psMuDecl</td> <td>FLOAT</td> <td></td> <td>mas/yr</td> <td>pos.pm</td> <td>Proper motion (decl) for the Point Source model.</td> </tr> <tr> <td>psMuDeclSigma</td> <td>FLOAT</td> <td></td> <td>mas/yr</td> <td>stat.error;pos.pm</td> <td>Uncertainty of psMuDecl.</td> </tr> <tr> <td>psParallax</td> <td>FLOAT</td> <td></td> <td>mas</td> <td>pos.parallax</td> <td>Stellar parallax, for the Point Source model.</td> </tr> <tr> <td>psParallaxSigma</td> <td>FLOAT</td> <td></td> <td>mas</td> <td>stat.error;pos.parallax</td> <td>Uncertainty of psParallax.</td> </tr> <tr> <td>uPsFlux</td> <td>FLOAT</td> <td></td> <td>nmgy</td> <td>phot.count</td> <td>Calibrated flux for Point Source model for u filter.</td> </tr> <tr> <td>uPsFluxSigma</td> <td>FLOAT</td> <td></td> <td>nmgy</td> <td>stat.error;phot.count</td> <td>Uncertainty of uPsFlux.</td> </tr> <tr> <td>gPsFlux</td> <td>FLOAT</td> <td></td> <td>nmgy</td> <td>phot.count</td> <td>Calibrated flux for Point Source model for g filter.</td> </tr> <tr> <td>gPsFluxSigma</td> <td>FLOAT</td> <td></td> <td>nmgy</td> <td>stat.error;phot.count</td> <td>Uncertainty of gPsFlux.</td> </tr> <tr> <td>rPsFlux</td> <td>FLOAT</td> <td></td> <td>nmgy</td> <td>phot.count</td> <td>Calibrated flux for Point Source model for r filter.</td> </tr> </tbody> </table>	name	type	not null	unit	ucd	description	objectId	BIGINT	y		meta.id;src	Unique id.	parentObjectId	BIGINT				Id of the parent object this object has been deblended from, if any.	procHistoryId	BIGINT	y			Pointer to Processing-History table.	psRa	DOUBLE		deg	pos.eq.ra	RA-coordinate of the center of the object for the Point Source model at time 'psEpoch'.	psRaSigma	FLOAT		deg	stat.error;pos.eq.ra	Uncertainty of psRa.	psDecl	DOUBLE		deg	pos.eq.dec	Decl-coordinate of the center of the object for the Point Source model at time 'psEpoch'.	psDeclSigma	FLOAT		deg	stat.error;pos.eq.dec	Uncertainty of psDecl.	psMuRa	FLOAT		mas/yr	pos.pm	Proper motion (ra) for the Point Source model.	psMuRaSigma	FLOAT		mas/yr	stat.error;pos.pm	Uncertainty of psMuRa.	psMuDecl	FLOAT		mas/yr	pos.pm	Proper motion (decl) for the Point Source model.	psMuDeclSigma	FLOAT		mas/yr	stat.error;pos.pm	Uncertainty of psMuDecl.	psParallax	FLOAT		mas	pos.parallax	Stellar parallax, for the Point Source model.	psParallaxSigma	FLOAT		mas	stat.error;pos.parallax	Uncertainty of psParallax.	uPsFlux	FLOAT		nmgy	phot.count	Calibrated flux for Point Source model for u filter.	uPsFluxSigma	FLOAT		nmgy	stat.error;phot.count	Uncertainty of uPsFlux.	gPsFlux	FLOAT		nmgy	phot.count	Calibrated flux for Point Source model for g filter.	gPsFluxSigma	FLOAT		nmgy	stat.error;phot.count	Uncertainty of gPsFlux.	rPsFlux	FLOAT		nmgy	phot.count	Calibrated flux for Point Source model for r filter.
name	type	not null	unit	ucd	description																																																																																																														
objectId	BIGINT	y		meta.id;src	Unique id.																																																																																																														
parentObjectId	BIGINT				Id of the parent object this object has been deblended from, if any.																																																																																																														
procHistoryId	BIGINT	y			Pointer to Processing-History table.																																																																																																														
psRa	DOUBLE		deg	pos.eq.ra	RA-coordinate of the center of the object for the Point Source model at time 'psEpoch'.																																																																																																														
psRaSigma	FLOAT		deg	stat.error;pos.eq.ra	Uncertainty of psRa.																																																																																																														
psDecl	DOUBLE		deg	pos.eq.dec	Decl-coordinate of the center of the object for the Point Source model at time 'psEpoch'.																																																																																																														
psDeclSigma	FLOAT		deg	stat.error;pos.eq.dec	Uncertainty of psDecl.																																																																																																														
psMuRa	FLOAT		mas/yr	pos.pm	Proper motion (ra) for the Point Source model.																																																																																																														
psMuRaSigma	FLOAT		mas/yr	stat.error;pos.pm	Uncertainty of psMuRa.																																																																																																														
psMuDecl	FLOAT		mas/yr	pos.pm	Proper motion (decl) for the Point Source model.																																																																																																														
psMuDeclSigma	FLOAT		mas/yr	stat.error;pos.pm	Uncertainty of psMuDecl.																																																																																																														
psParallax	FLOAT		mas	pos.parallax	Stellar parallax, for the Point Source model.																																																																																																														
psParallaxSigma	FLOAT		mas	stat.error;pos.parallax	Uncertainty of psParallax.																																																																																																														
uPsFlux	FLOAT		nmgy	phot.count	Calibrated flux for Point Source model for u filter.																																																																																																														
uPsFluxSigma	FLOAT		nmgy	stat.error;phot.count	Uncertainty of uPsFlux.																																																																																																														
gPsFlux	FLOAT		nmgy	phot.count	Calibrated flux for Point Source model for g filter.																																																																																																														
gPsFluxSigma	FLOAT		nmgy	stat.error;phot.count	Uncertainty of gPsFlux.																																																																																																														
rPsFlux	FLOAT		nmgy	phot.count	Calibrated flux for Point Source model for r filter.																																																																																																														

<http://ls.st/s91>



## Images

Persisted: ~**38 PB**  
Temporary: ~ $\frac{1}{2}$  **EB**

~3 million “visits”  
~47 billion “objects”  
~9 trillion “detections”

Largest table: ~**5 PB**  
Tallest table: ~50 trillion rows  
Total (all data releases,  
compressed): ~**83 PB**

Ad-hoc user-generated  
data  
Rich provenance





## ➤ Database

- **Real-time Alert DB.**

  - No-overwrite updates between Data Releases

  - Real-time replica of Alert Prod DB for analytics. No long-running analytics here

- **Immutable Database (+user workspaces)**

  - Released annually. Immutable

  - 2 most recent releases on disk

## ➤ Images

- raw: 2 most recent visits for each filter
- coadds and templates: for 2 most recent releases
- raw calibration: most recent 30 days
- science calibrated: most recent 30 days
- observatory telemetry: all
- cutouts for alerts: all
- EPO full-sky jpeg: one set



- **Aiming to enable majority of analytics via database**
  - **Aiming to enable rapid turnaround on exploratory queries**
- 

- **In a region**

- get an object or data for small area - <10 sec

- **Across entire sky**

- Scan through billions of objects - ~1 hour
- Deeper analysis (Object\_\*) - ~8 hours

- **Analysis of objects close to other objects**

- ~1 hour, even if full-sky

- **Analysis that requires special grouping**

- ~1 hour, even if full sky

- **Time series analysis**

- Source, ForcedSource scans - ~12 hours

- **Cross match & anti-cross match with external catalogs**

- ~1 hour

Sizing the system for ~100 interactive + ~50 complex simultaneous DB queries. Same for images



# APIs



- Metadata
  - RESTful WebServ
- Images
  - RESTful ImageServ
- Databases
  - **RESTful DbServ**
  - SQL92 +/-, **MySQL-like DBMS**
  - Next-to-database python-based
- Query volume controlled by Resource Mgmt







## ➤ **Spatial constraints**

- `qserv_areaspec_box(lonMin, latMin, lonMax, latMax)`
- `qserv_areaspec_circle(lon, lat, radius)`
- `qserv_areaspec_ellipse(semiMajorAxisAngle, semiMinorAxisAngle, posAngle)`
- `qserv_areaspec_poly(v1Lon, v1Lat, v2Lon, v2Lat, ...)`

```
SELECT objectId
FROM   Object
WHERE  qserv_areaspec_box(2, 89, 3, 90)
```

# Current Restrictions (SQL92 +)



## Only a SQL subset is supported

For example:

- Spatial constraints (must use User Defined Functions, must appear at the beginning of WHERE, only one spatial constraint per query, arguments must be simple literals, OR not allowed after `area_qserv_areaspec_*`)
- Expressions/functions in ORDER BY clauses are not allowed
- Sub-queries are NOT supported
- Commands that modify tables are disallowed
- MySQL-specific syntax and variables not supported
- Repeated column names through `*` not supported



# Selected Common Query Types



- SELECT sth FROM Object
  - massively parallel
- SELECT sth FROM Object WHERE qserv\_areaspec\_box(....)
  - selection inside chunks that cover requested area, in parallel
- SELECT sth FROM Object JOIN SOURCE USING (objectId)
  - massively parallel without any cross-node communication
- SELECT sth FROM Object WHERE objectId = <id>
  - quick selection inside one chunk

Common queries – see <http://ls.st/ed4>



# QServ Under the Hood



# Implementation Strategy

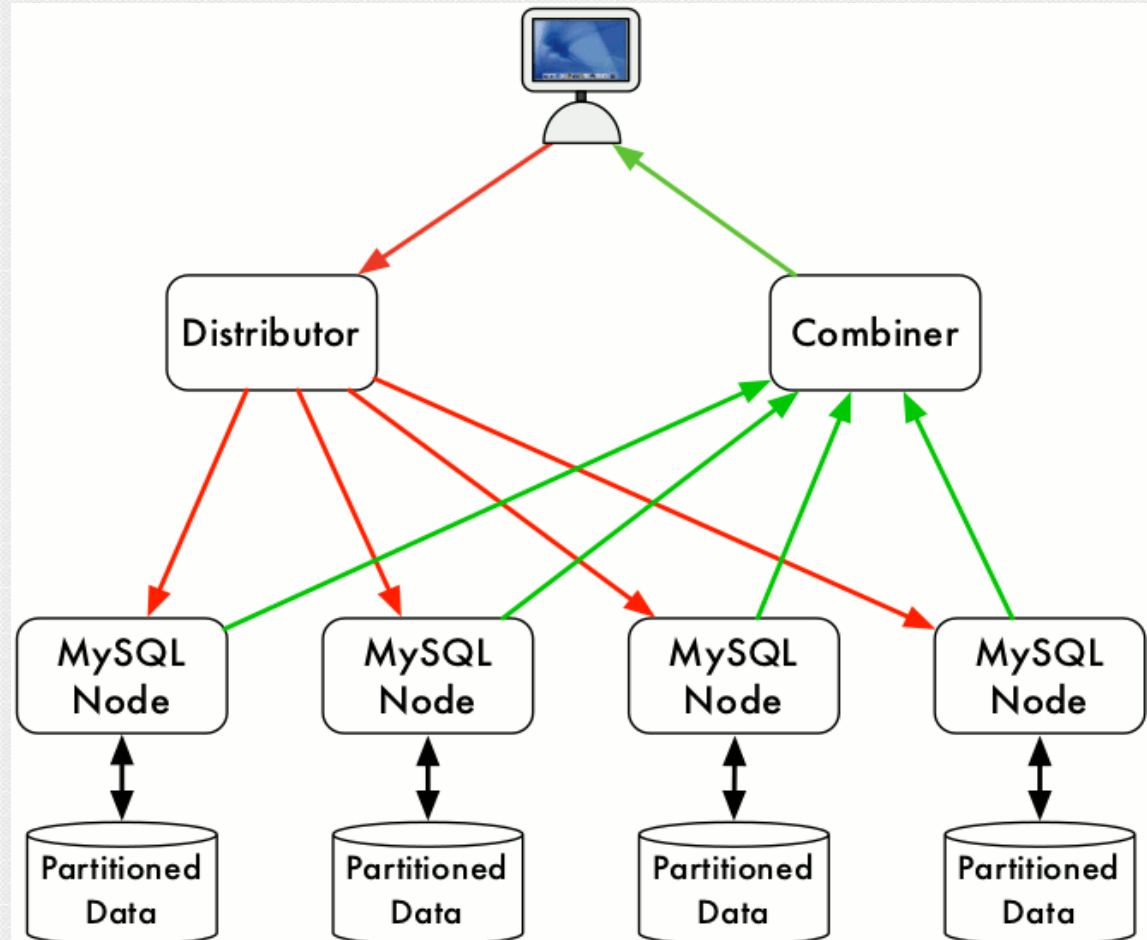
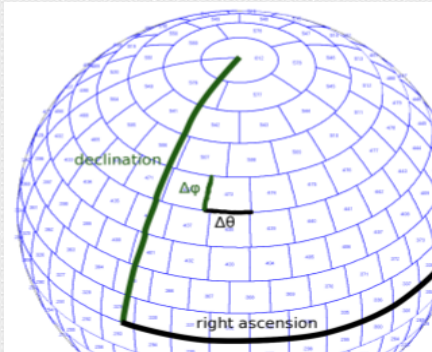


- 100% Open source
- Keep it flexible
- Hide complexity
- Reuse existing components:
  - MariaDB, MySQL Proxy, XRootD, Google protobuf, flask
- Plus custom glue
  - C++ + a bit of python. Some ANTLR
  - Lots of multithreading, callbacks, mutexes and sockets
- And custom UDFs

# QServ Design



- Relational database, spatially-sharded with overlaps
- Map/reduce-like processing







## ➤ **Scalable spherical geometry**

- 0/360 RA wrap around, pole distortion, convex polygons,
- accurate distance computation, functions for distance (angle),
- point-in-spherical-region tests (circle, ellipse, box, convex polygon)
- Custom (HTM-based) UDFs (<https://github.com/wangd/scisql>)

## ➤ **Optimized spatial joins for neighbor queries, cross-match**

- Spherical partitioning with overlap
- Director table, secondary index
- Two-level, 2nd level materialized on-the-fly

## ➤ **Shared scans**

- Continuous, sequential scans through data, including L3 distributed tables
- (Non-interactive) queries attached to appropriate running scan

## ➤ **All internal complexity transparent to end-users**

# Tests and Demonstrations



300 nodes, 10 TB data set

- 1-4 sec easy queries, 10 sec-10 min table scans, ~5 min complex joins

Running now: 2x 25 nodes, ~35 TB data set @IN2P3

+ LVM-express machine with ~2TB memory

In the near term: **Prototype Data Access Center at NCSA**



# Scale testing to date @IN2P3



S15 large scale tests:

Data: replicated SDSS Stripe 82

~10% DR1 (~2B Object, ~35B Source, ~172B F. Source)

Hardware: 24 nodes @ IN2P3, 2 x 1.8GHz 4 core, 16G RAM

Simul. 50 low-volume queries + 5 high-volume queries:

<1s for low-volume queries

~15m for high-volume Object scans

~1h for Source scans

See confluence page “S15 Large Scale Tests”



Official LSST  
code repositories



Developers  
workstations



Docker Hub



Private registry mirror

**CC-IN2P3**



Deployment  
scripts

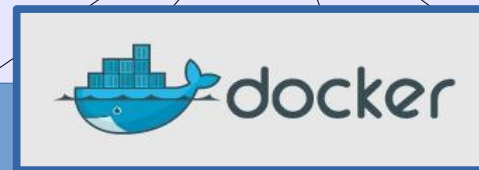


Input data



Kerberos

Build  
node



Master



Worker\_1



Worker\_i



Worker\_49

Private subnet



# CI multi-node integration tests



Official LSST  
code repositories



Developers  
workstations



## Travis CI

**SAAS CI server**

Automatically:

- build
- configure
- start cluster
- launch tests



**Ephemeral and virtual fresh Qserv cluster**



master



worker 1

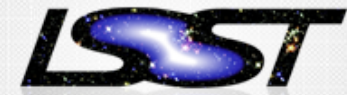


worker 2



worker 3

# Automated Qserv deployment in OpenStack



Official LSST code repositories



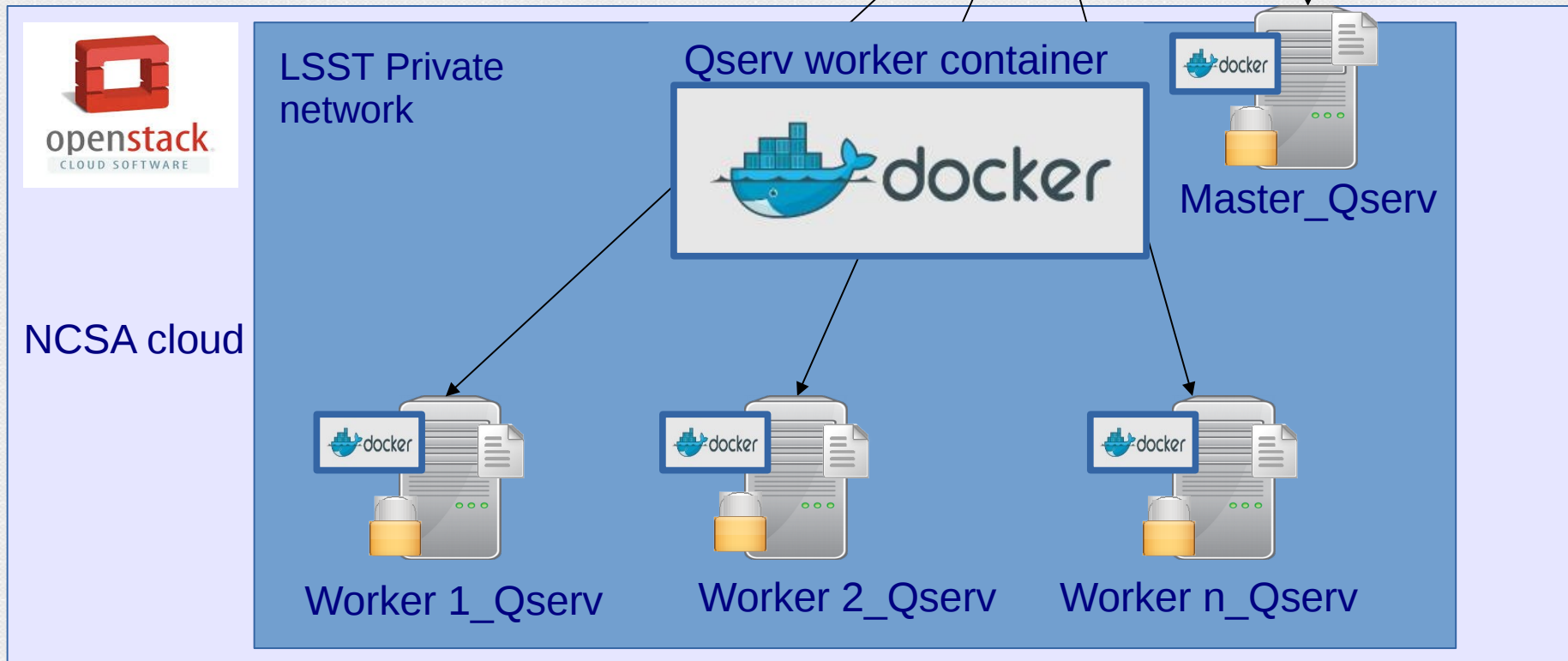
workstation

**shmux:**  
- containers management  
- integration tests



Docker Registry

Qserv master container





# Summary



- Big Data with Complex Analytics
- Spatially-sharded, map/reduce-like RDBMS
- Open source + custom glue
- Optimized for astronomical data sets at scale
- Have working prototype
- Turning it into a production system
- Want to learn more?
  - <http://ls.st/4gh> (Database Design doc)
  - <http://ls.st/6ym> (User Manual)
- Are you an adventurous super early adopter? You can try it now
  - <http://ls.st/89y> (Qserv Documentation)

# Implementation Details

